DOI: 10.16943/ptinsa/2018/49413

Computing for Science, Engineering, and Society: *A Strategic Roadmap*

RK Shyamasundar, Fellow INSA

and

Vipin Chaudhary, Ashwin Gumaste, Inder Monga,
Ankur Narang, Vishwas Patil, Prabhat

Contents

Preface

Year 2011 was the centenary year of Homi Bhabha – the founder of Tata Institute of Fundamental Research (TIFR) and other key Science & Technology establishments in India. To commemorate the centenary, I had organized a conference on *Computing for Science Discovery and Innovations: A Roadmap* at TIFR, Mumbai. During the conference, some of my friends and colleagues persuaded me to take up writing of a report *Computing for Science, Engineering, and Society: Challenges, Requirement, and Strategic Roadmap* under the solicited programme of "Commissioning of well-researched in-depth reports on topics of scientific and societal importance" from INSA (Indian National Science Academy). It was certainly a good idea but I underestimated the challenges having accepted the task. It is indeed gratifying that many of my friends and colleagues rescued me by identifying several authors who could potentially contribute to the task, as it was not the task of a single person. I am indeed extremely grateful to my co-authors who have contributed in bringing out various challenges by discussing the basis of prioritization in their sub-areas.

Providing a comprehensive roadmap is almost an impossible task due to the dynamic and the ubiquitous nature of computing and communication. In this report, we shall highlight some of the areas like HPC challenges for science, engineering and society, big data applications, HPC requirements and impact on public infrastructures and glimpses to some of the possibly disruptive research areas of computer science including AI, Cyber security, MOOCs, and computational thinking. Each chapter is structured to bring out a main challenge and a few takeaways to meet the challenge or get the best for the society from the investment in the sector. The report is organized as follows:

- In Chapter 2, RK Shyamasundar provides a brief reflection on evolution and future of computer science; a brief account of some of the research challenges of AI, Cyber security, MOOCs are also highlighted.
- In Chapter 3, Inder Monga and Prabhat explore the impact on Information and Computing Technology (ICT) infrastructure on science and highlight their experiences from the Energy Sciences Network of USA.
- In Chapters 4 and 5, Ashwin Gumaste brings out the requirements of network computing, challenges and the impact it can generate in Indian context from his experiences of building SDN routers at IIT Bombay and deployed across India. It also provides a lead for one to think as to how the local innovations in frontier technologies need a support base from the industry, business, and the Government to take it forward.
- Vipin Chaudhary explores the role of Big Data in Government in Chapter 6, in which he brings out his research experiences that include his findings from the role Big Data is playing in USA.
- Ankur Narang who was with IBM India Research Laboratories while writing the report explores HPC applications in smart-grid, oil and natural gas in Chapter 7.
- In Chapter 8, RK Shyamasundar and Vishwas T Patil explore the supposedly disruptive blockchain technology in finance and governance.
- In Chapter 9, RK Shyamasundar briefly highlights Computational Thinking – a paradigm that is building bridges across various science disciplines through the computing paradigm.
- The report concludes in Chapter 10 with a very broad takeaways that policy-makers can have on "the impact of computing paradigm on science, and society."
- The report also has an Appendix that highlights the recommendations of various experts in the context of a conference held in connection the centenary celebration of Homi Bhabha at TIFR.

There has been a delay in arriving at the report owing to a spectrum of challenges including my movement from TIFR to IIT Bombay in early 2015. In spite of the delay, I hope that this report will be useful to a large section of the communities involved in R&D activities relying on computing infrastructure.

Having got the manuscripts in various formats from the co-authors, to place the reports in an uniform book format was really a gigantic task. Vishwas Patil (CSE, IIT Bombay) took the task and typeset it using LaTeX. I am indeed grateful to him. Initial reading, suggestions and proof reading were done by NV Narendra Kumar (TIFR/IIT Bombay, currently with IDRBT Hyderabad) and Vishwas Patil. I am extremely thankful to both of them for their suggestions and careful reading of the manuscripts. It is indeed a pleasure to thank STCS at TIFR and Department of CSE at IIT Bombay for all the facilities provided while writing this report.

R.K. Shyamasundar, *Fellow INSA*
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, Mumbai
October 2017.

Current Affiliations of Authors

Professor Vipin Chaudhary
Department of Computer Science
SUNY Buffalo
(Former CEO, Computational Research Lab. (CRL), Tata Sons. Ltd.);
Currently duty at NSF and also on HPC advisory board of the US Council of Competitiveness
vipin@buffalo.edu

Professor Ashwin Gumaste
Department of Computer Science and Engineering
IIT Bombay
ashwin@cse.iitb.ac.in

Dr. Inder Monga
Lawrence Berkeley National Lab
University of Berkeley
imonga@es.net

Dr. Ankur Narang
Associate Vice President – Data Science, Chief Data Scientist at Mobileum
(Formerly with IBM India Research Labs)
annarang2@gmail.com

Dr. Vishwas T Patil
Department of Computer Science and Engineering
IIT Bombay
ivishwas@gmail.com

Dr. Prabhat
National Energy Research Scientific Computing Center
USA
prabhat@lbl.gov

Professor R.K. Shyamasundar
Department of Computer Science and Engineering
IIT Bombay
rkss@cse.iitb.ac.in

Chapter 1

Executive Summary

Computing, communication, and technology have not only made an amazing progress in the past couple of decades but also made a huge impact on the evolution of science and society. During the last half-century, we can broadly say that the pace & direction of this evolution is due to the rapid growth in theoretical & technological advancements in Computer Science (hardware & software), networks, electronics, photonics – each one catalyzing the other's growth & thus reducing their costs and in turn their accessibility. In fact, some of the outcomes were a result of funding from NSF, DARPA (USA), and advancements realized at Bell Labs, AT&T, IBM, Xerox, SRI, et al. One of the most striking observations since the beginning of the century has been that the pace and growth is largely dictated by the market[1]. Some of the important characteristic observations are:

1. There is an information avalanche as the digital universe is expanding at a rapid rate. It is estimated that by year 2020, the digital data produced will exceed 40 zettabytes. This corresponds to saying for every human there is approximately 5200 gigabytes of data[2].

2. Disruptive, efficient infrastructures have shrunk the world making it possible to produce and consume regardless of the location. This, in fact, is showing a tendency of huge growth; thus evolution of society will be dictated by the people and social communities who are able to coalesce in a very short time and make their presence and influence felt.

   In these days, with the availability of massive computations, there has been successful attempts to shift *parts of decision phase of applications, that had been the forte of humans, to machines.* This is often referred to as AI/Machine Learning in various glorious terms in the media. Thus, the machines are not necessarily just number crunchers but also a decision makers. It is important to note that the compound annual growth rate (CAGR) of investment in such applications is more than 47 billion USD. This naturally reflects the need to invest in such massive computing infrastructures that would drive innovations and discoveries.

3. Blockchains and the distributed ledger technology have been making inroads in finance and governance for transparency, efficiency and trust management.

4. Another very significant point to be noted is that scaling up of scientific discovery has become dependent on the computing power both in theory (paradigms) and practice (applications.)

5. Even disjoint areas of science & technology are influencing one another, for instance photonics, quantum mechanics, smart-grid, CRISPR, containment of communicable disease outbreaks through social networks (on GPS.)

6. Apart from the foundational areas of Computer Science including ICT, some of the recent works on cognitive computing have shown an enormous potential for healthcare, education. Keeping these fallouts in mind, there is a need to consider how ICT can accelerate human decision making, creativity, and innovation, etc., in a variety of scenarios of health and other impactful societal applications.

7. Power of computing lies in its open-endedness and its understanding rather than constraining it. Thus, there is a dire need to find ways by which cooperation, collaboration, synergy, and consilience be provided on a platform that provides momentum to science and engineering discoveries and innovations. Such a need becomes amply clear when we look at the scenario of *precision medicine* that has become possible only due to the cooperative work of doctors, patients, biomedical researchers, engineers, and computer scientists.

[1] IEEE COMSOC 2020 Report, June 2012.
[2] Zhiwei Xu and Guojie Li, Computing for the Masses, CACM, 54, 10, pp. 129-137.

Thus, policymakers need to strategize policies for investment in ICT for a spectrum of purposes ranging from societal needs to science discoveries, keeping in mind at least a few crucial facts like:

1. A number of main crises that are facing the world, environmental pollution, scarcity of basic resources (like water and food), energy (in terms of availability and cost) shall matter.

2. The proliferation of Internet of Things along with the underlying sensors, and converging data standards are all combining to provide new possibilities for the physical management and the socio-economic development of cities. For instance, in the context of smart cities, it is necessary to keep in mind[3]: *Technologies influence patterns of behavior. Digital and mobile technologies are making the connections between service providers and users, tighter, faster, more personal, and more comprehensive. Sharing-economic business models are emerging that enable more efficient use of physical assets, such as cars or real estate, and provide new sources of income to city residents.*

3. Forecasting the future of ICT is hard and risky due to dramatic changes in technology and limitless challenges to innovation. In fact, it is only a small fraction of the innovations that truly disrupt the state of the art. Some are not practical or cost-effective, some are ahead of their time, and some simply do not have a market. There are numerous examples of superior technologies that were never adopted because others arrived on time or fared better in the market.

4. There is a significant digital divide: developing countries with a shortage of technology and education are still "technology dependent", but developed countries are already market driven, with technology having become a commodity, and with expectations covering more than just technology. Industry and institutions have broadened their interest and now focus very much on meeting market needs.

Keeping the above rationale in view, the report is only an attempt to better understand the impact of the disruptions the disciplines and technologies may lead to. Such an understanding enables us in arriving at strategies for ICT investment to promote innovations[4] in science, technology and societal applications, to meet the spectrum of requirements and aspirations of the country. **It is necessary to keep in mind that while the country is a giant in software, it has a long way to gain that level of supremacy in hardware and products/systems. Such a reality makes it necessary for the country to get supremacy in building and evolving complex intelligent systems that could have an impact on society and science.** Needless to say that the overall strategy for serious innovations for scientific and technological discoveries/inventions require a thorough re-organization of higher education keeping in view the role of ICT. The report provides a broad overview to cater to an effective ICT investment for innovations for research, higher education, and societal applications.

The important broad takeaways are:

1. Initiate and invest in strategic computing centers to promote scalable science discoveries keeping in view the expected requirements rather than just the raw power.

2. Establish scalable strategic centers for cyber security (including big data analytics, blockchain/crypto-currencies, etc.), e-governance, public infrastructures, etc.

3. Invest in innovative approaches for network and system design that would promote scalable architectures leading to large scale systems research.

4. Invest in adapting quality human resource developments by architecting computational thinking in science and engineering disciplines.

5. Invest in a strong research above threshold/critical strength in core computing disciplines that include frontier areas, including Deep Learning (AI), cognitive computing, quantum computing, blockchain applications that have demonstrated societal impact, that could be game changers in the years to come. In fact, the recent leap into quantum computing processors (IBM, Google, Intel), or neuromorphic chips from HP needs to be kept in mind, to address the needs of building strong computing infrastrtures for pushing innovations and inventions for science and engineering as well as societal applications.

6. The most important aspect is to monitor progress with constructive feedback through independent evaluations.

[3]Technologies and the Future of Cities, Feb 2016, `https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_cities_report___final_3_2016.pdf`

[4]The country that wants to out-compete must out-compute (Suzy Tichenor, June 2007)

Chapter 2

Computer Science: Reflection and Future

RK Shyamasundar
IIT Bombay

2.1 Early Formative Era

There are various evidences of calculation as an ancient activity [Tedre, 2014] that dates back to Babylonian days used for varieties of navigational, astronomical and other day-to-day needs. In fact, there were methods of storing like Quipu of Incas and tools for calculating like Chinese counting rods. Computing as a discipline is a recent one even though the practice of using mechanical aids for calculation can have various dates based on the perspective of the reader like Blaise Pascal in 1600s, George Boole in the 1800s or the Babylonian dates of 1800BCE. It is only in the early 20^{th} century a firm foundation of Computing was laid while attempting to solve the problem referred to as the *Entscheidungsproblem*[1] posed by David Hilbert and Wilhelm Ackermann in 1928. Alan M. Turing – British mathematician established that there is no method to solve this problem through a formal definition of an abstract machine, now referred to as Turing Machine. This seminal work [Turing, 1937] laid the foundation of computing. It must be mentioned that while other contemporary logicians like Alonzo Church, Emil L. Post, and A. A. Markov had proposed logical formalisms to show that the *Entscheidungsproblem* was not solvable and in fact, it was later shown that these formalisms turned out to be equivalent to Turing Machine's. However, it was Turing's work that gave a firm momentum to the computing field from multiple dimensions. This becomes evident from the quote due to Kurt Gödel[2] from his Gibbs Lecture: "the greatest improvement was made possible through the precise definition of the concept of *finite procedure*, which plays a decisive role in these results. There are several different ways of arriving at such a definition, which, however all lead to exactly the same concept. The most satisfactory way, in my opinion, is that of reducing the concept of finite procedure to that of a machine with a finite number of parts, as has been done by the British mathematician Turing". Furthermore, Gödel accepted the earlier thesis of Church only after Turing's work. The thesis since then comes to be known as Church-Turing thesis. The thesis, which had a far-reaching impact on this field, is informally stated below:

> *Any algorithmic problem for which an algorithm can be found in any programming language on any computer (existing or that can be built in future) requiring unbounded amounts of resource is also solvable by a Turing Machine.*

In other words, the thesis implies that the most powerful supercomputer with the most sophisticated array of programming languages is no more powerful than a PC with a simple hardware and software up to polynomial loss in efficiency. Thus, the seminal paper can be treated as the birth of Computer Science. John von Neumann engineered Turing's ideas of programs as data (the concept often referred to as Stored Program concept) to realize the first stored program computer – often referred to as von Neumann machines. These ground breaking theoretical and practical realizations essentially launched the field of Computer Science and Computer Systems that have had a great impact on science and society.

Not only did Turing invent a machine capable of computing all effectively computable functions, he formulated a test, which has come to be known as Turing Test for testing *normal human intelligence* – that initiated

[1] The question posed was: Is it possible to have a method that takes a proposition in first-order logic as input which will decide in a finite number of well-defined steps, whether the proposition is true or not?

[2] Note that by establishing that there is no complete and consistent set of axioms for all of mathematics, Gödel shattered the dream of Bertrand Russell and A. N. Whitehead.

the area termed Artificial Intelligence. Turing's digital forecast done in his paper "Computing Machinery and Intelligence" [Turing, 1950] gives a reflection of where the field has reached. To quote from Turing:

> *I believe that in about fifty years' time it will be possible, to program computers, with a storage capacity of about* 10^9*, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.* – Quote (A)

An analysis of Alan Turing's prediction due to Alan Turing by Jim Gray (A Turing Laureate) provides a good assessment as to where the field has reached:

> *With the benefit of hindsight, Turing's predictions read very well. His technology forecast was astonishingly accurate, if a little pessimistic. The typical computer has the requisite capacity, and is comparably powerful. Turing estimated that the human memory is* 10^{12} *and* 10^{15} *bytes, and the high end of that estimate stands today. On the other hand, his forecast for machine intelligence was optimistic. Few people characterize the computers as intelligent. You can interview Chatter Bots on the Internet (http://www.loebner.net/Prizef/loebner-prize.html) and judge for yourself. I think they are still a long way from passing the Turing Test. But, there has been enormous progress in the last 50 years, and I expect that eventually a machine will indeed pass the Turing Test. To be more specific, I think it will happen within the next 50 years because I am persuaded by the argument that we are nearing parity with the storage and computational power of the mind. Now, all we have to do is understand how the mind works(!).* – Quote (B)

2.1.1 Impact of Computer Science

One of the hallmarks of Turing was that he was seeing computation everywhere: from abstract mathematics to developmental biological observations like stripes of a tiger or a zebra. He firmly established a variety of computational methods for the concrete understanding of traditional mathematical concepts specified by finitely definable approximations, such as measure or continuity. Some of his notable contributions of significance explicitly in this direction are:

- LU decomposition,
- Finite approximations of continuous groups,
- Computation over reals,
- Chemical basis for morphogenesis and non-linear dynamic simulation.

Around the time of this great intellectual revolution in computing as briefed above, the second World War had begun and was in full swing. Naturally, the military establishments of USA and UK became seriously interested in automatic computations of ballistic and navigation tables as well as the cracking of ciphers. One of the most successful projects in this direction was the UK's top-secret project at Betchley Park that cracked the German Enigma cipher using several methods devised by Turing. These efforts had the good side-effect of providing a boost to the spread of computing resulting in universities offering new fields of study.

In the early stages of computer usage, the emphasis was on making computers useful. The research, education and development efforts could be broadly divided into four parts as shown in Table 2.1:

As the use of computers reached a reasonable level of maturity, the areas of specialization like theory of computation, algorithmic analysis, data structures, numerical analysis, compiler construction, operating systems, programming methodology, artificial intelligence, software engineering, etc., evolved. It is of interest to note that the mathematical foundations pursued for the above studies, happened to be not classical analysis as is the case in science studies; it was rather logic (mathematical, computational, philosophical), universal algebra & ordered sets, discrete structures & combinatorial theory. These topics could be termed "mathematics of weak structures" where "weak" is used in an axiomatic sense like semi-groups vs. groups, distributive lattices vs. Boolean algebras, projective vs. Eucledian geometry. These topics, perhaps due to lack of stimulating applications, have always existed as topics of peripheral interest within mathematics. The requirements of Computer Science completely changed the situation. Computer Science needed ideas from these topics and in turn stimulated the development within these topics by posing questions which would not have been posed otherwise.

Right from the days of germination of ENIAC/EDSAC, John von Neumann had been advocating that computers would not be just a tool for aiding science but a way of doing science. With the *computing* reaching a stage

Characterization of computable functions/problems, intrinsic complexities of algorithms, logic of programs:	These areas got bunched under the broad name theoretical computer science that developed the underlying mathematical foundation to support this direction of research that lead to creation of automata theory, formal languages, computability theory, algorithm analysis, logic of programs, semantics of programming languages.
Languages for specifying algorithms and data so that they could be automatically computed in an effective manner:	These goals developed areas like programming languages, compilers, databases, etc.
Building reliable systems that can realize computations efficiently:	The underlying goals developed areas like computer architectures, operating systems, software engineering (intellectual manageability of large programs), etc.
Artificial Intelligence:	The efforts were to see how best the computer could mimic a human and build systems to aid human reasoning.

Table 2.1: Research, Education, and Development efforts during early stages of computer usage

of robustness in terms of hardware, software and user interface by early 1970s (time around which Computer Science germination happened in India – thanks to TIFR and IIT Kanpur) and the use of computers in science & engineering gained momentum. Ken Wilson, a Nobel Laureate in Physics, promoted an idea that simulation on computers was a way to do science and scale-up discoveries and inventions. It may be noted that Wilson's breakthroughs were realized through computational models whose simulations produced radical understanding of phase changes in materials. In fact, he championed the promotion of computational science saying that grand challenges in science could be cracked through computers. He went on to call that ***computation has become a third leg of science.*** His promotions lead to formal streams under "Computational Sciences" and also government funding for building computers increased quite substantially leading to further technological advancements. It is to be noted that these initiates lead to graduate programmes in computational sciences worldwide and the area of "High Performance Computing" took shape in academia, industry and business.

With the gearing up of science, engineering, and technological advances, areas like databases, visualization, graphics and image processing, etc., became important. The importance of human machine interface for both computer science experts and non-experts for productivity as well as varieties of applications, including business and media, lead to the invention of personal computers at Xerox PARC. These developments galloped at high momentum and developed the area of human computer interfaces (HCI) with vast applications that made computing ubiquitous.

The success of ARPANET leading to birth of Internet, the advances in mobile technology and computing and communication coming together stimulated areas like mobile computing, security, network science, etc. With the invention of world-wide-web in the early 1990's, computing spread widely even to areas which one had not imagined and in particular e-commerce. The growth of e-commerce, use of Internet for infrastructures (online store, online payment), innumerable ubiquitous applications has lead to the new field of network and information security – that has been immensely challenging from various perspectives including national and public life.

Widespread developments along with the technological advances that brought together computing and communication on one platform has lead to vast set of unimaginable applications to entertainment that includes: live music, video conferencing, virtual reality, online games, 360° photos, etc. These developments have been driving a revolution in Computer Science. The principle drivers of this revolution are:

- Integration of computing and communication,
- Huge digital data,
- The deluge of networked devices and sensors.

These developments have further triggered ways of looking at networks of people and organizations, and their integration into management, law, and policy. Concretely, the developments have given rise to a vast variety of social networks for entertainment, business, and societal governance. Needless to say, these trends have carved an entirely multi-disciplinary spectrum of challenges for integrating information systems for societal requirements.

Scaling up these computing technologies (hardware and software) with high productivity has been a huge impact on discoveries and inventions in science and engineering disciplines. In fact, we have reached to a point, wherein significant progress either in science, engineering or society is dependent on the computing power, for example, antibiotic drug discovery, study of gravitational waves or predicting next solar flares so that satellites and critical ground electronics can be safeguarded from burning. Some of the areas wherein computing has made a huge impact and expected to have disruptive impact are: smart materials, epidemiology, genomics and molecular modeling, astronomy, computational chemistry, biology, e-commerce, e-governance, health-care, robotics, earthquake engineering, disaster management, national security, public infrastructures, large scale societal systems, etc.

The current information age is a revolution that is changing all aspects of our lives. Those individuals, institutions, and nations who recognize this change and position themselves for the future will benefit enormously. Thus, we need to position ourselves in order to drive the potential benefits to the society. The magnitude of impact made by computing to science & society can be gauged by the highly convincing argument in the report [Tichenor, 2007], where Suzy Tichenor, Vice President, U.S. Council on Competitiveness, argues that:

> ***the country that wants to out-compete must out-compute***

In the report, it is argued that to drive the growth of innovations (hence the growth of the country) it is necessary to gain competitiveness with computational modeling and simulation. The main reasons behind that being (again quoting):

- *High Performance Computing (HPC) is an innovation accelerator*
- *HPC shrinks "time-to-insight" and "time-to-solution" for both discovery and invention*

The key takeaway argued for USA at that time, was:

> *enable companies, entrepreneurs, individual inventors to: innovate anywhere, with anyone, using any domain specific application running at any available High Performance Computing Center.*

Given that we are still to gain competitiveness in hardware and scalable computing is capital intensive – *we should concentrate on building large scale systems using innovative architectures and make available to stake-holders such as companies, entrepreneurs, researchers, and individuals and give momentum in driving innovations.* Some of the specific findings in terms of HPC, Big Data analytics as well as infrastructure takeaways will be elaborated later.

2.1.2 Shaping of Computing Discipline

As computing is omnipresent, it has benefitted from the best of the talents from all disciplines. Just to mention a few in the early days of computing like, Alan Turing, John von Neumann, Claude Shannon, Alonzo Church, etc., each a towering personality in a multiple disciplines of the day. Computer Science as a discipline is not even a century old, and furthermore, due to the application strides being made by the computing, the field attracted very many people from several areas of mathematics, electrical engineering, physical sciences, economics, law, and business. Due to such a large spectrum of interests, there have been a large number of dizzying arguments about the core features of computing as an academic discipline. Thus, it is but natural that various views arise depending on pioneers of the field, the background training of the persons etc. Some of the common viewpoints are:

1. Computer Science is just a technological application of mathematics, electrical engineering or science.
2. Computer Science is an independent discipline with a sound body of knowledge with its own set of challenges and ultimately is the foundation of Art of Thinking.
3. Computing is primarily a technical field that aims at cost-efficient solutions.
4. Computing is an empirical science of information processes that are found everywhere.

Several early computing pioneers have argued about the nature of Computer Science keeping in view their key perspectives. An excellent discussion of these are given in Matti Tedre [Tedre, 2014]. Some of the views are briefed below:

- *Programming* is computer science (Edsger W. Dijkstra)
- *Algorithmic* analysis is the unifying theme (Donald E. Knuth)

- Juris Hartmanis in his FSTTCS 1993 address discusses the nature of Computer Science as a science by analyzing it and comparing or contrasting it with other physical sciences. He argues that Computer Science differs from the known sciences so deeply that it has to be viewed as a new species among the sciences. This view is justified by observing that theory and experiments in Computer Science play a different role and do not follow the classic pattern in physical sciences. The change of research paradigms in Computer Science are often technology driven and simulations can play the role of experiments. Furthermore, the science and engineering aspects are deeply interwoven in Computer Science, where the distance from concepts to practical implementations is far shorter than in other disciplines.

- Herbert Simon, an economics Nobel prize winner and a Turing Laureate, called "computing" – *The Sciences of the Artificial.*

Over the past few decades, vast streams of insights on foundational aspects of algorithms, programming, representations of problems and languages of representation have been achieved. Feats of integrating computing and communication to build large complex, reliable systems have been realized and further, Artificial Intelligence (AI) techniques (like Deep Learning) have shown enormous potential in building real intelligent systems that mimic human intelligence (as forecasted/envisaged by Alan Turing) like driver-less cars, robots for medicine administration or aid in disasters like earthquake, systems that can challenge and defeat human experts who play games like Chess, Go, Jeopardy!, etc. Computer modelling and simulation has made a huge impact in computational chemistry, genomics/biology analysis, smart materials, etc.

In summary, computing has been a driver in different traditions of physical sciences, engineering, mathematics and also building societal systems in the digital era. It is almost impossible to draw a line between them as the intellectual endeavors/pursuits they represent/impact are not definable. The nature of Computer Science has evolved at a rapid pace in theory, practice and applications. For instance, the relationship between Computer Science and Mathematics is nicely captured by Knuth (1994) quoted below:

> *Like mathematics, computer science will be somewhat different from the other sciences, in that it deals with man-made laws which can be proved, instead of natural laws which are never known with certainty. Thus the subject will be like each other in many ways. The difference is in the subject matter and approach – mathematics dealing with more or less theorems, infinite processes, static relationships and computer science dealing with more or less with algorithms, finitary constructions and dynamic relationships.*

While above sets the stage for evolution, the following quote from Knuth (1985) shows the limitless nature of evolution:

> *I suppose the name of our discipline isn't of vital importance, since we will go on doing what we are doing no matter what it is called; after all, other disciplines like Mathematics, and Chemistry are no longer related very strongly to the etymology of their names.*

The table showing the range of topics during 1968-2008 taken from [Tedre, 2014] is given in Figure 2.1.

There have been several views saying that Computer Science dealt with laws of nature, as well as computing is natural science [Denning, 2003] and a thorough analysis of these aspects is explored in [Tedre, 2014]. With the maturing of the discipline and the huge impact it has made, one can conclude that it has provided a way of thinking in almost all branches of science, engineering and society; the latter abstraction can be succinctly seen in the coining of the phrase "Computational Thinking" by Jeannete Wing of CMU (we shall look a bit more into this aspect in the sequel). The elucidation of such an impact along structures of science and engineering frameworks has been captured nicely by Peter Denning [Denning, 2003] in Figure 2.2. One inference you can see, why the notion of "experiments" has also an important role nowadays and also fits well in the significant contributions of machine learning being played along for societal applications. In fact, these arguments and happenings are reflected in the following quote from Forsythe (1969):

> *The question: "What can be automated" is one of the most inspiring philosophical and practical questions of contemporary civilization*

While *Computing* has penetrated all areas as discussed already, in the following we shall take a broad look at the challenges of some of the areas that have arisen from computing for science and engineering. In particular, we focus on some of the research challenges in select areas of relevance.

Our focus of the report is to broadly highlight the role of computing science and engineering in various areas of science, engineering and societal applications. We shall discuss:

1. Computing for scaling up discoveries in science

2. Computing as a disrupter in building societal systems and the implications to the human society.

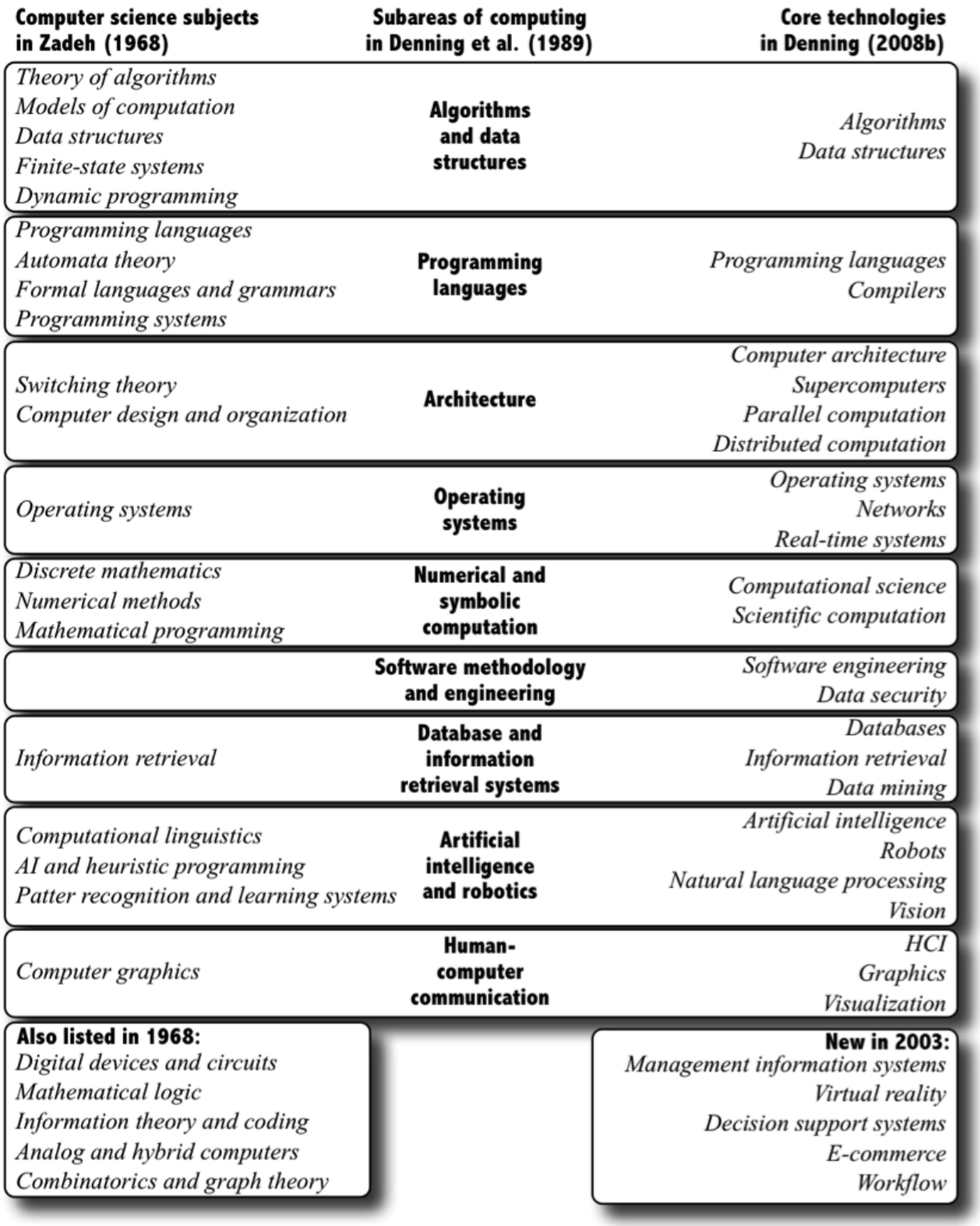


Figure 2.1: The Evolution of Computer Science Topics

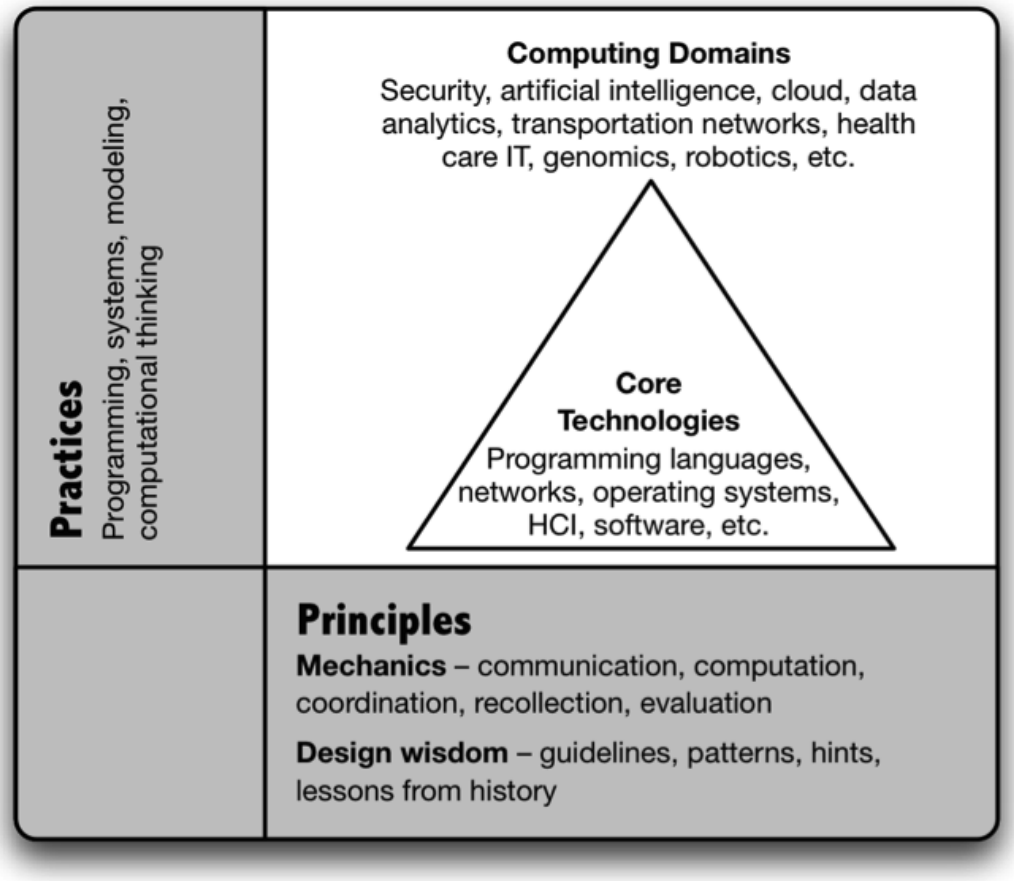


Figure 2.2: Computing Frameworks

3. Broad understanding on the challenges in computing.
4. Broad view to the funders and government to organize appropriately to meet the challenges.

Keeping this in view, the report discusses the following topics:

1. A glimpse into Computer Science research challenges
2. HPC significance for science and engineering
3. Societal impact: blockchain as the protocol for trust in Internet Era
4. Exploiting network computing towards such requirements
5. Big Data applications for governmental needs
6. HPC in public infrastructures
7. Reorganizing computing education for various disciplines; in this connection we append a report that was recently arrived on the sideline of a conference dedicated to Homi Bhabha.

Rest of the section provides glimpses of research challenges and developments in computing *w.r.t.* some select areas like: AI, MOOCs, security & privacy.

2.2 A Glimpse into Research Challenges in Computing

In these days, with the availability of massive computations, there has been a shift of *parts of decision phase of applications, that had been the forte of humans, to machines.* This is referred to as AI/Machine Learning in various glorious terms in the media. It is important to that the compound annual growth rate (CAGR) of investment in such applications is more than 47 billion USD. Thus, the machines are not necessarily just number crunchers but also a decision makers. For instance, *driving* by machines is being largely explored and indeed

has demonstrated a reasonable success[3]. Thus, it is important to keep this in mind while addressing the future challenges[4].

While there are the classical scientific research challenges like the "P=NP" problem, in this section, we shall look at a broader perspective from the viewpoint highlighted already. We shall briefly discuss areas that are covered explicitly as separate chapters.

The address by John Hopcroft (another Turing Laureate) at the Heidelberg Laureate Forum 2013, is a good starting point, as it articulates the shift in focus of Computer Science becoming more application-oriented in the years to come. He argues that the following topics would be some of computational challenges for the decade:

- Tracking evolution of communities in social networks
- Extracting information from unstructured data sources
- Processing massive data sets and streams
- Extracting signals from noise
- Tracking the flow of ideas in scientific literature
- Dealing with high dimensional data and dimension reduction

It is apparent from the topics that challenges relate to inferences on data and in a sense correspond to the Big Data analytics or machine learning. Natural outcome of the challenges is the need of building a theory to support new directions. This naturally calls for a re-look into computer science education. As extrapolated by John Hopcroft, Computer Science would need to include topics like large complex graphs, spectral analysis, high dimensions and dimension reduction, clustering, collaborative filtering, learning theory, sparse vectors, signal processing, etc.

From the above perspective, an immediate broad take-away is:

> Introduce computing paradigms and methodologies in schools and colleges, and revisit curricula of science, engineering and humanities (UG and PG) to provide the needed ICT paradigms.

Keeping in view, that various challenges in areas like HPC, SDNs, Big-Data, smart-grid, etc., are covered in other chapters, in this section, we shall provide glimpse of some of the challenges in areas like AI, MOOCs, and security & privacy that could have a disruptive impact on science and society.

2.2.1 Artificial Intelligence

"... if a machine is expected to be infallible, it cannot also be intelligent"
A. M. Turing, London Mathematical Society Address, 20 Feb 1947.

"Artificial Intelligence" was coined by Alan M. Turing through the formulation of a test which has come to be known as Turing Test for testing *normal human intelligence.* The Turing Test is an imitation game, played by three people. In this game, a man and a woman are in one room, and a judge is in the other. The three cannot see one another, so they communicate via e-mail (letters/notes). The judge questions them for 5 minutes, trying to discover which of the two is the man and which is the woman. This would be very easy, except that the man lies and pretends to be a woman. The woman tries to help the judge. If the man is a really good impersonator, he can fool the judge 50% of the time. But, it seems that in practice, the judge is right about 70% of the time. Now, the Turing Test replaces the man with a computer pretending to be a human. If it can fool the judge 30% of the time, it passes the Turing Test. The rationale of the test may be seen in Turing's own words in his BBC interview:

> *The idea of the test is that the machine has to pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing ... We had better suppose that each jury has to judge quite a number of times, and that sometimes they really are doing with a man and not a machine. That will prevent them from saying 'It must be a machine' every time without proper consideration.*

The underlying arguments against the test can again be seen in his own words:

[3] Two points to be kept in mind: (i) Is driving an intellectual activity?, (ii) Can machine handle ethics - for instance, when it comes to choosing an action between self protection vs an external person (say pedestrian) in an emergency, what will be the basis for the machine's decision? These are questions, for which there are no easy answers.

[4] In such a framework there has been many forecasts while predicting the future. For the scientists/technologists who invent the future, it is nevertheless important to understand correct scenarios of the real world. This is particularly important for the AI discipline, as it has had a roller coaster role. A recent article by the Turing Laureate Prof. Rodney Brooks, *The seven deadly sins of AI Predictions, MIT Technology Review, 6 October 2017.*

Figure 2.3: Turing Test for Normal Human Intelligence

> *The game may be criticized on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.*

Even though it is not formally defined, it is a practical test applied to an existing entity that is "running". It consists of a conversation over a period of time between the tester and the entity being tested. This demands an ability to learn and adapt the contents and the structure of the sayings of the tester. Note that the testing becomes harder the longer it goes on. The point of the test is that if some entity passes it, it is hard to deny that it is intelligent and hence throws up the possibility of judging artificial entity to be intelligent. The basis for this is based on Turing's view that *"thinking is singularly and critically indicated by verbal behavior indistinguishable from that of people as determined by a blinded experiment."* In summary, Turing Test shares important properties with interactive proofs such as exponentially rare false positives, non-composability, non-transferability, etc. Turing's seminal contribution was in enabling blinded controls. While the Test can provide an interactive proof of intelligence, it is not particularly useful as a research goal itself. While several Internet sites offer Turing Test chatterbots, none pass and still stands as a long-term challenge. The movement of AI becomes clear if we re-look at the Quotes (A)-(B) on page 6 and 6, due to Alan Turing and Jim Gray respectively.

Implicit in the Turing Test, are two sub-challenges that in themselves are quite daunting:

- Read and understand as well as a human,
- Think and write as well as a human.

Both of these appear to be as difficult as the Turing Test itself. Due to advances in computing technology, there has been tremendous progress in speech recognition[5], understanding, speech synthesizers, limited language translation, visual recognition, visual rendering, etc. While one may say the conceptual progress in these areas is limited, it is still a boon to the handicapped and in certain industrial settings. There is no doubt that these prosthetics have helped and will help a much wider audience and shall revolutionize the interface between computers and people. In fact, it has made a tremendous progress in the above (Google Glass is an example) as well as in smell measurements and odor reproduction (cf. Harel [R. Haddad and Sobel, 2008, Harel, 2016]). When computers can see and hear, it will break communication barriers. It should be much easier and less intrusive to communicate with them. In a sense, it will allow one to see better, hear better, and remember better. In the past couple of years, we have been seeing successes in these aspects.

In the following, we briefly highlight several of the rapid strides in some of these areas through the eyes of Artificial Intelligence.

Whither Artificial Intelligence?

While there are several examples wherein computers have assisted in arriving at proofs of several open problems in mathematics including the four color conjecture, it is the defeat the human experts in games like Chess, Go or Jeopardy! by computers that have ignited Artificial Intelligence in the eyes of public and business [Harel, 2016].

[5]It is of interest to look at the recent claim of Microsoft (https://goo.gl/CD9wHD) that claims its speech recognition has attained human parity.

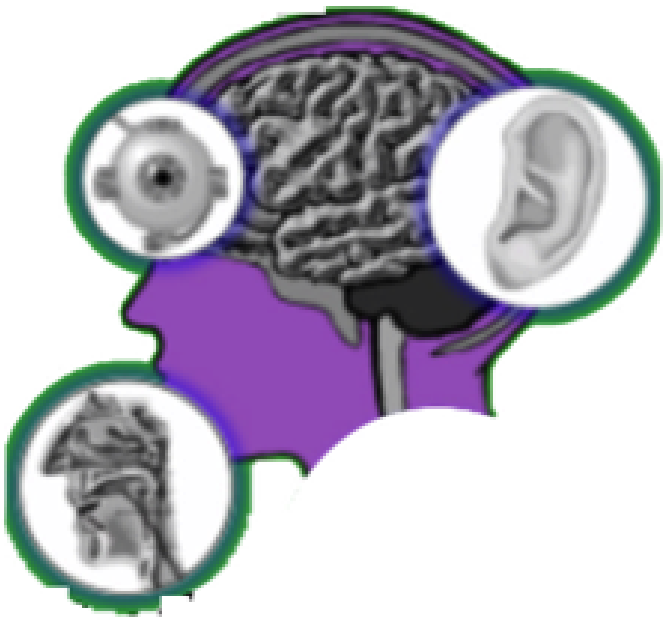


Figure 2.4: Three Prosthetic Challenges: Vision, Hearing, and Speech

It began with IBM's Deep Blue computer beating Gary Kasparov, the then reigning world chess champion. It is of interest to note that Kasparov was at a significant disadvantage during the match as the designers of Deep Blue had the opportunity to tweak Deep Blue's programming between matches to adapt to Kasparov's style and strategy. Further, they had access to full history of his previous public matches. However, Kasparov has no similar record of the machine performance as it was being modified between matches. Further more, Kasparov and other chess masters blamed the defeat on a single move made by the IBM machine. In that move, the computer made a sacrifice that seemed to hint at its long-term strategy. Kasparov and many others thought the move was too sophisticated for a computer, suggesting there had been some sort of human intervention during the game. In respect of that move grand-master Yasser Seirawan told WIRED in 2001, "It was an incredibly refined move, of defending while ahead to cut out any hint of counter-moves," and further he added "it sent Gary into a tizzy". Later, one of the designers of Deep Blue admitted that it was a bug that made Deep Blue make a random move.

While it established that a machine could play Chess like a champion, several people expressed whether the same strategy would work for games like Go as the choices at each of the points were horrendously large. Certainly the achievement was laudable and significant, philosophically there were apprehensions whether it really solved the problem of *intelligent chess programs* that had been the goal right from the early days of CS/AI. In this connection, we quote a remark of John McCarthy – a pioneer of Artificial Intelligence.

> *Alexander Kronrod a Russian AI researcher, said* Chess is Drosophilae of AI. *He was making an analogy with geneticists' use of that fruit fly to study inheritance. Playing chess requires certain intellectual mechanisms and not others. Chess programs now play at grand-master level, but they do it with limited intellectual mechanisms compared to those used by a human chess player, substituting large amounts of computation for understanding. Once we understand these mechanisms better, we can build human-level chess programs that do far less computation than do present programs...*

Interestingly, newspaper interview of David Harel, another distinguished scientist, with the title *Why is it easier to beat Kasparov than to beat Turing?*[6] in response to the news "Deep Blue Beats Gary Kasparov" speaks of the inventiveness of Alan Turing!

While the achievements of beating the chess champion by a computer program (or infrastructure) were not the "end goals" themselves, it brought out the power of massive computer infrastructure and the learning/feedback/inference from behavior patterns. Needless to say in this new millennium this has made big impact on science and society.

Moving from Chess, let us look at the next achievement again by IBM that built a "cognitive" system, Watson, that debuted in a televised Jeopardy!, challenged and defeated the show's two greatest champions.

[6] D. Harel, "Why is it easier to beat Kasparov than to beat Turing?" (in Hebrew), in Z. Yannai, ed., *The Infinite Search: Conversations with Scientists*, Am Oved Publishers , Tel Aviv, 2000, pp. 48-56

The challenging goals for Watson were to answer varieties of questions such as puns, synonyms and homonyms, slang, and jargon posed in possible subtle uses of natural language. As it was not to be connected to the Internet for the match, there was a need for it to amass knowledge through years of persistent interaction and learning from a large set of unstructured knowledge. Using machine learning, statistical analysis or natural language processing, it was required to understand the clues in the questions, compare possible answers, by ranking its confidence in their accuracy, and respond – all in about three seconds. Indeed, a challenging feat. The Watson indeed conquered Jeopardy! in 2011.

As highlighted already, conquering Jeopardy! was not the goal. It was the start to initiate cognitive applications that would be welcome in the society. IBM is using the realized technology to build newer generations of Watson so that it can be effectively used in oncology diagnosis by health-care professionals, and in varieties of customer services as a support representative. Currently, it is spread across the cloud with different "avatars" that can serve simultaneously a spectrum of customers across the world accessing it via phones, desktops, or data servers. As the AI improves with the feedback and hence with the usage, one should see Watson becoming smarter; anything it learns in one instance can be immediately transferred to the others. Thus Watson is now an aggregation of diverse software engines – its logic-deduction engine and its language-parsing engine might operate on different code, on different chips, in different locations – all cleverly integrated into a unified stream of intelligence. IBM provides access to Watson's intelligence to partners, helping them develop user-friendly interfaces for subscribing doctors and hospitals. Alan Greene, chief medical officer of Scanadu – a startup that is building a diagnostic device inspired by the *Star Trek* medical tricorder and powered by a cloud AI, says:

> *I believe something like Watson will soon be the world's best diagnostician – whether machine or human... At the rate AI technology is improving, a kid born today will rarely need to see a doctor to get a diagnosis by the time they are an adult.*

One other important research that is pursued at IBM that has been under the broad umbrella of "Cognitive Computing" is the brain inspired computers [Preissl et al., 2012], lead by Dharmendra Modha. The multi-disciplinary, multi-institutional effort lead by Dharamendra Modha, has lead to architectures, technology, and ecosystems that break paths with the prevailing von Neumann architecture and constitutes a foundation for energy-efficient, scalable neuromorphic systems.

The next milestone perhaps has been the mastering of the game GO with Deep Learning technology. Deep Learning techniques allow a computer system to connect the dots from different areas of knowledge akin to how the brain works to arrive at the best possible. The game of GO has long been viewed as the most challenging of classical games for AI, owing to its enormous search space and difficulty of evaluating board positions and moves. Recently, in a full-sized fame of GO, a (human) professional GO player was defeated by a *neural network*; the point to be noted is that the neural network was trained by a novel combination of supervised-learning from human GO experts & reinforcement-learning from ALPHAGO [Silver et al., 2016] self-play games. It is a feat previously estimated to be at least a decade away!

As predicted by Alan Turing, AI has reached to a significant level of language, image, and speech understanding systems (even smell measurements) that have shown enormous applications in the society – truly reflecting actions by a human of a good intellect. This has shown enormous potential for societal applications. To get a view of the status of "Machine Learning" (in a true sense Artificial Intelligence) that has taken deep roots in science and engineering, a brief discussion is given below.

Geoff Hinton [LeCun et al., 2015] highlights the underpinnings of the successes in Natural Language Processing (language translation), Image Classification, etc. In the real world, there is a range of learning tasks starting at a typical statistical analysis or inference to Artificial Intelligence. For instance, typically, statistical analysis is characterized by:

- Low-dimensional data (e.g., less than 100 dimensions).
- Lots of noise in the data.
- There is not much structure in the data, and a fairly simple model can represent what structure there is.
- The main problem in the context is distinguishing *true structure from noise.*

On the other end of the spectrum, the task typically has the following characteristics:

- High-dimensional data (e.g., more than 100 dimensions).
- The noise is not sufficient to obscure the structure in the data if we process it right.
- There is a huge amount of structure in the data, but the structure is too complicated to be represented by a simple model.
- The main problem is figuring out a way to represent the complicated structure so that it can be learned.

With the remarkable capability of these Deep Learning neural networks, one has made remarkable advances in speech and image recognition, natural language translation, driver-less cars, etc. What has really stunned AI experts, has been the magnitude of improvement in image recognition. Google has indeed become a center for Deep Learning and related AI talent.

While the above discussion shows what the pioneers of Computer Science were looking for in the building of Chess playing machines, one of the remarkable inferences that can be drawn from the various successes of computing machines beating human champions are the demonstrations of:

- Excellent engineering and experimentation with a deep knowledge of the domain,
- Capability of building a massive computing infrastructure to realize the goal.

While the prophecies of Alan Turing have come true, extending Deep Learning into applications beyond speech and image recognition will require more conceptual and software breakthroughs, not to mention many more advances in processing power (reflect on the computing power of Google).

While the achievements on speech and image understanding, natural language translation have stunned the scientists and public alike, another exciting area of work has been the interactive learning through computing related to evolution of life, pioneered by Leslie Valiant – another pioneer of Computing. Valiant [Valiant, 2013], proposes the notion of *ecorithms*, which unlike most algorithms, can be run in environments unknown to the designer, and learn by interacting with the environment how to act effectively in it. Thus, after sufficient interaction they will have expertise not provided by the designer, but extracted from the environment. The model of learning they follow, known as the *probably approximately correct model* [Valiant, 2013], provides a quantitative framework in which designers can evaluate the expertise achieved and the cost of achieving it. Valiant argues that these *ecorithms* are not just a feature of computers but imposition of such learning mechanisms, determines the character of life on Earth. The course of evolution is shaped entirely by organisms interacting with and adapting to their environments. This biological inheritance, as well as further learning from the environment after conception and birth, have a determining influence on the course of an individual's life. Thus, such a line of study shall lead to a unified study of the mechanisms of evolution, learning, and intelligence using the methods of Computer Science.

Takeaways

1. There is a need to address the basic conceptual challenges in AI by core researchers.
2. Supporting the use of AI for varieties of societal benefits (hence, could use a PPP model); a recent report on "*The First Report of the 100 Year Study on Artificial Intelligence (AI100)*" has been released very recently [AI100, 2016].
3. Machine Learning and Deep Learning has been offered as a standard package on varieties of systems. In fact, India should push to bring a viable HPC-Deep Learning/Machine Learning for a varieties of applications like cyber security, DNA analysis, translation, service delivery for illiterate/layman, and also science and engineering applications that require from skeletal to deep computing. India is one of the poor investors in technology, research & education even when considered among the G20 countries. A serious push should be made for deriving the benefits through innovations and discoveries from such an investment.

2.2.2 Massively Online Open Courses (MOOCs)

MOOC is the result of the hypothesis that Internet has the potential of becoming the touchstone of education, disruptively changing the face of education. MOOCs offer free, high quality, university course content to anyone with an Internet access. These courses have been drawing tens of thousands of students to a single section. As it requires only a computer and Internet access to enroll, MOOCs can be used for continuing education courses and credit-bearing under-graduate courses, leading to degree programs and even graduate education. Such a technology is indeed attractive from two perspectives:

- Huge scaling up of education at all levels leading to huge economic advantages in particular for developed countries,
- It is naturally an attractive option for countries like India that has a huge population residing in rural areas with an acute shortage of qualified faculty/instructors (this is true even for urban elite centers).

MOOC hit the headlines through an online course on Artificial Intelligence offered from Stanford, instructed by Peter Norvig and Sebastian Thrun, with a worldwide enrollment of 165000. New MOOC centers like Khan Academy, UDACITY (www.udacity.com), Coursera (www.coursera.org) have the success rate of completion at

just 8%. While these "universities without walls" have the potential to transform literacy, awareness of public education, and formal education, there are significant unresolved issues relating to their educational quality and financial sustainability.

Challenges in MOOCs

- Evaluation:
  - Frequency – frequent appraisals are needed to make sure that the students have understood the material presented.
  - Presentations augmented with laboratories, plausibly virtual laboratories:
    * Application of concepts learnt from the lecture presentations in a virtual laboratory environment.
    * An effort in India called *Colama* (www.coriolis.com) has been able to provide virtual laboratories.
    * Raspberry programming systems have become widely used to support practical experiments on theories learned online.
    * One of the interesting experiment has been a course on *Design and Analysis of Cyber Physical Systems* offered at UC Berkeley (https://www.edx.org/course/uc-berkeleyx/uc-berkeleyx-eecs149-1x-cyber-physical-1629#.U_sTUICSxBM). A major characteristic of the course is on the interplay of practical design with formal models of systems, including both software components and physical dynamics. Students applied concepts learned in lectures on programming a robotic controller in a specially-designed virtual laboratory environment with built-in automatic grading and feedback mechanisms.
  - The factors discussed above play a vital role in deciding how students can be given credit and graded.
- A comparison of effectiveness of MOOC in comparison with that of traditional structure:
  - Devise ways to compare performance of students' learning via MOOCs as against those taking traditional courses?
  - While MOOC would serve the paradigm of "Life Long Learning", as it stands the traditional or the universities with traditional teachers shall remain main contributor, at least, for higher education.
  - Noting that faculty-student interaction plays a vital role in traditional learning, it is not clear whether that affects at different levels of MOOC learning.
- In the MOOC world, detection of cheating by students (and thereby their assessments) is quite a challenge.

Status of MOOCs

With no tuition fees required, the convenience of online learning, and access to world-class faculty, MOOCs have the potential to draw vast numbers of students away from traditional bricks-and-mortar universities. The sheer economics of MOOCs attracts a large number of students, and several organizations are investing to build viable systems to cater to the requirements. While current MOOC offerings are targeted to the undergraduate market, there shall be a limited number of professional-, graduate-, and even doctoral-level MOOCs. While even in India, one sees signs of reluctance and disappointment on behalf of students, instructors, and universities, there is a growing feeling of being useful for skill development and training. Certainly, as we proceed, all universities shall use MOOCs in some way or the other – to provide prerequisites or some interdisciplinary training requirements.

A significant migration of students to MOOCs would threaten the viability of some MOOCs and also threaten to change the role of faculty, student, and teaching assistants and the nature of the university. For example, one quality metric for traditional universities is the average number of students per class, with a lower ratio considered desirable. Automated course delivery and grading allows for immense up-scaling of course enrollments. Does the growth of MOOCs mean we will need fewer professors but more teaching assistants? We believe that there may be pressures on traditional universities to scale course sizes by adopting partial MOOC attributes (e.g., more automated grading) but still preserving a high level of instructor-student interaction.

Takeaways

MOOCs have the potential to transform the higher educational landscape, but it is too soon to tell how significant this impact will be. MOOCs will likely play a future role predominately in continuing education, course prerequisites, and, on a limited basis, credit-bearing courses. It is unlikely, but possible, that complete credit-bearing courses from accredited universities will be available through MOOCs before 2022.

2.2.3 Security & Privacy

We never are definitely right
We can only be sure we are wrong
Richard Feynman: Lectures on the character of Physical Law.

Dependence on inter-networked computing systems in this ubiquitous world has been growing in leaps and bounds. The dependence on such systems is true for all entities: be it business, corporation, government, military, infrastructure (communication, energy, health-care, transportation, elections, finance, et al.,) not even the common man is excluded. The most interesting observation is: *none of the systems of such networked systems are indeed trustworthy by themselves.* More than that, all of them are under continuous active and deliberate attack from attackers ranging from a single individual to a nation-state (government). The loss of property, business, or life due to the attacks in cyber space is enormous and ever-growing.

The over dependence of the communities and the society at large makes it mandatory to secure these inter-networked systems and defend them from attacks. A secure system must defend against all possible attacks – including those unknown that could come out in future. As defenders, having limited resources, they develop defenses only for attacks they know about. The result is new kinds of attacks are then likely to succeed. While the costs of securing these IT systems have grown overwhelmingly over the years, the direct/indirect losses, due to attacks have grown in a significant way. Thus, our adopted engineering practices as well as defenses have not succeeded; in fact, they have failed. Thus, the challenge is to provide holistic approach to attack/fraud prevention to realize a safe inter-networked world.

As highlighted in [Schneider and Savage, 2009], the core of the problem of failure is inherent in the nature of security itself. Security is not a commodity like computer and communications hardware and software. It cannot be scaled simply by doing more. ***Security is holistic – a property of a system and not just of its components.*** Even a small change to a system or a threat model can have catastrophic consequences to its security. The familiar and predictable technology curves by which computer processing performance, storage, and communication, scale over the time, cannot be applied to security. Security does not follow such a model. In fact, security is characterized by following asymmetries:

- Defenders are reactive, attackers are proactive.
- Defenders must defend all places at all times, against all possible attacks (including those not known to the defender); and
- attackers need to only find one vulnerability, and they have the luxury of inventing and testing new attacks in private, as well as, selecting the place and time of attack at their convenience.
- New defenses are expensive, new attacks are cheaper.
- Defenders have significant investments in their approaches and business models, while attackers have minimal sunk costs and thus can be quite agile.
- Defense cannot be measured, but attacks can be.
- Since we cannot currently measure how a given security technology or approach reduces risk from attack, there are few strong competitive pressures to improve these technical qualities. So vendors frequently compete on the basis of ancillary factors (e.g., speed, integration, brand development, etc.)
- Attackers can directly measure their return-on-investment and are strongly incentivized to improve their offerings.

Research on cyber security can be summed up by: ***Security never settles on a claim. Every security claim has a lifetime***. This fact provides a basis for setting an agenda for cyber security research. To quote Fred Schneider [Schneider, 2012]:

> *Medicine is an appropriate analogy, since despite enormous strides in medical research, new threats continually emerge and old defenses (e.g., anti-biotic) are seen to lose their effectiveness. As the nation pursues opportunities for sustainability, health-care, and commerce, there will be on-going needs for cyber security research or else the trustworthiness of these systems will erode as threats evolve.*

Takeaways

A broad take-away from this broad perspective is summarized below. Further, cyber security is not purely a technology problem, nor it is purely a policy (economic or regulatory) problem. The basis of security is *building trustworthy systems*, which requires combining technology and policy. In fact, it is for this very reason there is a need of articulating a ***Cyber security Doctrine*** that would specify the goal and means of realizing cyber security in India. A synergy in the understanding of technology, law, economics and investment policies is needed to set up a clear *Cyber security Research* agenda that has appropriate research, development, assessment, measurement, and deployment components.

A Broad Landscape of Cyber security Issues

There has been a wide range of significant contributions by the scientific community across academia, industry, business, government, etc., to realize trustworthiness in terms of varieties of parameters like authentication, access-control, availability, confidentiality, privacy, etc. Towards this, there has been works in areas like: (1) cryptography and PKI (public-key cryptography), (2) analysis of code for vulnerabilities, (3) malware/virus patterns via data mining, machine learning, (4) hardware/software firewalls, intrusion detection systems, etc.

In the following, we shall highlight a few of the general class of problems that need to be addressed to overcome the evolution, maturation and diversification of threats, attacks and fraud strategies to realize a secure cyber space.

1. Static Defense Mechanisms: Most of our approaches are reactive that have severe limitations. Instead, it would be a challenge to transform systems into safely protected systems.

2. Governed by slow and deliberative processes: security patch deployment, testing, episodic penetration exercises, and human-in-the-loop monitoring of security events.

3. Adversaries do greatly benefit from the above situation.

4. Attackers may continuously and systematically probe targeted networks with the confidence that those networks will change slowly if at all.

5. Adversaries have the time to engineer reliable exploits and pre-plan their attacks. And, once an attack succeeds, adversaries persist for long times inside compromised networks and hosts.

6. Hosts, networks, software, and services do not reconfigure, adapt, or regenerate except in deterministic ways to support maintenance and uptime requirements:

   (a) Malware Trends: Infection mechanisms (malware) are on the rise either due to the vulnerabilities in the environment or due to creation of new infection mechanisms. It is quite evident that malware is still the most dangerous threat to enterprises, governments, defense, financial institutions, and the end-users. While catastrophes caused by it have lead to better preventive technologies, cyber theft has stayed ahead of these technologies due to the un-decidability of the general problem of prevention, by discovering new loopholes in the underlying hardware/software systems, and arriving at new mechanisms to evade the existing detection methods. This becomes clear if we look at the general trend of malware in 2014 (http://www.slideshare.net/ibmsecurity/the-top-most-dangerous-malware-trends-for-2014):

   (b) The source code for a crime kit, CarberpTrojan (widely used by the underworld) became an open, leading platform to develop similar crop of new Trojans and crime-ware kits. The new invariants would have characteristics that can be quite new and makes it very difficult to be detected by the prevalent virus detectors. In other words, malware is being commoditized.

   (c) Mobile SMS forwarding malware are becoming prevalent. Thus, SMS – the basic 2FA authentication – that is widely used in financial sector gets completely compromised.

   (d) Malware attacks the victim's device itself rather than remote devices.

   (e) Evasion of malware analysis developed by researchers.

   (f) Security of infrastructures: Due to technological advances, it has been a common practice for quite some time to use embedded computers for monitoring and control of physical processes/plants. These are essentially networked, computer-based systems consisting of application-specific control-processing systems, actuators, sensors, etc., that are used to digitally control physical systems (often in a federated manner) within a defined geographical location such as power plants, chemical plants, etc. Different terminologies like distributed control systems (DCS), cyber physical systems (CPS), supervisory control and data acquisition systems (SCADA), etc., are used for denoting similar usages.

SCADA have evolved from special purpose closed system of the early era to a network of components-off-the-shelf systems consisting of computers and communication components using TCP/IP. While it has greatly enhanced the flexibility and usability, it has also exposed itself at several vulnerable points.

7. Technology has further made it possible to federate/integrate heterogeneous (built by different manufacturers) systems. While such capabilities have provided the needed flexibility and usability, it has also created challenges for system designers/integrator, not only from the correctness point of view but also from the point of view of security and protection of the underlying physical plants. With the arrival of complex malware (APT - advanced persistent threat), it has become very challenging to secure network and information systems from intruders and protect the systems from attackers. Recently, complex malware like Stuxnet, Flame etc., have specifically targeted SCADA of public infrastructures like power grids/plants, and thus, bringing to the forefront the challenges in securing and protecting SCADA. The above mentioned malware are horrendously complex and hence, need a wholesome approach for detection and protection.

   (a) Internet of Things (IoT): IoT has emerged as a global Internet-based technical architecture that has deeply facilitated the exchange of goods and services in global supply chain networks. If one uses the broad definition of IoT, it encompasses home automation, industrial SCADA, connected vehicles, smart meters, implantable medical devices, etc., and in a sense, the backbone of **smart cities**. This is quite a large coverage [Mirai DDoS attack that used IoT devices to produce DDoS traffic of 620 GBps!] and hence, security and privacy assurance in IoT is quite challenging given that it covers not only domains of IT but also other specific application domains.

   (b) Privacy issues: Rapid advances in digital technologies and communication have lead to modern systems such as varieties of social media networks like Facebook, Twitter, etc., mobile computing platforms, and wearable devices (Google Glass, Oculus Rift), which in turn have brought new benefits to almost all aspects of our lives. For example, personalized content or service recommendations like Netflix, AdSense, NewsFeed that are dependent on collection of users' data (inferred preferences) through various direct/indirect channels, often with users' consent. Users get relevant content during their search, relevant match of service while searching on the web. It saves users' time and money. Such an immensely attractive benefit is also plagued by both conventional and emerging threats to security and privacy. In the context of web, for example, a large amount of personal information about individuals that is being collected, used, and shared across organizations, is a threat to privacy thus undermining trust, with potential to surveillance by foreign players – at times influencing democratic elections. This is a serious issue in regulated sectors like health, finance, insurance, etc. In these cases, the organizations need to assure the compliance of privacy even when the data traverses from one social/web media to another one that may have distinct privacy policies.

   (c) Usable Privacy: The de facto standard to address expectations of "notice and choice" on the Web is natural language. The users usually agree to the policies even before reading the policies as these are neither easy to understand nor the user finds it relevant. Initiatives to overcome this problem with machine-readable privacy policies or other solutions that require website operators to adhere to more stringent requirements have run into obstacles, with website operators showing reluctance to commit to anything more than what they currently do. One of the challenges is to combine machine learning, natural language processing and crowd-sourcing to semi-automatically annotate privacy policies in order to provide users with succinct privacy notices like the one used for energy efficiency of electric appliances – 5-star rating.

Broad Challenges for Cyber security

To realize a firm cyber security is to provide a holistic approach to fraud prevention. This requires disruptive approaches to handle infections and cyber crime. In the following, we briefly outline some general approaches to address the issues discussed in the previous section on a similar structure:

a) **Dynamic Cyber Defense:** The basic approach is to move to proactive defense. Two of the widely used strategies are "moving targets" or "cyber kill chain" (cf. [Okhravi et al., 2013] for details). In the former, the idea is to protect various entities like applications, OS, machine, network, session, traffic or data through various techniques including coding. In the latter that is cyber kill chain, various phases like reconnaissance, access, attack launch or persistence are identified and moved. It must be noted that the above mentioned strategies have several limitations [Okhravi et al., 2013] that need to be addressed effectively. One example, covert channel's prevention needs dynamic strategies as such channels are almost unbounded. The recent story of the cloud hosting giant Akamai Technologies that dumped journalist Brian

Krebs from its servers after his website came under a "record" cyber attack [Mirai botnet; DDoS attack] is an eye opener (http://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/)

b) **Guaranteed Leak-Free Information:** Flow in Multi Level Security (MLS) Systems: In MLS, there is a need for correct integration of Mandatory Access Control (MAC) and Discretionary Access Control (DAC) and assure the flow of information as per the hierarchy of trusted and untrusted objects/subjects. Typical systems wherein this is important are that of operating systems (which is vital for almost any application) and cloud service brokers. In other words, the challenge is to build systems wherein a trusted party interact with an untrusted party without getting infected.

c) **Malware, Fraud, and Crime Detection:** Big Data analytics has emerged as a key player in the security arena with several applications in areas like homeland security and cyber security. Big Data applications are being deployed to identify the most critical and actionable items of intelligence in near real-time. It is now considered a crucial element in detecting and deterring emerging threats. Big Data analytics in this field includes proactive data mining, data fusion, and predictive analytics techniques that are applied to all available data to gain useful insights.

d) **Secure Infrastructures:** In these scenarios, apart from the classical IT security, there is a need to look at other plausible new attacks considering the domain of the physical systems in conjunction with the capabilities of the embedded computers, and arrive at methods of protection and risk evaluation.

e) **Security of IoT:** IoT architectures resilient to attacks, data authentication, access control and client privacy are the need of day.

f) **Privacy:** In the two reports published in November 2014 [Alkhatib et al., 2014, Mandiant, 2014], analysts estimated that the IoT (Internet of Things) will represent 30 billion connected *things* by 2020, growing from 9.9 billion in 2013. These connected *things* are largely driven by intelligent systems (including organizations incorporating BYOD policy – Bring Your Own Device) – all collecting and transmitting data. This connectivity is changing the way we live and creating new questions about personal privacy, marketing and Internet security, as the *things* are manufactured and sold to consumers. A couple of challenges are:

   (a) To have controlled privacy over web and social media there is a need to arrive at distinct privacy policies whose compliances can be verified either statically or dynamically.

   (b) With the growth of Big Data and Analytics, there is a need to arrive at a tradeoff between security and privacy among varieties of stake-holders that include people, businesses, government, and malevolent actors so that each of the groups decide about releasing certain information to government, merchants, and even other citizens and to consider the consequences of every activity in which they engage.

g) **SNS as tool for Information Weaponization:** Social Network System, have shown potential to rapidly disseminate unverified news information by nodes in the network. This has serious potential to swing the public opinion in either directions.

h) **Cyber security Doctrine:** Succession of doctrines advocated in the past for enhancing cyber security like prevention, risk management, and deterrence through accountability have not proved effective. There is a need to learn from failed doctrines and study the possibility of viewing cyber security as a public good similar to that of public health and see the viability to adopt mechanisms inspired by those used for public health.

As mentioned before, the areas of computing or theory to practice is too vast to be covered in any one report. In the following chapters, the authors discuss areas like HPC and its cutting-edge applications, blockchain as a distributed trust management system, Big Data Analytics and its impact, ICT Infrastructures and its applications, Network Computing (SDN) and its importance, future potential, etc. The report also provides a glimpse into the challenges and suggestions (key takeaways) for innovative applications in science and society.

Bibliography

[AI100, 2016] AI100, S. (Sep. 2016). One Hundred Year Study on Artificial Intelligence (AI100). Technical report, Stanford University.

[Alkhatib et al., 2014] Alkhatib, H., Faraboschi, P., Frachtenberg, E., Kasahara, H., Lange, D., Laplante, P., Merchant, A., Milojicic, D., and Schwan, K. (Dec. 2014). IEEE CS 2022 Report. Technical report, IEEE Computer Society.

[Denning, 2003] Denning, P. J. (2003). Great Principles of Computing. *Communications of the ACM*, 46(11):15–20.

[Harel, 2016] Harel, D. (2016). Niépce-Bell or Turing: How to Test Odor Reproduction? *CoRR*, abs/1603.08666.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521:436–444.

[Mandiant, 2014] Mandiant (2014). M-Trends Threat Report.

[Okhravi et al., 2013] Okhravi, H., Rabe, M. A., Mayberry, T. J., Leonard, W. G., Hobson, T. R., Bigelow, D., and Streilein, W. W. (Sep. 2013). Survey of Cyber Moving Target Techniques. `http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA591804`.

[Preissl et al., 2012] Preissl, R., Wong, T. M., Datta, P., Flickner, M., Singh, R., Esser, S. K., Risk, W. P., Simon, H. D., and Modha, D. S. (2012). Compass: A Scalable Simulator for an Architecture for Cognitive Computing. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 54:1–54:11, Los Alamitos, CA, USA. IEEE Computer Society Press.

[R. Haddad and Sobel, 2008] R. Haddad, H. Lapid, D. H. and Sobel, N. (2008). Measuring Smells. *Current Opinion in Neurobiology*, 18:438–444.

[Schneider and Savage, 2009] Schneider, F. and Savage, S. (Feb. 2009). Security is not a Commodity: The Road Forward for Cybersecurity Research, Computing Research Initiatives for the 21^{st} Century. `http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Cybersecurity.pdf`.

[Schneider, 2012] Schneider, F. B. (2012). Blueprint for a science of cybersecurity.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–503.

[Tedre, 2014] Tedre, M. (2014). *The Science of Computing: Shaping a Discipline*. Chapman & Hall/CRC.

[Tichenor, 2007] Tichenor, S. (June 7, 2007). Out-Compute to Out-Compete: Driving Competitiveness with Computational Modeling and Simulation.

[Turing, 1937] Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX:433–460.

[Valiant, 2013] Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books, Inc., New York, USA.

Chapter 3

Big-Data Science: Infrastructure Impact

Inder Monga and Prabhat
Berkeley and NERSC

3.1 Introduction

The nature of science is changing dramatically, from single researcher at a lab or university laboratory working with graduate students to a distributed multi-researcher consortiums, across universities and research labs, tackling large scientific problems. In addition, experimentalists and theorists are collaborating with each other by designing experiments to prove the proposed theories. 'Big Data' being produced by these large experiments have to verified against simulations run on High Performance Computing (HPC) resources.

The trends above are pointing towards

a. Geographically dispersed experiments (and associated communities) that require data being moved across multiple sites. Appropriate mechanisms and tools need to be employed to move, store and archive datasets from such experiments.

b. Convergence of simulation (requiring High Performance Computing) and Big Data Analytics (requiring advanced on-site data management techniques) into a small number of High Performance Computing centers. Such centers are key for consolidating software and hardware infrastructure efforts, and achieving broad impact across numerous scientific domains.

The trends indicate that for modern science and scientific discovery, infrastructure support for handling both large scientific data as well as high-performance computing is extremely important. In addition, given the distributed nature of research and big-team science, it is important to build infrastructure, both hardware and software, that enables sharing across institutions, researchers, students, industry and academia. This is the only way that a nation can maximize the research capabilities of its citizens while maximizing the use of its investments in computer, storage, network and experimental infrastructure.

This chapter introduces infrastructure requirements of High-Performance Computing and Networking with examples drawn from NERSC and ESnet, two large Department of Energy facilities at Lawrence Berkeley National Laboratory, CA, USA, that exemplify some of the qualities needed for future Research & Education infrastructure.

Most scalable deep-learning implementation

National Energy Research Scientific Computing Center (NERSC) reported in their communication dated 28 August 2017, that a collaborative effort between Intel, NERSC and Stanford has delivered the first 15-petaflops deep learning software running on HPC platforms and is, according to the authors of the paper (and to the best of their knowledge), currently the most scalable deep-learning implementation in the world. The work described in the paper, Deep Learning at 15PF: Supervised and Semi-Supervised Classification for Scientific Data (https://arxiv.org/abs/1708.05256), reported that a Cray XC40 system with a configuration of 9,600 self-hosted 1.4GHz Intel Xeon Phi Processor 7250 based nodes achieved a peak rate between 11.73 and 15.07 petaflops (single-precision) and an average sustained performance of 11.41 to 13.47 petaflops when training on physics and climate based data sets using Lawrence Berkeley National Laboratory's (Berkeley Lab) NERSC

(National Energy Research Scientific Computing Center) Cori Phase-II supercomputer. The group utilized an amalgamation of Intel Caffe, Intel Math Kernel Library (Intel MKL), and Intel Machine Learning Scaling Library (https://github.com/01org/MLSL) (Intel MLSL) software to achieve this scalability and performance.

3.2 High-Performance Computing

As one of the world's premier supercomputing centers, NERSC supports perhaps the largest and most diverse research community of any high-performance computing facility, providing large-scale, state-of-the-art computing for DOE's unclassified research programs. More than 6,000 scientists worldwide use NERSC to conduct basic and applied research in energy production and conservation, climate change, environmental science, materials research, chemistry, fusion energy, astrophysics and other areas related to the mission of the DOE Office of Science. Figure 3.1, provides a brief overview of the hardware resources at NERSC. Two major supercomputing

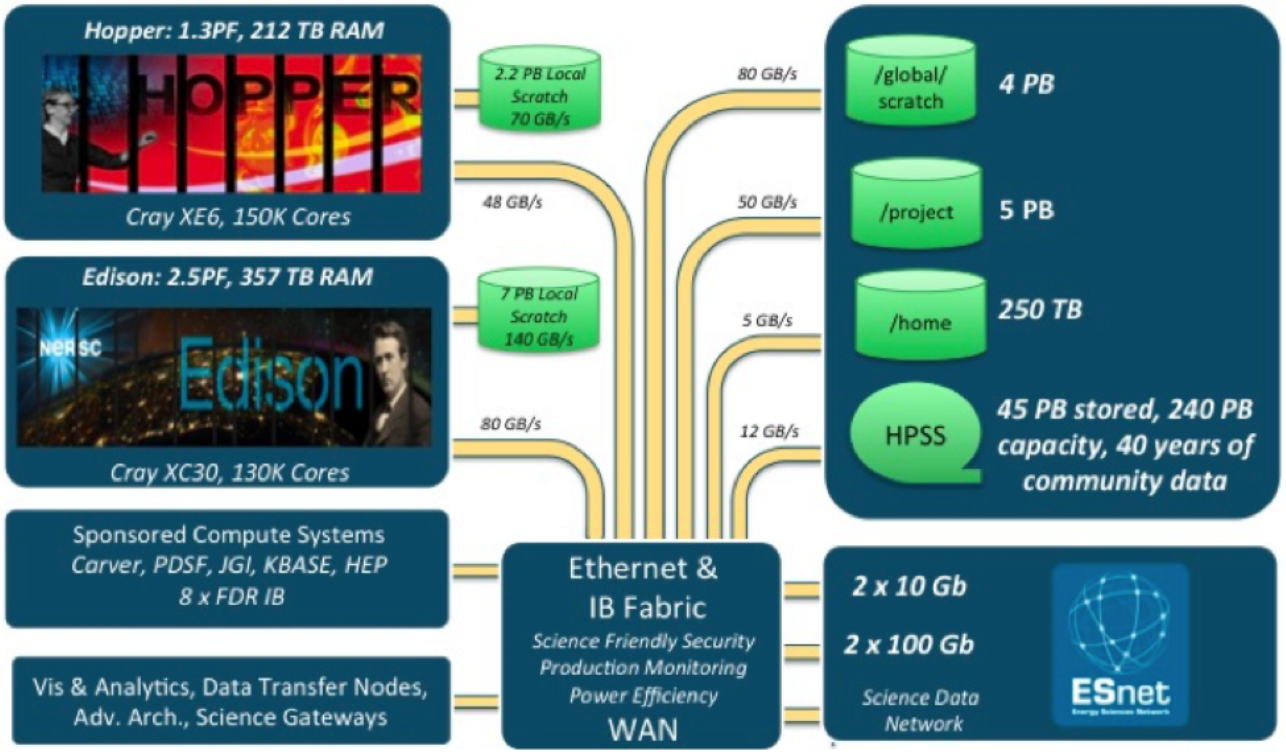


Figure 3.1: Overview of NERSC systems (circa October 2015)

platforms are operational at any point in time; currently we host two petaflop class systems: Cray XE6 system (Hopper) and Cray XC30 (Edison). These tightly coupled systems feature a high performance interconnect, and a fast, distributed parallel filesystem. Relatively higher capacity, but lower bandwidth project and archival systems are available to users for longer term retention of data. NERSC has experimented with installing dedicated, smaller-scale clusters for handling data-intensive workloads of specific domain science communities. Finally, in order to move data efficiently between supercomputing centers, dedicated data transfer nodes provide an end-point for both 10G and 100G ESnet connections.

In 2016, NERSC will install Cori, a Cray XC40 system. This system will provide unified resources for handling both HPC, as well as data-centric workloads. In the HPC space, NERSC is making a major push to port applications to energy-efficient, many-core architectures with the NESAP (NERSC Exascale Science Application Program); teaming up talented post-docs with strategic applications. In the Big Data space, NERSC is innovating on a number of fronts: we are utilizing the Datawarp technology to provide users with access to extremely high bandwidth and low-latency NVRAM storage; we are configuring our batch system to provide real-time, interactive, serial and high-throughput queues; we are enabling compute nodes on the system to have external connectivity, and we are enabling custom user-environments through Docker-like containers.

In terms of HPC software, NERSC provides a broad portfolio of compilers, code development tools, domain-specific application codes, programming libraries, performance and debugging tools. In the Big Data space, NERSC provides software capabilities in the areas of Data analytics, Data Management, Work-flows, Data Transfer, Data Access and Visualization. Documentation on all of these capabilities are provided at

www.nersc.gov and regular outreach and training events are conducted to keep the scientific community abreast of latest technologies.

NERSC gathers HPC, data and services requirements from the science community in many ways. Chief among them are the program requirements reviews held with each of the six offices within the DOE Office of Science. This ongoing series of reviews brings together DOE program managers, leading domain scientists and NERSC staff to derive each scientific community's future HPC needs. The results of the reviews include requirements for computing, storage and services five years out. Each review report also contains a number of significant observations, topics of keen interest to the review participants. These results help DOE and NERSC plan for future systems and HPC service offerings.

NERSC is much more than just a collection of computers, servers, routers and software tools. One of its most valuable attributes is its staff, a talented group of computer scientists, mathematicians, engineers and support personnel. More than 50 percent of NERSC staff hold advanced degrees in a scientific or technical field. And collaboration–aka "team science," a concept pioneered by Berkeley Lab founder Ernest O. Lawrence in 1931–is a cornerstone of NERSC's philosophy, both internally and through its engagements with the broader science community.

3.2.1 Impact of Computing on Science

Theory, Experiment, Simulation and Data-Driven Discovery are now widely accepted as the four paradigms of modern science. Simulation and High Performance Computing go hand-in-hand; all natural or man-made systems require higher fidelity, either in terms of the spatial/temporal resolutions, or in terms of the physical processes being modeled. HPC has had a broad impact across a number of domains, as highlighted by the following brief examples:

- NOAA routinely use HPC resources for making regional weather forecasts over the US and UK respectively
- Major aircraft manufacturers (Boeing, Airbus) use HPC to create, and simulate digital models of planes before fabrication
- NASA utilizes HPC to explore space shuttle and spacecraft design for both robotic and manned missions
- DOE utilizes HPC to simulate next generation Tokamak reactors for exploring the promise of fusion energy
- NSF utilizes HPC to conduct simulations of earthquakes along various fault lines in California, and impact on local economies
- Several firms on the the Wall Street utilize HPC resources to enable high frequency trading
- Intelligence agencies utilize HPC resources to find patterns and anomalies in unstructured data

Big Data has its origins in the commercial world. Internet-driven companies such as Google, Facebook and Twitter need to be able to analyze massive amounts of user data, and find mechanism to add value to their user's online experience, as well as monetize user behavior to generate a revenue stream. Major investments have been made by these firms in data-centers throughout the globe, and an associated software stack.

In the remainder of this article, we will focus on success stories from NERSC, that highlight the kind of progress that can be achieved in basic sciences through investments in HPC and Big Data resources.

3.2.2 Scientific Success Stories

Over the past 40 years of NERSC's history, we have witnessed the evolution of High Performance Computing from Terascale to Petascale and now en route to Exascale class systems. HPC has been successfully applied to simulate the evolution of the universe, model supernova explosions, model climate change, simulate carbon sequestration, perform quantum mechanical simulations of various materials and simulate experiments such as the Large Hadron Collider on its search for sub-atomic constituents.

Big Data Analytics is a relatively recent trend at NERSC, having gained prominence over the last 5 years. Major projects include high throughput pipelines for genome assembly, automated candidate identification in astronomy images, 3D reconstruction of light source data, interactive exploration of high energy physics experiments and so on. We also observe the trend of the integration of observational data with simulations: a classic example is the production of climate 'reanalysis' datasets, which use a climate model to interpolate satellite and weather station datasets.

Annually, NERSC users produce over 1900 publications in top-tier scientific venues such as *Nature, Science, PNAS,* etc. NERSC also has a rich history of contributions to a number of Nobel Prizes:

- 2015 Nobel Prize in Physics on discovery of neutrino oscillations

- 2013 Nobel Prize in Chemistry on development of multi-scale models for complex chemical systems
- 2011 Nobel Prize in Physics on measuring the acceleration of cosmic expansion
- 2007 Nobel Peace Prize on characterization of climate change
- 2006 Nobel Prize in Physics on Cosmic Microwave Background Radiation

In the next two subsections, we briefly comment on science stories which show the successful application of HPC, as well as Big Data Analytics methods to further scientific discovery.

Characterizing Extreme Weather in a Changing Climate

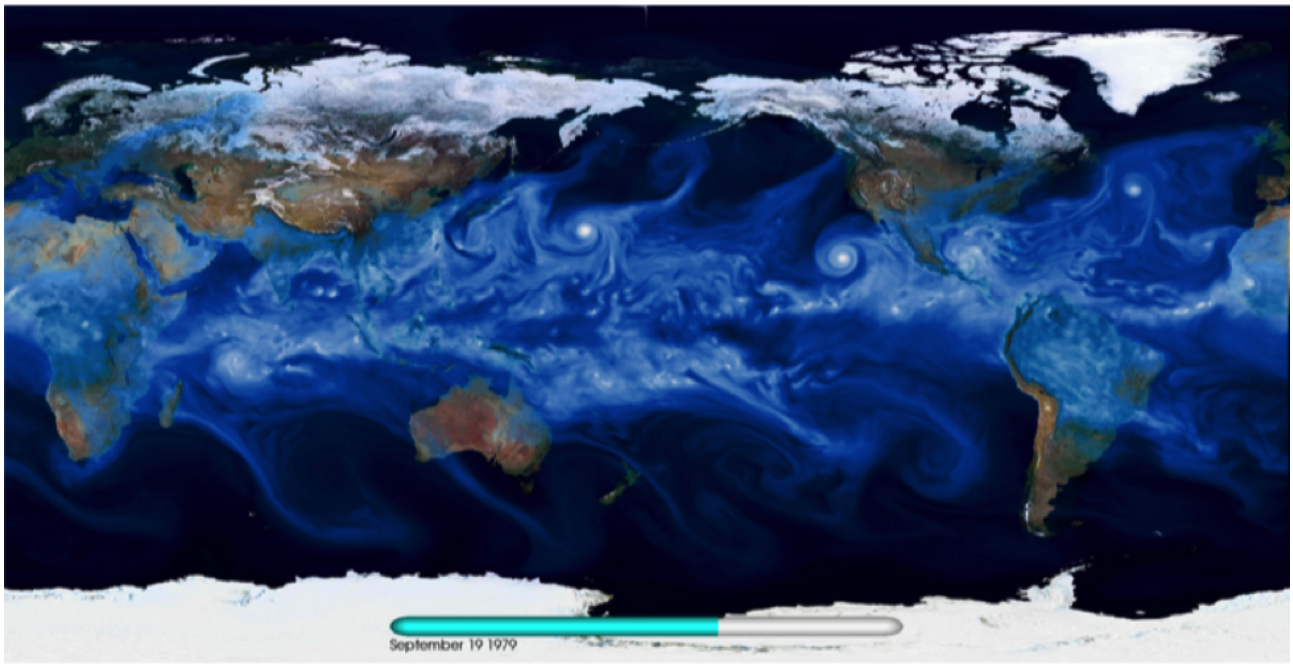


Figure 3.2: Snapshot of CAM5.1 25-km global climate simulation

Not long ago, it would have taken several years to run a high-resolution simulation on a global climate model. But using supercomputing resources at NERSC, in 2014 Berkeley Lab climate scientist Michael Wehner was able to complete a run in just three months. What he found was that not only were the simulations much closer to actual observations, but the high-resolution models were far better at reproducing intense storms, such as hurricanes and cyclones. The study was published in the Journal for Advances in Modeling the Earth System.

"I've been calling this a golden age for high-resolution climate modeling because these supercomputers are enabling us to do gee-whiz science in a way we haven't been able to do before," said Wehner, who was also a lead author for the recent Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). "These kinds of calculations have gone from basically intractable to heroic to now doable."

Using version 5.1 of the Community Atmospheric Model, developed by the DOE and the National Science Foundation for use by the scientific community, Wehner and his co-authors conducted an analysis for the period 1979 to 2005 at three spatial resolutions: 25 km, 100 km and 200 km. They then compared those results to each other and to observations. One simulation generated 100 terabytes of data. Wehner ran the simulations on NERSC's Hopper supercomputer; the CAM code was optimized for parallel execution and scaling on the Cray system, special emphasis was also laid on the parallel I/O strategy for the code.

"I've literally waited my entire career to be able to do these simulations," Wehner said. The higher resolution was particularly helpful in mountainous areas since the models take an average of the altitude in the grid (25 square km for high resolution, 200 square km for low resolution). With more accurate representation of mountainous terrain, the higher resolution model is better able to simulate snow and rain in those regions.

"High resolution gives us the ability to look at intense weather like hurricanes," said Kevin Reed, a researcher at the National Center for Atmospheric Research and a co-author on the paper. "It also gives us the ability to look at things locally at much higher fidelity. Simulations are much more realistic at any given place, especially if that place has a lot of topography."

The high-resolution model produced stronger storms and more of them, which was closer to the actual observations for most seasons. "In the low-resolution models, hurricanes were far too infrequent," Wehner said.

The IPCC chapter on long-term climate change projections concluded that a warming world will cause some areas to be drier and others to see more rainfall, snow and storms. Extremely heavy precipitation was projected to become even more extreme in a warmer world. "I have no doubt that is true," Wehner said. "However, knowing it will increase is one thing, but having confidence about how much and where as a function of location requires the models do a better job of replicating observations than they have."

Wehner says the high-resolution models will help scientists to better understand how climate change will affect extreme storms. His next project is to run the model for a future-case scenario. Further down the line, Wehner believes scientists will be running climate models with 1 km resolution. To do that, they will have to have a better understanding of how clouds behave.

"A cloud system-resolved model can reduce one of the greatest uncertainties in climate models, by improving the way we treat clouds," Wehner said. "That will be a paradigm shift in climate modeling. We're at a shift now, but that is the next one coming."

Figure 3.3: Big Data Analytics on CMIP-5 data facilitated by ESnet and leadership computing resources at NERSC and ALCF

In a related exercise, Michael Wehner teamed up with Prabhat (NERSC), Suren Byna (CRD) and Venkat Vishwanath (ALCF) to process the massive CMIP-5 archive at scale on Mira, ALCF's flagship BG/Q system. The team downloaded over 60 TB of climate data from a world-wide repository using the Earth System Grid Federation, pre-processed the data on NERSC's Hopper system, and then transferred 6 TB of data over ESnet to ALCF in 2 days. Prabhat then developed and scaled the TECA (Toolkit for Extreme Climate Analytics) framework, to run on 750,000 cores on ALCF's Mira system. The entire CMIP-5 archive was processed in 1 hour and produced a summary of the expected change in extra-tropical cyclones in future climate change scenarios. It is estimated that a similar task on standalone workstations would take over a decade. This is one of DOE's leading examples of what Scientific Big Data Analytics can accomplish on HPC resources.

Changes in the seasonality of Indian monsoon, availability of fresh water supply through snowpacks in Himalayas and rising sea levels are some examples of the regional impact of global climate change. Characterization of climate change, and adaptation to a changing weather will be key issues for the Indian economy in the 21^{st} century. High resolution simulations through HPC resources, and the application of Big Data Analytics methods on the resulting massive datasets will be key for the scientific community to inform policymakers.

The Life and Death of Stars and the Evolution of the Universe

In recent years, astronomy and cosmology have been transformed from data-starved endeavors into data-intensive sciences. Three key factors have propelled this revolution. First is exponential growth in detector resolution, sensitivity, scale, and reliability. Second is the proliferation of remote, semi-robotic, and fully-robotic telescope operations enabled by powerful networks and intelligent machine scheduling. Third is the fusion of HPC, large-scale databases, and parallel file systems for real-time data analysis and large-scale simulation of astrophysical phenomena.

Today's high-impact astronomical surveys routinely generate >100 GB of digital sky images per night, transfer those images to HPC centers and use automated software pipelines to process the data. These data are then used to generate catalogs of hundreds of millions of objects, and identify time-varying phenomena minutes or hours after they have been observed. Scientists uses these products to identify phenomena for more intensive study using more specialized instruments on the largest telescopes in the world.

The future landscape of astronomy is dominated by the Large Synoptic Survey Telescope (LSST), a facility being constructed in Chile that will generate upwards of 100 PB of data over its lifetime starting in 2022. One of the major surveys in operation today that is paving the way to LSST is the Intermediate Palomar Transient Factory (iPTF, PI Shrinivas Kulkarni, Caltech). The iPTF is an example of how leveraging high-performance networking and computing resources like ESnet and NERSC opens vast new territory in our understanding of the Universe, in this case in the physics of stellar death (supernovae). Understanding supernovae is important because they test our theories of the behavior of matter under extreme conditions, they create and disperse chemical elements heavier than helium, and are useful as tools for measuring distances to study the fundamental physics of Dark Energy. But making progress in this space requires maximizing both data velocity and volume, as iPTF has successfully demonstrated time and again.

Figure 3.4: A star in a distant galaxy explodes as a supernova: While observing a galaxy known as UGC 9379 (left; image from the Sloan Digital Sky Survey; SDSS) located about 360 million light-years away from Earth, the team discovered a new source of bright blue light (right, marked with an arrow; image from the 60 inch robotic telescope at Palomar Observatory). This very hot, young supernova marked the explosive death of a massive star in that distant galaxy. Images: Avishay Gal-Yam, et al., Weizmann Institute of Science.

For example, iPTF scientists were the first to demonstrate that Type IIb supernovae arise from a kind of massive star called a Wolf-Rayet star. This was the first direct confirmation of the theory, even though the Type IIb supernova phenomenon was first identified some two decades ago. Researchers at Israel's Weizmann Institute of Science were able to identify supernova SN 2013cu within hours of its explosion using the iPTF pipeline running at NERSC. Mere hours after photons from the cataclysmic explosion reached Earth, iPTF was able to trigger telescopes both on the ground and in space to follow the evolution of the supernova more intensively at all wavelengths. These follow-up observations enabled iPTF scientists to determine what elements were present on the surface of the star and in its immediate environment prior to explosion (the findings appeared in the May 22, 2014 edition of *Nature*). The ability to make such discoveries depends on time-critical processing of large volumes of data with HPC, and the ability to identify patterns in the data for scientists to exploit to make new discoveries.

HPC resources not only help scientists transform massive amounts of raw astronomical image data into new knowledge about the life-cycle of stars, it can result in Nobel Prize worth science. In the 1990's the Supernova Cosmology Project (SCP, PI Saul Perlmutter, LBL) used observations of distant supernovae to map out the expansion history of the Universe. Instead of finding that the expansion of the Universe was slowing down, they found that it was speeding up. To eliminate potential sources of systematic error, simulations of the SCP supernova survey were undertaken at NERSC and confirmed the result. Ultimately this discovery led to a Nobel Prize for Saul Perlmutter. This combination of computational science and cosmology led to other projects and established LBL and NERSC as key players in the emerging field of observational cosmology.

There is only one sky, but astronomers are looking deeper into the Universe, opening up new regimes of the electromagnetic spectrum, and examining changes on the timescales of minutes and seconds. "All the low-hanging fruit has been picked by the previous generation of astronomers," notes Berkeley Lab data scientist

and astrophysicist Rollin Thomas, "HPC and Big Data are new and essential ladders that lift us up to reach the highest branches."

3.2.3 Key Takeaways

1. Setting up a national resource in HPC and Big Data will require major, sustained investments in hardware. It is recommended that India not embark on the race for flops, but rather focus on well-balanced systems that emphasize compute, memory, storage and networking.

2. In conjunction with investments in hardware, special consideration should be given to system software and applications. *Productivity* of the scientific user community is key, hence investing in purchasing and developing software, and more generally being in sync with the broader open source community is highly recommended.

3. Finally, in our experience, the quality of operational and research staff at such centers is fundamental to the eventual success of such initiatives. Staff needs to be highly qualified, motivated and collaborative; they also need to be compensated appropriately.

3.3 Democratization of Data

Modern science is inherently collaborative, and collaborations produce ever more data. Many large-scale instruments being planned and built that will serve tens of thousands of scientists. These facilities will create petabyte-scale data sets to be analyzed and archived, in many cases using distant computational resources. Though it might seem logical and efficient to house these centers close to their data repositories and computational facilities, this is not always the likely scenario. Distributed solutions – in which components are scattered geographically – are much more common at this scale, for a variety of reasons; the largest collaborations will likely depend on distributed architectures.

The LHC, the most well-known high-energy physics collaboration, was a driving force in the development and adoption of such advanced network services. Early on, the LHC community understood the challenges the experiment would present in terms of data generation, distribution, and analysis. In response, the community pioneered a tiered data-distribution model that enables tens of thousands of physicists around the world to access and analyze experimental data. This model is now changing to be more of an 'on-demand' model, where data is moved to the computation, wherever resources are available, and the high-speed networking capabilities are leveraged.

Not just Physics, but many research disciplines are facing the same challenge and marching towards similar solutions. The cost of genomic sequencing is falling dramatically, for example, and consequently, the volume of data produced by sequencers is rising exponentially. In climate science, researchers must analyze observational and simulation data sets located at facilities around the world. Climate data is projected to top 200 petabytes by 2020. The need for productive access to such data led to the development of the Earth System Grid (ESG), a global work-flow infrastructure giving climate scientists access to data sets housed at modeling centers on multiple continents, including North America, Europe, Asia, and Australia.

New detectors being deployed at X-ray synchrotrons generate data at unprecedented resolution and refresh rates. The current generation of instruments can produce 300 or more megabytes per second and the next generation will produce data volumes many times higher; in some cases, data rates will exceed DRAM bandwidth, and data will be preprocessed in real time with dedicated silicon. Large-scale, data-intensive science projects on the drawing board include the International Thermonuclear Experimental Reactor (ITER, the international fusion energy prototype) and the Square Kilometer Array (a massive radio telescope that will generate as much or more data than the LHC).

3.3.1 Role of Networking in big-data science

The structure of large-scale science now assumes the availability of high-bandwidth, reliable, feature-rich networks that can interconnect globally-distributed instruments, facilities and collaborators. The Large Hadron Collider at CERN may have been the first experiment for which reliable global networking was a design premise, but it certainly will not be the last.

Within the research and education (R&E) community, instruments, facilities and collaborators are normally served by administratively separate networks. Each of these networks has its own policies, funding models, and technical capabilities. As a result, R&E networking is inherently a multi-domain endeavor. Members of this

[0]This section is based on the network research and operational knowledge in ESnet, Energy Sciences Network especially the science requirement reviews, Science DMZ and wide-area data movement

community spend much of their time coordinating and communicating, in an effort to assure that the global R&E network ecosystem functions optimally from end-to-end.

While this hierarchical, multi-domain, multi-scale model links research facilities no matter where they are located, it is far from seamless. To help address its challenges, various collaborations of R&E networks – including ESnet, Internet2, Regional Optical Networks (RONs), GÉANT, the NRENs of Europe, and other networks from the Americas and Asia – have worked together for decades to develop and standardize technologies and services to assure high performance for scientific data flows from end-to-end.

This infrastructure is only helpful if it can be used effectively by moving data between resources that generate, store and process data. Even if it is within the same supercomputer center, sometimes data movement and data sharing is impeded by architecture and design patterns that are not thought through end-to-end. **The democratization of data i.e. data 'for' all, and shared 'by' all, will be extremely critical in ensuring scientific progress of India.**

3.3.2 Key issues with data sharing over the network

One of the challenges of building shared information systems for scientific data is that funding models make them actually extensions of existing projects, often going back decades, which have embedded logic and work practices that are highly resistant to change. In this section, we do not talk about the social causes of data hoarding or business drivers in resisting this, but the infrastructural impediments which prevent users from easily sharing their data.

1. **Inadequate infrastructure at the campus or the data-hosting facility**
   Local area networks are usually general-purpose networks that support multiple missions, the first of which is to support the organization's business operations including email, procurement systems, web browsing, and so forth. Second, these general networks must also be built with security that protects financial and personnel data. Meanwhile, these networks are also used for research as scientists depend on this infrastructure to share, store, and analyze data from many different sources. As scientists attempt to run their applications over these general-purpose networks, the result is often poor performance, and with the increase of data set complexity and size, scientists often wait hours, days, or weeks for their data to arrive, often times they give up getting data over the network.

2. **End hosts not optimized for wide-area data transfers**
   Systems used for wide area science data transfers perform far better if they are purpose-built for and dedicated to this function. When the systems are not designed for data transfer, typically there is a mismatch between the network interface speeds of the end-system, say 10 Gbps, and the capability of the wide-area network, say 1 Gbps. This mismatch overwhelms the WAN connection, and causes packet loss and performance issues for the entire site.

3. **Wide-Area networks are not architected for 'zero packet loss' regardless of their bandwidth capabilities**
   The Transmission Control Protocol (TCP) of the TCP/IP protocol suite is the primary transport protocol used for the reliable transfer of data between applications. TCP is used for email, web browsing, and similar applications. Most science applications are also built on TCP, so it is important that the networks are able to work with these applications (and TCP) to optimize the network for science.

   TCP is robust in many respect–in particular it has sophisticated capabilities for providing reliable data delivery in the face of packet loss, network outages, and network congestion. However, the very mechanisms that make TCP so reliable also make it perform poorly when network conditions are not ideal. In particular, TCP interprets packet loss as network congestion, and reduces its sending rate when loss is detected. In practice, even a tiny amount of packet loss is enough to dramatically reduce TCP performance, and thus increase the overall data transfer time. When applied to large tasks, this can mean the difference between a scientist completing a transfer in days rather than hours or minutes. Therefore, care must be taken when designing networks, with attempts to make it loss-free, so that TCP-based data-intensive science applications perform ideally.

4. **Establishment of a trust-model**
   Data transfer between Science DMZ works within the DOE context since government funding mandates

[0]Dart, E.; Rotman, L.; Tierney, B.; Hester, M.; Zurawski, J., "The Science DMZ: A network design pattern for data-intensive science," in *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for*, vol., no., pp.1-10, 17-22 Nov. 2013

[0]Meyer, E.T. "Moving from small science to big science: Social and organizational impediments to large scale data sharing", *In Jankowski, N. (Ed.), e-Research: Transformation in Scholarly Practice (Routledge Advances in Research Methods series)*. New York: Routledge, pp. 147-159, 2009.

sharing of data at least between other DOE funded scientists and facilities. The centers leverage the grid model to establish trust between the end-systems or a common trusted data movement provider like Globus. In order to facilitate data sharing in the Indian context, it is important to establish the drivers and the trust model, so data sharing is not impeded by issues of trust and verification - such mechanisms must be established by the facilities that host the data and the funding organizations that fund research that produce the data-sets.

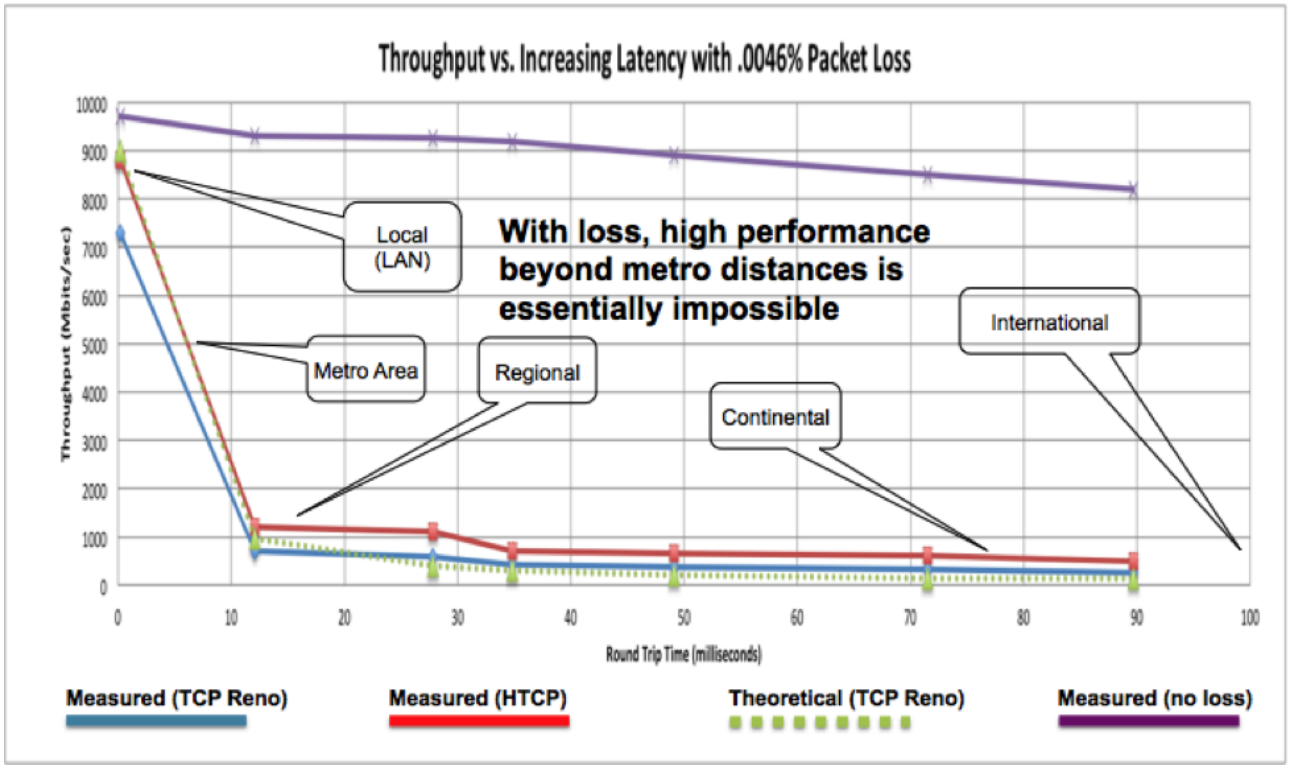


Figure 3.5: TCP performance as a function of throughput for different sized science networks (note that the performance degradation is quite unique to heavily granular science flows as opposed to classical telco traffic)

3.3.3 Science DMZ Infrastructure

As we discussed above, networks optimized for business operations are neither designed for nor capable of supporting the data movement requirements of data intensive science. When scientists attempt to run data intensive applications over these so called "general purpose" networks, the result is often poor performance – in many cases poor enough that the science mission is significantly impacted and/or the data is shared among the many researchers.

Since many aspects of the campus networks are impossible to change in order to improve performance for everyone, an architecture must be adopted to allow the networks to support science applications without needing to change or impact the general purpose campus network.

The Science DMZ[1] design pattern accomplishes this by creating an enclave in the campus network that is engineered for science applications. By separating the data-intensive portion of the network from the general purpose network, it can be assured that the science users get optimal performance to conduct their research while the general-purpose network can be tailored to meet its own purpose.

Scientific collaboration, like any other network-enabled endeavor, is inherently end-to-end. The Science DMZ can easily incorporate wide area science support services, including virtual circuits and software defined networking, and new technologies such as 100 Gbps Ethernet. Developed by ESnet engineers, the Science DMZ model addresses common network performance problems encountered at research institutions by creating an environment that is tailored to the needs of high performance science applications, including high-volume bulk data transfer, remote experiment control, and data visualization.

[0]For detailed information: http://fasterdata.es.net/science-dmz/

[1]Dart, E.; Rotman, L.; Tierney, B.; Hester, M.; Zurawski, J., "The Science DMZ: A network design pattern for data-intensive science," in *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for*, vol., no., pp.1-10, 17-22 Nov. 2013

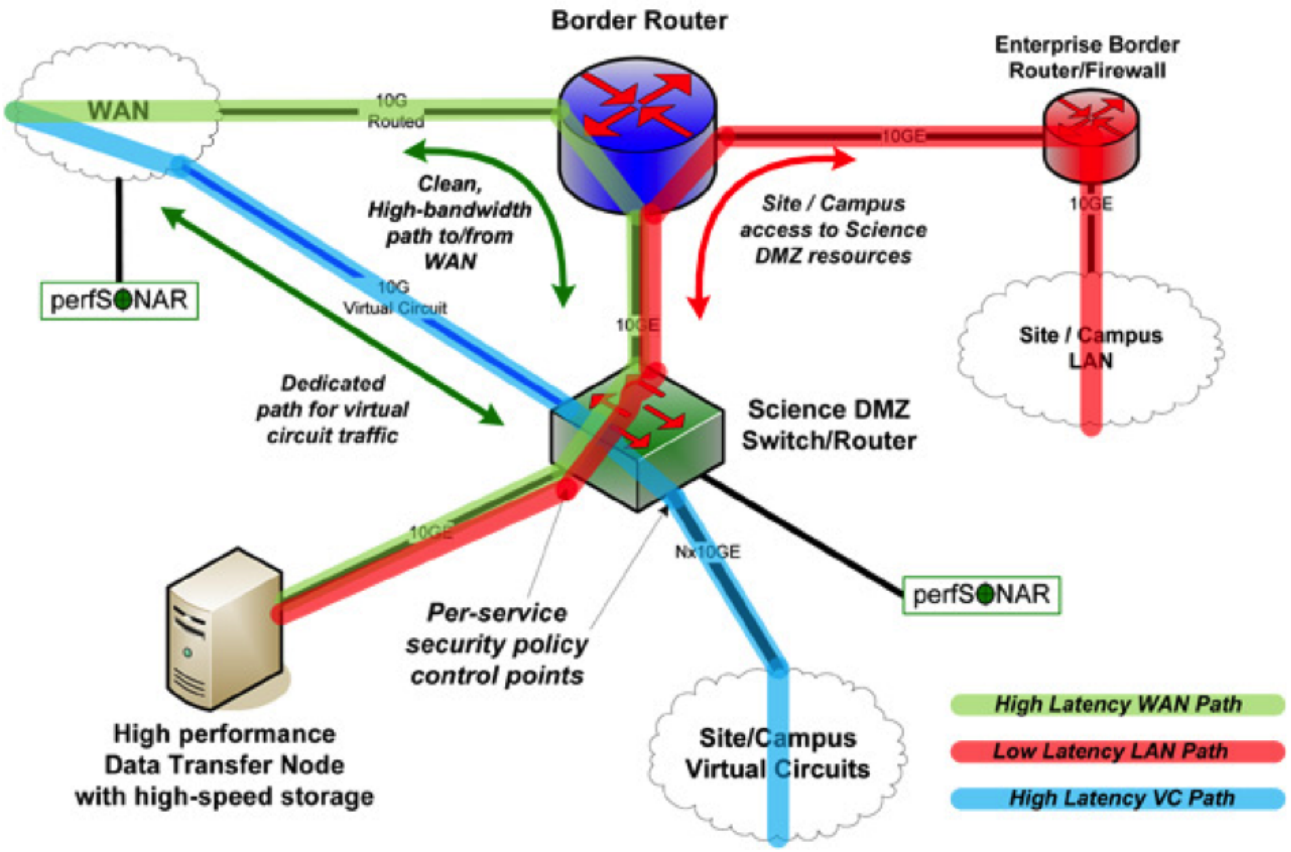


Figure 3.6: A basic Science DMZ pattern

A Science DMZ design pattern integrates four key concepts that together serve as a foundation for this model. These include:

- A network architecture explicitly designed for high-performance applications, where the science network is distinct from the general-purpose network
- The use of well-tuned, dedicated systems for data transfer
- Performance measurement and network testing systems that are regularly used to characterize the network and are available for troubleshooting
- Security policies and enforcement mechanisms that are tailored for high performance science environments

3.3.4 Key Takeaways

1. It is not enough to just have bandwidth in the local or wide-area network. Care should be taken on how the network is operated and for science big-data movement, a loss-free network architecture and design should be encouraged. Throughput is an end-to-end quality and providing systems and network that are well matched will enable the research effectiveness of scientists in the modern century.

2. Data throughput is an end-to-end problem, and as such only investments in wide-area network will have minimal impact on the data movement ability of scientific data unless similar architectural improvements are undertaken in the campus network. One approach through which campuses can limit their overhaul of the campus network is to build Science DMZ's, enclaves for big-data science and optimized for wide-area data movement.

3. Policies must be set for researchers to share data freely. Such policies need to be backed with funding for data management platforms at well connected institutions so that the data may be shared without any infrastructure impediment.

3.4 Vision for HPC and Data in India

Modern science relies heavily on theory, experiment, simulation and data-driven discovery. Any long term investment in scientific infrastructure should consider these modalities. In particular, facilitating HPC and data-intensive workloads to run efficiently will be key for future progress.

3.4.1 Key infrastructure investments recommended

a. The Government of India should invest in 2-3 strategic computing centers. These centers should have world-class hardware and software resources. The computing centers should accommodate both super-computing workloads, as well as data-intensive workloads. Addressing both strong and weak scaling workloads for HPC is important.

b. Investing in developing and growing manpower resources is vital to the long term success and sustainability of a national computing initiative. Hiring and retaining the best national and international talent should be the top priority of such computing centers. Attempts should be made to establish deep connections with leading academic institutions domestically (IITs, IISc) and internationally. Collaborations with leading industry vendors (Intel, Cray, IBM, HP) is highly recommended to keep the center abreast of latest developments. The centers may choose to create an International Advisory Board to keep them abreast of latest developments, and to seek independent evaluation of progress.

c. During the inception phase, these centers should identify key science partners, and develop close working relationships with the relevant scientific institutions and/or communities. We would recommend that the centers align their mission with national scientific priorities.

d. Policies must be established for researchers and facilities to share scientific experimental and simulation data freely. Such policies need to be backed with funding for data management platforms at well-connected[2] institutions (like the strategic computing centers) so that the data may be shared without any infrastructure impediment.

e. Data throughput is an end-to-end problem, and as such only investments in wide-area network will have minimal impact on the data movement ability of scientific data unless similar architectural improvements are undertaken in the campus network. One approach through which campuses can limit their overhaul of the campus network is to build Science DMZ's, enclaves for big-data science and optimized for wide-area data movement.

f. Care should be taken on how the network is architected and operated for science big-data movement. A loss-free network architecture and design should be encouraged. Since application throughput, end-to-end, is the most important metric - providing compute systems, storage systems and network that are well matched will enable the research effectiveness of scientists in India within this modern century.

[2]By 'well-connected' - we mean good network infrastructure at the campus and sufficient bandwidth connectivity to the wide-area network.

Bibliography

[Dart et al., 2013] Dart, E., Rotman, L., Tierney, B., Hester, M., and Zurawski, J. (2013). The Science DMZ: A Network Design Pattern for Data-intensive Science. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC 2013, pages 85:1–85:10, New York, NY, USA. ACM.

[ESnet-WPa, 2015] ESnet-WPa (2015). ESnet Science Requirement Reviews Reports. `https://www.es.net/science-engagement/science-requirements-reviews/requirements-review-reports/`.

[ESnet-WPb, 2015] ESnet-WPb (2015). Fasterdata by ESnet. `http://fasterdata.es.net/`.

[Meyer, 2009] Meyer, E. T. (2009). Moving from Small Science to Big Science: Social and Organizational Impediments to Large Scale Data Sharing. In *Jankowski, N. (Ed.), e-Research: Transformation in Scholarly Practice (Routledge Advances in Research Methods Series)*, pages 147–159. Routledge, New York.

[NERSC, 2016] NERSC (2016). NERSC Annual Reports. `https://www.nersc.gov/news-publications/publications-reports/nersc-annual-reports/`.

[NERSC-WP, 2017] NERSC-WP (2017). HPC Requirement Reviews. `https://www.nersc.gov/science/hpc-requirements-reviews/`.

Chapter 4

Networks for Computing Needs

Ashwin Gumaste
IIT Bombay

4.1 What does it mean?

As computing power grows and follows Moore's law, material science is unable to meet such computing requirement within a single processor. Extremely large scale integration and device technology that enables us to go to sub-20 nm processes and hence pack millions of logic gates in a single chip. Even such integration is not enough to meet the ever growing needs of users, especially with the world-wide-web throwing a plethora of applications year-on-year. The compounded-annual growth rate (CAGR) of data-traffic in the Internet is almost doubling every other year, and this huge amount of data requires processing that needs to be done across various data-centers in the Internet. Combine that with the disparate requirements of the data such as voice, video based services and crunching for numbers to provide real-time analytics that are quintessential to providing advanced services. The goal of this chapter is to understand what does it mean when processing entities need to be connected to create a virtualized environment to meet user needs. The underlying network becomes a tactical glue that binds many processing entities into a virtualized environment that cumulatively adds the computing power of disparate entities. Networking began as a way to transport data and voice and is now the key enabler for the Internet and all the applications subtended by the growth of the Internet. Networking technologies have progressed through wired and wireless mediums from bit-rates of a few Kbps (kilobits per second) to several hundred Gbps (gigabits per second) in the optical fiber. The deployment of the network as an aid to computing marks an interesting and important revolution in the next generation of computing. In effect, the onus of making processing more rigorous has been replaced by making processors behave as a large single unit across a network. This makes next generation computing effective and pragmatic from cost, performance and usability standpoints.

Networks have transgressed from copper cables, to shielded twisted pair based cables, to networks in the air (wireless) and to use of the optical fiber. Networks have become smarter from just transporting bits from one place to another to providing application awareness that is quintessential to next generation applications and computing. High Performance Computing (HPC) cannot become pragmatic without the underlying network. The network is used between processors, processors and memories, processor blades and server farms, and between data-centers. Each type of interaction between computing entities (processors, servers, memories) requires a different type of networking flavor – tailored to meet the key aspects of the interaction. Networks for computing needs have transformed processor development and related data-centers theatre seen as the brain of the Internet.

Perhaps the most important feature that a network can provide towards HPC facility is that of latency. Latency determines the usefulness of a network, especially from the perspective of virtualizing computing entities. Lower the latency, between the interaction between computing resources across a network. Low-latency architectures can be achieved in two ways: (1) by providing large bandwidth between computing entities to move big amounts of data from one place to another and (2) creating novel architectures that facilitate low-latency interconnection across the entities in a network. Network architecture plays a strong role in facilitating low latency in a HPC environment. The choice of technology, protocol and the scalability of the system all eventually determine latency requirements in an HPC environment.

A second important feature that networks provide to HPC is the agility provided due to reconfiguration of the network interconnection graph. Significant research has gone into creating networks that provide reconfigurability. Regular network architectures were initially deployed that had predictable routing indices. These are now been replaced with customizable irregular network fabrics that can create extremely reconfigurable network

fabrics – essential for HPC applications.

HPC applications have by themselves become very complex requiring rapid-back-and-forth communication between computing entities such as servers across an interconnection network. Apart from agility there is also the issue of providing one-to-many service across a network backbone. Popularly called as data multicasting the service is critical for easy and fast replication of data from one computing entity to several others in a parallel manner.

Scalability of providing a multitude of network services across an HPC environment is another key feature that measures HPC performance. Another figure of merit for HPC systems is as to how many computing entities (servers or storage devices) can be connected without loss of performance. The trade-off is that when we have N entities that need to be connected, we classically face the $N2$ problem – that of creating a non-blocking paradigm with $N2$ cross-bar switches. As N increases, the size of such an interconnection paradigm becomes unmanageable, expensive, and hence difficult to implement. So how do we create next generation HPCs and data-centers with several 10s of thousands of computing entities while meeting the requirements of disparate services? This problem has received substantial attention in the recent past, and we will examine the various approaches towards solving such interconnection paradigms.

From a protocol perspective, there is an interesting yet problematic trade-off that one encounters. Simply put, a protocol data unit (PDU) that can scale, requires significantly larger header implying that for processing the header large amount of time is lost thus compromising on latency. It is no wonder that protocols like Infiniband that have excellent latency do not scale very well. In contrast, protocols such as IP, especially in IPv6 format scales very well but is plagued by its overhead and control processing implying very poor latency. Much effort has been devoted to the design of scalable yet latency-sensitive protocols. To this end, we will outline the correct protocol requirements that would aid towards the design of next generation protocols for HPC applications.

Application development is what is dictating future HPC requirement. Applications are becoming exceedingly parallel in behavior with both symmetric parallelism and asymmetric parallelism. In symmetric parallelism, entities communicate with each other in a homogeneous parallel structure, while in asymmetric parallelism, a group of M entities are continuously used for communication and computation by a group of N-M entities in an N-node HPC structure. In the latter case, there is tremendous stress on the interconnection fabric, especially when we consider that due to application behavior it is impossible to predict the stochastic behavior of the interconnection pattern between the computing entities.

The last pieces of the HPC interconnection puzzle are security and energy consumption. HPC environment requires higher level of security on account of sensitive applications as well as to conserve processing power on legitimate tasks. Usually, physical security is a first step towards attaining HPC security, but when the HPC machine is connected to the outside world, the degree of security preparedness becomes a challenge. Securing HPC systems in the context of next generation applications and pertinent cyber-threats is of paramount importance in very large HPC clusters. How does one continue to be ahead in the security framework when attacks can be launched by users both legitimate and otherwise?

Processing entities consume huge amounts of energy. Energy consumption is often cited as a limitation factor to the size of an HPC system. There are three aspects of energy consumption associated with a data-center/HPC system: (1) Energy required by the processing entities; (2) Energy required for cooling the entities and (3) Energy required for the networking protocol that facilitates communication between the processing entities (servers). Our main focus is on the third aspect of energy consumption but it is also believed that an efficient communication protocol also optimizes the energy consumption at the servers.

4.2 Networks for Computing

As HPC system grow, the underlying interconnection fabric i.e., the network becomes important. The architecture of the network, its connection methodology and ability to adapt to HPC requirement are all figures of merit to eventually judge the HPC system.

Networks in an HPC system can be classified into three types: on-board interconnects that are capable of connecting chip-sets among each other; server to server interconnects that facilitate intra-HPC interconnection; and data-center to data-center interconnect, as is prevalent to create a cloud computing environment. Each of these classifications requires a different type of network architecture, protocol and connection methodology. We will examine these in detail now.

4.2.1 Chip Interconnection

Chip interconnection has become an important issue in recent times. One can say that chip interconnection technology has made progress in discrete steps. Wire based interconnection on breadboards and rudimentary PCBs were perhaps the first interconnection technology. The number of discrete wires, the line-rate that

the wires could support, and distance between chips were all limitations to scaling such an interconnection system. Then came PCB with multiple layers, whereby PCB tracks were used in some of the many layers as an interconnection pattern. The tracks in the PCBs also had limitations in terms of distance and bandwidth they supported. It must be noted here, that as the line-rate between chips increased, the behavior of the tracks had to be carefully analyzed to support such increase. PCB tracks which were regular conductors of the bit-stream running between chips will have extreme waveguide properties exhibited by them, as the line-rate increases. The frequency domain analysis of a high-speed signal (essentially now an RF-signal) will have harmonics that will create a frequency domain response of the track. The track layout and the material used to build the PCB will determine how severe the response would be. As a rule of thumb, higher the permittivity constant of the material used, better its ability to transport high-speed signals. High-speed signal transfer is extremely crucial to achieve the economies of scale in HPC sub-systems. Processors are becoming very fast and all the associated chip-sets that are connected to the processors also must be able to communicate at very high speeds with the processors. For example, it is common today to run memories at several 100 MHz. Note that at such speeds, the response of the line-rate is not just a digital waveform, but also exhibits microwave characteristics. Communication between processors and server IOs can be in fact at much higher line-rates such as at 1 Gbps or 10 Gbps. At 10 Gbps, the pulse width is 100 picoseconds and the probability of error can be high. Often a 10 Gbps line between a processor and an IO interface is divided into 4 parallel lines with error coding on each line. Such a mechanism is called XAUI. XAUI allows for slower-speed communication using parallel lines to achieve overall high-speed throughput. Complementing XAUI is the PCI-express standard that also facilitate communication at 10 Gbps speeds. PCB design at such high-speeds is an intricate affair and involves pre-layout and post-layout signal integrity analysis. Schematics are first designed that enable the interconnection pattern between the chips that are to be connected. Then, we create a layout of the schematics so that the exact placement of the chips on the PCB is evaluated. The layout is followed by routing of the signals among the various chips used. Impedance matching is one of the most important tasks during signal routing in the PCB. Most of the chips are specified to 50 Ohm impedance matching. This is a right amount of impedance to drive current to create the necessary potential difference between chips to enable signal flow. Impedance matching techniques vary. At higher-speed the technique assumes microwave like characteristics. One has to model the trace as a waveguide. A waveguide is a medium that assumes flow of microwaves and is a boundary condition to the well-known electromagnetic propagation equations called Maxwell's Equations. Upon modelling a waveguide, we are able to perform signal integrity analysis to finalize if the signal will indeed correctly be transported between the chips. As part of routing we also need to ensure that the farthest route is well within the specified maximum for that particular line-rate. Another factor to consider is clock-skew. On a PCB there are various clocks, each of which determine the clocking of different chips as well as are used as drivers on the same chip. On large PCBs, clock skew can be an issue due to differential delay between the same clock signal reaching two different chips. Another aspect of differential delay that must be considered is when there are multiple parallel lanes between two chips. Such designs are common between memories and processors or processors and IOs. For example, all RAMs have multiple address and data-lines that are interconnected to the processor. These lines must have exactly the same length on the PCB. If this is not attained then there is the issue of differential delay. Differential delay can potentially lead to loss of synchronization, eventually causing irrecoverable errors. Such errors are very difficult to be rectified in bulk. Lane matching is a well-known technique used to ensure that length of all PCB traces are same that run across two chips and need to act as parallel lanes. The process of layout is hence iterative and involves routing the traces and then ensuring that system parameters are met.

In some cases, the routing and layout problem is done using software in an automated fashion. In most cases, the CAD software for routing and layout is augmented by human intervention. The latter is generally the default industry practise. Simulation models are available to model the signal integrity on traces. There is also a temperature dependent factor that should be considered in HPC environments when PCBs are designed. In most cases, the temperature considered is up to 50°C, but in some industry/military applications we should go up to 70°C to check if the signal integrity is as per what is desired. A very simple way to check signal integrity is to run the simulation model and manually observe the "eye pattern". If the eye "opens" well enough, then it is quite clear that the signal integrity is intact. If on the other hand the eye opening is negligible then one can assume that it will be difficult to isolate the "0"s from the "1"s.

Future of on-board communication technology: as line-rates increase with processing power, there is an absolute need for on-board technology to change. This change is about to happen. There is a strong research push towards inculcating photonics technologies as an enabler for chip-to-chip communication, popularly called as optical interconnects. Optical interconnect technology is today in its infancy, but is slated to be an important breakthrough for HPC applications. There are two clear advantages of using optical interconnects: optics through fiber provide for a low loss medium and secondly there is seemingly near infinite bandwidth offered by the optical fiber. Optical fiber is the default choice for communication paradigm for the Internet especially in the core of the Internet. The reliability of the optical fiber as a communication medium is second to none and it serves also as a low-cost medium. Optical fiber is able to provide about 30 THz of bandwidth in its default communication band i.e., when light is transmitted through the fiber at 1.5 micrometers. This translates to 30

Tbps of bandwidth when we deploy simple ON-OFF keying (OOk) techniques. However, current electronics are unable to create a switched bit-stream at such high line-rates. This difference between high-speed optics and low speed electronics is called the opto-electronic bottleneck. To absolve this opto-electronic bottleneck, a solution is to divide the bandwidth into frequency specific channels. Such kind of frequency division multiplexing is commonly deployed to make good use of the optical fiber. Since the frequencies used are of the form of 193 THz (corresponding to 1.5 micrometers), it is more convenient to state the multiplexing pattern as wavelength division multiplexing as opposed to frequency division multiplexing. On each wavelength, we can modulate a slower speed electronic signal and several such wavelengths with individual signals modulated create a composite wavelength division multiplexed (WDM) signal. WDM technology for high-speed communication has significantly matured and it is now a question of time when it would be used as an interconnect technology. When the channels in the WDM signal are spectrally close, then the composite signal is called Dense Wavelength Division Multiplexing (DWDM), while if the channels are far apart in the spectral domain, then the resulting signal is called Coarse Wavelength Division Multiplexed. Multiplexing technology uses optical components such as fused fibers, liquid crystal on silicon substrates, digital lightwave processors (DLP), and interferometers.

The key to adapting the optical technology within the domain of interconnects for HPC application is in the ability to miniaturize the components. Miniaturization must be such that the components must "fit" within the chips on-board a PCB. Of critical importance is the ability to induct optical sources that can generate data. Classically, lasers are used to generate a coherent source of light. Miniaturization of lasers is a difficult task – it requires substantial semiconductor enhancements and the yield could be substantially low. A key difference between lasers required for commercial optical communications for transmission purposes (long-distance) versus lasers required for optical interconnects in a chip is the power of the laser within the PCB environment is significantly less.

In this regard, an important breakthrough is that of the Vertical Cavity Surface Emitting Laser (VCSEL). Unlike transmission lasers, that require a fiber to be fused into the laser almost perpendicular to the lasing action, VCSELs can be built with fibers parallel to the surface so that the assembly of fibers into the chip can be realistically achieved. VCSELs produce substantially lesser power than lasers, but this is perfectly fine in the ambit of HPC environments where distances are small and the lesser power produced is sufficient to achieve optical interconnections. VCSEL technology is now appearing for directly interconnecting chips in large scale integration. A large number of VCSELs can be grown on the same substrate to create a parallel transmission medium whereby each VCSEL can support a particular frequency, and together they can be coupled together to form a composite WDM signal. Such an arrangement can then be supported by another transmission innovation called plastic fibers, whereby instead of using silicon based fibers, less expensive plastic is used to create the fibers. Plastic fibers are easier to manage and are also more durable to the consumer-centric server interconnect application. It is envisaged that the combination of plastic fibers with VCSEL arrays will form the backbone of optical interconnect applications. The role of photonics is likely to go beyond transmission and communications within HPC environment. It is perhaps possible in the future to have optical processors, whereby wavelength interaction using non-linear effects such as cross-phase modulation and four-wave mixing can create logical gates that work at speeds significantly faster than what is achieved with silicon technology. Newer materials such as Indium Phosphide used for creating monolithic lasers and embedded photonic components would potentially change the interconnection pattern in an HPC environment. Recently, Graphene has been added to the list of materials with smart photonic properties, and the abundance of Graphene combined with its optical properties is likely to be a game-changer in the HPC interconnection scenario.

4.2.2 Server Interconnection

A second aspect of network interconnection architecture is to connect servers to each other to create an HPC cluster. Server interconnection is perhaps the most important interconnection architecture within the HPC environment. The key challenge to server interconnection is to create a scalable interconnection paradigm. Scalability for network interconnection exists in two forms: (1) scalability in number of servers that can be connected and (2) scalability in terms of performance. The network architecture plays a strong role in determining the performance of the HPC system. There are many possible network architectures to choose from and each architecture has its pros and cons. Typically in contemporary HPC environments or data-centers, servers are represented as blade servers or stand-alone rack-mounted servers. In either case, such servers have front mounted or back mounted interfaces for interconnection. Such IOs, typically are supported inside the server with network interface cards (NICs) that are pluggable modules into the server. The back-end of a NIC is connected via a PCI-express bus. NICs can support a multitude of protocols, but typically the line-rate is either 100 Mbps or 1 Gbps or 10 Gbps. Off late, 10 Gbps server ports are appearing in the market and it is anticipated that there would be an exponential surge towards such adoption in the very near future. Server NICs support two kinds of PHY – either an optical PHY or a copper PHY (physical interface). The optical PHY typically comes with a pluggable optical module, which could be fitted into a PHY-slot that can encompass the module.

Servers need to be interconnected to each other to create a large HPC environment. The interconnection fabric must scale and also provide a non-blocking cross-connect functionality. Creating a large non-blocking interconnection fabric is a challenge. Typically, as the number of servers increases, the number of interconnection points (cross-bar switches) increases exponentially (square of the number of servers to be interconnected). Creating such an interconnection fabric is not feasible.

There are hence multiple methods of creating modularly scalable interconnection fabrics. Such network architectures usually compromise in some feature of the interconnection fabric.

A commonly used methodology for interconnection fabrics is to use regular graphs. Graph structures such as ShuffleNet, De Bruijn graph, Torus and Hypercube are well known parallel processor environments. The problem with these regular structures is that these do not scale very well. A common mechanism towards interconnection within the HPC environment hence is to stack up switches (generally commodity switches), that facilitate a tree like structure. Usually multiple tiers of nodes are connected across these switches. The advantage of a tree-shaped interconnection topology is that it can be grown quite well with possibility of incremental growth.

The manner in which HPCs are developed using tree shaped interconnection is that a rack of servers is connected to a top-of-the-rack switch (TOR-Switch). Many TOR switches are further back-connected to a root switch. This type of a design requires the root switch to have a large number of ports – equal to the number of racks that are part of the HPC environment. A slightly more scalable design enables multiple levels of root switches so that there is less restriction on the number of ports of a root switch. Yet another efficient design uses multiple paths between any two racks by using a Clos interconnection network between TOR switches. To obtain very large HPC clusters and data-centers, we may break the clusters into pods whereby full non-blocking switching is available within the pod, but across pods only partially non-blocking switching is possible. Such kind of network architecture requires some degree of under-provisioning, whereby the bandwidth in the higher tiers of the architecture is less than the cumulative bandwidth of the servers in the lower layers of the hierarchy. This means that if all the servers in the lower layers of the hierarchy were to communicate with some servers in another branch across the hierarchical TORs, then the TORs would not be able to provision full non-blocking bandwidth between the discrete branches. In fact as the number of servers in an HPC increases, under-provisioning cannot be avoided.

An Example: if we have a 16 client port switch and each servers' interface is at 1 Gbps, then one TOR can support 16 servers. We assume that the switch has 2-4 network ports to back-haul the traffic from the servers towards the root of the tree. Now assume that the TORs are further back-hauled into an aggregator switch which has 48 ports each of say 10 Gbps. That means the network ports of the TOR switch will also be at 10 Gbps, and 48 TORs can be connected together. This means that the under-provisioning factor is 10/16=62.5% Now, further let us assume that the HPC just described is part of one *module*, and several such modules are connected via a core switch. Then what would be the number of ports and line-rates that this core switch would support? Such questions are difficult to answer without choosing the protocol and technology. Each module would cumulatively generate 480 Gbps of data. Even if we assume the aggregator switch has 48 client ports (connected to TORs) and another 48 network ports used for connection to other module, then we have a scalability limitation of 48 modules and an overall under-provisioning ratio of 62.5%. The only way to grow such a system is to replace the core and aggregator switches with larger port-count switches. This may not be possible with current cost and protocol limitations.

4.2.3 Performance

Another factor of importance is to consider the performance of such an HPC system. Within a module, the longest route is of 4-hops long, while for the larger multi-module HPC system, the number of hops is 6. In general, we need 2logD hops where D is the diameter of the HPC cluster. The performance of such a system degrades in terms of both throughput and latency. Latency will be described in detail in the next section.

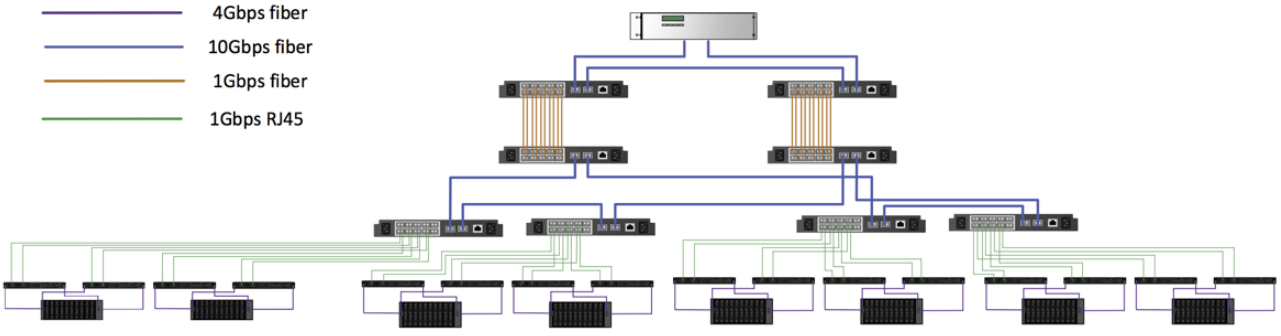


Figure 4.1: Data-center architecture arranged as a fat-tree

Servers within an HPC environment can have optical interfaces or copper interfaces. Similarly, TOR, aggregator and core switches can also have copper or optical interfaces. However, for the purpose of reliable transmission at higher line-rates it is always desirable to have optical interfaces. Optical interfaces are crucial from the perspective of provisioning large amount of bandwidth within the HPC cluster.

The core or aggregator switches in a tree or a multi-Clos network design become the bottleneck in an HPC environment. One aspect of scalability limitation is that of providing scalability in terms of number of supported server systems. Another aspect is in terms of providing mechanisms for in-situ addition of servers. Both these approaches require the bandwidth of core switches in an HPC to be upgradeable. With increasing line-rates this can be a serious challenge. Copper interfaces have rate-limitations and cannot scale beyond 10 Gbps, and that too can be supported over a few feet. A large HPC environment can be more than few tens of feet, implying that copper based connections will not work. The solution is to use optical interfaces for both reach and support of larger bandwidth.

4.2.4 Optimal Backpane

Another aspect of server interconnection is the recent use of the optical backplane. A backplane is either a switching card that connects many servers together or can be a mating connector set that facilitates any-to-any connectivity. In typical HPC environments, backplanes can be designed using stand-alone switches or mating cards. The job of the backplane is to provide connectivity between servers. It may be passive, in the sense that it may not support switching and connectivity occurs by a broadcast and select architecture. The backplane may be active, in the sense that it may actually support switching. In case of a passive backplane, there need to be enough connectors that provide one-to-many connectivity for each mating server. The idea is that a server that desires to communicate with other servers, does so by sending the data on one of the traces of the backplane. Other servers can all listen to this data and will have to select whether to pick data by this server or any other servers. In this system, if we are to support K servers, then each server must have the capacity to send data into the backplane, but when it comes to receiving data, each server must be able to receive from any of the K-1 servers. There is typically no scheduling policy or efficient sharing of the backplane in such a design. The limitation of such a design is the number of receptors that can be architected into each server line-card, since the server backplane IO now has to process data that to decipher whether to select or not. An active backplane is shown in Figure 4.2 below, and consists of many server line cards connected to a switching card. The switching card is the backplane. The switching card could be a stand-alone pluggable unit or could be a mating stationary connector that comes with the HPC chassis. Scalability in such a case is restricted by the number of line-cards that can be connected to the switching card and creating a non-blocking switching fabric to support the line-cards.

In both cases, of backplane design there are design limitations in terms of scalability of the HPC fabric. Even when we connect multiple backplanes together, we are encountered with the same limitation of being unable to provide a non-blocking fabric without compromising on efficiency and performance. We could for example have a severely compromised under-provisioning ratio and scale up such a system. But then the performance would be acceptable to only certain traffic types. The problems of such a system would be significantly enhanced, if we assume that processors talk to memories or other processors in executing a task using the switching behavior of the larger system. Such a design is complex and must always be optimized for performance. A general figure of merit of the optical backplane is the amount of bandwidth that it can support for switching.

In this regard, recent research has focused on optical backplane design. In lieu of its tremendous bandwidth availability, an optical backplane is proposed as a scalable alternative to traditional electronic backplanes. An optical backplane can be of active or passive type. The active type of optical backplane can be engineered using all-optical switches. All optical switches can be fiber switches – switching signals between ports without analyzing whether the signals are of a particular wavelength; or could be wavelength-level switches. Wavelength-level switches are used in commercial core networking technology today and are expected to become more popular as network bandwidth grows. Such all-optical wavelength-level switches are called as Wavelength Selective Switches or WSS. These are primarily of a 1xN design, and a group of these can be collectively engineered to create a non-blocking cross-connect functionality. Shown in Figure 4.3, is an all-optical switch that can support up to 4 Tbps of bandwidth – better than any electronic backplane. In this design the 1xN WSS is further re-engineered to create an MxN architecture and several such WSS are used together to create an optical switch. Such a large optical switch with wavelength-sensitive properties is also called a Reconfigurable Optical Add-Drop Multiplexer. The architecture is based on the broadcast and select concept, where incoming signal is broadcast to the different arms of the switch, and individual servers can select a wavelength of their choice. The broadcast is done with couplers/splitters, while the selection is done via the MxN WSS. The inherent limitation of this architecture is the switching speed of the WSS. Typically WSS are built using MEMS or LCOS designs and have a switching speed of a few milliseconds which is clearly not acceptable in the HPC environment. Despite this limitation optical switching is being considered by a large number of HPC builders either as an exclusive backplane technology, or as a hybrid technology that has both optical backplane and an electronic

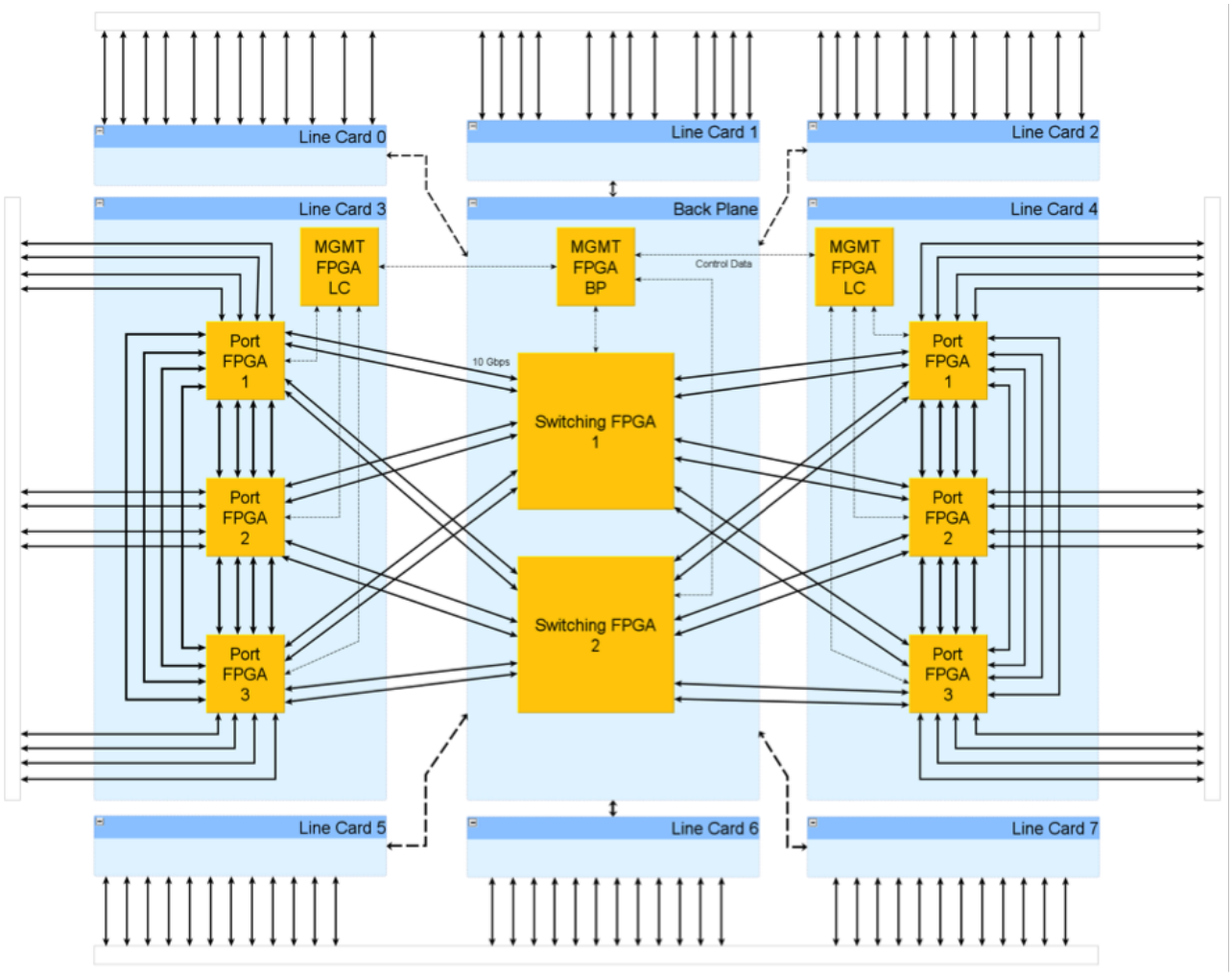


Figure 4.2: Example of an active backplane

backplane. The hybrid backplane is being considered for switching different kinds of flows. For example, flows among servers can be classified into mice and elephant. Mice flows are those that trickle between servers have low granularity and generally require fast switching – for them the electronic backplane is the best solution. Elephant flows are rare, but when they do happen, they require much coarser granularity and are present for large time-duration. Elephant flows can hence be switched by an optical backplane.

While WSS based backplanes are more scalable than electronic backplanes on account of the larger supported bandwidth – at higher levels there is a scalability limitation.

Yet a third design that is being considered in the backplane is a passive optical backplane. A recent study showed that it was possible to have a near infinitely scalable backplane using passive optics. Such a design also uses a broadcast and select architecture, but the selection process is made much easier using a recently proposed superchannel concept that involves creating OFDM modulated channels in the optical domain. Such a design absolves the need for fast reconfigurable optical switches relying primarily on the broadcast domain for efficient interconnection.

Many HPC systems, each located at different locations need to be connected across these locations to provide a cloud-like environment. Each HPC essentially could be a repository of information and a processing unit. Such an individual HPC environment is called a data-center, and many such data-centers together become a cloud. The architecture of the data-center is quite similar to the architecture of the server-to-server interconnect. Generally the design uses tree shaped topology to connect a server farm at the leaves, with switches at branch and root interconnection points. Such a design is commonly used in small and medium data-centers and HPC clusters up to several 100 nodes or even for 1000 nodes (servers). Apart from the scalability issue of the tree architecture, there is a second scalability issue that of providing protocol support – how do servers talk to each other, discover each other and create a monolithic entity that can be a virtualized environment for computing. Protocols will be discussed subsequently.

Like server interconnection architectures data-center architectures can also have alternate models using some modified form of the Clos switching architecture where by a combination of smaller port-count non-blocking

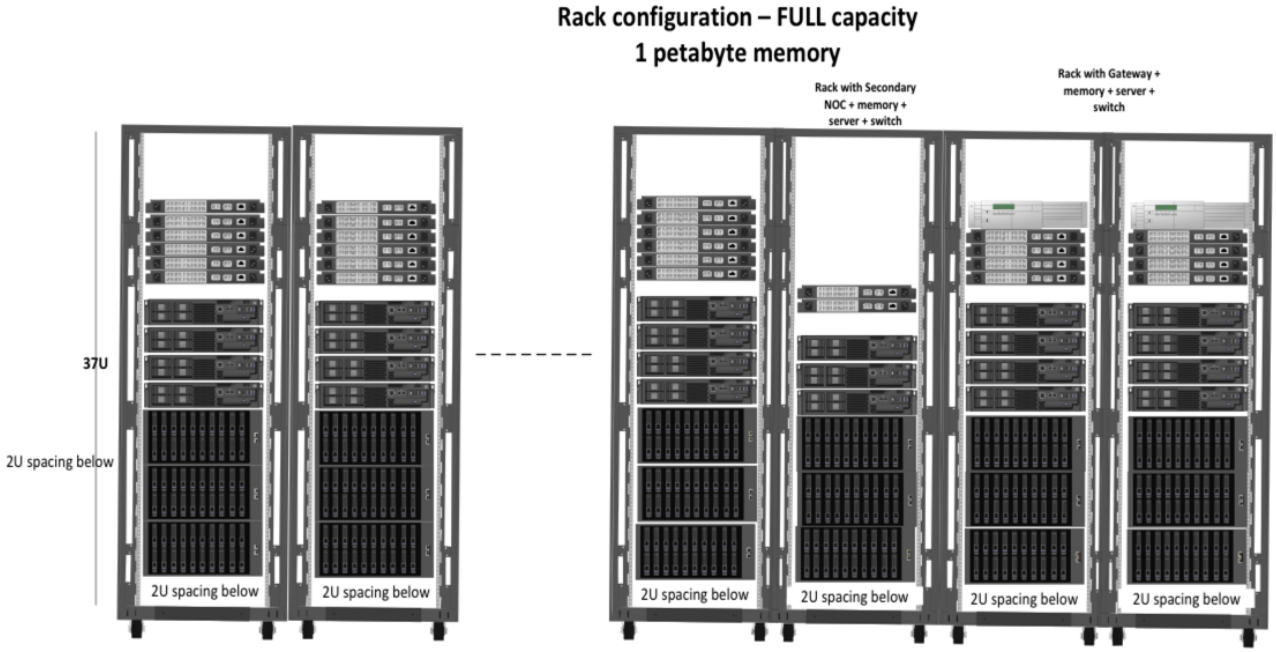


Figure 4.3: Stacked HPC architecture

switches can suffice for a larger switch design.

Data-center design can also be limited by physical space, Internet connectivity and physical location proximity issues. It is very common to deploy many data-centers for a service provider network or many HPC clusters spread across a Wide Area Network. For example, a group of research labs may each have its own HPC cluster with physical proximity to it. All these HPC clusters can eventually be interconnected to form a cloud like environment. Cloud network design is a complex engineering process that often involves managing bandwidth in the WAN that connects the clusters and creating a virtualized environment. Such an environment has to scale as well.

One of the key features of today's virtualized environment is the use of Virtual machines (VMs) that can be used to move across servers within a cluster so that compute resources can be optimally utilized. VM migration across servers within a cluster is a well formulated process. However, VM migration across servers which reside in separate data-centers and HPC environments can be a complex issue. One method to enable all the HPC environments behave as a cohesive unit is to interconnect all the HPC environments with a Layer-2 VPN. The advantage of a layer-2 VPN as opposed to a layer-3 MPLS VPN is in terms of cost and performance. As a rule of thumb, keeping data in the lower stacks of the Internet layering hierarchy is lower cost, more energy efficient and is less prone to vulnerabilities. Hence a L2VPN is preferred to aL3VPN to create a virtualized environment. The other significant advantage of a L2VPN is that L2-technologies are usually carrier-class. Examples of such technologies are SONET/SDH and Carrier Ethernet. SONET/SDH is a time-division multiplexing protocol that enables the creation of payload interspersed in the time-domain and then sent into the optical fiber. SONET/SDH based L2VPNs can be managed very well and provide the necessary operations, administration, maintenance and provisioning (OAMP) features required for service-provider networks, especially critical for creating cloud-like environments. SONET/SDH is essentially a circuit switched technology and the bandwidth granularities are quite coarse. This means that with such a technology, the advantage of statistical multiplexing of packets is not available. The dominant network protocol in the Internet is IP, and IP exists as packets or datagrams. IP-packet switching and routing is one of the foundation blocks of the Internet. However, IP as a service is best-effort and not carrier-class. Due to statistical multiplexing and best-effort behavior IP even with MPLS does not offer the same kind of OAMP features that SONET/SDH does. IP can hence be a good residing technology on a SONET/SDH based VPNs but such a solution is also expensive.

What is required is an efficient packet technology that provides good statistical multiplexing, yet is able to guarantee carrier-class OAMP support. Carrier Ethernet which is a carrier-class alternative of Ethernet is here a good packet alternative. Carrier Ethernet is quite different from Ethernet in the LAN, whereby there is no MAC learning and no spanning tree protocol. This avoids any probabilistic behavior of finding routes in a broadcast domain and creation of loops. Forwarding in Carrier Ethernet is accomplished based on backbone switch address in conjunction with a series of VLAN tags or labels. Two implementations of Carrier Ethernet exist: in the IEEE and the IETF. The IEEE implementation called PBB-TE or Provider Bridged Backbone-Traffic Engineering uses VLAN based switching by mapping incoming tagged or untagged services into network-specific ISID tags

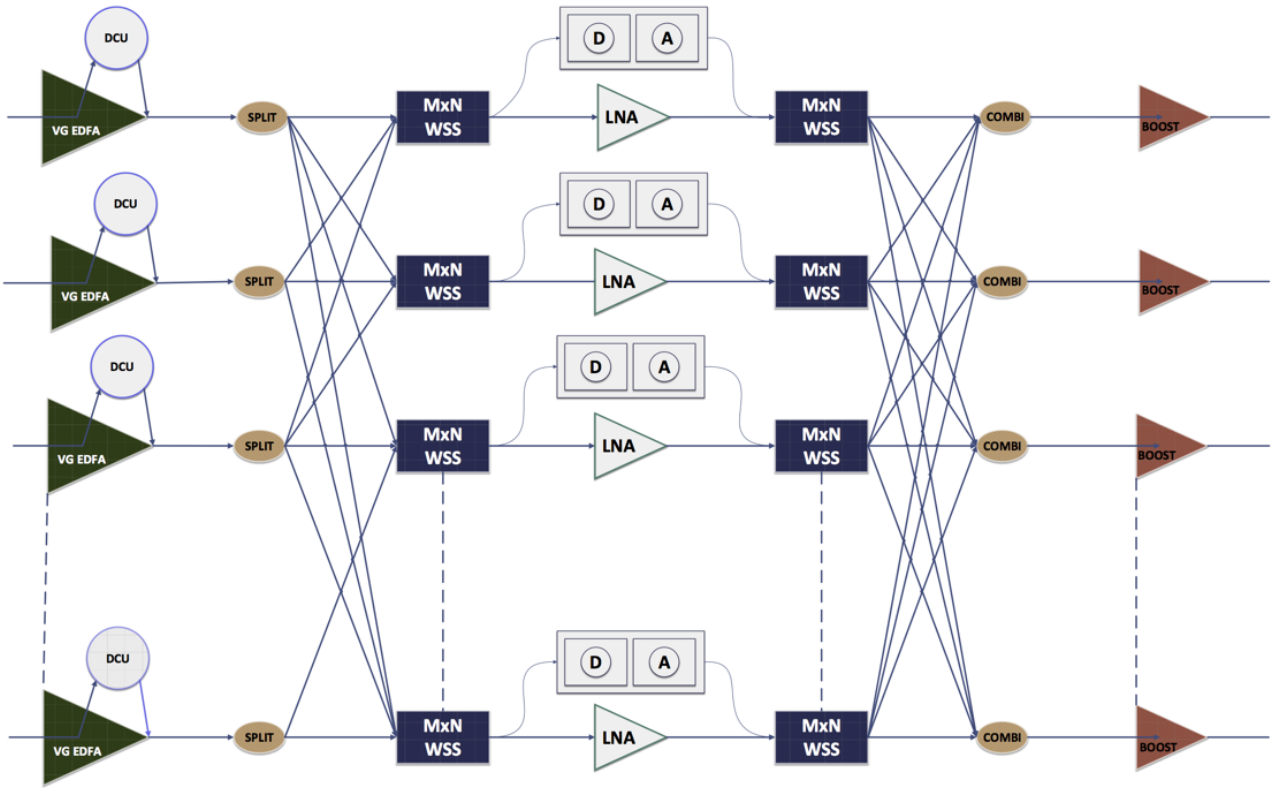


Figure 4.4: Data-center interconnection

that are service tags and which are further mapped to backbone MAC addresses and backbone VLAN tags. Forwarding is done exclusively using the 60-bit backbone MAC address and VLAN tags. In PBB-TE, paths are set up using a network management system (NMS) that communicates with core and edge bridges (PBB-TE switches) to assign the requisite MAC and VLAN identifiers. The IETF version of Carrier Ethernet is called Multi-Protocol Label Switching – Transport Profile or MPLS-TP. In this method, labels are used to forward packets and label to packet mapping is again explicitly done through the NMS. MPLS-TP is a scaled down version of MPLS with no merging capabilities as well as no unidirectional support. MPLS-TP also does not support equal cost multiple paths.

Three types of services are defined in the gamut of Carrier Ethernet – ELINE, ELAN and ETREE. An ELINE service is a point-to-point bidirectional connection that is created using switched Ethernet identifiers at nodes across the network. An ELAN service is an exemplification of a LAN environment in a core network, while an E-TREE service is one in which there are many leaves communicating directly to the root of the tree.

In addition to service definitions, Carrier Ethernet also uses the IEEE802.1ag Connectivity Fault Management standard (or sometimes the Y.1731 standard) to check for faults in the network and ensure a healthy network. As part of the 802.1ag standard, connections are demarcated by management points, and management points exchange information periodically through heart-beat messages. Loss of 3-consecutive heart-beat messages signals to the end points that there is a fault in the connection and hence a graphically alternate (pre-provisioned) path is chosen.

4.3 Software Defined Networking (SDN) for Cloud Environments

SDNs are recently being proposed as enablers to launch any service on a cloud like environment, by creating a "dumb" hardware platform into a user-defined switching elements through a control plane. The control plane and data plane interact through well defined Application Programming interfaces and these are used to describe the user requirements to the hardware. SDNs have the potential of changing the way we plan, route, traffic-engineer and evolve networks.

SDNs are different from traditional networks in the following ways: (1) separation of control and data plane; (2) centralization of the control plane; (3) programmability of the control plane; and (4) standardization of north-bound Application Programming Interfaces (APIs).

SDNs are being used to set up services in cloud like environments with the goal that previous service provisioning methods could only achieve so much and with SDN implementation those horizons are being

further extended. By bringing user programmability within the gamut of networking means that new services can be implemented in a network that could previously not be even envisaged.

SDNs have the potential of being a game-changer for HPC environment as the user base of HPC is quite diverse implying a strong role for user-defined service support. In such a scenario, SDNs can control the HPC environment and resources such as bandwidth management can be done through user-defined interfaces. The SDN contribution to HPC is that a traditional symmetric switching environment within an HPC domain can now be made to a user-defined, possibly asymmetric environment.

4.4 Scalability aspects of Network computing

In this section we delve upon the scalability aspects of HPC systems from the network computing perspective. There are two aspects of scalability to be considered:

- Switch architecture scalability
- Protocol Scalability and fault tolerance

The first aspect of switch architecture scalability has been considered in the previous section. Specifically, the N^2 connectivity problem is the first impediment towards switch architecture scalability in an HPC environment. Using modified Clos architecture and creating a system of conjugated cross-connects is one approach towards HPC scalability. Using multi-degree backplanes is yet another approach towards achieving switch architecture scalability. This has been described in the previous section. Newer scalable architectures involve the use of optical backplane which have also been shown in the previous section. An example of an optical backplane architecture is described below: In Figure 4.5 above, a 200 Gbps cross-connect fabric is created using an optical

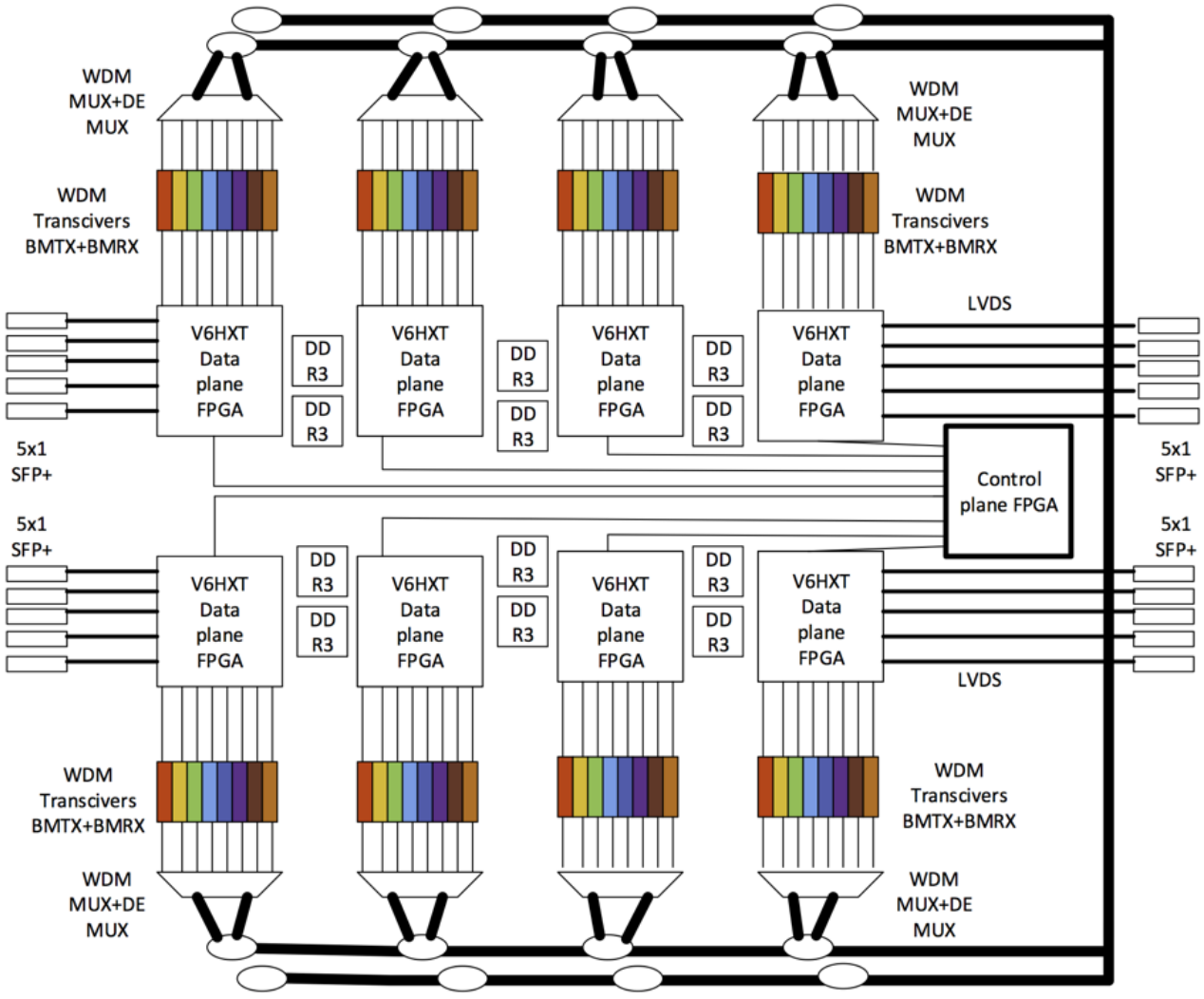


Figure 4.5: 200 Gbps cross-connect fabric

backplane. Each FPGA is part of a processing card, and it emits 10 Gbps data that is modulated onto an optical

transceiver (colored). The transceiver is further connected to an optical channel multiplexer that multiplexes in the frequency domain all the wavelengths, one each from every transceiver in the card. The output of the transceiver is connected to a passive coupler that facilitates adding data into the optical bus. Another coupler is used for drop side communication. The second coupler allows data to be tapped on from any channel. In this way the optical passive backplane facilitates communication between multiple computing IO cards. The FPGA is responsible for protocol support and scheduling data onto the backplane. A control card is used to monitor and maintain the health of the network.

4.4.1 Protocols for Network Computing

In this sub-section we will describe some of the prevailing protocols used in network computing and postulate a Roadmap to describe efficient protocols. Classically the impact of a protocol on a network fabric is quintessential towards making a network efficient. There are many protocols that do similar tasks, and it is important to choose the correct protocol. Unlike large scale networks in the wide-area or even in local areas, the HPC activity is largely within a cluster or a closed environment, implying that proprietary protocols are just as good as standard backed protocols. A protocol is a method that is agreed between two or more systems to achieve a common set of communication goals. Usually, a protocol is backed with some control mechanism that sets about the protocol in motion.

Feature	Multicasting	Dynamic bandwidth	Control abstraction	Bandwidth-Delay	Fault tolerance	Compatibility	Scalability	Low Cost
Consolidation	×	×					×	
Adaptability		×						
Virtualization			×					
Automation				×				
Latency				×				
Interoperability						×	×	
Economy								×
Reliability					×		×	

Figure 4.6: HPC Features - Network Specifics analysis matrix

4.5 Impact of Latency on HPC Environments

One of the critical aspects of HPC performance is the end-to-end latency that is experienced amongst HPC machines. Generally larger a cluster, worst the latency of the system. In fact, the latency increases non-linearly with the HPC size. Latency is due to protocol processing, queuing and lookups at intermediate nodes and switches. The impact of latency is that it adversely affects virtualization. Most applications cannot be made to work efficiently in a system that is impaired by latency. This is a cause of concern in modern HPC environments. Usually the delay incurred in processing a task is the combination of computational latency and communication latency. Computational latency can be reduced by efficient coding practises as well as parallelizing tasks. Communication latency can be reduced by appropriate choice of protocols and faster interconnect methods. To a large extent, communication latency depends on protocol processing. The larger the protocol overhead required to be processed, the higher the latency. Hence, it makes sense to have more efficient protocols that require minimal processing. To this end, one approach is to keep data in the lower layers of the Internet stack as the processing requirements here are lower. To do so, the granularity of processing in lower layers is much higher than the higher layers, which is not always optimal. A trade-off needs to be attained to process data so that latency is minimized while achieving objectives of communication. Latency also has an impact on virtualization. It is generally preferred that latency be deterministically computed in an HPC environment. It is very difficult to achieve a virtualized computing environment if the latency is probabilistic in nature. Many protocols that require complex forward of information are probabilistic in nature. Protocols such as Infiniband, Fiber Channel

Data Center Type	Network Fabric	Protocol	Redundancy
Small (storage centric)	L2 switches	Ethernet/Infiniband/Fiber Channel	1:1
Small (computational)	Dedicated	Infiniband	N:M
Medium (storage)	L2 switches	Ethernet/FCOE, Fiber Channel	1:1
Medium (computational)	L2/L3	Ethernet, modified Ethernet	1:1
Large (storage)	Ethernet	FCOE/Ethernet	1:1
Large (computational)	L3/L2.5	Smart Ethernet/IP/MPLS-TP	1:1
Cloud	L2-L3	Smart Ethernet	1:N

Figure 4.7: Data-center technologies and specifics

and some versions of Ethernet are able to maintain deterministic latency, while most other protocols such as IP and MPLS in cloud environments or Ethernet LANs (primitive Ethernet) are probabilistic in nature.

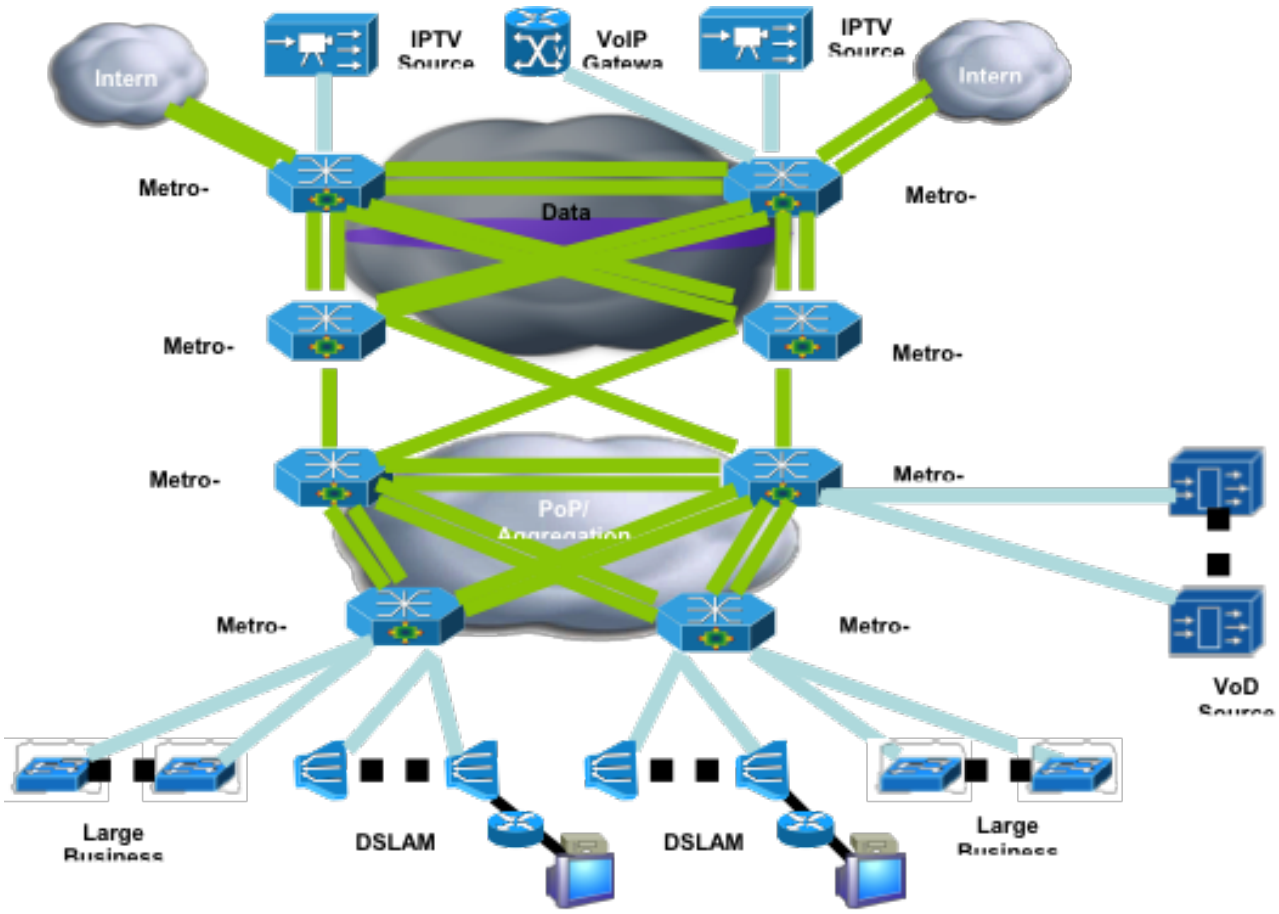


Figure 4.8: An example of a virtualized infrastructure

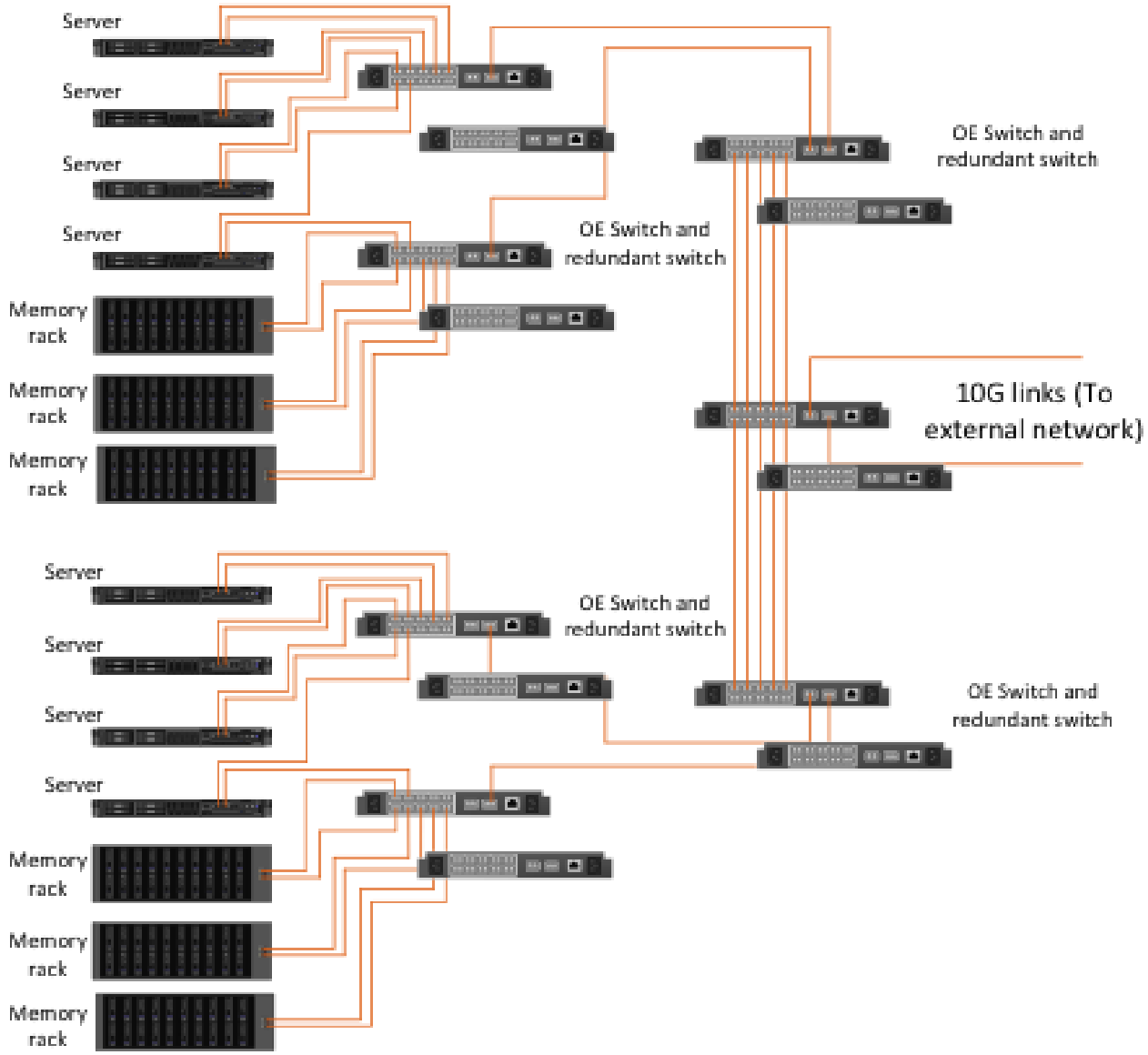


Figure 4.9: Typical Data-center implementation at MTNL using Indigenous Routers designed by IIT Bombay (OE-Switches and Routers)

Chapter 5

Network Computing: The role of Network Virtualization and Network Function Virtualization in the context of Service providers

Ashwin Gumaste
IIT Bombay

5.1 Introduction

Provider revenues are growing primarily based on provisioning next generation services such as video, cloud, mobile-backhaul and data-center. Applications that dominate provider revenues are becoming aggressive in their network requirements [Misra, 2015]. It can be said that if service providers do not reinvent themselves to meet application requirements, their revenue prospects will worsen due to over-the-top vendors capturing much of the newfound e-commerce revenue. For example, video distribution OTT vendors like Netflix, Amazon, Dropbox or Salesforce are cashing in on raw bandwidth pipes provided by network operators. This situation creates a constant feud between network providers and the application providers. In the worst case, a network provider could impede good quality service to application providers as they do not share revenues, given that the network is merely seen as a basic bandwidth pipe. This feud must be resolved for the larger sake of the ecosystem, as explained shortly. Another aspect of this feud is the drive to protect network neutrality. Shown in [FCC-WP, 2015] are multiple aspects of network neutrality. Not throttling someone's service is a given, however, a more important aspect is how to create a new service that facilitates the OTT operator better. It is not a question of how long would it take for service providers to support OTT services, but rather a question as to how to support such a service. Finally, it's a question about not routing packets, but about routing money [Misra, 2015], whether or not we like this ultimate situation.

The scope of this chapter is to study the interaction between the network provider and the application provider through the use of network virtualization as a tool. Network Virtualization (NV) manifests itself as an excellent way to resolve this feud by facilitating the partitioning of the network into qualitative domains that are especially responsible for providing specific service to the application provider.

Our proposal in this chapter is to use NV as an enabler towards solving the paradox between network operators and OTT application providers. Network operators reason that they have to invest in the network infrastructure, license and maintain the network while the application provider uses the network and earns revenue from consumers who are also customers of the network provider and at times misuse the liberties provided by the network provider. Application providers, on the other hand, treat the network as a bunch of bandwidth pipes that pre-exist and do not see the reason to share their revenue. There are merits in both arguments from both perspectives of network and application providers. The deadlock needs to be resolved for both parties to maximize profit as well as serve the end-user better.

This deadlock can be technically resolved by the use of NV implementation in the network. The idea is simple: by using NV in the network, a service provider can now customize services that suite the OTT application. An OTT application provider has now the incentive to share revenue or buy a specific related service that better drives his application to his end-user (the consumer).

The next obvious question is how to implement NV in a network operator. We begin by understanding application requirements at a broad level and mapping them to possible capabilities of networks to offer customized services. Webb et al. [Webb et al., 2011] described ways by which an application can communicate to the network in terms of customization required for a particular application. However, rather than real-time application level changes, it is quite obvious that most OTTs have very specific and well known requirements from the network [Mogul and Popa, 2012]. So can we model a network based on such requirements, mapping these requirements to NV partitions?

To do so, we first understand if it indeed is feasible to model OTT requirements over a service provider network, by isolating key services that would have: (a) strong business case for implementing NV, and (b) have key requirements that a provider can fulfill. To this end, Section 5.2 presents a table that manifests OTT requirements from the network including network technology choices [Gumaste and Akhtar, 2013]. For sake of brevity, we will focus only on the metro and core parts of the network, assuming that the access network pipes are essentially static entities with little scope for technology enhancement due to voluminous users. Section 5.3 presents a method for the Application Service Provider (ASP) to interact with the service provider and then shows how such a method can be implemented in four different technology classes of networks, namely IP/MPLS over WDM, MPLS over OTN with WDM support, Carrier Ethernet+OTN over WDM and IP over Carrier Ethernet+OTN over WDM, each using a software defined control plane. The section also shows how SDN can be made to function in such a scenario and the relationship between SDN and NV pertaining to the technology solutions. Section 5.4 captures results from a simulation model that validates our hypothesis. Section 5.5 discusses the aspect of Network Function Virtualization (NFV) and how it impacts the network. Use cases are also discussed.

5.2 Disparate Networking Requirements

In this section, we discuss application level requirements of various domains and how these can be mapped to network equipment through NV at a high level. Shown in Figure 5.1 is a table containing a list of OTT services that are becoming key revenue generators for application providers. In the second column of the table are network-centric specifics that an ASP desires for a particular service. The third column of the table points towards plausible technology options for provisioning the service while meeting the specific network-centric options. This table was prepared considering ASP businesses that are today valued at least 1 billion USD in revenue [SDN-WP, 2013]. The key driver towards ASP traffic is video and many of the domains shown in column 1 of Figure 5.1 shows services whose primary ingredient traffic type is video or heavy graphic content. We must note that since we are not focusing on the access part of the network, it is safe to say that the traffic is largely B2B in nature but can without loss of generality be extended to B2C model whereby the customer is an access aggregation point. In this table, we have characterized the domain or service by its specific requirement from the network in column 2. In column 3 probable technology solutions have been proposed that can adhere to the requirements proposed in column 2. Column 3 discusses only those technologies that are relevant to the metropolitan and core regions of the service provider network. For many of the applications, there are multiple technology solutions possible and the ones that are commercially viable in a tier-1 provider network have been illustrated in column 3. The key question that this table highlights is how a SP can provision a particular service requirement in the network. To this end, a system must be designed that undertakes interaction between the provider and the OTT ASP keeping in mind the tenets of network neutrality. This interaction must be mapped on to network hardware so that service provisioning is indeed possible. Our proposal is to create an SDN system that would facilitate interaction between incoming traffic requests from ASPs mapping these onto provider hardware that adheres to NV principles. The key challenge in this approach is to: (a) map the incoming demand into network specific parameters that can be used for traffic engineering, bandwidth brokering, provisioning and service support, and; (b) enable the network hardware to be able to provision new services with specific OTT needs.

The challenge in the latter is to be able to create services and differentiate them at the network layer. To this end, we propose in the next section a solution using NV principles to partition provider hardware to meet ASP service goals.

5.3 Building A Solution With VNEP

In this section we describe a method to implement NV to meet specific application provider requirements. We assume that a request for a service arrives into a service provider domain and a *network management system* (NMS) can talk to a SDN controller, which would further provision services. The NMS can abstract specific requests into network-centric parameters for the goal of provisioning services. The NMS maps a service request onto an abstracted network topology by considering specific service parameters that are required for the service. These parameters are then mapped onto all the network elements in the path to check the feasibility of provisioning the service. To check feasibility, there must be a parameterized relationship between incoming service requests and the equipment deployed. The SDN controller maps an incoming request to a network virtualized hardware resource. The idea is that every piece of hardware is further divided into service supporting modules that are parameter-driven. Virtualization happens by the creation of multiple (virtualized) instances of the data-plane at each network element. Each such instance of the data-plane enables OTT-service specific feature implementation.

Domain/OTT Service	Requirement from Network	Technology
Video Services	Guaranteed Bandwidth Low Jitter	IP/MPLS/ WDM/CE
Mobile VAS	Unconstrained Bandwidth Low Packet Drop	MPLS/ OTN/CE
Video Advertising & Merchandise Delivery	Bandwidth on Demand Low Jitter	MPLS/CE
Real-time Events & Entertainment Delivery	Extreme Multicast Bandwidth on Demand	MPLS/ CE/WDM
Healthcare and Telemedicine	Low Downtime, High Bandwidth, Security, Low Latency	MPLS/CE
Defence Networks	Minimal Downtime, Low Latency, Security, Virtualization, Multicast, Bandwidth	MPLS/OTN /CE/WDM
Finance and Banking	Virtualization, Minimal Latency, Security	IP/MPLS/ CE/OTN
Education	Multicast, High Bandwidth	WDM

Figure 5.1: Service-Technology Matrix

5.3.1 Method to implement NV in SP-ASP (OTT) interaction

In this section, we describe how to implement NV in a provider network that offer APIs for service based network virtualization. A request that enters the network is provisioned through the network interface supported by the NMS. For each new incoming request, the NMS computes the optimal network resources to be allocated and provisioned. To this end, the following steps are envisaged at the centralized NMS:

- A route is computed based on service requirements. Actual bandwidth allocation is computed along the route depending on the specified request and other requests at that point of time.
- Each element along the computed route is examined from a service support perspective, whether it can satisfy *specific* requirements of the service.
- To compute the specific requirements of the service request, we propose the concept of *VNEP* or *Virtualized Network Equipment Partitions* that enables a network equipment (such as a switch or router) to be partitioned according to it satisfying some basic parameters. An example on VNEP is provided subsequently.
- If VNEPs are possible along the path to provision the request then all the network equipment are provisioned to meet the new request by the NMS through the SDN controller.
- If however, a VNEP required for the service is not available in one or more network elements along the path, then an alternate path is chosen, such that all the elements along the alternate path have VNEPs that are available for provisioning the service request.
- If no path is possible, then a best effort path may be selected amounting to a lower type of service level agreement.
- It may also happen; that a VNEP created at a node may be moved to another node depending on resource availability over a period of time.

The above steps detail the VNEP computation, which is now described in detail.

5.3.2 VNEP Computation

A VNEP is represented by the physical partitioning of hardware by software such that the partitioned elements correspond to discrete fully functional elements each capable of performing all the functions as that performed by the larger hardware. The key to VNEP creation is to note that the overlaid software creates partitions by allocating hardware resources within a larger network element. Partitions could be created in switching elements, network processing units, buffers and packet classifiers. Partitions correspond to hardware resources as defined by the software and are made available strictly for a particular service or function.

Our conjecture is that a networking element E can be divided into partitions such that a partition E_i can act as a completely independent networking element producing $arg(i)$ properties. We further say that the sum of parts, i.e., the union of all E_i's does not necessarily add up to E for that particular property. Throughput, average latency, packet-loss rate are examples of a property. Let us consider an example: Assume a 60 Gbps switch fabric with finite VOQ (virtual output queued) buffers. Assume 6 input lines, each at 10 Gbps, and 6 output lines, also each at 10 Gbps to this switch fabric. Now, assume that one of the lines is sending data at 2 Gbps, and the average packet size is say 250 bytes and the VOQ memory for storage of packets while contention is being resolved is 3Mb. Obviously, the max ingress to egress latency then is of the order of 0.3 millisecond. However, our desire is to predict average latency. In this case, average latency is a function of the provisioned services at the other 5 ingress ports, the nature of the traffic (distribution), and the switching speed of the fabric (cut-through, store and forward, shared memory etc.)

Since the latency of a flow through a switch also depends on other flows, one way to control the latency is to bind the number of flows that go through a switch. Let us consider an example of this process. A simple 4x4 cross-bar with VOQ (making it a 12x4) switch (each port at 1 Gbps) can take 4 flows each with 250 byte average packet size at full line-rate (wire speed operation), resulting in 1.2 microsecond switching, while the same switch will result in 2.4 microsecond port-to-port latency if the average packet size goes down to 128 bytes [Bidkar et al., 2014]. Similarly, the switch will result in a latency of 3 microseconds if the average frame size is 64 bytes [Bidkar et al., 2014]. The switch behavior becomes more erratic when the standard deviation between flows across multiple ports increases [Das et al., 2013]. For example, the switch will result in a port-to-port latency of 12 microseconds for multicast traffic if the frame size is 64 bytes, and all other ports have provisioned flows with frame sizes 1500 bytes. The above discussion shows the complex relationship between frame-sizes, port counts, traffic distribution (random, unicast, multicast) etc. implying that for carrier-class services, i.e., those with deterministic networking parameters of delay and jitter, predicting switch behavior is important.

The paradox just presented also implies that predicting switch behavior is difficult in real-life. Even the most intricate queuing models are difficult to converge in terms of producing exact parameters. Frequent use of such G/G/1 models will prove even more difficult to solve and are unusable to provision services in real-time by the NMS. So our approach is to provision services without getting involved in the intricacies of computing switch-specific parameters in real-time.

In our approach, we partition the switch into VNEPs that can individually provision services. The idea is to dynamically create a VNEP that will adhere to all the system-wide parameters for a particular service, with the constraint that the sum of all the VNEPs in a network element (switch) are less than the total capacity of the switch. The union of VNEPs is not linear. This implies that the system leads to over-provisioning which though undesired, is required to maintain many of the carrier-class attributes necessary for OTT services.

VNEP creation and sizing involves the following steps:

1. We identify a set of parameters that would be crucial to a service request.
2. A network element is then viewed as a number of instances qi of a particular parameter i such that $f(qi) = a$ constant exemplifying total network equipment availability for that parameter.
3. When a service using parameter i is provisioned, its impact on other parameters is given by $qj(i)$, and $f(qj)$ includes $qj(i)$.

Let E_x indicate a service parameter for a networking element. We note that for every E_i there is a corresponding E_j i.e., when we instantiate parameter i it will also have an impact to how many instances of another parameter j we can have in the same switch.

Consider the last point with the support of an example. A 60 Gbps cross connect fabric can support say 6x10 Gbps connections with 250 byte average frame size and 3 microseconds port-to-port latency. The same fabric will have a 12 microsecond latency for the same number of flows if the average frame size reduces to 64 bytes. The delay can go up almost exponentially if the number of flows increases to 60 flows of 1 Gbps each. So, now if we have to provision a service of 1 Gbps with a latency requirement of 3 microseconds, and another service of 5 Gbps with a latency requirement also of 3 microseconds, then how do we do so given that the frame size of the first service is say 128 bytes and the second one is say 64 bytes? Obviously the second service will require more over-provisioning as compared to the first one, i.e., to say that though the second service is 5x of the first service, in order to achieve similar parameters, the second service may have to be provisioned through the switch with 12x resources (buffers primarily) so that the switch can meet the provisioning requirements. Now how do we arrive at the number 12x? This number is a function of both volume and quality – volume, as in how much more would the service take in every parameter, and quality, as in what would be the impact of the service provisioning in other parameterized domains.

Shown in Figure 5.3 is the actual process for creating VNEPs and allocation of VNEPs. We compute from an incoming request(i) the implication that a particular partition would imply to other partitions. The SDN controller first computes VNEPs for each service at each network element using the aforementioned process. The controller then sends specific information to each node to partition itself according to its VNEP computation. Partitioning happens based on the use cases discussed in Section 5.3.

VNEP Partitioning analogous to VM creation and migration: VNEP creation and NV using VNEPs is analogous to virtual machine creation and migration in data-centers/cloud environment across CPUs. Shown in Figure 5.3 is the analogy of the forwarding plane in a network element with a VM hypervisor system in a multi-core processor. VMs can be created in a processing environment on-the-fly. The same analogy is used for VNEP creation, where by VNEPs are like VMs – created on-the-fly and use the switch fabric resources in an independent way. As can be seen in Figure 5.3, VNEPs are created by the control plane (SDN-based), and implemented within the network element through NV.

5.3.3 Partitioning Network Equipment using NV

In this sub-section we discuss how partitioning of network equipment using NV from a technology perspective is achieved. We examine NV through 4 technology-centric use cases that in principle correspond to the OTT services shown in Figure 5.1.

Technology Case 1 – IP/MPLS over WDM

In this case we assume an IP/MPLS network overlay and a WDM ROADM based underlay. The IP/MPLS routers (label switched routers or LSRs) can be partitioned based on how many flows are supported. Similarly the WDM ROADM equipment can be partitioned to support non-blocking and reconfigurable connections. Full flexibility of VNEPs in the ROADM is possible when the ROADM supports, colorless, direction less and contention less (CDC) properties. A VNEP in an LSR is an MPLS tunnel, while a VNEP in a ROADM is a wavelength. MPLS as such has no label or identifiers to indicate granularity, and hence VNEP information

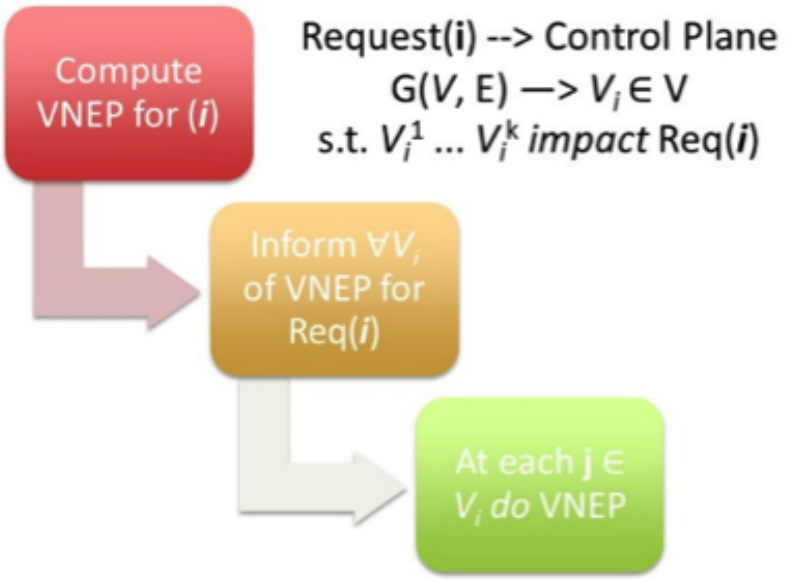


Figure 5.2: VNEP computation and information

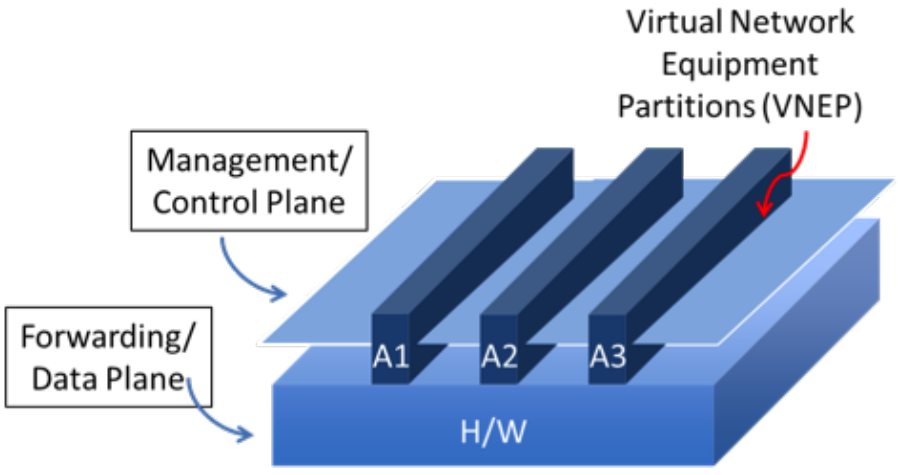


Figure 5.3: VNEP and VM migration analogy

pertaining to how much part of a data-forwarding plane should be reserved is local to the MPLS router (switch-fabric). The SDN controller indicates through the north-bound interface of the equipment the exact amount of partitioning of resources that is to be allocated for a particular service. In this section, we have proposed 3 simple policies to create VNEP partitions.

Technology Case 2 – MPLS over OTN with WDM

In the case of MPLS over OTN with WDM underlay, all VNEP partitions must take into consideration OTN pipes at MPLS LSR interfaces supporting the WDM network. We will assume that the wavelength assignment and lightpath creation policy is beyond the scope of VNEP creation for this use case, as most services are only sub-wavelength granular implying wavelength assignment as a multi-service aggregation and provisioning problem. Hence in case of MPLS over OTN, the partitioning happens at two levels: (1) partitioning the LSR forwarding plane into switching chunks for a particular service and (2) partitioning the OTN based ODU (optical data unit) switching fabric such that each MPLS LSP (label switched path) is mapped onto an OTN ODU tunnel in the egress port that is connected to the appropriate next hop MPLS LSR. Note that with ODU-flex we get extreme flexibility in terms of switching granularities.

Technology Case 3 – CE+OTN over WDM

In this case, we partition the Carrier Ethernet switch fabric into discrete switching chunks so that an Ethernet Switched Path (ESP) is mapped onto an OTN ODU port. In this use case, we do not assume the existence of an ODU-cross-connect. OTN technology in this case is used as distance enhancers with forward error correction.

Technology Case 4 – IP over CE+OTN over WDM

In this case, we have IP routers at select locations as an overlay over a CE network underlay, which is a further overlay on a ROADM based WDM network. Whenever a service has granularity that is near to a wavelength's full capacity, it is routed all-optically by the ROADM. Whenever, a service can be routed at layer-2 through the use of an ESP, it is done so using the CE network. Similarly, whenever a service cannot be routed all-optically or through the CE network or such provisioning is deemed ineffective by the controller, then it is routed through the IP layer. The IP layer may further be used for aggregation as well as provisioning multicast traffic. VNEP information created by the centralized controller may be for partitioning of switching resources at any or the entire CE/IP layers. A summary of VNEP examples is provided below.

Technology	Instance	VNEP Example
IP/MPLS over WDM	Flows (LSR) Wavelength (WDM)	MPLS Tunnel (LSR) ROADM (WDM)
MPLS over OTN with WDM	MPLS LSP	MPLS LSP mapped onto OTN ODU tunnel
Carrier Ethernet+OTN over WDM	Ethernet Switched Path (ESP)	ESP mapped onto OTN ODU port
IP over Carrier Ethernet+OTN/ WDM	ESP (CE layer) or ESP on a lambda	ESP mapped onto OTN ODU port or ESP on a lambda

Figure 5.4: VNEP Examples

Figure 5.5 presents a relationship diagram between SDN and NV. As can be seen there is significant amount of dependence as well as complementary relationship between SDN and NV in terms of what they can do in conjunction especially leading to network agility and achieving scalability. SDN is able to abstract a dedicated programmable control plane that can be used to control NV-compliant hardware. This section details how such interaction can happen in terms of policy mapping of SDN-centric features on NV-compliant hardware. It is important to note that in a provider domain some critical SDN-centric manifestations can be implemented through the use of NV-compliant devices. In addition, implementation of NV results in excellent CAPEX and OPEX reduction as well as revenue enablers.

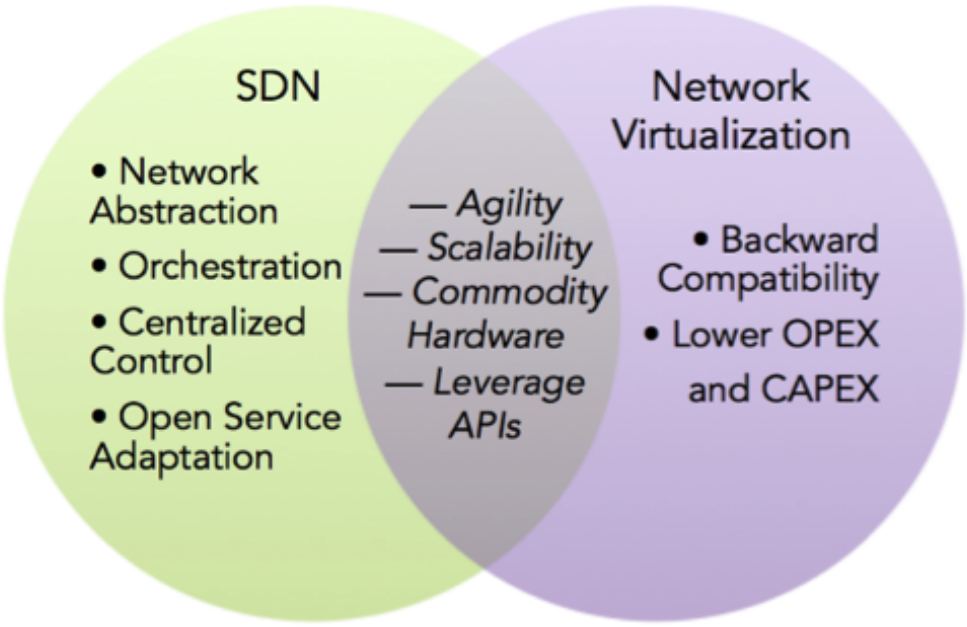


Figure 5.5: SDN and NV Relationships

Shown in Figure 5.6 is switch architecture to implement SDN with NV. A controller is shown that is connected to the north-bound interface of the switch. The switch could support one or more of the layer 2/3 protocols and the interfaces would be mapped onto wavelengths. Incoming flows are segregated at the input buffers (which are further segregated to support virtual output queues). Flow headers are worked upon by the control state machine that uses SDN tables. The tables are populated through the controller. Entire protocol functioning happens at the controller. To support scalability, we assume that the controller runs on a VM.

We now propose 3 policies for VNEP partitioning that are instructive for our simulation model in Section 5.5.

Policy 1: **Throughput maximization** – In this policy, VNEP computation involves maximizing the throughput at every network element. This is a non-carrier class policy implying that the port-to-port latency per network element is not deterministic. The policy implies an additive increase of throughput and hence, whenever a new request arrives at the SDN control plane, a VNEP is created with a view to maximize network-wide throughput.

Policy 2: **Latency bounded partitioning** – In this policy, a VNEP is created such that the corresponding service is guaranteed to meet its end-to-end latency requirement through every network element having bounded latency.

Policy 3: **Latency sensitive-service maximization (LSSM)** – This is a slightly more global policy compared to the previous policy. In the LSSM policy, the approach is to maximize the number of services through a network element. To do so, the controller creates VNEPs such that they balance each other in terms of parameterized requirement. For example, similar delay constraint and similar bandwidth requirement services may be load balanced over the network. Another strategy used by the controller is to provide for equal cost multiple paths (ECMP) to load balance the service further. In LSSM, care is taken to ensure that the number of services is maximized.

In our simulation model, we rationalize service requirements based on their utility to the network (revenue the provider makes), and normalize the utility over delay constraints. We then provision services such that the delay constraints are met while bundling as many services together. The LSSP policy is a heuristic greedy approach and its complexity is of the fourth-order in terms of number of links in the network. As a result, its functioning also depends on graph size.

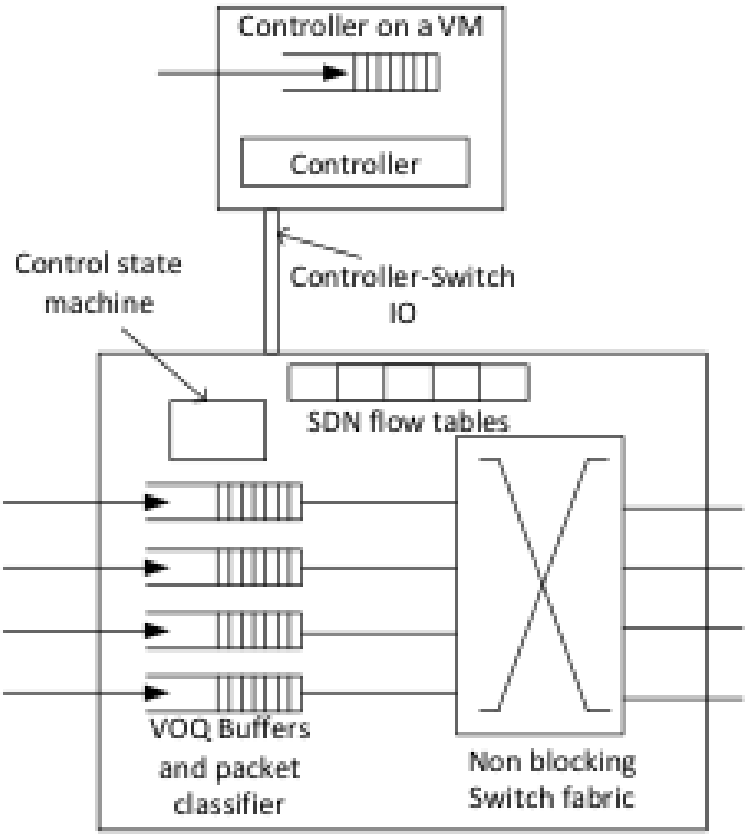


Figure 5.6: Switch Architecture to Implement SDN

5.4 Simulation Model and Hypothesis Verification

In this section we describe a simulation model that was built to test our hypothesis pertaining to the interaction between service providers and application providers using network virtualization. We model a large provider network with two autonomous systems (AS), 5 metropolitan regions, with each metropolitan region further divided randomly into 20, 40, 60, 80 and 100 access regions. The backbone and metro networks use fiber, while the access networks could be wireless, FTTC (fiber to the curb) or coaxial cable based. Our goal is to evaluate the impact of NV over different technologies by provisioning OTT services. To this end, the simulation model implements each technology solution using proposed VNEP creation policies.

Each access region has a random number of subscribers between 10,000 and 100,000. An access region is connected to a metro network, and multiple metros are backhauled in the AS to a core network. The point of presence (POP) connecting the access to the metro is a fiber termination point and can support ROADM equipment. The overlay depends on the technology being simulated and we study IP/MPLS, MPLS, OTN and CE technologies. The control plane is implemented as an SDN overlay that consists of controllers, one each for an AS of 10K users and hierarchically arranged thereafter.

The simulation model works as follows: service requests are generated randomly and are assigned parameters (also a random distribution). Services are organized into two levels – services and sessions. Services are exponentially distributed with a mean holding time of 6 month time-frame while session holding time is exponentially distributed with a mean time equivalent to a 100 MB video-file download session. The service is guided to the appropriate controller, which uses one of the three VNEP creation policies and computes if provisioning is possible. Services are lumped together through pre-assigned aggregation policies. Once a service is provisioned, we compute service and switch statistics with counters at each node. Load is computed as the average occupancy of all the incoming services to the maximum allowable input rate over all the ingress ports. MPLS LSRs are assumed to have 1 Gbps and 10 Gbps interfaces and a net switching capacity of 640 Gbps, while, CE switches are modular with 1 Gbps and 10 Gbps interfaces leading to 80 Gbps modules that can be stacked up together to create a 640 Gbps node. ODU switching where applicable is assumed at ODU0, ODU1 and ODU2e. Transport wavelengths can be generated by MPLS/CE/IP forwarding plane or a stand-alone muxponder and support 10 Gbps, 40 Gbps and 100 Gbps paths. Network cost is computed as in [Mathew et al., 2015] for both CAPEX and OPEX, while we assume that for provisioning OTT services, the OTT ASP shares 20% of its revenue with the service provider.

Shown in Figure 5.7 is a comparison of all the three policies used for VNEP computation using MPLS and

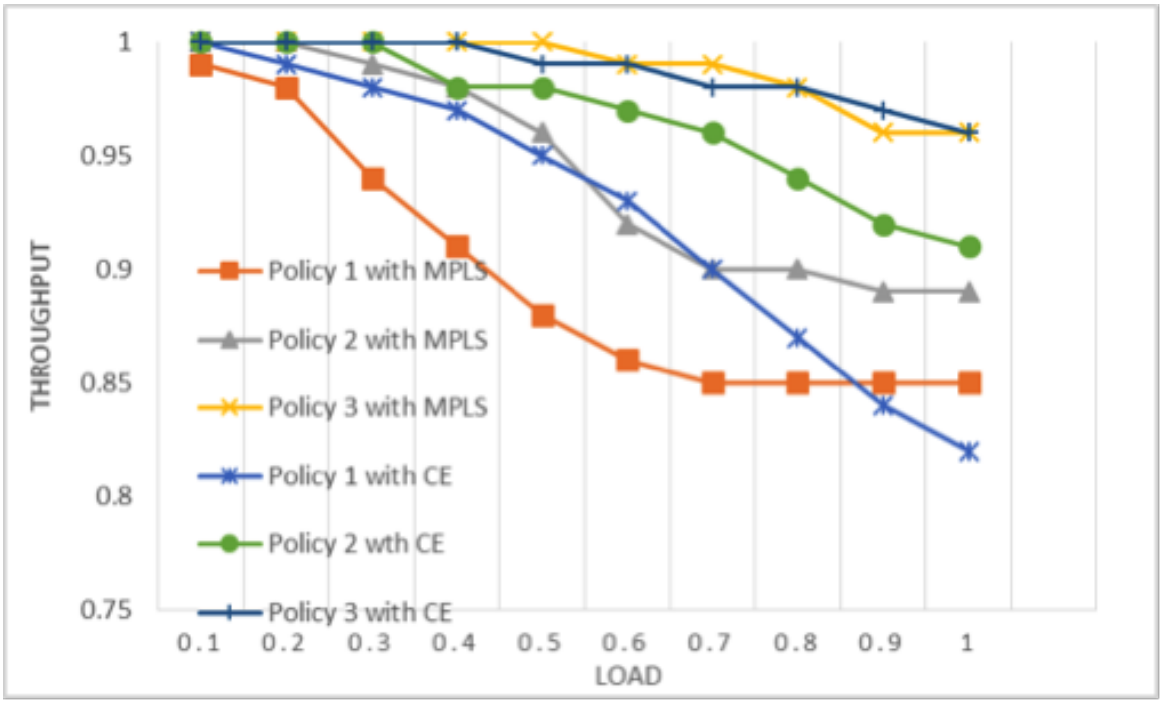


Figure 5.7: Throughput as a function of load for different policies

CE technologies. We show throughput versus load in the entire network. MPLS and CE were chosen because of cost considerations and hence likely to be deployed for NV. A peculiar behavior is that policy 3 has the best throughput though it also takes into account service latency.

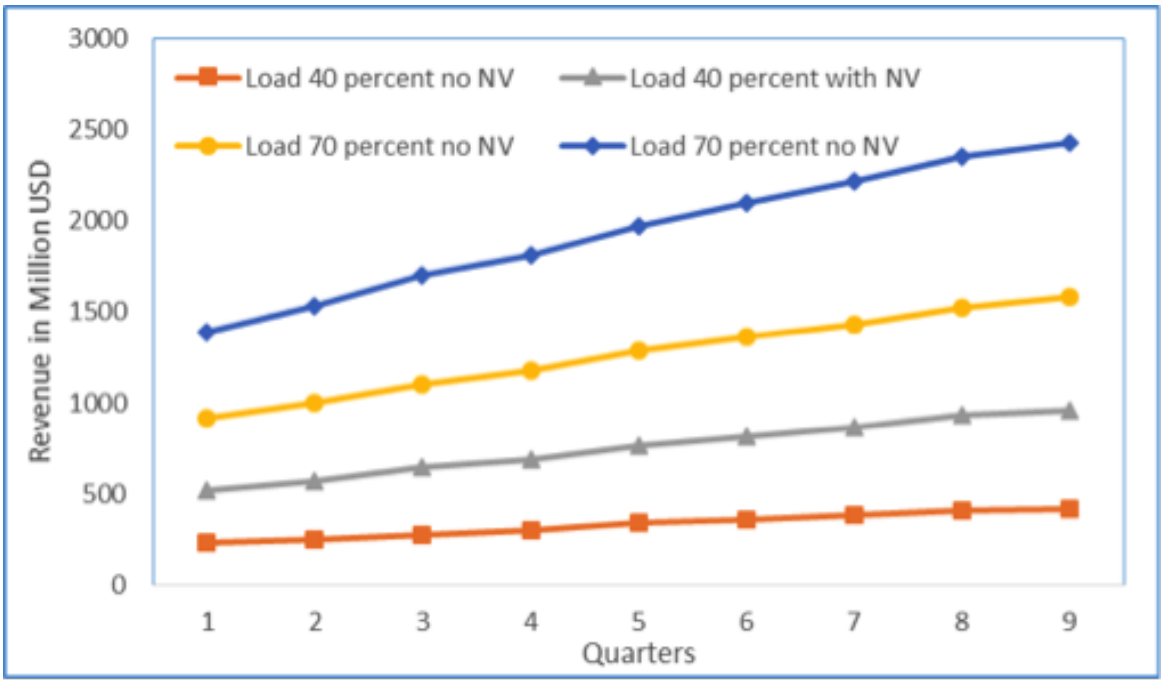


Figure 5.8: Service provider revenue with and without ASP revenue sharing through NV

Shown in Figure 5.8 is the revenue impact and reach of ASPs with and without revenue sharing with the ISP. This figure shows that there is sizable incentive for ASPs to share their revenue, as the providers would be able to grow the network thereby facilitating larger and qualitatively superior reach for the ASPs. Figure 5.8 is generated as follows: We first measure ASP revenue without NV, and no revenue sharing. Then we re-compute the revenue by stating that each service is now priced 20-40% higher than what it was priced before depending on the parameters being adhered to through NV. For example, a 12 Mbps HD-video pipe was priced 20 dollars per month with no sharing of revenue and hence no NV support. The same pipe with guaranteed bandwidth (no

packet loss) is priced at 24 dollars, while it is priced at 28 dollars with bounded latency and 50 ms restoration of service in case of fiber cut/equipment failure.

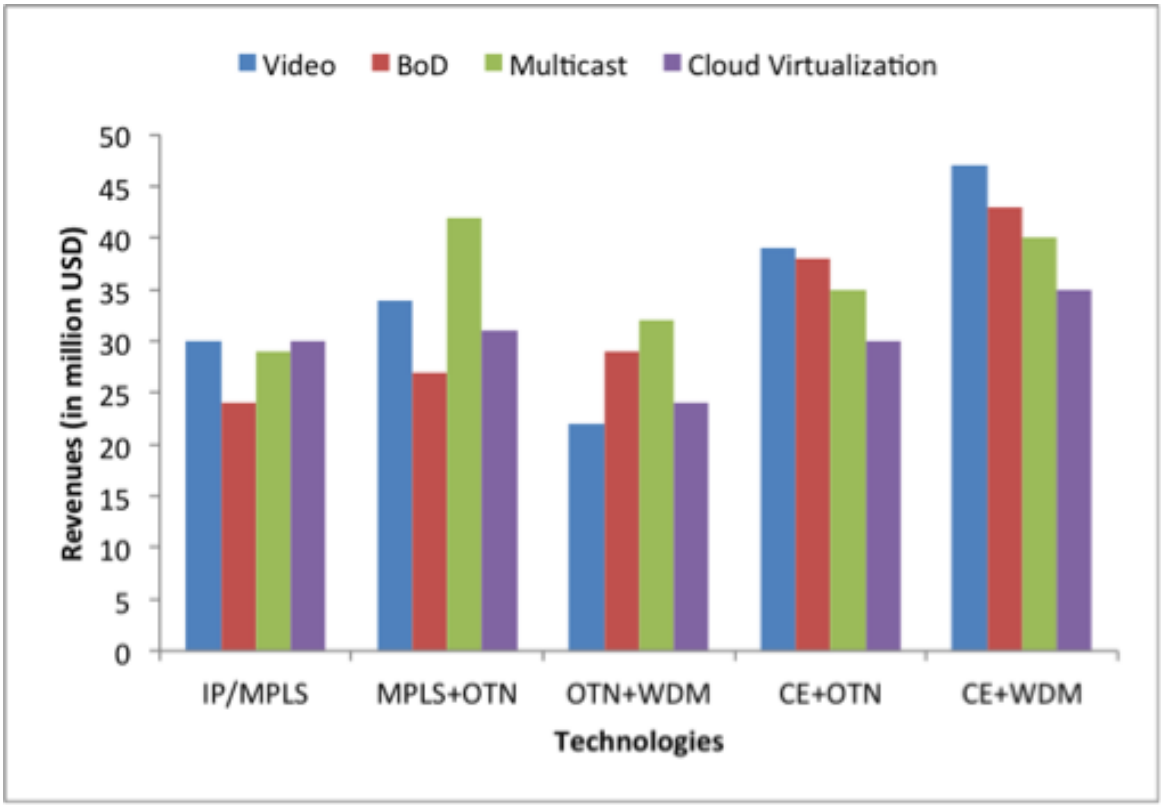


Figure 5.9: Technology comparison for increase in revenue

Shown in Figure 5.9 is a comparison of increase in revenue using different technologies for different services. This figure was created to observe which services will have an increase in revenue due to NV implemented by the SP. NV is implemented through policy 3. We consider 4 services, video, bandwidth-on-demand (BoD), multicast-video and cloud virtualization. Video requires guaranteed bandwidth and error-free delivery; BoD requires the ability to change the size of bandwidth pipes between 1 Mbps and 250 Mbps in increments of 1 Mbps; multicast requires guaranteed delivery to a number of users (100-10,000) in increments of 100 users, and pipe sizes from 12-24 Mbps (multi-HDTV streams); cloud virtualization requires multiple enterprise sites (4-40) connected with guaranteed bandwidth, bounded latency and bounded jitter. It is seen that the CE solution is best for video and BoD, while the multicast service is best provisioned through an MPLS solution (as expected). There is not much difference between MPLS, OTN and CE technologies for cloud virtualization service.

Shown in Figure 5.10 is the impact of VNEP on throughput in the network as a function of the ratio of point-to-multipoint traffic (MP2MP) to point-to-point traffic (defined by Omega). The graph is generated for the case MPLS. As Omega increases, i.e., percentage of MP2MP traffic increases, the throughput without VNEP decreases rapidly while the throughput with VNEP decreases much more gradually. This is a critical result showing the maximum benefit of the use of VNEP. The graph shows how VNEP can impact new service support such as multicast services, which are generally poorly handled at higher loads.

5.5 Network Function Virtualization

Network function virtualization is being heralded as the next big thing in service provider networks with a promise of changing the entire landscape of telecommunications from both a new service offering as well as a lower CAPEX/OPEX perspective. Network function virtualization is a transformative technique that facilitates better programmability in the entire networking domain. To be specific, NFV is a method by which various network functions that were earlier conducted by specific hardware can now all be virtualized. As part of this virtualization process, NFV allows the network to be lumped into processing entities – common servers and server farms in data-centers that interact with data streams to facilitate network functions to be virtualized. The base premise of NFV is that we can now have commodity hardware in the network that would replicate network functions, which were expensive to deploy using specific hardware as well as involved intricate traffic engineering for excellent resource utilization. With NFV the goal of cost reduction through low-cost commodity hardware

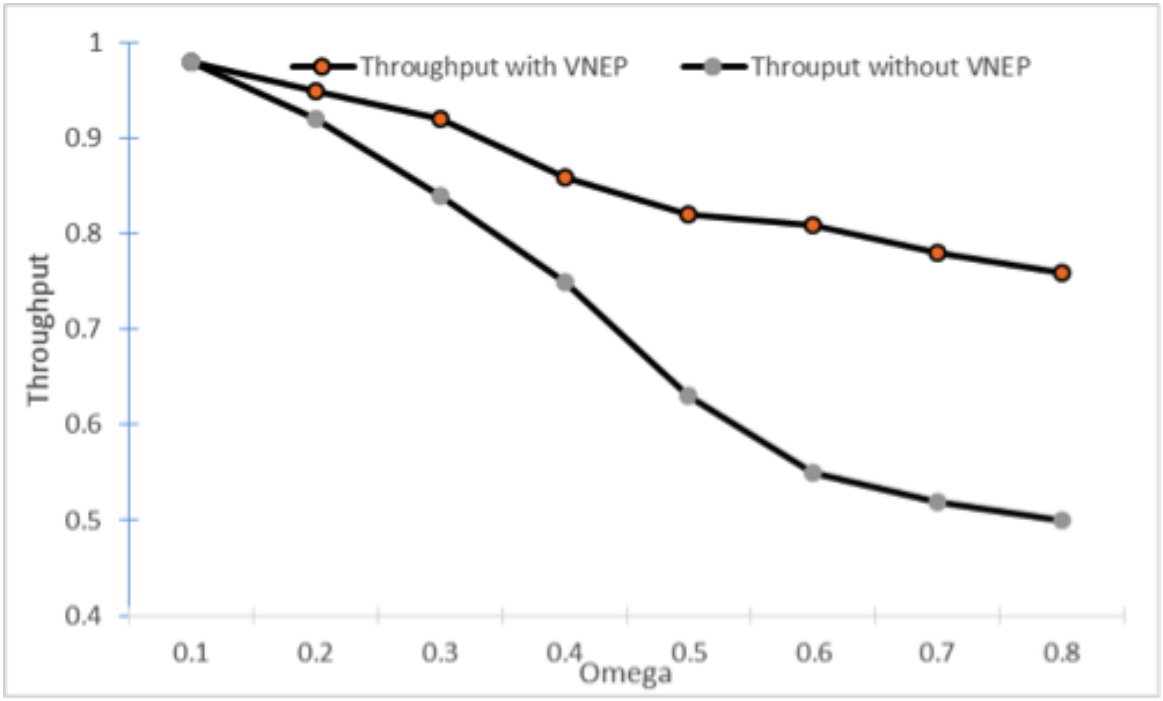


Figure 5.10: Throughput versus ratio of MP2MP/P2P traffic

is well achieved, however the aspect of traffic engineering does get further complicated. NFV allows for the efficient usage of data-centers which so far were primarily storage entities with enterprise functions embedded to now being used as entities that would facilitate all of service provider key functions such as switching, routing, security, policy management, control, service provisioning, automation, billing and fault management. With a centralized approach one of the key advantages that NFV beings about is massive consolidation of various resources. Consolidating resources is done by ensuring that network functions are not distributed and lie idle in a network but are instead acted upon at all times through the virtualization process. NFV allows *service chaining* which essentially implies that a particular service is now processed at different entities i.e., in a chain so that the entities processing the service are optimized.

NFV based network optimization is an essential role in network computing for future telecommunication applications. There are two kinds of NFV deployments that providers are dealing with – those that are stand alone with commodity gear (processors) in the network and those that are provisioned with the use of Software Defined Networked white boxes (SDN white boxes). NFV deployments using commodity processing units in the datapath of the network have the advantage of cost and programmability. Cost – because the gear is essentially a commodity entity, and programmability because the processor based entity manifests itself as an excellent programmable platform. The problem with programmable commodity hardware is that it does not support wirespeed operations especially at high line rates. The programmability is very useful but due to the limitation of line-rate it is not possible to make it work at all times for all applications and across all layers. One approach to deploying NFV based commodity hardware is to encapsulate network functions especially those at higher layers such as firewalls, intrusion detection and policy based session controllers in commodity hardware. The reason to do so is that many of these functions can be at lower line rates implying an ease of implementing these at wirespeed operations. Another reason why higher layer functions are better suited to such commodity hardware in the network is that they require constant updates and hence programmability which is readily possible using commodity processing units. A third reason of using such computing equipment is to also ensure distributed processing and efficient service chaining. The SDN based approach is a more stable approach to NFV implementing in the network. In the SDN approach, white boxes are used that allow population of SDN flow tables based on which forwarding is carried about. Flow tables in an SDN box enables a provider to program the whitebox as per its requirement of the service. New service definitions are now possible. The approach of implementing NFV without SDN mostly today implies putting commodity hardware in data centers that do the processing. The network then functions as a routing entity and routes data between peers through appropriate data-centers that house servers running appropriate software implying effective service chaining. NFV has the potential to bring down the network cost significantly. NFV based networking facilitates faster service velocity, i.e., facilitates the speed at which a new service can be enabled. In summary, NFV seems to be a good technique to revolutionize the use of computing for communication.

5.5.1 NFV use cases

We now discuss NFV use cases.

Use case 1 – Enterprise services:

Enterprises form almost 80% of provider revenue in the metropolitan region. There has always been a traditional gap between what enterprises desire and what providers can offer. This gap is primarily due to the difference in telecommunication technology from an application's standpoint. Most of the telecom services that a provider offers tend to be rigid and difficult to grow from the requirement of enterprises. Enterprises need to meet their customer and internal needs and need to be able to achieve these needs in quick time. The latter is especially important and constitutes service velocity aspects of provisioning connections and new services between the provider and the enterprise. Another important aspect of enterprises today is that they tend to not want to invest heavily into networking entities, especially specialized entities and hardware. A service provider that deploys NFV can be of good use to an enterprise seeking to deploy specialized services or seeking to outsource its networking needs. Now a provider that connects to the enterprise maps all the enterprise network functions that the enterprise wants to offload on to the providers' data-center. The provider data center is now almost like an extension of the enterprise. Network functions are implemented in the data-center on servers and these functions reside on VMs, which imply that the enterprise is allocated a bunch of VMs in accordance with the service chains that the enterprise subscribes to. Service chaining is dynamic and hence a new service can be provisioned just by aligning the VMs in accordance with the requirements of the new service and this leads to much faster service delivery. Many traditional networking functions that would have been deployed in the enterprise are all now offloaded to the data-center. This leads to optimization of software licenses as well as effective usage of resources in the data-center which otherwise would be suboptimally used within an enterprise. Network functions now can be tailor-made to suit enterprise needs, and these functions are embedded as software applications on VMs. One of the key engineering decisions to be made is whether the price and service velocity benefits of NFV is good enough as compared to the suboptimal routing and performance using commodity gear. This trade off needs to be solved at an individual enterprise level, and this trade off is of great significance from a planning perspective. Another important aspect of deploying NFV in provider networks is to understand the time and network state at which point should a provider make a transition to deploy NFV. Further, how much part of a provider network should be transitioned from a conventional hardwired specific function oriented network to a network function virtualization based computationally sensitive network. These questions would have to be dealt with at a meta level as well as on a case by case level to understand how NFV can become an important weapon in the arsenal of providers to meet enterprise needs. Smaller enterprises may outsource their entire IT operations to the provider, while larger enterprises may outsource only those applications that have a direct benefit from a providers perspective onto the enterprise network.

Use case 2 – Access CPEs:

The access part of the network is critical to a provider as it is most customers facing and interconnects the core network to voluminous users. Access networks deploy a plethora of technologies from wireless, to wire-line and which often tend to get upgraded. Access networking also involves tremendous variations of billing and management functions. Deploying such transport entities along with software is a herculean task and the cost of such deployments tend to severely bleed a service provider. NFV comes in as an attractive low cost option that can facilitate service migration and delivery in the access area. The key to deploying NFV in the access is to facilitate virtualization of access devices. To this end customer premise equipment or CPE are virtualized using NFV. CPE though low cost is severely impacted by new technology adoption and leads to large cost cycles for upgrade. Hence by facilitating virtualized CPEs we can migrate from an older technology version to a newer one in quick time and saving CAPEX. Examples of access devices that are currently been virtualized include optical line terminals or OLTs, ONUs, DOCSIS3.0 modems, Wi-Fi access points, LTE backhaul equipment and gateways. The virtualization happens in commodity hardware connected to layer 2, layer 2.5 switches.

Use case 3:

NFV in the Data center: Another important use case for NFV is the use in a data-center. NFV is done through applications on servers (on VMs) that are housed in a data-center. Key to note is that now the data-center becomes a repository as well as a computational entity. NFV in the data-center also enables the provider to consolidate licenses severely.

5.6 Discussion

We have presented an approach to integrate OTT application providers with service providers by the use of network virtualization. A comprehensive approach to adapting NV in pragmatic service provider networks is discussed. We propose the concept of virtual network equipment partitions (VNEPs) that enable a network element to be partitioned as per service requirement, thereby benefiting from programmability of the control plane. Policies to partition a network element are discussed. Results from a simulation study show the benefit for ASPs in a provider network using NV-compliant hardware. We have also discussed about the emerging role of network function virtualization or NFV and how it would impact network computing. We have precisely understood NFV through 3 dominant use cases that would have a strong impact on network computing in the years to come.

5.7 Takeaways: Software Defined Networks

Software defined networking (SDN) is heralded as the next big thing in the field of high-speed networking in core, metropolitan and access networks. The premise of SDNs is based on bifurcation of the data plane and control plane to create a service centric networking environment. SDN leads to programmability in the networking domain thus facilitating new services to be provisioned without the need for specialized hardware. SDN in a way re-looks at the way we have been doing networking. SDN has taken the industry by storm, and there is hardly any major network vendor in the world who is not migrating their product portfolio to SDN based products. Almost every service provider in the developed world is migrating towards an SDN enabled networking core. For example AT&T has declared in its domain 2.0 that 75% of its network would be SDNized by the year 2020. SDN implies using a hierarchy of controllers that facilitate service provisioning in commodity white boxes. This means that the control plane provides for all the intelligence and runs protocols that were so far run using distributed apparatus. Such distributed apparatus in the past included routers and MPLS equipment which was specialized for certain services only. Such equipment was not programmable from a new service requirement.

Indian networks have so far primarily aped other large enterprise networks – for example, the NKN is an IP/MPLS network, when it is well known that IP/MPLS does not cater to carrier-class technologies that are required for achieving the robustness of service support. To this end, it may perhaps be prudent for Indian networks to begin adapting SDN type technology.

In that context at IIT Bombay we have designed a series of SDN capable routers called Carrier Ethernet Switch Routers, that can support any type of services by deploying programmability at the data plane through a centralized controller. Such SDN centric technology can meet the service needs of the country using indigenous technology and at price points of our choice.

Bibliography

[Bidkar et al., 2014] Bidkar, S., Mehta, S., Singh, R., and Gumaste, A. (2014). On the Design, Implementation, Analysis, and Prototyping of a 1-μs, Energy-Efficient, Carrier-Class Optical-Ethernet Switch Router. *Journal of Lightwave Technology*, 32(17):3043–3060.

[Das et al., 2013] Das, S., Parulkar, G., and McKeown, N. (2013). Rethinking IP core networks. *IEEE/OSA Journal of Optical Communications and Networking*, 5(12):1431–1442.

[FCC-WP, 2015] FCC-WP (Nov. 2015). Protecting and Promoting the Open Internet. https://www.fcc.gov/document/protecting-and-promoting-open-internet.

[Gumaste and Akhtar, 2013] Gumaste, A. and Akhtar, S. (2013). Evolution of packet-optical integration in backbone and metropolitan high-speed networks: a standards perspective. *IEEE Communications Magazine*, 51(11):105–111.

[Mathew et al., 2015] Mathew, A., Das, T., Gokhale, P., and Gumaste, A. (2015). Multi-layer high-speed network design in mobile backhaul using robust optimization. *IEEE/OSA Journal of Optical Communications and Networking*, 7(4):352–367.

[Misra, 2015] Misra, V. (2015). Routing Money, Not Packets. *Commun. ACM*, 58(6):24–27.

[Mogul and Popa, 2012] Mogul, J. C. and Popa, L. (2012). What We Talk About when We Talk About Cloud Network Performance. *SIGCOMM Comput. Commun. Rev.*, 42(5):44–48.

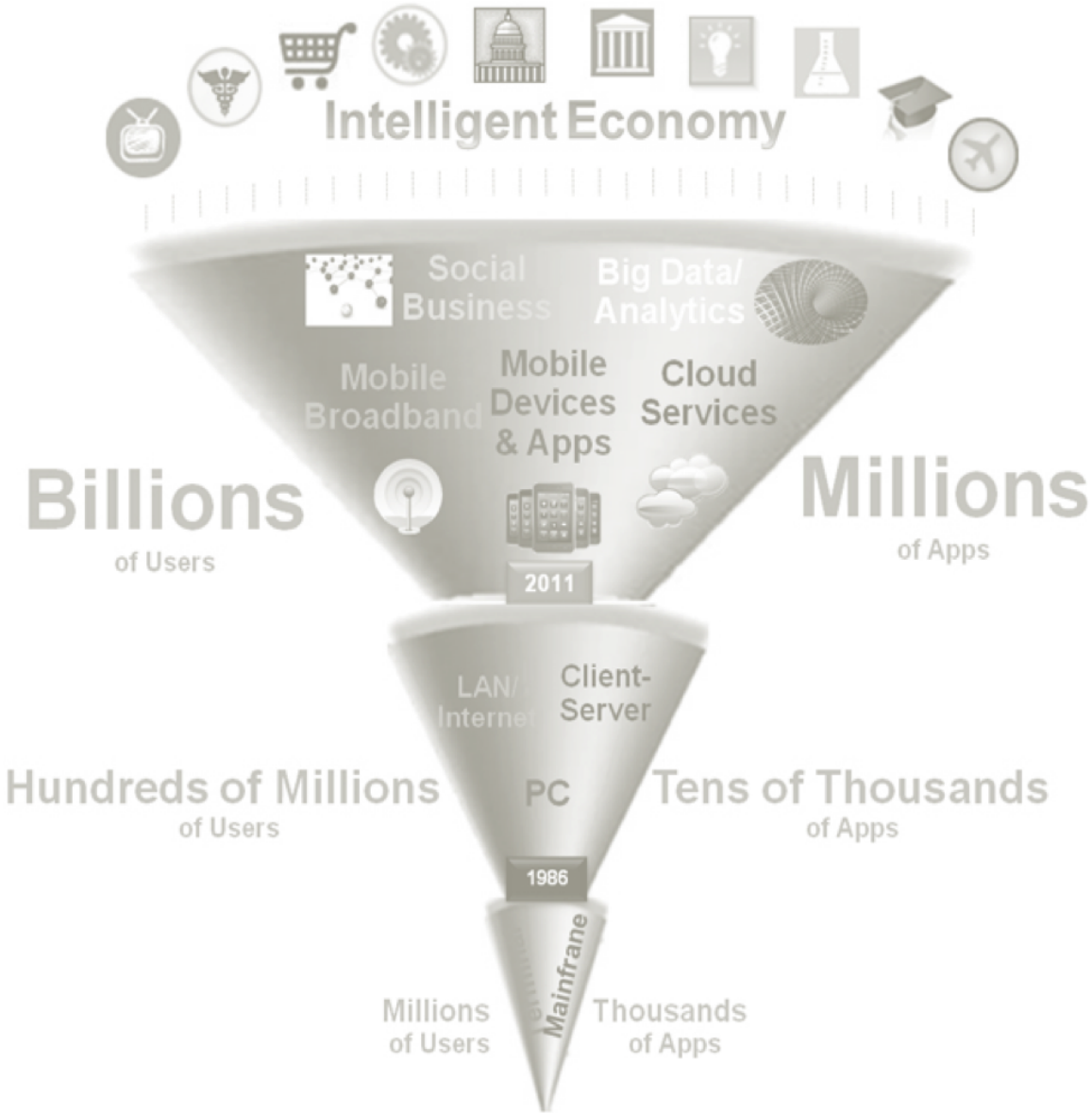
[SDN-WP, 2013] SDN-WP (2013). Data Center and SDN Market Highlights. http://www.infonetics.com/pr/2013/Data-Center-and-SDN-Market-Highlights.asp.

[Webb et al., 2011] Webb, K. C., Snoeren, A. C., and Yocum, K. (2011). Topology Switching for Data Center Networks. In *Proceedings of the 11th USENIX Conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services*, Hot-ICE'11, pages 14–14, Berkeley, CA, USA. USENIX Association.

Chapter 6

Big Data in Government

Vipin Chaudhary
SUNY Buffalo



6.1 Background

In this modern digital era, gleaning uncommon insights from Big Data has become a crucial element for innovation, scaling science discoveries, excellence, and survival in a global economy. Today Big Data is prevalent in almost every stratum of the world ranging from Science to Society and has opened up new horizons in a wide variety of fields ranging from health-care to security.

In this chapter, we uncover the facts about Big Data and present representative examples of the advancements that have been made with regards to Big Data in four key areas of national interest, specifically Education, Energy, Health-care, and Security.

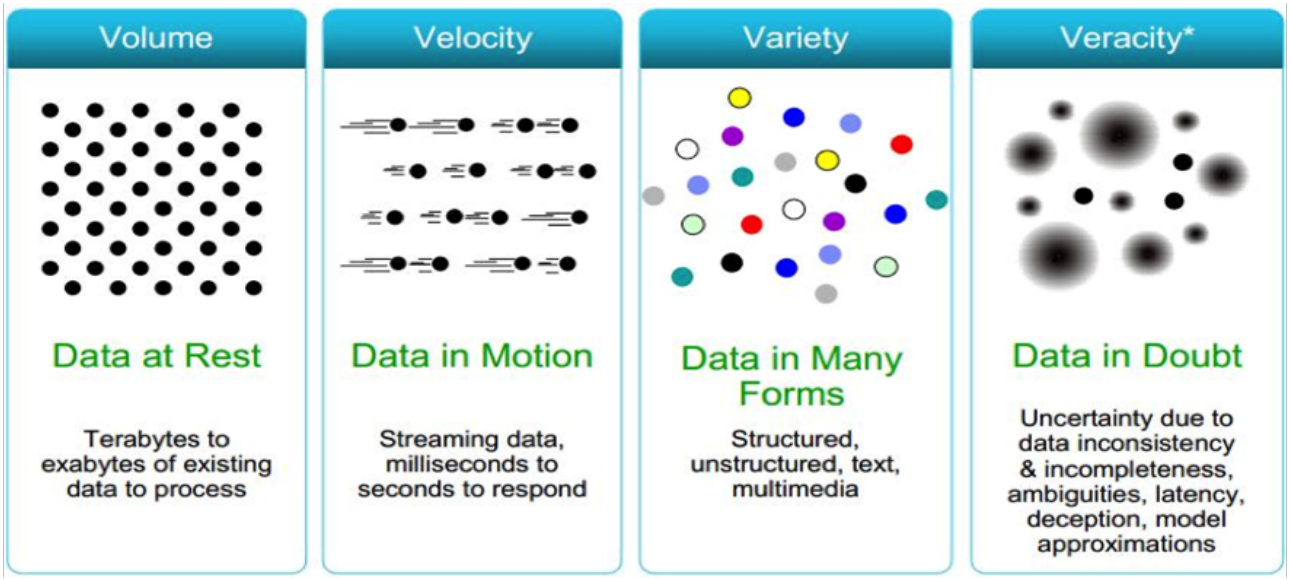


Figure 6.1: Characteristics of Big Data – The Four V's

6.2 Big Data Overview

The term 'Big Data' is used to refer to large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information [CTO Labs WP, 2012].

As shown in Figure 6.1, Big Data is typically characterized using four V's namely, Volume, Variety, Velocity, and Veracity. Although uncommon, some other descriptors like Variability, Value, and Visibility are also associated with Big Data. The first "V" or Volume refers to the amount of data that is produced. The second "V" or Velocity indicates the frequency at which the data is generated or arrives at an organization. The third "V" or Variety signifies the heterogeneity associated with the data and the fourth "V" highlights the uncertainty associated with the data. These four V's essentially differentiate regular or traditional data from Big Data. Unlike regular data, Big Data involves extremely large data volumes scaling from hundreds of terabytes to exabytes or even higher amounts. Big Data entails fast collection, processing and consumption of data. This property characterizes data with different arrival rates e.g., batch, periodic, near real-time and real-time data as illustrated in Figure 6.2. Such data is heterogeneous and comes in multiple data formats such as structured, semi-structured, and unstructured. This feature represents data captured from various sources like text data from blogs, news, or other social media, streaming data from web, audio, and video, and image data, as well as structured data. Big Data is also often characterized by Veracity, which is a measure of the uncertainty that arises due to data quality issues like incompleteness, ambiguities, latency, deception, and model approximation. In recent years, several technological advancements have resulted in the proliferation of a myriad number of sources that generate vast amounts of data. Today, many new sources of data are in use which include but are not limited to web sources such as the social media, news, sensors, audio, video, interactive applications, logs, emails, and other traditional sources of structured data. The ever-increasing utilization of these sources has led to the generation of the so-called Big Data. According to a recently published report, internal sources of Big Data result in 88% transactions, 73% log data, and 57% emails whereas external sources such as social media, audio, photos, and video result in approximately 43%, 38%, and 34% of data respectively. Utilization of such large number of sources has resulted in the generation of fast, semi-structured, and unstructured data, besides the conventional structured data, in prolific amounts. The variety associated with the data generated

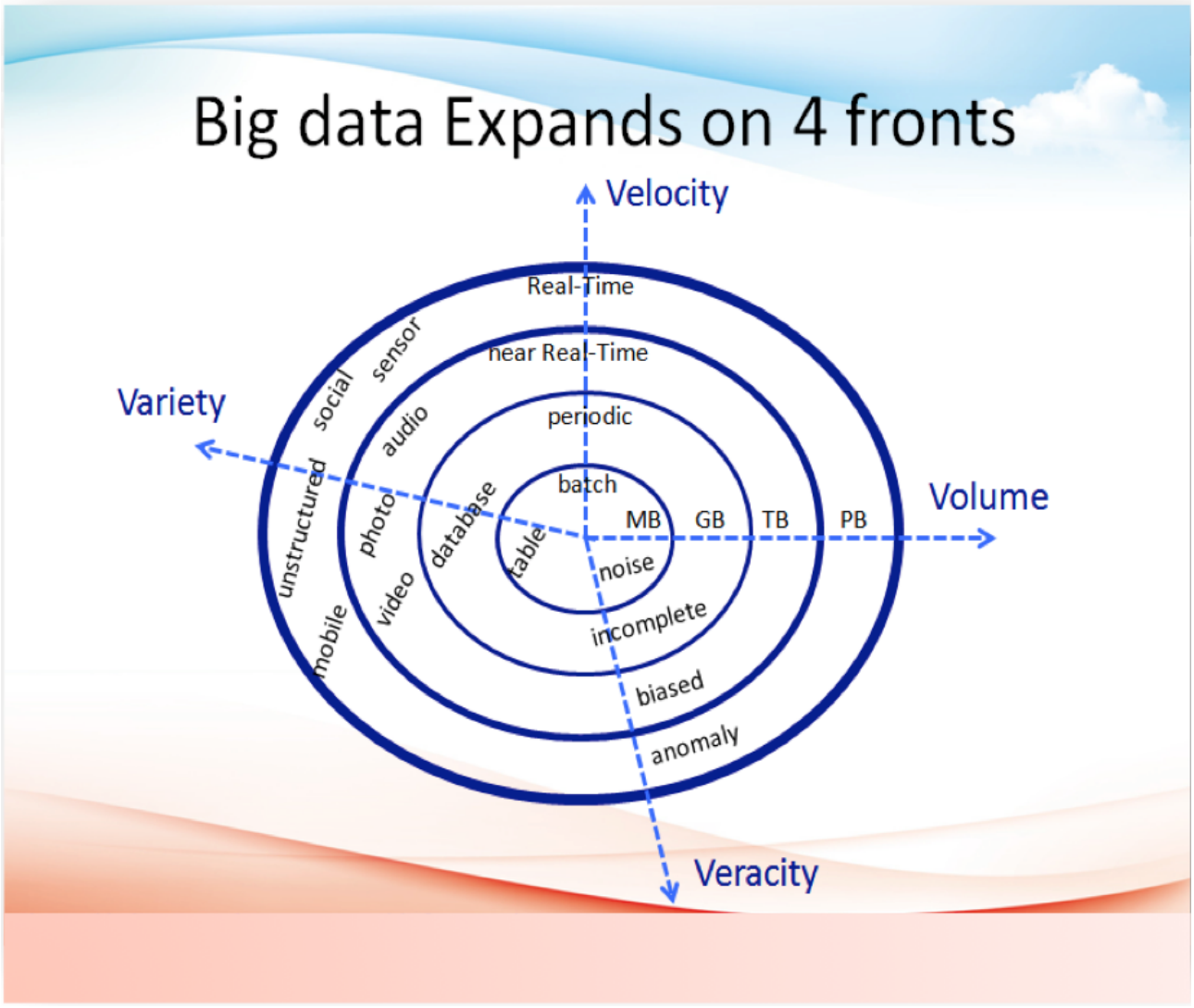


Figure 6.2: Expansion of Big Data on four key fronts

by these sources along with the massive volumes and speeds, at which most of the sources generate data, make big data really big. The tremendous data growth we witness today further exemplifies this fact. For instance, the world's information is doubling every two years. In the year 2011 alone the total generated information was estimated to be around 1.8 zettabytes. This is equivalent to 3 tweets every minute by every person in the USA for a period of 26,976 years continuously. Storage requirements of this information alone equate to about 57.5 billion 32 GB Apple iPads [Catone, 2011]. The exponential growth is expected to get even stronger in years to come and cross the 8 zettabyte mark by 2015.

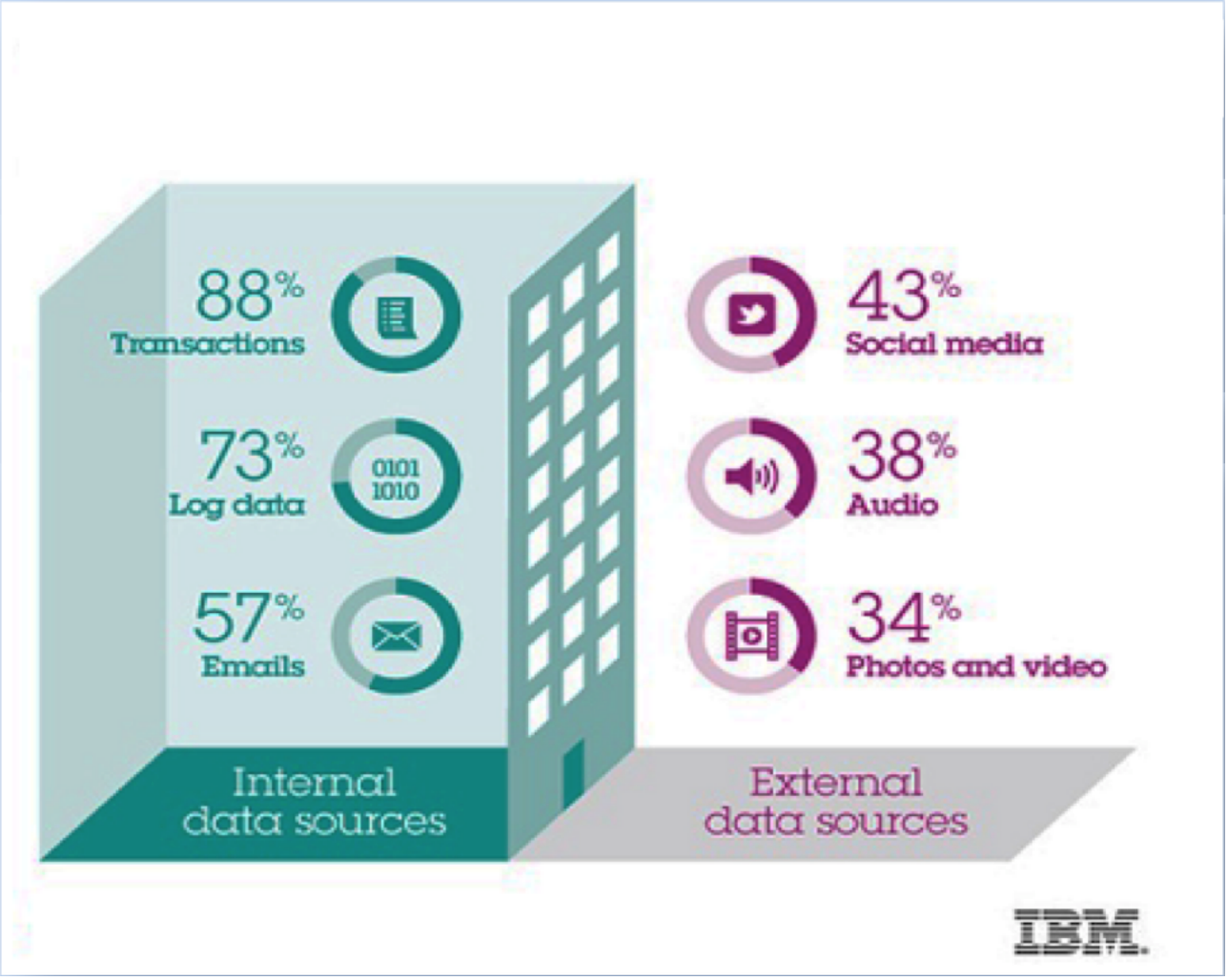


Figure 6.3: Common sources of Big Data

In the following, we provide a high level comparison between the traditional data and the Big data.

Traditional Data	Big Data
Megabytes to Gigabytes to Terabytes	Terabytes to Petabytes to Exabytes
Centralized	Distributed
Structured	Structured + Semi-structured + unstructured
Stable Data model	Flat schemas
Known Complex relationships	Known + unknown complex relationships

Some of the characteristics demanded by Big Data are:

1. Cost efficiency in processing the growing volume of data.
2. Response to the increasing velocity of the data streaming. It is of interest to note that the sensors are increasing ever (30 billion RFIDs and counting)
3. Collective analysis of broadening varieties of the data. It must be noted that 80% of the world data is unstructured.
4. Establishing Veracity of the data. It is a challenge to convince the data being used for making decisions.

6.3 Emerging Big Data Landscape

In the last couple of years a new data model has begun to evolve as a result of different efforts to harness the power of Big Data. Specifically, more and more organizations have started to utilize Big Data to cater to their organizational specific needs such as fraud detection, news curation, customer relationship management, and forecasting. Today, new scalable data management solutions are being explored and deployed to efficiently handle rapidly increasing data volumes. The adoption of non-traditional data management technologies is being driven not just by the volume, variety, velocity, and veracity of data, but also by changes in the manner in which users want to interact with their data. Some of the key players in this domain include commercial systems like InfoBright and Greenplum which utilize variants of open source technologies like MySQL and Postgres respectively, Hadoop/MapR, and a new generation of data intensive super computing (DISC) appliances like Netezza and Xtremedata. The latter represent a paradigm shift in computing wherein computing is moved to the location of the data. Figure 6.4, depicts some of the key elements in the services, analytics, and data management tiers of the Big Data domain.

In the past, organizations relied on traditional data that spanned no more than a few terabytes, stored centrally, with a well-defined structure that was sufficient to represent and extract the known complex interrelationships that existed among the various data entities. In contrast, much of the data that is being produced by the numerous existing interactive applications and data-generating machines could easily exceed Exabytes. Nature of that data is heterogeneous i.e., could be structured, semi-structured, or unstructured, and can no longer be efficiently handled through centralized storage based options. In other words, the data lacks a structure to make it suitable for storage and analysis in traditional relational databases and data warehouses. In addition to this variety, the data is also being produced at a velocity that is often beyond the performance limits of traditional systems. Moreover, unlike traditional data in which the interrelationships among data are known a priori, Big Data often involves unknown complex relationships that can only be discovered by exploring such data in its entirety. As a result it is often impossible to predefine database schemas for exploratory analysis of Big Data because the nature and type of the queries are usually ad-hoc.

To harness the real value out of Big Data it is crucial to employ effective solutions that can cost efficiently process the vastly growing volumes of data that are expected to exceed 35 ZB by the year 2020, handle streaming and “at rest” data volumes, and can facilitate collective processing and analysis of structured, semi-structured, and unstructured data. According to a recent study about 1 in 3 business leaders do not trust the information they use to make business decisions. This further emphasizes the fact that an effective solution must handle the uncertainty associated with data to improve its trustworthiness and ultimately its value.

Traditional architectures based on relational database management systems generally rely on a sampling approach to process and analyze all available data and inductively extrapolate the results obtained from small samples to make decisions. These systems were primarily developed to address simple analytic needs and can only be used to handle structured data. The database management in such systems is separated from the analytics processing. Structured data is maintained centrally in the traditional RDBMS based data management systems and a representative sample of the dataset is moved to a separate analytics environment for processing. So, in such models the analytics is done in its own infrastructure decoupled from the data management system

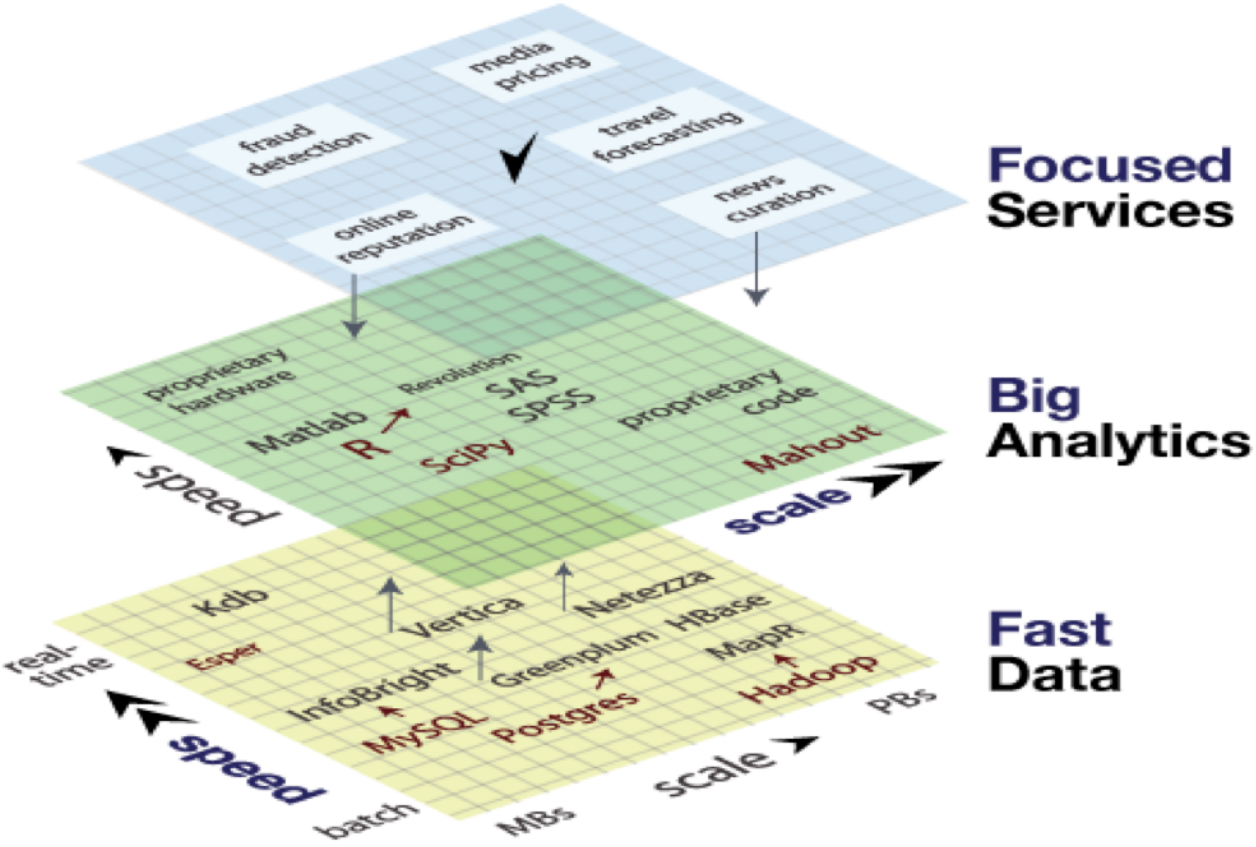


Figure 6.4: Few key elements of the emerging Big Data stack

and they mostly rely on extract, transform, and load (ETL) mechanisms as intermediaries to store only the well defined, structured data into the database systems. In these models the back-end data environments are optimized only for basic data management operations that ensure faster access to data but are not usually optimized for performing complex analytics in the data. Figure 6.5, illustrates a typical data flow pipeline in a conventional RDBMS based architecture. The traditional architectures offer the basic ACID – atomicity, consistency, integrity, and durability properties for data volumes but lack the breadth and depth to handle the storage, integration, and complex analytics requirements needed to extract valuable information from increasing volumes of structured, semi-structured, and unstructured data. These limitations could be attributed to the fact that such systems were designed for online analytical processing (OLAP) tasks, performed on small or sampled data sets, which are very different from the unknown or ad-hoc data exploration methods and complex analytics required for Big Data. Moreover, such architectures are further paralyzed by the I/O constraints. In the conventional systems, because of the disintegrated data storage and analytics environments, data has to be moved from storage to analytics systems for analytical processing. This requires movement of data from the disk over the network for processing which becomes almost impractical for larger data sets because of the limited network bandwidth and disk I/O constraints. Figure 6.6, illustrates some of these constrictions. Despite several advancements or optimizations to alleviate these bottlenecks the problems remain largely unsolved today. Even though techniques like caching, partitioning, and indexing have been explored extensively in an attempt to mitigate these bottlenecks, they have not been able to make any significant improvements that can make such systems usable to cater to the Big Data needs of today. For instance, caching schemes are often effective when the data access patterns are somewhat known. However, the kind of ad-hoc queries or exploratory analysis required with Big Data often takes this predictability factor out of the data usage patterns thereby rendering such schemes ineffective. Moreover, most of the available tools are not scalable and become unrealistic for extremely large-scale data sets, especially involving unstructured data. Today it is no surprise that the limitations of the conventional architectures are uncovered or discovered to a large extent through a confluence of the four Vs that essentially characterize Big Data. For example, the enormous volumes that characterize Big Data can no longer be accommodated through traditional data storage mechanisms. According to a recent study, the global information created in year 2011 was estimated to far exceed the available storage capacity and the trend is going to continue in the future.

The inadequacies mentioned above clearly necessitate a move towards architectures that can overcome these

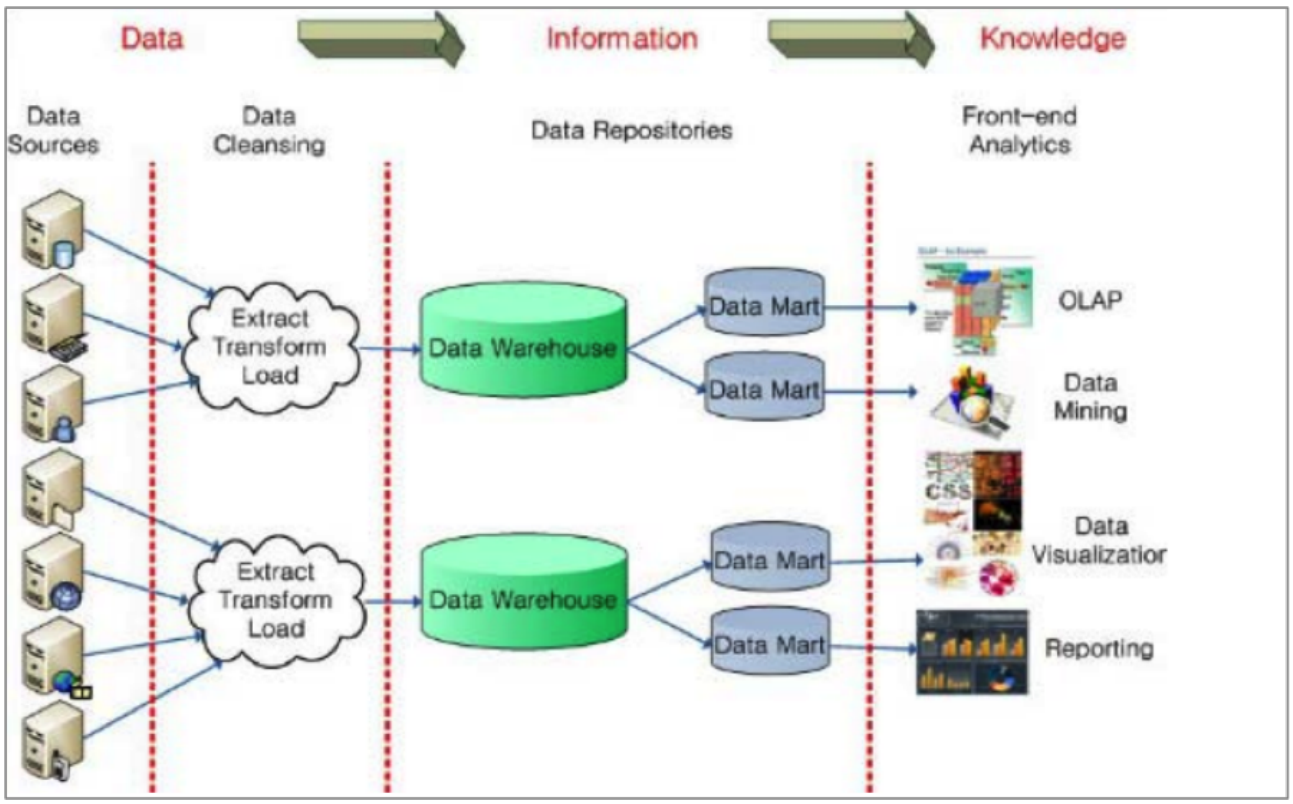


Figure 6.5: Traditional RDBMS based architecture

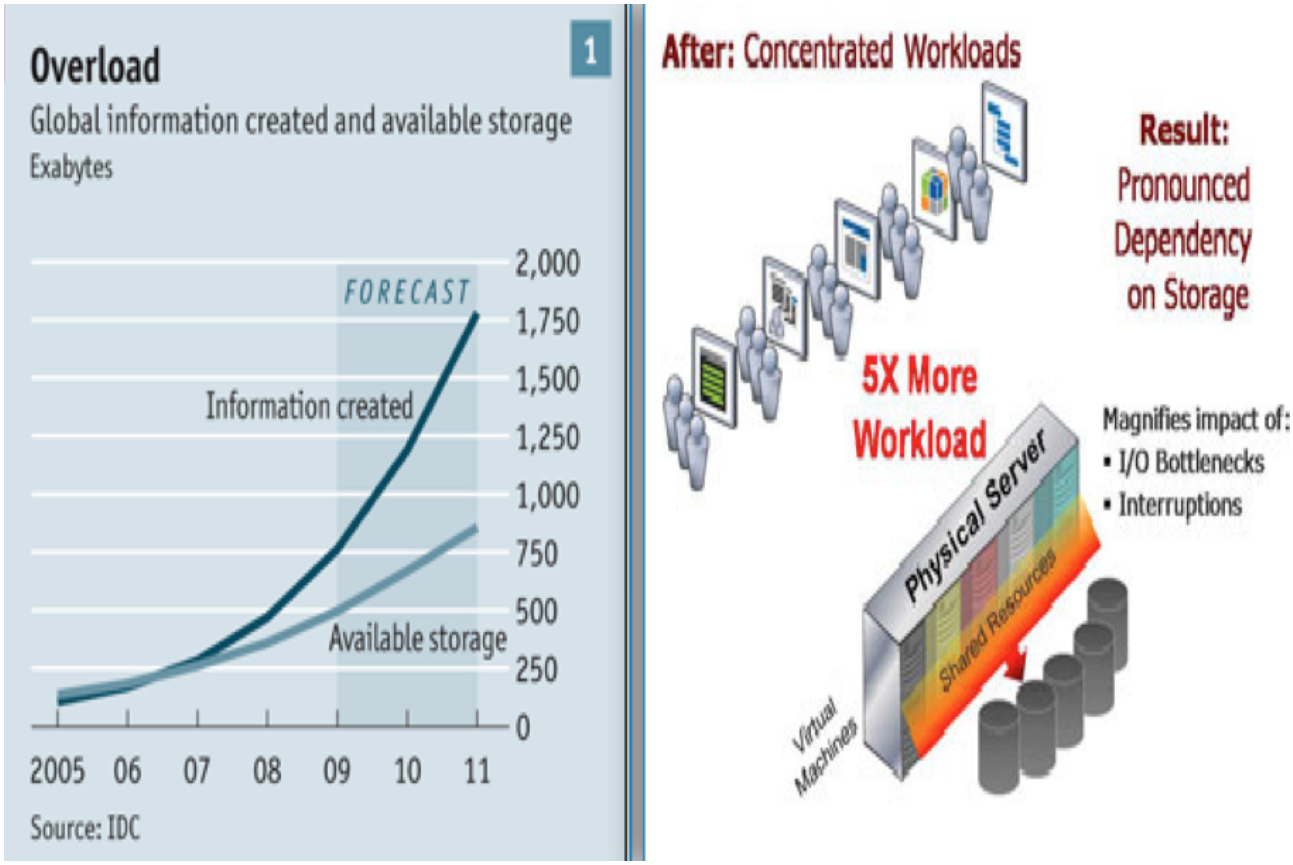


Figure 6.6: Limitations of conventional systems

limitations by offering alternative solutions such as scale-out storage, increased parallelism, heterogeneous data integration and aggregation, high I/O throughput, and advanced analytics engines for large scale data analysis. In recent years, several advancements have been made in this domain ranging from the development of strictly relational systems, non-relational data systems like Versant, and Operational systems like Starcounter to NoSQL systems like Cassandra and HBase, Analytic systems like Hadoop/MapR, and DISC systems like Netezza and Xtremedata. Some of these innovations such as the DISC appliances have introduced completely new paradigms in computing. Unlike the traditional systems which move data to analytics engines for processing, this new generation of appliances performs analytics at the location of the data thereby eliminating the need to move large sets of data across networks and hence, reducing the network latencies or I/O bottlenecks. Moreover, such appliances make it possible to analyze Big Data in its entirety without resorting to the traditional ways of sampling.

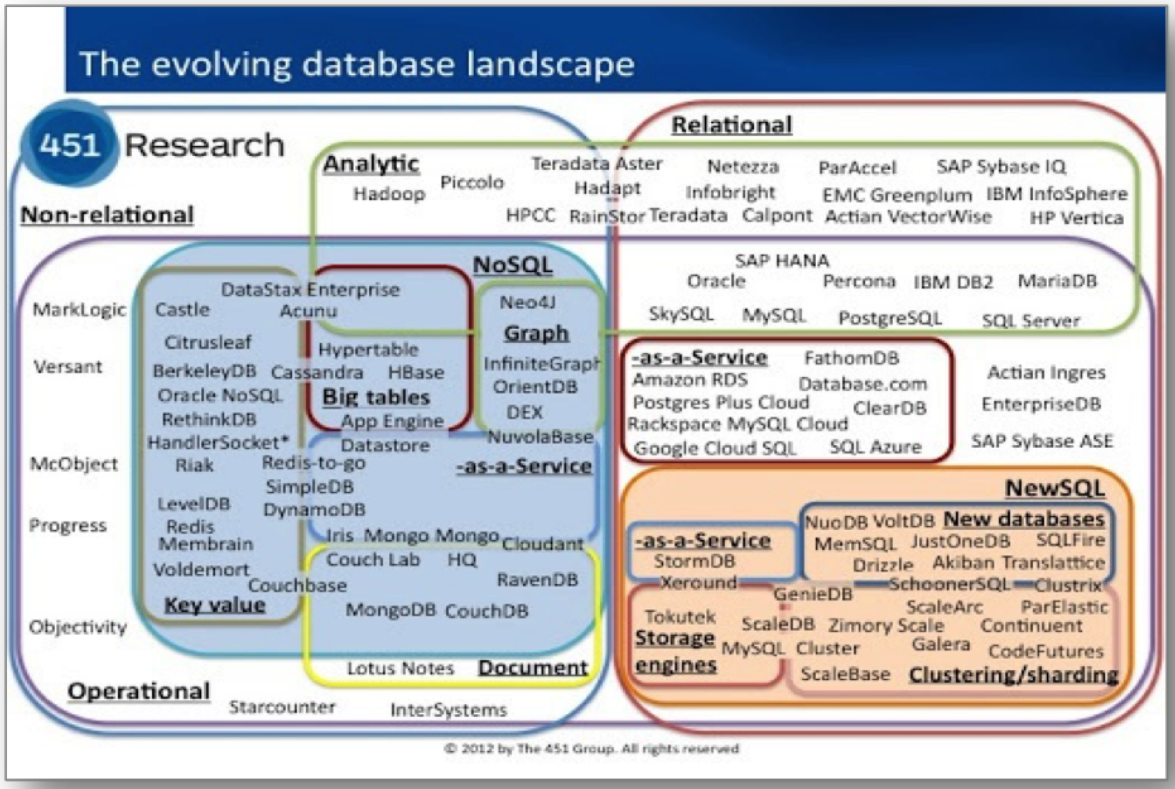


Figure 6.7: The emerging data landscape

Evidently, today there are many potential solutions available at different levels of the Big Data spectrum. Figures 6.7 and 6.8 exemplify some potential Big Data solutions in terms of data management, infrastructure, analytics, and applications. These include hardware solutions that range from shared everything and shared disk to shared nothing architectures, data management and data analytics systems that range from open source and custom-developed solutions to commercial solutions, and Big Data domain specific data access, mining, and other application solutions. Clearly, to extract value out of Big Data it is necessary to select and deploy the right technologies that can exploit Big Data efficiently. Obviously no one solution can fit all the Big Data needs that exist in this digital era and it is critical for organizations to adopt solutions that can best service their specific Big Data application needs. However, it is important to recognize that Big Data has not much value by itself and its true value can only be realized through analytics. Even though it is critical to capture all available data but equally critical is to deploy tools that can efficiently filter the vast amounts of captured data to extract useful signals and derive actionable intelligence from it. Today, the availability of Big Data has made it feasible to accomplish things that could not be done previously and has opened up new horizons for innovations. Following sections illustrate some of the representative examples of Big Data solutions that have been deployed in different Government sectors in the USA.

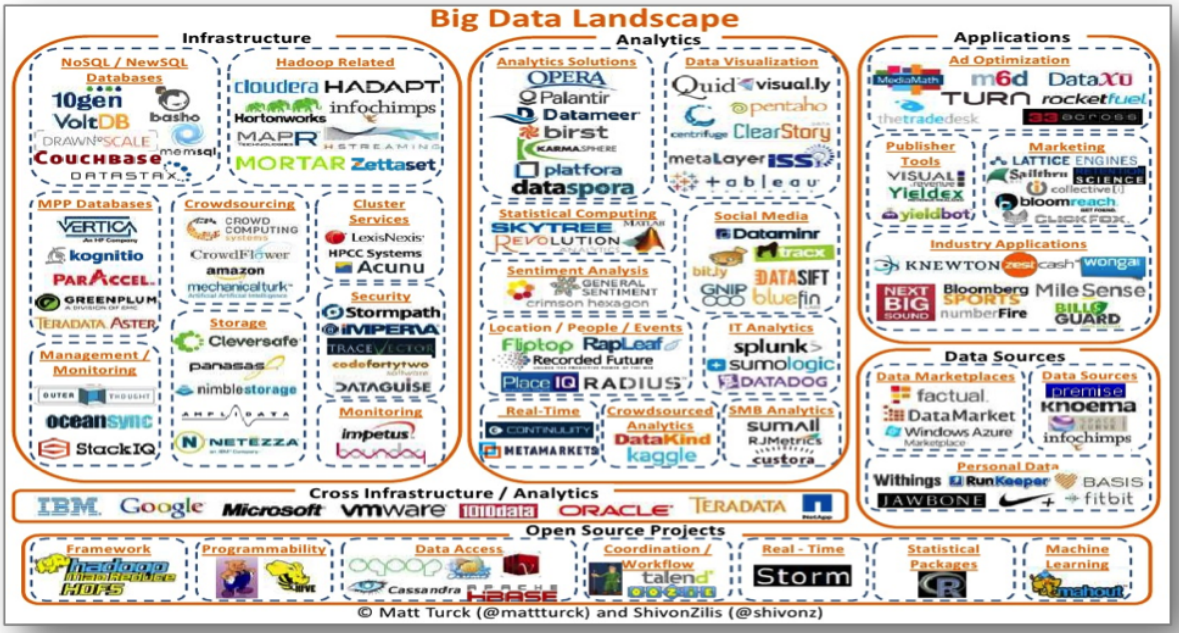


Figure 6.8: The evolving Big Data Landscape

6.4 Big Data Analytics

Analysis of Big Data has found a variety of applications. In the following, we shall highlight some of the analytics that have found applications in various government sectors. We shall illustrate some examples in the context of USA.

6.4.1 Education

Today, higher education institutions in the USA to improve education services at various levels are adopting Big Data technologies and specifically analytics. There are different kinds of heterogeneous data in the education sector which include, but are not limited to, data such as: (i) identity data like demographics, school district etc., (ii) user interaction data like engagement metrics, (iii) inferred content data like evaluation data that gauges how well a piece of content affects proficiency of a group or subgroup of students, (iv) inferred student data which comprises of statistical data like the probability of students' performance above a certain percentile, and (iv) system wide data like school rosters, grades, records etc. The data in these categories can usually be sub-categorized further depending upon its sources of origin. For instance, assessment data in the educational system can come from interim and formative assessments, teacher assessments of student engagement, and student work, behavior, and attendance reports, grade histories, test scores, etc. Notably, teachers in the education systems are not just the consumers of data but also the generators of data and have the most detailed first hand data within the school systems [Marie Bienkowski, 2012]. The educational institutions in the USA recognize the applicability and potential of Big Data in several different areas of the domain. They have begun to exploit the capabilities of Big Data analytics to enable the use of new learning technologies, identify barriers to kids' education that could be solved by technology, and provide standardized assessment data electronically, with an ultimate goal to improve educational services, increase student grades, retention outcomes, and productivity of K-12 education systems e.g., through personalized learning scenarios and adaptive learning systems [Marie Bienkowski, 2012]. For example, to improve learning performance of students during online study or test sessions, analytics can help identify patterns of boredom based on the patterns of key clicks and redirect the students' attention accordingly. This prospect opens up several possibilities of continuous improvement via multiple feedback loops in various temporal settings. For example, such feedback can instantaneously help a student improve attention or focus and hence, expected outcome, on the next problem. Similarly, such feedback can be provided daily to teachers for improving next day's teaching, monthly to the principal for judging progress, and annually to the district and state administrators for overall school improvement. Likewise, analytics can be used for domain modeling which can help understand, identify, and discover the best ways to deliver course topics and effective ways to promote learning. Specifically, analytics in this domain can help with the formulation of innovative and adaptive ways

of teaching, prediction, clustering, and relationship mining through user knowledge, behavior, and experience modeling, user profiling, domain modeling, learning component analysis and instruction principle analysis, and trend analysis. Insights obtained from such analytics can help teachers quickly assess and address clear learning gaps and opportunities for acceleration. This knowledge can be shared with students to motivate them and help chart their own progress as well as with parents to help them better understand the needs of their children. Moreover, the extracted information from such analytics can greatly help with the development of personalized learning scenarios and adaptive learning systems. Figure 6.9, highlights the potential of data available in this domain and further emphasizes the need for Big Data based solutions in the education domain. Since the last

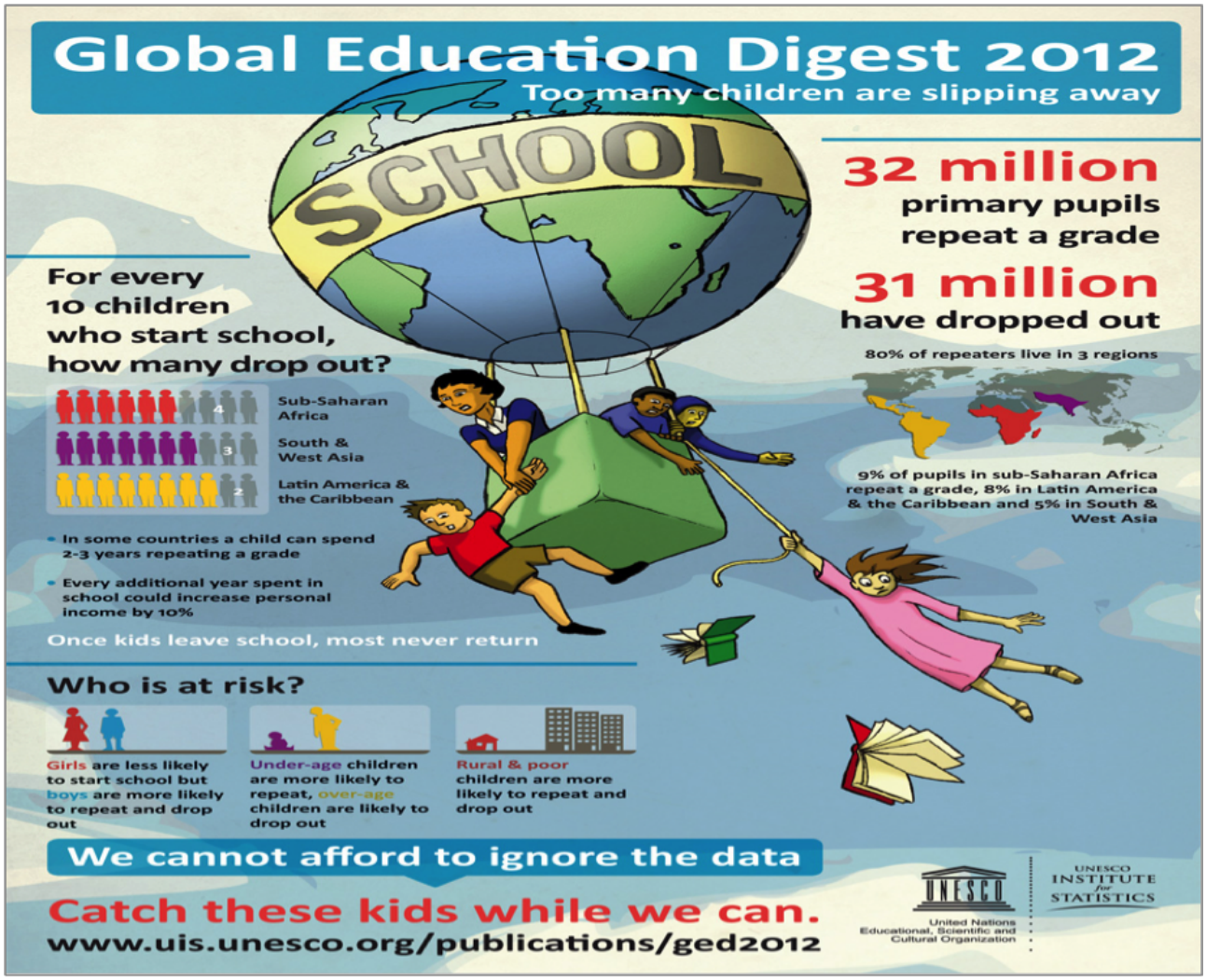


Figure 6.9: Potential of Data in the education domain and the need for innovative solutions to improve education and learning

decade, states have been collecting millions of data points about students for compliance purposes. Despite the availability of all this data, educators couldn't glean any insights from it simply because the educators lacked easy access to nonstandard student data in siloed systems. Moreover, transforming the non standard data into actionable knowledge had been almost infeasible within realistic time frames. Today, emerging technologies are offering new ways to make such feats feasible and convert the learned insights into actionable knowledge for improving the educational system and outcomes. Figure 6.10, depicts a graphical representation of the goals that can be attained using Big Data solutions.

Big Data Examples in the Education Sector

In the USA several systems are being developed and deployed to realize the goals discussed above. For example, a New York based adaptive learning company, KNEWTON[1], has built an adaptive learning platform that uses advanced analytics to offer personalized education content [Upbin, 2012]. Their unique adaptive platform provides efficient means to deliver basic math, writing, and science concepts. The platform analyses every move

[1] https://www.knewton.com/platform

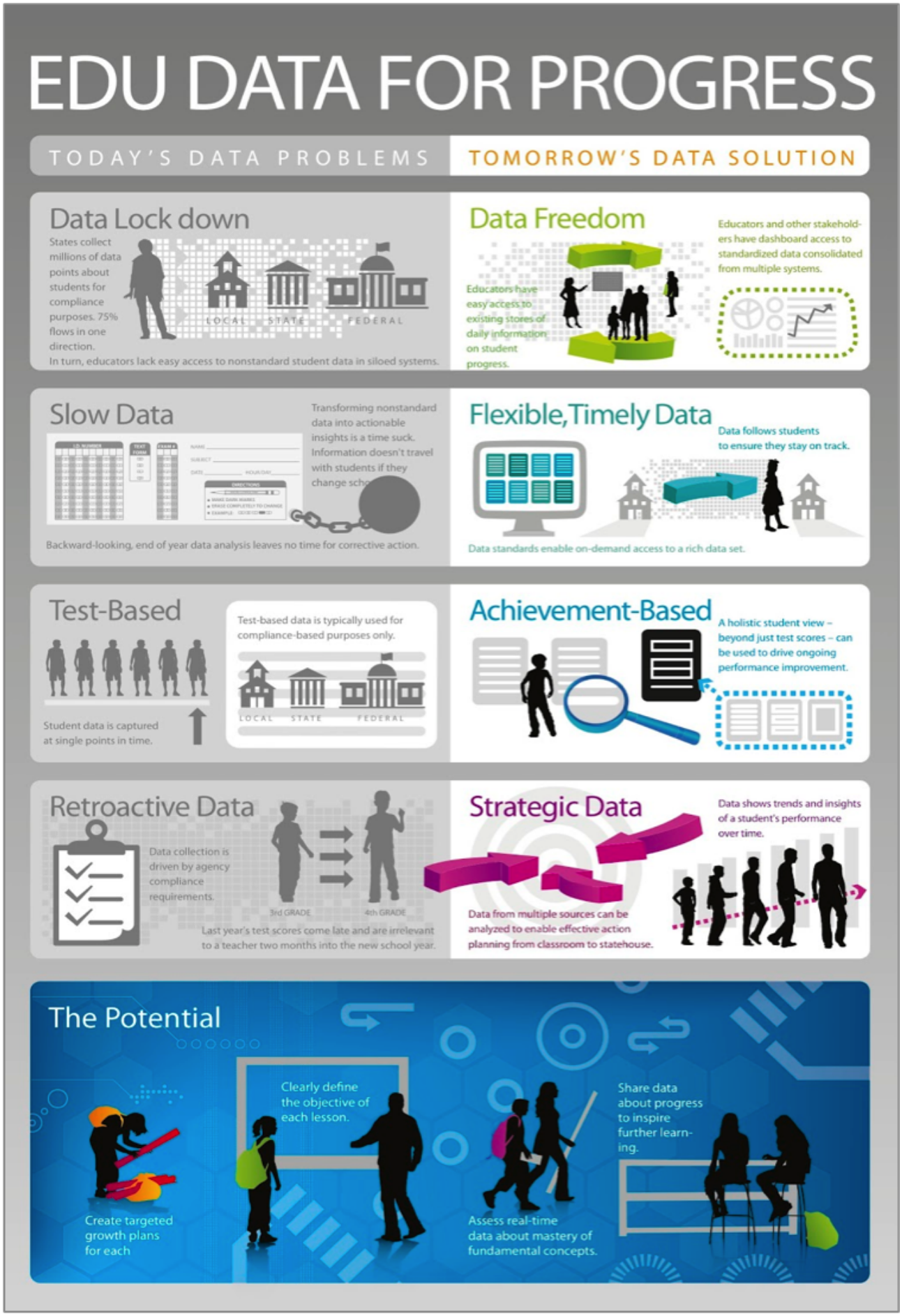


Figure 6.10: Potential data solutions to address some of the problems prevalent in the education domain

made by students including scores, speed, accuracy, delays, key-strokes, click-streams, and drop offs. It uses the insights obtained from such analysis, to adapt the course accordingly by challenging and cajoling students to learn based on their individual learning styles. It includes prominent partners like Pearson Education, and Macmillan Education, etc. Recently, Arizona state university also partnered with KNEWTON to offer developmental math and other learning courses using KNEWTON's adaptive virtual learning platform. Their collaboration resulted in improvements in the pass rates, from 66% to 75%, and drop in course withdrawal rates, from 13% to 6%.

Another representative example comprises of Western Interstate Commission for Higher Education (WICHE[2]) Cooperative for Educational Technologies. The organization developed a predictive analytics reporting framework (PAR) that has allowed the organization to gain insights into the relationships between retention rates, performance, and demographics. Their data mining initiative can effectively utilize approximately three million student records and help identify points of student loss and find effective practices that improve student retention in USA higher education. With sixteen WCET member institutions, over 1.7 million anonymized student records, and 8.1 million institutionally de-identified course level records, the PAR Framework offers educational stake-holders a unique multi-institutional lens for examining dimensions of student success from both unified and contextual perspectives.

For the Indian Sector: A majority of the above analysis is quite relevant for Indian Data (of course, data collection is a task that cannot be under estimated). Further, the analysis requires additional parameters like rural, semi-urban, urban as well as health and other student oriented Govt. schemes. Such analytics will also be beneficial to assess Govt. schemes, their adaptations to requirements as well as introduction of new schemes as appropriate.

6.4.2 Energy Sector

Today unprecedented volumes of complex, fast growing data are rampant in almost all areas of the energy sector. The information gained from analytics on these vast amounts of data is increasingly being utilized in sectors like the oil and gas industry and the utility sector to perform a large number of operations such as enable use of new clean energy technologies, identify barriers to using clean energy, effectively utilize electronic availability of building energy usage data, improve recovery rates, and reduce operating costs. In this digital age, extremely large-scale data is often collected and analyzed to judge the health of an oil-rig. For instance, a typical oil drilling platform can have 20,000 to 40,000 sensors on board with each sensor producing streaming data about the health of the oil rig, quality of operations, and so on. Even though all the sensors may not broadcast at all time, some sensors may report new information many times per second. Moreover, various key players in the oil and gas industry are increasingly moving towards digital oil field. A typical digital oil field scenario is shown in Figure 6.11. Their main line of business is to search for potential underground and underwater oil and gas fields, drill exploratory wells, and subsequently operate the wells that recover and bring the crude oil and raw natural gas to the surface. They use many types of captured data to create models and images of the Earth's structure and layers 5,000-35,000 feet below the surface and to describe activities around the wells themselves, such as machinery performance, oil flow rates and pressures. Other data such as data from sensors installed in subsurface wells and surface facilities is also utilized to provide continuous, real-time monitoring of assets and environmental conditions. Data from sensors and other semi-structured and unstructured data, ranging from high-frequency drilling and production measurements to daily written operations logs, business data e.g., internal financial results, news on energy and petroleum competitors bidding on leases and making major capital investments, easily scales up to several petabytes. With approximately one million wells currently producing oil and/or gas in the USA alone, and many more gauges monitoring performance, this dataset is growing daily. To extract value out of this data, companies are using distributed sensors, high-speed communications, and data mining to monitor and fine-tune remote drilling operations. Many companies have already started to use real-time data from these sources to make better decisions and predict glitches e.g., use real-time data to make collaborative decisions in drilling operations, or managing wells and imaging reservoirs for higher production yields. For example, Shell, is already collecting up to a petabyte of geological data per well using its advanced seismic monitoring sensors (co-developed with Hewlett-Packard), and plans to use the sensors on 10,000 wells in the near future [Farris, 2012a]. In this industry, Big Data technologies can help with the (i) collection, management, and analysis of large and rapidly growing volumes of data, such as petabytes of production data generated by oilfield sensors, (ii) analysis of a wide variety of data types including numerical data streaming in from drilling-rig sensors and unstructured data from logs, micro seismic and other sources (iii) rapid searches through massive volumes of historical reservoir data or unstructured information, such as text-based drilling reports or competitive press releases, that take weeks or months depending on the resources at hand. However, one of the key Big Data questions in the oil and gas industry is to determine effective ways to utilize and explore

[2] `http://wcet.wiche.edu/`

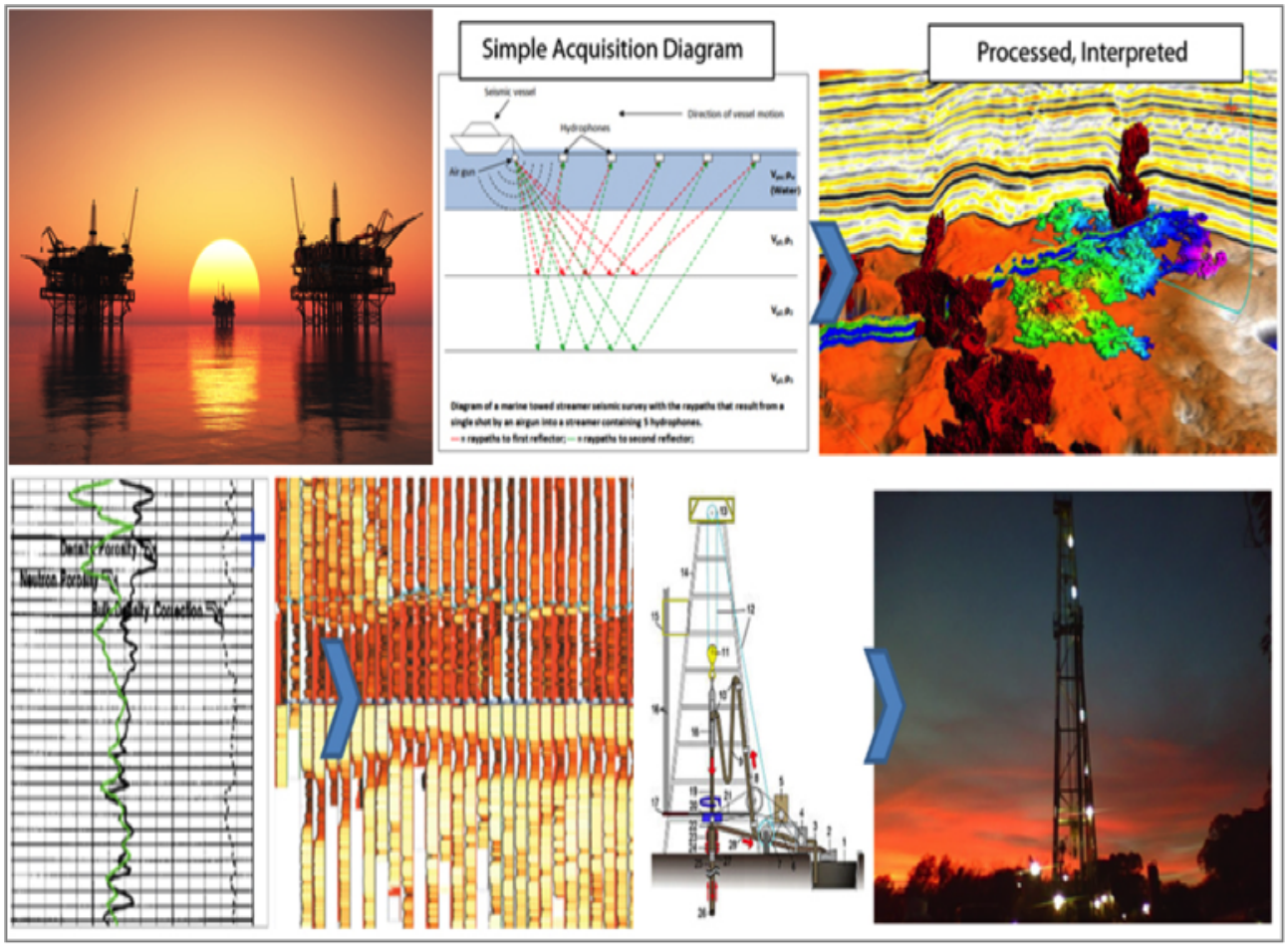


Figure 6.11: Drilling Digital Oil Fields with Big Data

data to address questions like determining the location of more oil and means to get substantially more out of the ground safely, with minimal environmental impact [Farris, 2012a, Farris, 2012b, IBM-WP, 2013, Nicholson, 2012, Leber, 2012].

Big Data Examples in the Energy Sector

A leading oil and gas company, Chevron[3], highlights the capabilities of Big Data analytics in this domain. It utilizes Big Data technologies to optimize oil and gas exploration or drilling operations with capabilities that enable analysis of large scale data streaming in the drilling of wells or the operating of surface-facility equipment without having to first store the data. Chevron has implemented a Netezza DISC based Big Data solution that is helping the company deliver faster performance than traditional ETL tools. For instance, Chevron was able to run a record 25,000 business queries in one day and accelerated key processes e.g., their traditional GIS reporting framework used to take 12 months but with their adopted solution the company can now generate reports in real time. The company has eliminated 50% of the cost of legacy ETL solutions saving USD 2 million per year. Recent estimates anticipate 8% higher production rates and 6% higher overall recovery from a "fully optimized" digital oil field. Similarly, Drillinginfo, a leading data and intelligence provider of upstream data for oil and gas decisions, has begun to break the barriers of geography and discipline to consider many potential variables and, based on thousands of wells, create a statistically predictive model for a given area's overall production[4].

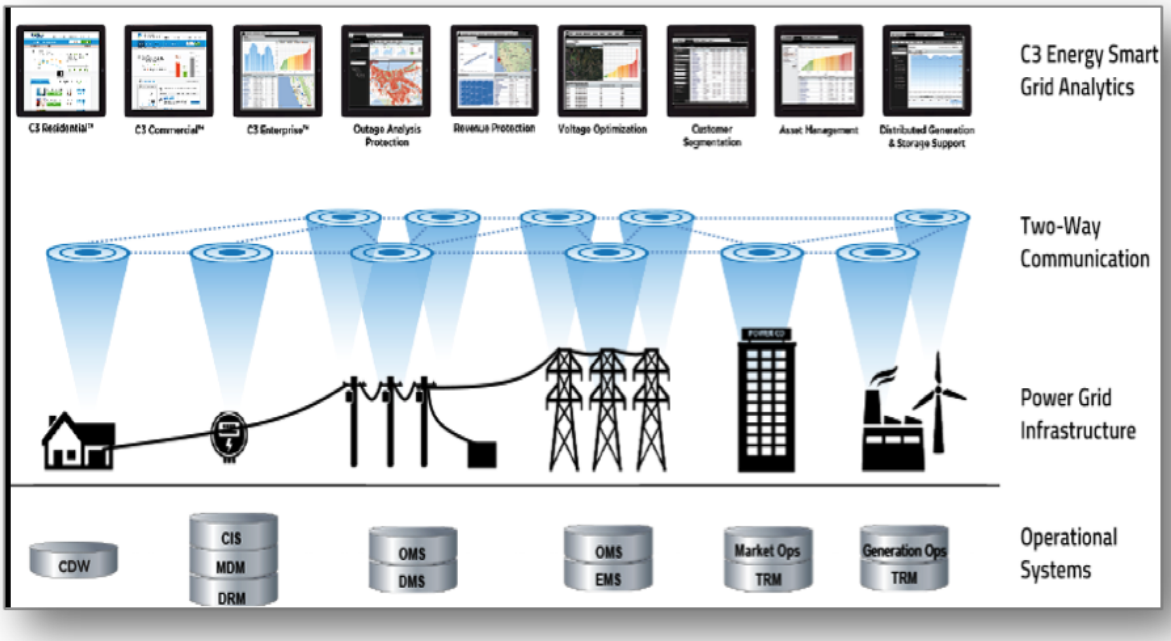


Figure 6.12: Big Data sources and analytics in the utility sector

Similar advances can be seen in the utility sector where more and more companies are recognizing the potential of Big Data analytics to predict consumption more effectively, which in turn is helping them manage energy procurement with greater precision. The US Department of Energy estimates that commercial facilities account for up to 50% of electricity use in the USA, nearly 30% of which is wasted through controllable inefficiencies. However, to paint a full picture of energy consumption and forecast demand, several variables e.g., billions of electronic transactions per day, procurement, weather, client portfolio, and business productivity data are required to be analyzed simultaneously[5]. Clearly, such tasks require Big Data analytics capabilities.

Today, smart meters and grids are generating an unprecedented volume, speed, and complexity of data. For example, going from one meter reading a month to smart meter readings every 15 minutes works out to 96 million reads per day for every million meters. The result is a 3,000-fold increase in data that must be managed. Data gathered from smart meters can provide better understanding of customer segmentation, behavior, effect of pricing on usage, can help transform the network, and dramatically improve the efficiency of

[3] http://www.chevron.com/
[4] http://info.drillinginfo.com/
[5] http://www.ecova.com/media/101352/ecova_whitepaper_a_big_data_look_at_energy_trends.pdf

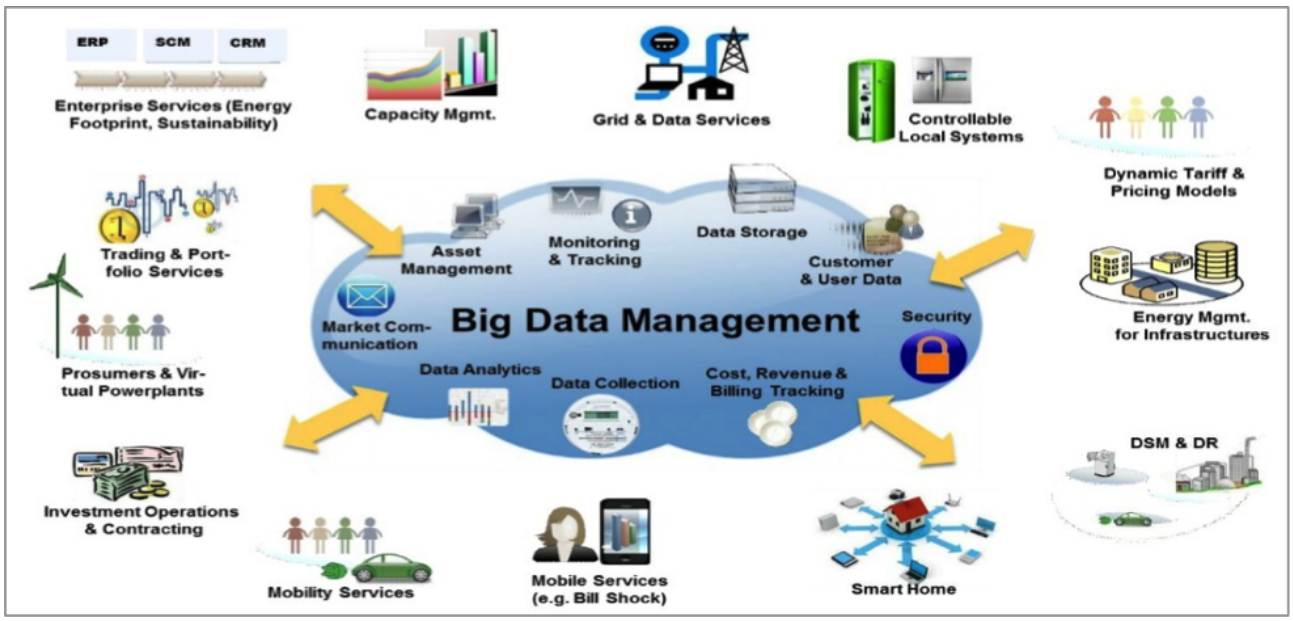


Figure 6.13: Big Data ecosystem in the utility sector

electrical generation and scheduling. For example, the collected data can provide information about the network operations such as information about the most stressed components of the network, identification of the best areas for future investments, proactive response to conditions indicative of future outages, and gain insights into price fluctuations, energy consumption profiles, and grid health assessment. Moreover real-time meter information can help discover areas where there is consumption that happens when energy is being diverted and stolen. Figures 6.12 and 6.13 illustrate some of the Big Data generation and usage scenarios in the utility sector. Today, utilities companies are analyzing Big Data derived from the power grid to optimize energy delivery and improve the use of renewable energy sources. With the help of Big Data, existing networks are becoming much more efficient energy grids and are gaining insights into finding ways to reduce energy consumption and consumer prices, among other things like IoT[6], Auto-grid[7], Smart-grid [John, 2013].

For instance, a regional transmission organization, PJM, one of the world's largest competitive wholesale electricity markets, utilizes a state-of-the-art grid management system today. Its Advanced Control Center program integrates very large-scale energy management systems and real-time market pricing systems using a scalable, Siemens/PJM[8] Shared Architecture integration platform [Siemens Case Study, 2012]. Among other things, the system uses high performance computing technologies to provide efficient and reliable services to its clients.

6.4.3 Health-care

Studies show that around 90% of the world's data was created in the last couple of years alone and the data is predicted to grow, by a factor of 50, to 25,000 petabytes by the year 2020. 50 petabytes of the data is expected to be in the health-care sector with estimates indicating that the IT systems in this industry will soon have more data than they would be able to handle. Like in other sectors, data in the health-care sector has tremendous value and the effective utilization of this data could result in savings of about $300 billion per year. Moreover it can help reduce expenses by 8% through reductions in administrative and clinical inefficiencies, fraud and abuse[9], and poorly coordinated care resulting in savings of about $175-250 billions, $125-175 billions, and $25-50 billions in each of these areas respectively [Roger Foster, 2012].

Big Data in the health-care industry comprises of data from a variety of sources, which include health data creators, clearinghouses, technology vendors, public health agencies, health information organizations, consumers[10]. As shown in Figure 6.14, these entities generate different kinds of data e.g., pharmaceutical companies and academia mostly generate research and development data, health-care services providers generate clinical, activity, and cost data and consumers and other stake-holders generate patient behavior and sentiment data. Recent reports suggest that around 80% of the patient information data today is unstructured but highly

[6] http://www.c3energy.com/
[7] http://www.nttdata.com/global/en/news-center/pressrelease/2013/082200.html
[8] http://www.pjm.com/Default.aspx
[9] https://oig.hhs.gov/reports-and-publications/hcfac/index.asp
[10] https://patientio.com

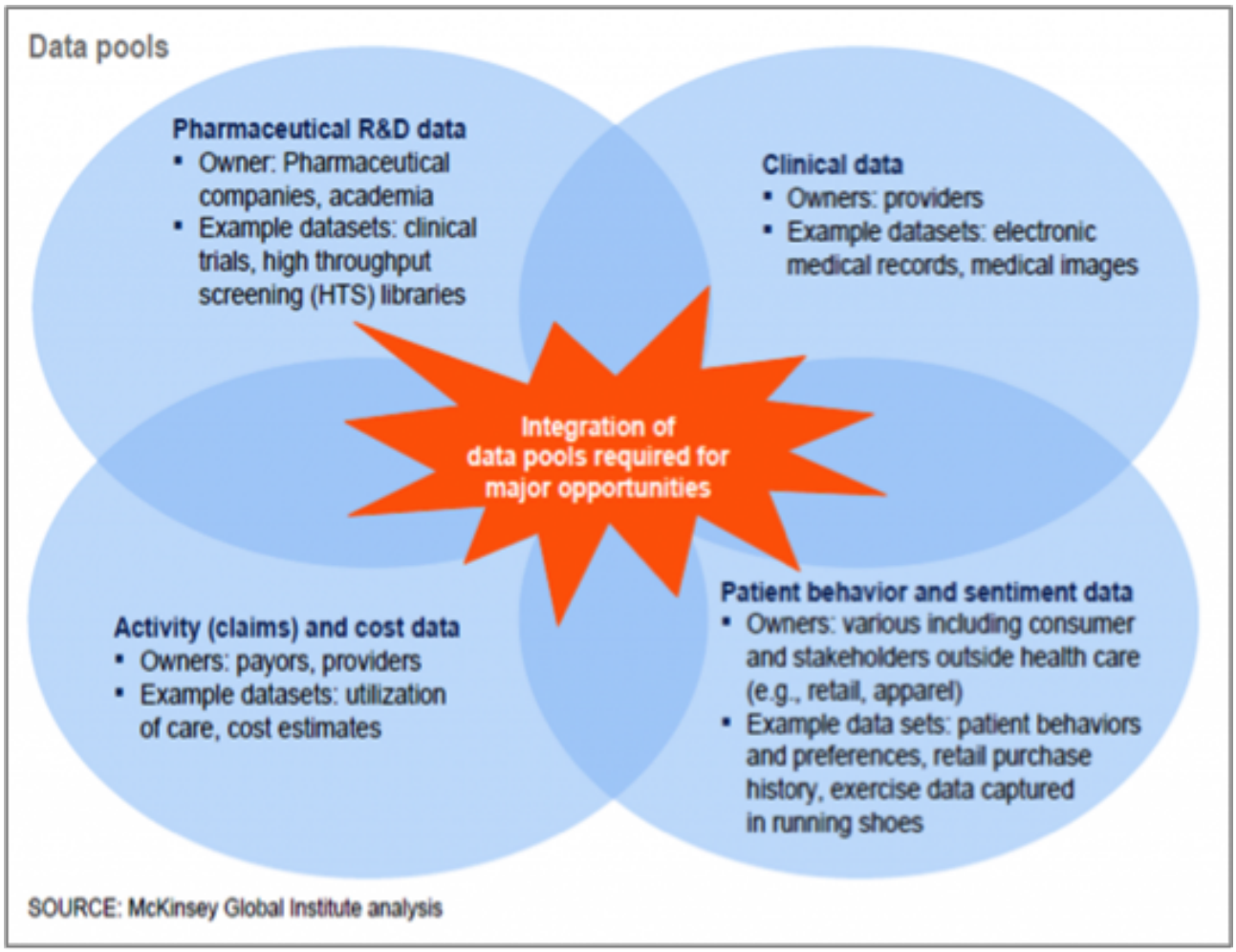


Figure 6.14: Sources of data in the health-care sector

valuable. It is suggested that patient mortality can be reduced by 20% by analyzing streaming patient data. Today health-care data is being utilized to support research e.g., research in genomics and proteomics, transform data into information, support self care, support providers, increase awareness, and uncover useful factors or variables that can help enhance treatment methods. Figure 6.15, showcases one such Big Data application area in the health-care sector.

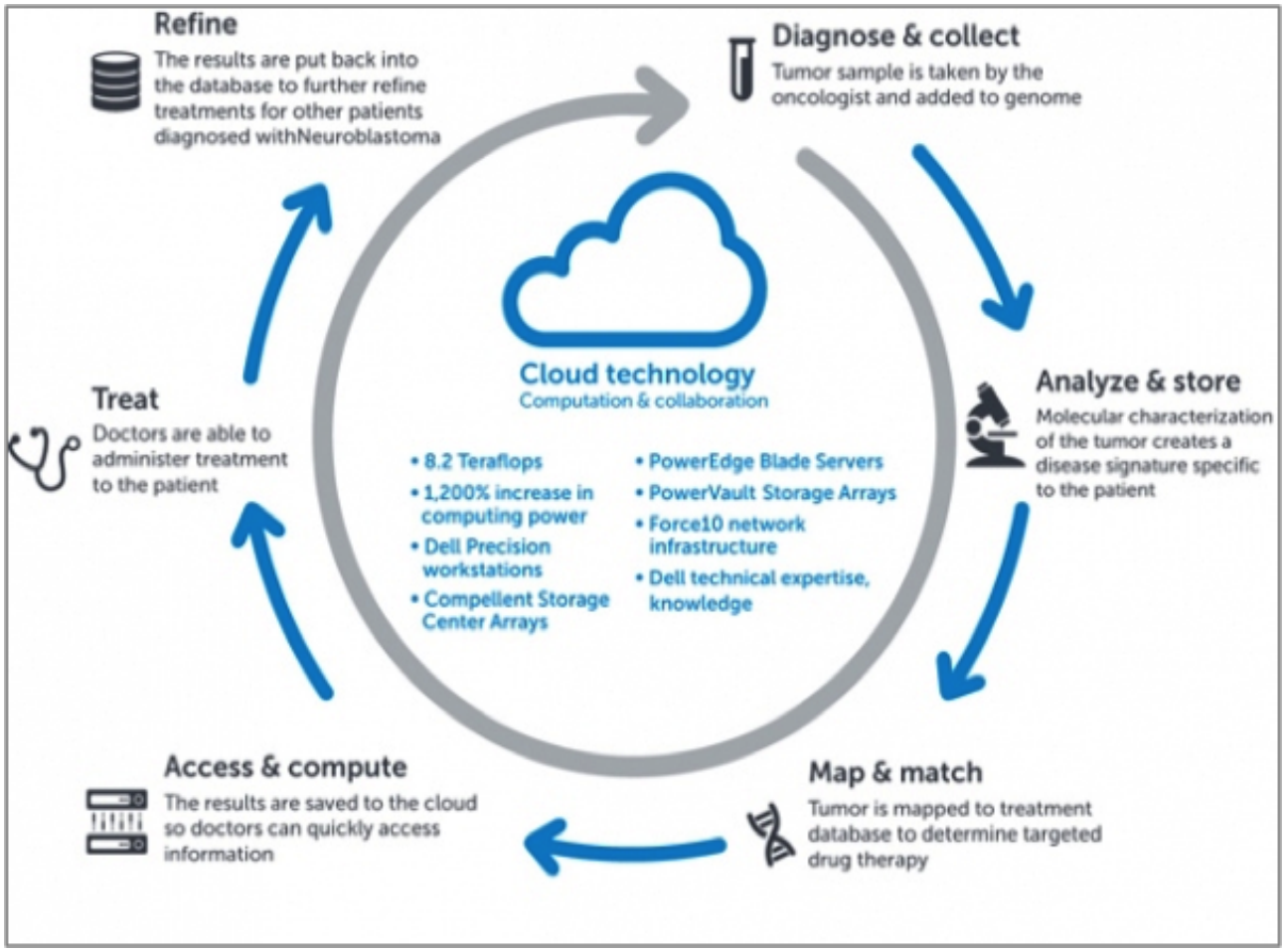


Figure 6.15: An application of Big Data in the health-care domain

Big Data Examples in the Health-care Sector

Today, Big Data analytics is being increasingly deployed in several areas ranging from evidence based medicine to health-care fraud detection. For instance, as part of the Affordable Care Act in the USA, patient-centered outcomes research institute is gearing up to gather data on as many as 12 million patients over long periods to determine which treatments are the most efficacious for a given ailment. Several agencies in the USA today are resorting Big Data for health-care fraud detection. In the USA, Strike force teams are already using advanced data analysis techniques to identify high-billing levels in health care fraud hot spots. They are utilizing analytics to gain insights into areas like: (i) assessment of payment risks associated with each provider, (ii) over-utilization of services in very short-time windows, (iii) patients simultaneously enrolled in multiple states, and (iv) up-coding claims to bill at higher rates etc. Furthermore, Big Data is having important implications in several other areas e.g., it is being used to uncover important correlations between adverse health events and various activities and develop personalized and evidence based medicine [Roger Foster, 2012, Kayyali et al., 2013].

Several examples of Big Data usage can be seen in the health care industry. For example, a Wisconsin based startup, Asthmapolis, established in 2010, leverages the advances in sensor technology and mobile data monitoring to help people manage their asthma more effectively, thereby reducing the costs for those suffering from asthma and for the USA health-care systems itself[11]. It uses Bluetooth-enabled sensors that track how often people use inhalers along with location and time-of-day, analytics, and mobile apps to help people visualize and understand their triggers as well as trends while receiving personalized feedback. The collected data helps doctors identify patients who are at risk or need more help controlling their symptoms thereby helping them

[11] `https://www.propellerhealth.com/`

potentially prevent attacks before they happen. Similarly, a California based company, Glooko[12], utilizes mobile and digital technology with analytics to help better manage diabetes. Other examples include Ginger.io[13], a Massachusetts based behavioral health analytics startup that supports proactive care. It uses passively collected data from mobile phones to detect and help treat health conditions. Specifically, it interprets passively collected mobile data related to location, movement, phone calls, text messages, and time of day to predict an patient's health status. In summary, it uses individual's smart phones to identify changes in behavior that may be warning signs, especially when monitoring people with chronic issues like diabetes, depression, and cardiovascular diseases. Several health-care organizations like the Cincinnati Children's Hospitals' chronic collaborative care network and the Carolinas based hospital system Novant have teamed up with the startup to reap the benefits of its analytics platform. Another Massachusetts based health-care analytics company, GNS HEALTHCARE[14] applies industrial-scale data analytics to empower key health-care stake-holders to solve complex care, treatment, and cost challenges. It develops solutions for biopharmaceutical, diagnostic, consumer product, and medical device companies. Its analytics platform helps provide a wide range of services which help: (i) identify predictive biomarker signatures to stratify patients in clinical trials utilizing genotype, gene expression, and patient outcome data; (ii) unravel mechanisms of drug efficacy and combination therapies by building models from pre- and post-treatment data to discover the pathways and networks, through which drugs drive clinical endpoints and simulate the effects of combination therapies; and (iii) discover new therapeutic intervention points for conditions representing unmet medical needs using genetic, gene expression, and clinical outcome data.

6.4.4 Security

In recent years, Big Data analytics has emerged as a key player in the security arena with several applications in areas like homeland security and cyber security. Big Data applications are being deployed to identify the most critical and actionable items of intelligence in near real-time. It is now considered a crucial element in detecting and deterring emerging threats. Big Data analytics in this field includes proactive data mining, data fusion, and predictive analytics techniques that are applied to all available data to gain useful insights.

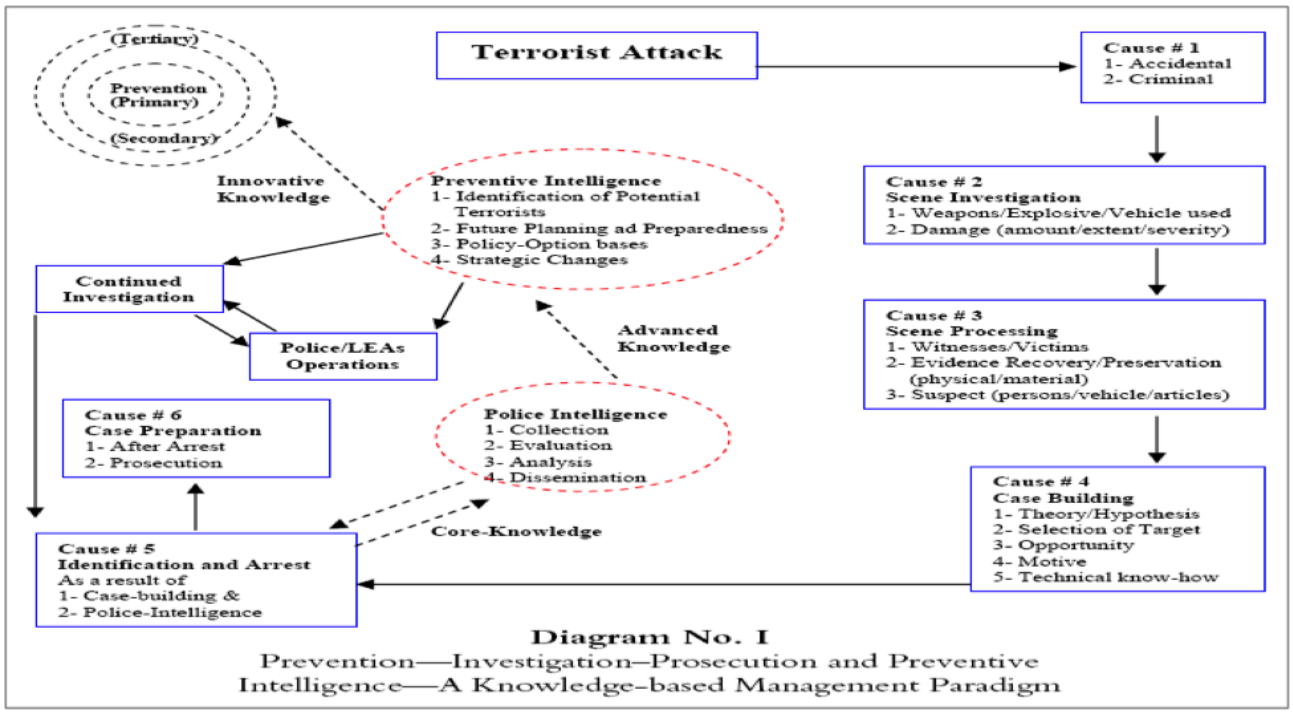


Figure 6.16: Role of data analytics in preventive intelligence

Today, it is being utilized to detect hidden relationships and attack patterns to stamp out security threats, analyze unstructured data sources like customer transactions, email, network, and flow data, for evidence of security breach, and discover and investigate high-risk behavior across variety of communication channels to avoid or proactively handle incidents [Tse, 2013, Hoffman, 2013, Curry et al., 2013].

[12] https://www.glooko.com/
[13] https://ginger.io
[14] http://www.gnshealthcare.com/

Analytics is also being increasingly used in the intelligence community to spot anomalies and subtle indication of attacks based on historical data, uncover fraud by correlating real-time and historical account activity to spot abnormal user behavior, and provide Activity Based Intelligence (ABI) and Geospatial Intelligence. Geospatial intelligence involves acquisition of high caliber Geo-spatial data from remote sensing systems, their mapping, modeling, and analysis to derive essential intelligence so as to inform decisions needed for public safety and security [Tse, 2013]. For example, advanced analytics on such high volume and velocity data helps identify and locate areas where actions need to be taken to ensure public safety and security [Hoffman, 2013].

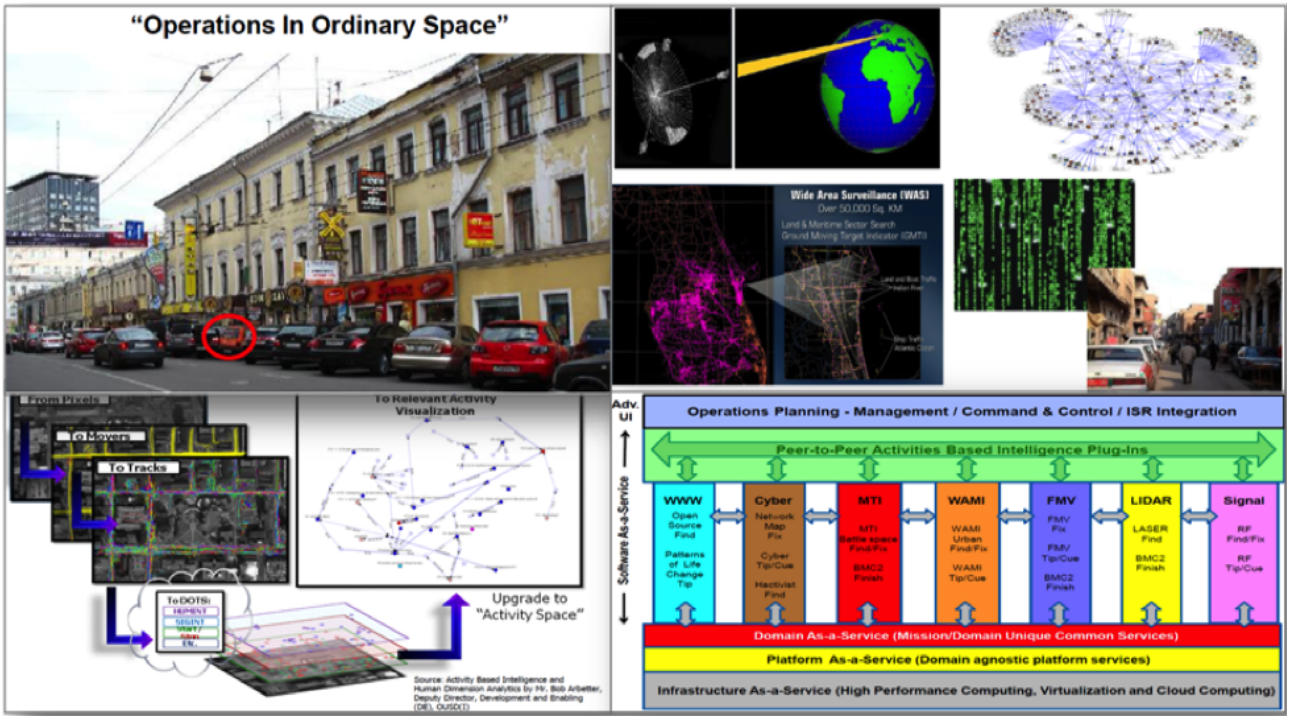


Figure 6.17: Activity Based Intelligence

Figure 6.17 showcases one ABI scenario. Here the entities are defined based on a possibly suspicious vehicle stop and include a restaurant, IT Security company, watch repair shop, internet café electronics store, travel agency, gym, youth club, bank, and apartment. The person of interest is suspected of involvement with one of the groups like state-sponsored hactivist, WMD proliferator, illegal arms dealer, bomb maker, or terrorist financier. Circled region in the figure represents suspect vehicle stop number of times: e.g., multiplication factor of 1, 2, 3, or more and duration in the range spanning from seconds to hours. Domain knowledge or heuristics is applied to link certain entities e.g., restaurant could be regarded as a meeting place, internet café could be considered as a place for communication, electronics store could be linked to the IED components, and banks could be linked to the money transfer activity. The goal of ABI is then to apply advanced techniques to discover truth from the available knowledge [Tse, 2013].

The basic principle utilized in such scenarios is that in the environments where there is no visual difference between friend and enemy, it is by their actions that enemies are visible - motion is the first indication of activity, temporal and visual patterns of change provide the context for intent. It requires an efficient methodology for analyzing how people, events, actions, and their activities interact instead of looking for an occurrence of a single stand-alone event. The problem is clearly multifaceted. It involves several steps such as detecting of the threats in real time, tracing the threats backward in time to extract supporting and contrary evidence, projecting the threats forward in time and re-tasking collection to gather evidence to verify or deny hypothesis, and reporting the threats as actionable intelligence. Its potential is also being explored to extract and correlate global intelligence for threat assessments, discover trends and monitor incoming data with fused actionable knowledge to diffuse threats, detect and deter money laundering, terrorism financing, and other illicit activities, make rapid decisions through high-resolution, mission-critical capture and in-stream processing of sensor data, rapidly detect threats through fast and efficient biometric data analysis, and deliver real-time synchronized and consistent information, which can be used to develop a common operating picture across navy, military, and other security services.

Examples in the Security Sector

Big Data solutions are already in use by the USA military [Hoffman, 2013]. The military collects too much data e.g., data from unmanned vehicles and sensors mounted onto them has spurred a wave of data collected on the battlefield. For instance their ARGUS-IS sensor system exemplifies one of the major advances being made in the world of intelligence sensors. It can stream up to a million terabytes of data and record 5000 hours of high definition footage per day. It can do this with the 1.8 gigapixel camera and 368 different sensors all housed in the ARGUS-IS sensor that can fly on an MQ-9 Reaper. Presently, the analysis of the data collected by those sensors cannot keep up with the extremely large rapidly generated data volumes and more than one person is required to operate its unmanned fleet. Their current focus is on PED (processing, exploitation, and dissemination) to gain useful insights from all the available data. Cyber4Sight[15] is another example that provides cyber threat intelligence services [Curry et al., 2013]. It uses multiple data sources to identify and monitor an organization's unique cyber security profile, determine its "attack surface", and deploy military grade predictive intelligence to anticipate, prioritize, and mitigate cyber threats. It employs an intelligence driven dynamic defense framework that includes threat intelligence (cyber4sight), incident response, preemptive response, and integrated remediation. Similar Big Data developments can be seen in the US army [Cruz, 2013]. For instance, the army has been exploring a Distributed Common Ground System-Army (DCGS-A). The system is based on cloud technology. In collaboration with High Performance Technologies Inc. the federal government agency aims to design and implement a private cloud that will convey the latest intelligence information to US troops in Afghanistan, in real or near-real time. The system will take 9 different IT systems for intelligence gathering and analysis and deploy them into a single cloud-based architecture. The system is expected to replace the other, traditional stove-piped systems that the military has been using to collect and analyze intelligence data. Unlike older systems, the new system will have cross platform interoperability.

Financial Forensics: On-going project at UB

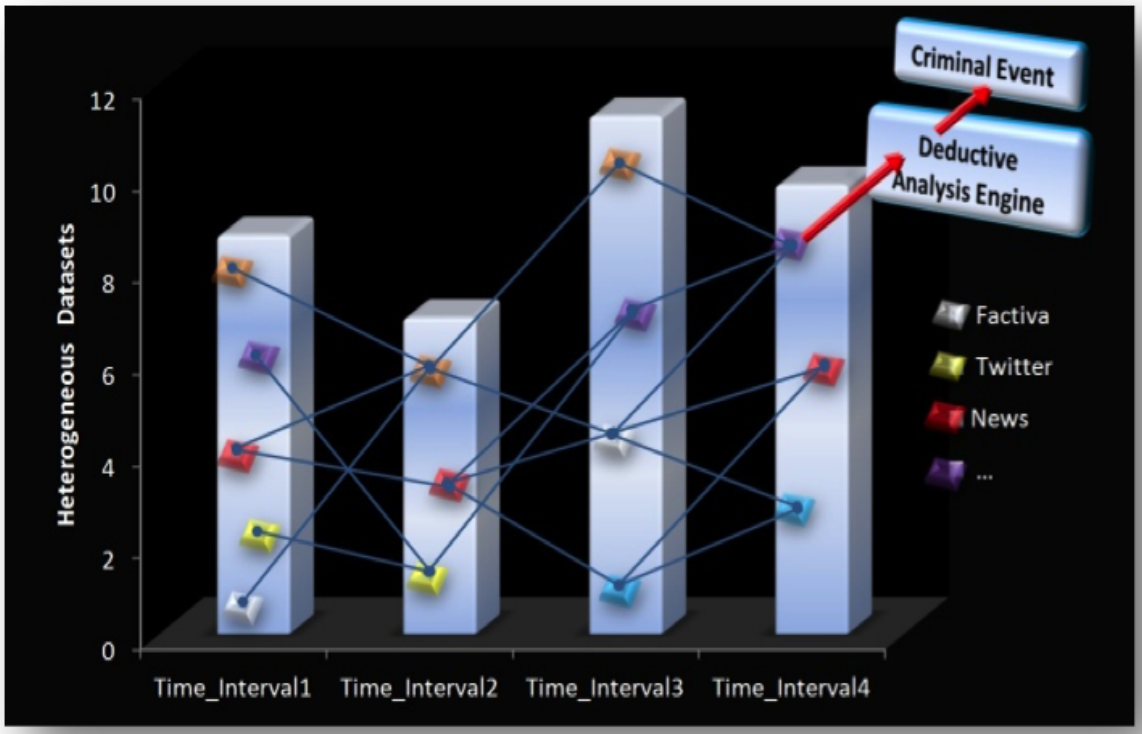


Figure 6.18: Big Data utilization in financial forensics

The goal of this research project is to develop a deductive analytics platform to detect criminal activities leading to market destabilization by establishing the underpinnings of association or co-referencing for complex attributed or labeled graph structures that are derived from large heterogeneous data sources such as trading

[15] https://www.boozallen.com/content/dam/boozallen/documents/2014/11/Cyber4Sight-brochure.pdf

quotes, litigation releases, and public news. Figure 6.18 illustrates a very high level view of our research initiative.

Despite different efforts in the data mining community the task of learning from oceans of data in finance remains challenging mainly due to the following reasons. First, existing methods developed for general usage ignore unique characteristics of the financial field and fail to take domain knowledge into the data mining procedure. Second, the speed of existing knowledge discovery algorithms still cannot keep up with the pace of data collection, which calls for more efficient algorithms to analyze noisy, heterogeneous, scattered, and large-scale data sets in real time.

Through this research we seek to establish the possibility that a data mining algorithm associates the occurrence of certain illegal events with abnormalities in market activity preceding the event. For example, in the case of the terrorist organization bombing oil wells, a data mining algorithm will detect an outlier in the joint distribution of volatility and equity prices. Such an outlier may not be separately identifiable if the terrorist organization elects to implement its insider information by shorting some stock and buying some volatility simultaneously without pushing the envelope of either strategy.

Bibliography

[Catone, 2011] Catone, J. (June 2011). How Much Data Will Humans Create and Store This Year? http://mashable.com/2011/06/28/data-infographic/.

[Cruz, 2013] Cruz, X. (May 2013). How the US Army Leverages Cloud and Big Data Technologies. http://cloudtimes.org/2013/05/14/how-the-us-army-leverages-cloud-and-big-data-technologies/.

[CTO Labs WP, 2012] CTO Labs WP (2012). Demystifying Big Data. http://ctolabs.com/wp-content/uploads/2012/10/techamericabigdatareport.pdf.

[Curry et al., 2013] Curry, S., Kirda, E., Schwartz, E., Stewart, W. H., and Yoran, A. (Jan 2013). Big Data Fuels Intelligence-Driven Security : Rapid growth in security information creates new capabilities to defend against the unknown. www.emc.com/collateral/industry-overview/big-data-fuels-intelligence-driven-security-io.pdf.

[Farris, 2012a] Farris, A. (Dec 2012a). How big data is changing the oil and gas industry. http://www.analytics-magazine.org/november-december-2011/695-how-big-data-is-changing-the-oil-a-gas-industry.

[Farris, 2012b] Farris, A. (Dec 2012b). How big data is changing the oil and gas industry. http://www.analytics-magazine.org/november-december-2011/695-how-big-data-is-changing-the-oil-a-gas-industry.

[Hoffman, 2013] Hoffman, M. (Feb 2013). Big data poses big problem for Pentagon. http://www.defensetech.org/2013/02/20/big-data-poses-big-problem-for-pentagon/.

[IBM-WP, 2013] IBM-WP (2013). Tapping the Power of Big Data for the Oil and Gas Industry. http://www-935.ibm.com/services/multimedia/Tapping_the_power_for_the_big_data_for_the_oil_and_gas_industry.pdf.

[John, 2013] John, J. S. (Jun 2013). C3 Energy: Smart Grid's Biggest Big Data Contender. http://www.greentechmedia.com/articles/read/c3_ smart_grids_biggest_big_data_contender.

[Kayyali et al., 2013] Kayyali, B., Knott, D., and Kuiken, S. V. (Apr 2013). The big-data revolution in US health care: Accelerating value and innovation. http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care.

[Leber, 2012] Leber, J. (May 2012). Big Oil Goes Mining for Big Data. https://www.technologyreview.com/s/427876/big-oil-goes-mining-for-big-data/.

[Marie Bienkowski, 2012] Marie Bienkowski, Mingyu Feng, B. M. (Oct. 2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf.

[Nicholson, 2012] Nicholson, R. (2012). Big Data in the Oil and Gas Industry. https://www-01.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/RICK%20-%20IDC_Calgary_Big_Data_Oil_and-Gas/$file/RICK%20-%20IDC_Calgary_Big_Data_Oil_and-Gas.pdf.

[Roger Foster, 2012] Roger Foster (May 2012). Top 9 fraud and abuse areas big data tools can target. http://www.healthcareitnews.com/news/part-3-9-fraud-and-abuse-areas-big-data-can-target.

[Siemens Case Study, 2012] Siemens Case Study (2012). PJM Interconnection: Integrating state-of-the-art, large-scale energy management systems. http://w3.usa.siemens.com/smartgrid/us/en/transmission-grid/products/ems-toolkits/Documents/PJM_CaseStudy_LoRes.pdf.

[Tse, 2013] Tse, E. C. (Mar 2013). SC Spring Retreat Activity Based Intelligence Challenges. http://imsc.usc.edu/retreat2013/iWatch-Ed.pdf.

[Upbin, 2012] Upbin, B. (Feb 2012). KNEWTON Is Building The World's Smartest Tutor. http://www.forbes.com/sites/bruceupbin/2012/02/22/knewton-is-building-the-worlds-smartest-tutor/#469a7bf3d251.

Chapter 7

High Performance Computing Applications in Smart-Grid, Oil & Gas

Ankur Narang
Mobileum

In this chapter we present the computational challenges prevalent in Smart Grid and Oil & Gas domains and how high performance computing is enabling key innovations in these domains. We present an overview of throughput challenges in multi-physics, multi-scale, multi-dimensional simulations, non-linear stochastic optimization, non-linear dynamics and control, large-scale inverse problems and large-scale data handling and visualization. High Performance Computing techniques including parallelism optimizations for throughput improvement and latency reduction for real-time applications are presented in some detail.

7.1 Introduction

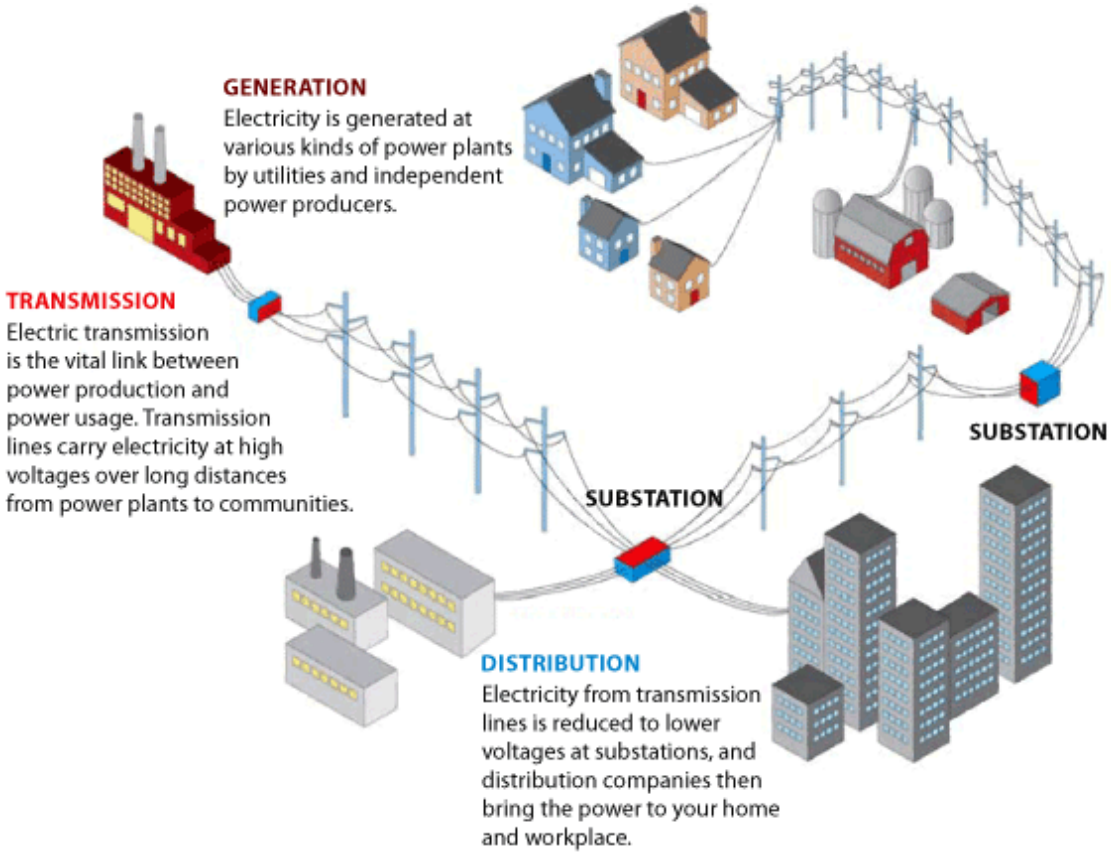
Traditional Power Systems have been designed to generate power at a single location, routing electricity through high voltage transmission lines to substations where voltage is stepped down and distributed over several feeders to additional transformers that step voltage down further for delivery to homes and businesses (Figure 7.1). Traditional power systems provide less information to utility operators and less control as one moves away from the generation source. Transmission lines are carefully monitored, real-time substation monitoring may or may not occur and limited metering is typically applied to distribution feeders or circuits to monitor electricity distribution voltage and other values. Some utilities can control some substation and line voltage levels and switches remotely from central stations; however, many utilities use remote equipment that senses voltages and other power characteristics and makes predefined adjustments.

Engineering calculations or models are used to determine required system characteristics and to evaluate distribution system modifications required to meet new loads or other changes in the distribution systems. Inputs to these models are periodically measured to recalibrate the models. The limited metering and communications from substations and points on feeders provide limited visibility into the current operating status of the distribution system and consequently provide only limited information on transformer loading, line losses, voltage sags and swells and other distribution system characteristics such as outage detail.

Traditional power systems face challenges integrating distributed energy resources including solar, wind and combined heat and power. Difficulty in monitoring and controlling distributed electricity generated from these sources and their intermittent nature can de-stabilize the grid. Increasing use of electric vehicles also contributes to concern over the ability of traditional power systems to adapt to future electricity demands. Most traditional power systems use electro-mechanical meters collecting readings manually once a month providing utility customers with little detail on how or when they use electricity. While some commercial and industrial utility customers are billed for their electricity on an hourly or 15-minute basis with rates that vary by time-of-day and season, most residential customers and smaller/medium-sized commercial and industrial customers face flat or simple block rates that reflect little if any of the time-of-day and seasonal variation in the cost of providing electric service.

Smart Grid technologies and applications, support a very different power system model than represented by the traditional power system. Five basic characteristics define the new smart grid power system model:

- *Extensive metering and communication throughout the distribution system:* Smart grids meter individual customers and individual grid equipment throughout the distribution system including transformers,



Source: U.S. Department of Energy. "Benefits of Using Mobile Transformers and Mobile Substations for Rapidly Restoring Electric Service: A Report to the United States Congress Pursuant to Section 1816 of the Energy Policy Act of 2005." 2006.

Figure 7.1: Traditional Power System Design

switches, capacitor banks, voltage regulators and other equipment. This information is relayed back to the utility typically through a combination of communications systems.

- *Two-way communication and power flows:* Instead of a traditional system that sends power in one direction (to the customer) and returns information in the opposite direction (back to the utility) at monthly intervals, the smart grid accommodates frequent and on-demand two-way information and power delivery.
- *Utility Customer Participation:* Utility customer participation is one of the most important smart grid system characteristics. Not only do customers provide electric production with solar, combined heat and power and other technologies, they can actively respond to signals from the utility to reduce electricity use during peak period times or during situations where the power system is stressed.
- *Increased control:* Smart grids increase utility control of distribution system equipment and operating characteristics and increase control of customer demand response (reduction in customer hourly loads at peak hours).
- *Coordination and integration:* Smart grids coordinate and integrate new metering, communication, control and customer engagement technologies and strategies, leveraging technologies and programs to achieve objectives across the entire utility system.

Smart grids take advantage of many of the dramatic changes in communications and solid state electronics that have occurred over the last several decades. Smart grids apply metering, communications and control strategies across the entire distribution system to optimize the delivery of electricity, integrating distributed energy resources and engaging customers with technologies and incentives to accommodate cost and efficiency considerations. Figure 7.2 illustrates the main components of the Smart grids:

1. *Smart Infrastructure System:* that includes Smart energy subsystem (Power generation, Transmission grid, Distribution grid), Smart Information subsystem (Smart meter, Sensor, Phasor measurement unit, data modeling and information analysis/integration) and Smart communication subsystem (wireless and wired networks),
2. *Smart Management System:* that includes Management methods and Tools such as Optimization, Machine Learning, Game Theory and Auctions, and,
3. *Smart Protection System:* that includes System reliability and failure protection along with security and privacy.

The development of smart grids for electrical power distribution is essential for making the most efficient use of precious resources. Infusing intelligent systems into the energy infrastructure, energy companies can create smart grids that incorporate new, renewable energy sources, optimize distribution to meet changing needs and reduce outages. To realize the full potential of smart grids, energy organizations will need to look beyond Terascale and Petascale computing systems. Organizations need the computational power to analyze tremendous volumes of data, streaming in from thousands of sensors, and deliver real-time models with billions of individual elements at millisecond-scale precision.

Exascale systems can provide the computational performance required for the planning and test of smart grid elements while laying the foundation for real-time, dynamic grid operations. Grid architects and engineers could use Exascale systems to plan integration of new technologies or renewable energy sources into electric power generation. They also could develop models that anticipate customer demand and analyze the possible impact of certain catastrophic events on electrical grid function. Ultimately, Exascale systems could be incorporated into electrical grid operation, providing the compute power for real-time management of generation capacity that takes into account constantly changing supply and demand variables.

7.2 HPC Application Areas in Smart Grid

7.2.1 Optimization

Optimization problems in smart grid range from the classical economic dispatch, which can be modeled as non-linear programming problem and solved by the gradient techniques, to the stochastic dynamic programming formulation of the multi-reservoir optimization problem. Other interesting and complex problems are the transmission and distribution network expansion planning and contingency constrained optimal power flow amongst many others [Eto and R.J., 2011]. In most of these problems, a realistic formulation leads to highly non-linear relationships, non-convex functions, integer and continuous mixed variables, and many other challenging cases of the mathematical models. Here, one could also encounter combinatorial optimization problems with exponential increase in computational requirements. Additionally, the number of constraints could reach tens of thousands or higher. By utilizing decomposition techniques, parallel solutions could be effectively utilized.

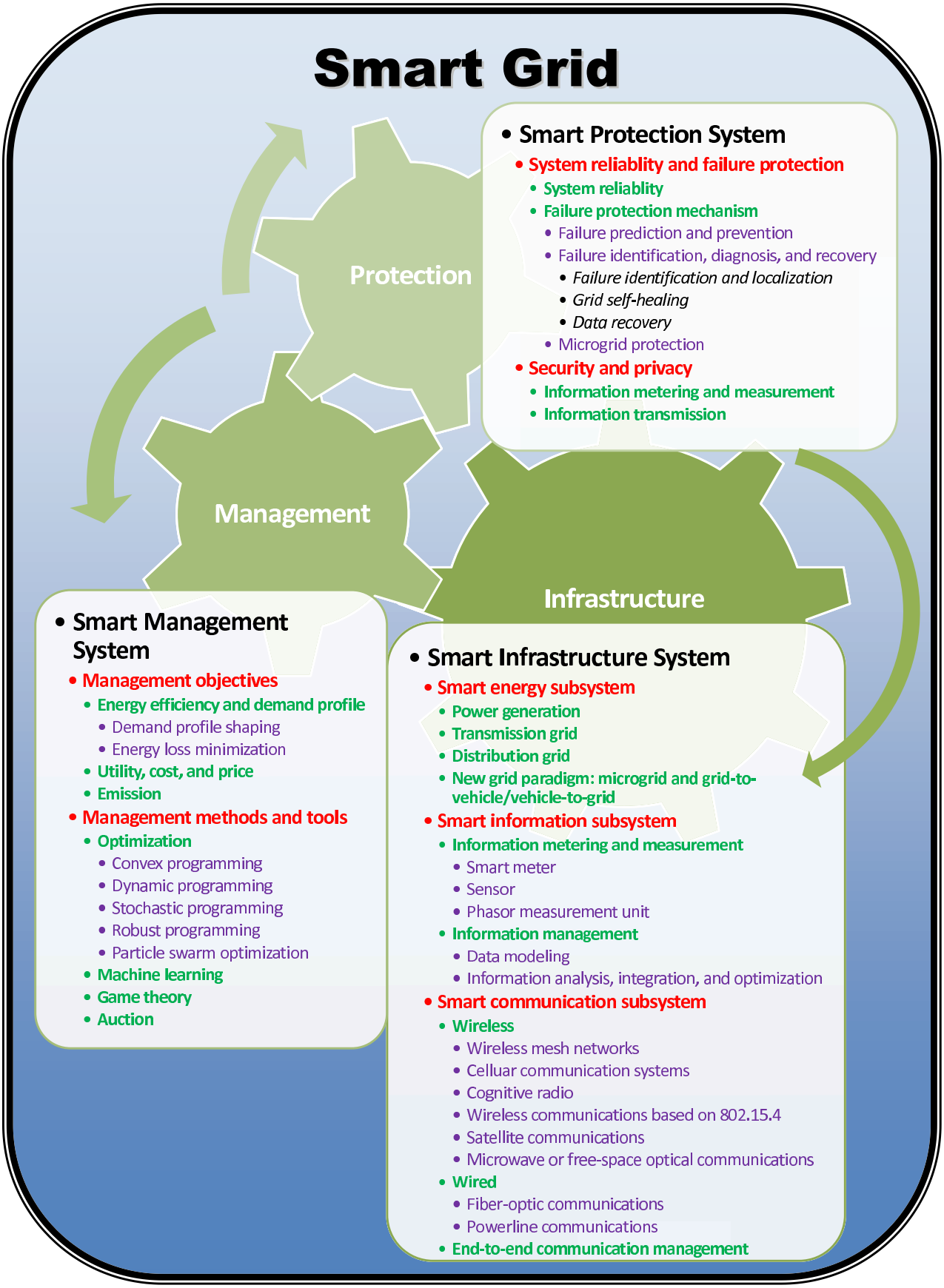


Figure 7.2: Smart Grid: High Level Overview

Multi-scale Optimization

The area of *multi-scale optimization* has been the subject of considerable research [Eto and R.J., 2011] among the finite element, multi-grid, and combinatorial optimization communities. Hierarchical optimization, which is a type of multi-scale optimization, has been the subject of research by the power systems community in the last decade [Cheng, 2009] [Torres-Hernandez and Velez-Reyes, 2008]. Hierarchical optimization is useful for handling large-scale systems that exhibit scale or organizational properties that can be modeled using a hierarchical arrangement. Computational solutions are obtained by decomposing a system into subsystems and optimizing them while considering the coordination and interactions between the subsystems. Hierarchical optimization takes the interaction into account by placing a coordinator above the subsystems to manage it. As a result, the entire system is optimized while considering the interaction between the scales by means of coordinated information. Most of the work related to power system hierarchical control and optimization addresses two-level problems. From research perspective, one needs study of the relevant industry subsystems that would participate in multi-level hierarchical control including staged coordination from ISO to utility to micro-grid to end consumer and even to individual devices. Applications such as EV supported frequency regulation would require such a level of coordination. Phenomena of interest include modeling the heterogeneity of the actors at the various layers, the coordination signals that need to be passed, and the protocols for information passing.

Scale decomposition is an important concept in multi-scale optimization, which consists in obtaining the solution to loosely coupled optimization problems at various scales through relaxation of interdependencies among scales [Chakraborty and Arcak, 2007] [Attinger and Koumoutsakos, 2004] [Muralidharam et al., 2008] [Liu et al., 2010]. A set of similar methods fall under the category multi-scale relaxation, which has been applied extensively in multi-grid algorithms. Scale decomposition in power systems is relevant when dealing with spatial and temporal dimensions. While little literature on multi-scale decomposition for power networks has been identified, multi-scale relaxation methods could be applied to problems such as wind generation including reserve considerations. The basic idea is to use wavelet decomposition of wind variability to identify and model wind behavior coupled to system control at various frequency ranges. Then the outer coordination control would utilize scale decomposition to solve the global coordinated problem.

Multi-scale decision making is a growing area in multi-scale optimization, which fuses decision theory and multi-scale mathematics. The multi-scale decision making approach draws upon the analogies between physical systems and complex man-made systems. On the theoretical side, the aim is to develop a generalized statistical formalism that describes a large variety of complex systems in an effective way. Rather than taking into account every detail of the complex system, one seeks for an effective description with few relevant variables. An example of the need of such formalism is decisions associated with transmission expansion investment for a portfolio of large scale distributed wind resources. Formulating and selecting transmission expansion alternatives is a function of not only the expected amount of power produced and its variance, but also the overall system conditions, including simultaneous operation of other renewable and conventional generation, the existing and planned transmission, $N-k$ security, and load profiles. The level of required transmission expansion also depends on the combined requirements of transmission for reserve provision from other locations and system production changes in the minutes range, as well as the effect of load growth and policy, acting up to the multi-year range. The application of rather simple transmission reliability metrics has demonstrated that the existing transmission planning methods do not fully capture the optimization problem complexity and reliability constraints. As a result, several systems in the U.S. are less reliable today than in previous years [Grijalva et al., 2007]. Large scale installation of renewable energy can compound this problem considerably if better methods to determine optimal transmission capacity and expansion needs are not developed.

Unit Commitment and Economic Dispatch with Stochastic Analysis

At the heart of the future smart grid lie two related challenging optimization problems: *unit commitment* (UC) and *economic dispatch* (ED). When operational and physical constraints are considered not only under normal operating conditions, but also under contingency conditions, the UC and ED problem becomes the security constrained UC and ED problem [Eto and R.J., 2011]. Both these problems can leverage HPC to meet real-time operational requirements.

UC is the problem of finding an optimal ramp up and down schedule and corresponding generation amounts for a set of generators over a planning horizon so that the total cost of generation and transmission is minimized, and a set of constraints, such as demand requirement, upper and lower limits of generation, minimum up/down time limits, ramp up/down constraints, transmission constraints, and so forth, are observed [Padhy, 2004], [Sheble and Fahd, 1994]. Economic dispatch is the problem of determining the most efficient, low cost and reliable operation of a power system by dispatching the available electricity generation resources to the load on the system. The primary objective of economic dispatch is to minimize the total cost of generation while satisfying the physical constraints and operational limits. The economic dispatch problem plays an important role in power system analysis [Wood and Wollenburg, 1996], [Glover et al., 2008], especially for planning,

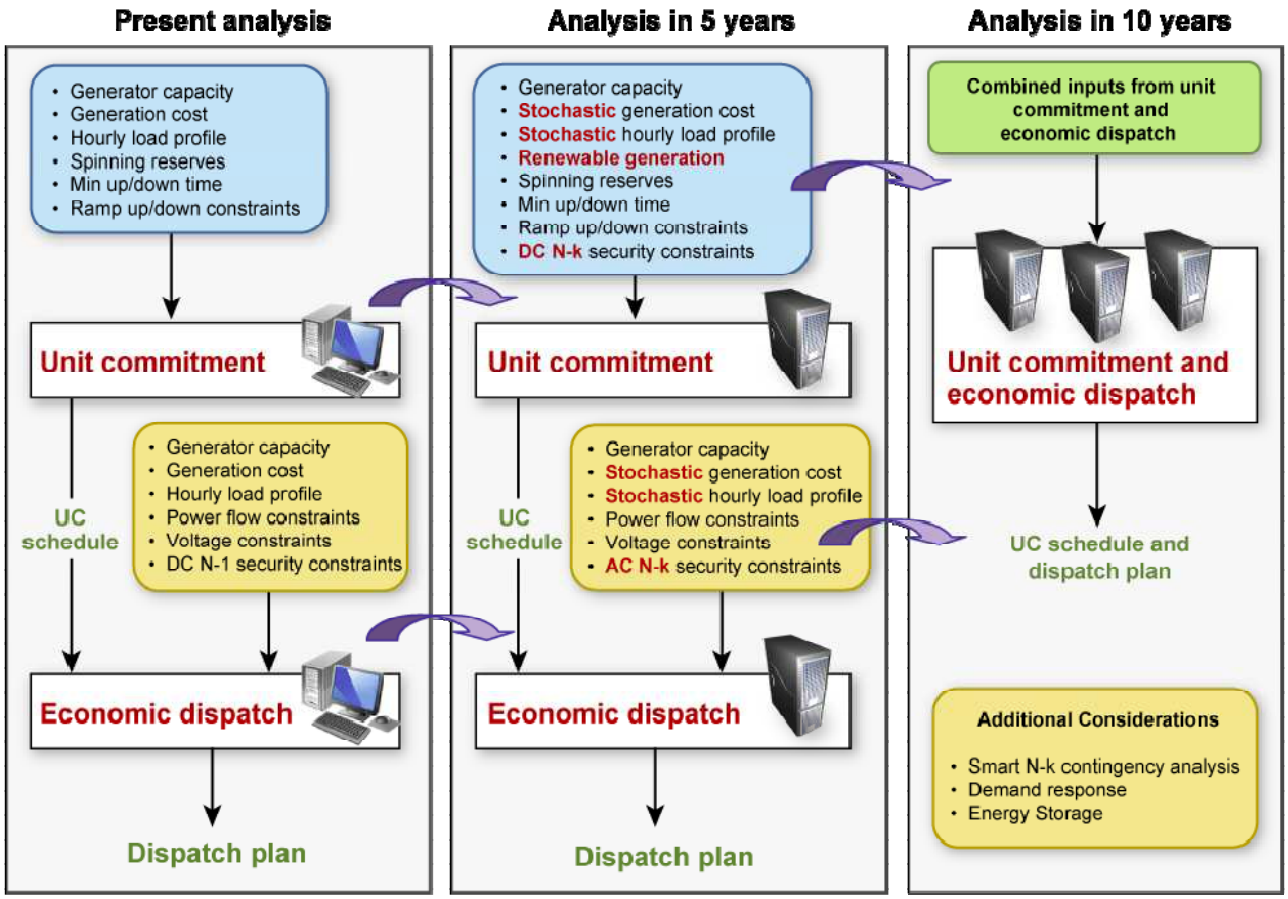


Figure 7.3: Unit Commitment & Economic Dispatch: Present and Future

operation and control of power systems. Although these two problems are intermingled with each other, most of the current theoretical and practical effort treats them separately, due to the computational difficulty of solving a single unified problem. As Figure 7.3 illustrates, present solutions to the unit commitment problem consider only a direct current (DC) approximation of the alternate current (AC) transmission constraints. This problem observes any generator related constraints, demand constraints, and linear transmission constraints. The output of this is an optimal schedule for generators in a twenty four hour time horizon, and is given as input to the economic dispatch problem. Economic dispatch problem then handles the original AC power flow constraints and outputs a dispatch plan: how much power to produce from each generator, and how to transmit the power over the network. To account for unexpected failure of generators and transmission lines, current unit commitment practices enforce spinning reserve requirements, allocating a fraction of a generator's capacity to reserves. Similar contingency analysis is also performed in economic dispatch, making sure that the load at each node of the network can be satisfied in the case of a failure of one of the generators, transmission lines, or other devices, which is also called $N-1$ contingency analysis.

The ultimate goal is to solve practical instance of these two problems together, considering other relevant issues such as demand response and energy storage. As will be shown, this definitely requires more algorithmic advancement that uses high performance and parallel computing environments. Today, a typical commercial integer programming solver can handle a unit commitment problem with 100 units, 24 time periods, and 50 uncertainty scenarios. Real life instances consist of several thousand buses, more than one thousand generators, 48 to 72 time periods, more than one hundred contingencies, and a few hundreds of scenarios. Solving such a large scale real life instance of the unit commitment and economic dispatch problem together, along with all other relevant issues, is a grand challenge that will reinforce need for fast and parallelizable decomposition algorithms.

Both UC and ED can be formulated as nonlinear optimization problems (NLP) and mixed-integer NLPs that are, in general, non-convex and nonlinear. Existing industrial solutions to these two problems have been traditionally dominated by the Lagrangian relaxation methods, and only recently they have switched to general purpose integer programming solvers [Eto and R.J., 2011]. Academic solutions are more diverse, but they are typically demonstrated on much smaller IEEE bus cases than the real life scenarios.

There are a number of limitations to existing solutions: (1) the inability to solve large scale real-life power system operating problems in a given time, (2) the sub-optimality of the solutions; globally optimal solutions

are seldom attained, (3) the lack of guarantee of convergence to a feasible solution, (4) limited consideration of contingency scenarios (typically focusing on $N-1$ but not on $N-k$ contingencies), (5) the use of deterministic formalism for handling unpredictable loading and generation, and (6) the lack of support for real-time operation.

To achieve significant integration of intermittent renewable energy sources and enable demand response will require that the formulations of both unit commitment and economic dispatch for system operation be enhanced. Advanced optimization techniques targeting globally optimal solutions and with guaranteed convergence need to be developed. To support real-time secure operations, $N-k$ contingencies with renewable integration must be considered. We believe future solutions to both unit commitment and economic dispatch problems need to be implemented in a hybrid computing environment that supports:

1. Parallel evaluation of multiple scenarios (resulting from different contingences or loading/generation profiles)
2. Parallel execution of decomposable formulations for large scale optimization problems
3. Large scale optimization with guaranteed convergence to high quality solutions

Below, we provide details on the Security Constrained Economic Dispatch problem and various solution approaches and scalability challenges.

Security Constrained Economic Dispatch (SCED)

One type of SCED formulation is the preventive SCED [Alsac and Stott, 1974], i.e., it minimizes some generation cost function by acting only on the base case (such as, contingency-free) control variables subject to both the normal and abnormal (with one of the $N-1$ contingencies) operating constraints. For k contingency scenarios, the problem size of the preventive SCED is roughly $k+1$ times larger than the classical (base case) economic dispatch problem. Solving this problem directly for large-scale power systems with numerous contingencies would lead to prohibitive memory requirements and execution times [Eto and R.J., 2011].

The second type of SCED problem is called the corrective SCED, with the assumption that post contingency constraint violations can be endured up to several minutes without damaging the equipment [Monticelli et al., 1987]. The corrective SCED allows post-contingency control variables to be rescheduled, so that it is easier to eliminate violations of contingency constraints than the preventive SCED. The optimal value of corrective SCED is often smaller than that of preventive SCED, but its solution is often harder to obtain, since it introduces additional decision variables and nonlinear constraints.

The third type of SCED problem is an improvement of the aforementioned corrective SCED. Capitanescu et al. [Capitanescu and Wehenkel, 2007] recognized that the system could face voltage collapse and/or cascading overload right after a contingency and before corrective action is taken. Therefore, their improved formulation imposes existence and viability constraints on the short-term equilibrium reached just after contingency occurrence and before corrective controls are applied. There are very few solutions to this formulation, but one exception is Yi [Li, 2008] that utilized the Benders decomposition method to solve this problem.

The inclusion of contingencies beyond $N-1$ for future power grid operation may increase the complexity and scale of the problem by several orders of magnitude. Therefore, great effort has been devoted to the development of parallel algorithms for large-scale problem formulations. In this case the SCED problem is decomposed and distributed on a number of processors with each one independently handling a subset of the post-contingency analysis.

There are currently two promising approaches for parallelism: one related to interior-point methods [Qiu et al., 2005] and one using Benders decomposition [Li, 2008], [Li and McCalley, 1994]. For interior-point methods, at each primal dual iteration, we need to solve a large scale system of linear equations. Because the matrix associated with these linear equations has a blocked diagonal bordered structure, by exploiting this fact, researchers have shown that the system of linear equations can be solved efficiently in parallel. For example, more than 10× speedup can be obtained on a system with 16 processors [Li and McCalley, 2009]. On the other hand, Benders decomposition is a two-stage solution method consisting of a base-case problem and a list of contingency subproblems. Since the evaluation of different contingencies can be done independently, this formulation is amenable for parallelism. One obvious benefit of exploiting parallelism in solving these problems is that it makes the complexity linearly dependent on the size of the problem as opposed to the quadratic growth for sequential computation.

7.2.2 Massive Data Processing

Electricity grid operators have traditionally had to contend with large amounts of data, from sources such as SCADA systems, disturbance monitoring equipment, outage logs, weather logs, and meter readings, to deduce useful information about the condition of the grid that would help them to better understand the grid and make correct decisions. Future electricity grids will, however, take the volume of data up by orders of magnitude both

in terms of size and frequency of measurement as high resolution data sources are integrated into the system. Some of the key sources of this future massive data are:

- Synchrophasor data from increased phasor measurement unit (PMU) deployment [Mahendra Patel, 2010]
- Energy user data from the millions of smart meters [FERC, 2008]
- High rate digital data from increasingly networked digital fault recorders (DFR)
- Fine grained weather data [Li et al., 2010]
- Energy market pricing signals and bid information [Niimura, 2006]

This high resolution data on grid, environmental and market conditions has the potential to tremendously improve our understanding of the interaction of the grid with its operating environment and our situational awareness during grid operation, enabling much closer to optimal planning and operation of the grid. One key link required to realize this vision is the ability to process this large amount of heterogeneous data, both offline and in real-time, to extract useful information from it; information such as trends and patterns in massive historical data, anomalies in grid behavior, the likelihood of imminent disturbances and models for load forecasting.

A scalable high performance computational infrastructure that can handle these high volumes of data and efficiently perform the data analytic tasks required for extracting relevant information from the data, must satisfy special processing requirements of these massive data analytic tasks. In general, there are three kinds of data processing needs:

- Distributed and parallel processing of large amounts of stored historical data (*data at rest*). An example of this type of processing would be post-event diagnosis.
- Real-time low latency processing of streaming data, or stream computing (*data in motion*). An example of this type of processing would be in a wide-area monitoring system (WAMS) where high frequency real-time PMU and other grid data are continuously processed to compute system reliability indicators.
- Along with these, there will also be a need for highly scalable database systems that can quickly perform a variety of queries on petabyte scale data.

An example for distributed processing of massive data at rest is MapReduce [Dean and Ghemawat, 2010]. It is particularly attractive because of its ease of development and deployment, and the availability of the open source implementations [Huan and Orban, 2011], [White, 2009]. For data in motion, there are several implementations of stream computing, including the open-source S4 [Neumeyer et al., 2010]. A third relevant technology is hardware accelerated data warehousing [Francisco, 2009], [Bakkum and Skadron, 2010], which typically exploits field programmable gate array (FPGA) based special purpose processors or graphics processing units (GPUs) to implement extremely fast database queries.

Two key enablers of a responsive, resilient and self-healing smart grid are wide-area monitoring system (WAMS) and situational awareness. Wide-area monitoring would require management and processing of detailed monitor data from a large number of data source spread out across the grid geography, substantially increasing the volume of data from today. Situational awareness would require management and near instantaneous processing of data streaming from the WAMS, substantially increasing the rate of data processing that is usually supported today. An illustrative example for synchrophasor data is available from [Mahendra Patel, 2010]: 100 PMUs sampling at 60 samples per second will generate 4,170 MB/sec of data and the tolerable latency of transporting and processing the data for real-time situational awareness is in the order of only $10s - 100s$ of milliseconds. Furthermore, the PMU data collected by the Tennessee Valley Authority from 90 PMUs over 2009 amounted to roughly 11 terabytes (TBs). With projections of increasing PMU sampling rate and the number of PMUs deployed [NASPI, 2009], just the PMU data across the North American ISOs can amount to 100s of terabytes to a petabyte (PB) over a year [Mahendra Patel, 2010]. Similarly large volumes of data are expected from smart meters in the Advanced Metering Infrastructure. This is orders of magnitude more in volume and processing rate compared to today's SCADA systems: TVA's SCADA system accumulated only 90 GB of data in 2009 from 105,000 points of measurement [Mahendra Patel, 2010]. These high resolution data sources are being deployed with a vision to enable new applications that will make the grid significantly smarter. A key requirement for enabling this vision is a data processing infrastructure that can scale gracefully to handle the increasingly large volumes and variety of data, and increasingly complex and numerous applications that will consume this data in parallel.

Some of the *killer* applications enabled over the next decade by high resolution synchrophasor data, as foreseen by the North American SynchroPhasor Initiative (NASPI) [NASPI, 2009], and the corresponding type of data processing needed, are:

- Dynamic state estimation: high rate streaming data
- Oscillation monitoring: high rate streaming data
- Real-time controls: high rate streaming data
- Post disturbance analysis: high volume stored data

Although these are examples for synchrophasor data, other sources of data, such as the AMI, digital fault recorders and fine-grained weather data, can also be exploited to support the needs of a smart grid. The other sources of data will also pose challenges with handling high volume stored data or low latency processing of streaming data.

A large number of smart grid applications will look for patterns or features in the data that provide specific and useful information regarding the grid. One example is the class of methods use to detect anomalies in the measurement time series data that may be streaming in from PMUs, like the event detector from Oak Ridge National Lab (ORNL) that uses k-median clustering [Bank et al., 2009]. Within a sliding window, PMU data points are grouped into two clusters: points before and after a hypothetical *event*. If the distance between the two clusters is large enough then the event is flagged as having the potential to propagate and cause cascading failures. A more general time-series anomaly detector, like the one in [Monedero et al., 2007], may be one that uses classifiers from machine learning, for example neural networks or support vector machines. The classifier is trained offline to predict with high confidence whether streaming time-series data has anomalies. This may be done by creating thousands of waveforms with disturbances injected into them, which are then used to train the classifier to detect if an incoming signal has a voltage, frequency or harmonic disturbance.

Stream programming is typically done by creating a data-flow graph [Gedik et al., 2008] of operators, which performs the computation required by the application. The inputs to the data-flow graph can be data from a variety of sources, such as internet sockets, spreadsheets, flat files, or relational databases, and may be consumed by one or more input operators in the graph. A synchronization primitive is an example of an operator which consumes data from multiple data streams and then outputs data only when data from each stream is read. This can be a useful operator for PMU data which can arrive at different times from different PDCs due to variable network delays and sampling times. Other relevant operators would be a fast Fourier transform (FFT) operator or a moving average operator. Each data element arriving at the input to an operator is typically treated as an event and the operator takes appropriate action when events occur at its input. Operators may be pipelined to perform in a sequential manner: output data from one operator is consumed by the next downstream operator. Operators may also perform in parallel if they do not have data dependencies: output from one operator may be consumed by two or more different operators working in parallel.

These operators are typically contained within containers called stream processing elements. For fast, parallel execution, the processing elements are automatically partitioned onto parallel processors and/or machines. The optimal partitioning depends on factors such as the amount and type of data streaming through different processing elements, the resource requirements for each of them and the dependencies between them. Hiding the details of parallel programming from the user greatly improves productivity and efficiency of streaming application deployment. The flexibility of input formats, the ease of developing and connecting the operators, and the automatic compilation onto parallel processors makes stream processing attractive.

Simulation free grid instability monitoring methods can also be viewed as performing pattern discovery at an abstract level. These methods attempt to perform the important task of detecting and/or predicting undesirable instabilities, but by only analyzing the incoming measurement data, without simulating any model of the grid. SVD was used in [DeMarco, 2010] on streaming PMU data to find modal frequencies and damping ratios and to detect poorly damped oscillations. The real-time evaluation of large SVD matrices with very low sub-second latency is challenging and requires efficient data stream processing frameworks which exploit parallel processing on large clusters.

Apart from the real-time applications, pattern discovery is also performed on high volume stored historical data. A good example is post disturbance analysis where the goal is to identify events and trends in historical data from a variety of sources and even regions that help explain the cause of the disturbance. Such analyses provide deep, useful insight into the behavior of the grid and are critical for developing mechanisms to avoid future occurrences of the same type of disturbances. The availability of high resolution, time-synchronized PMU data brings the potential for fast and detailed forensic analysis, but also the challenge of efficiently processing high volumes of data on distributed storage systems. High performance processing systems such as Map Reduce can scale well with increasing data volumes and can perform highly parallel and distributed processing while maintaining data reliability with heterogeneous storage media and avoiding data bandwidth bottlenecks.

Given a data set, perform the *map*() function on each record to compute (key, value) pairs for each record. In other words, the user defined *map*() function is an application specific mapping from the domain of data records to the domain of keys and values. The (key, value) pairs are grouped together by key and each group is processed by the user-defined *reduce*() function to compute corresponding outputs. Figure 7.4 outlines this data

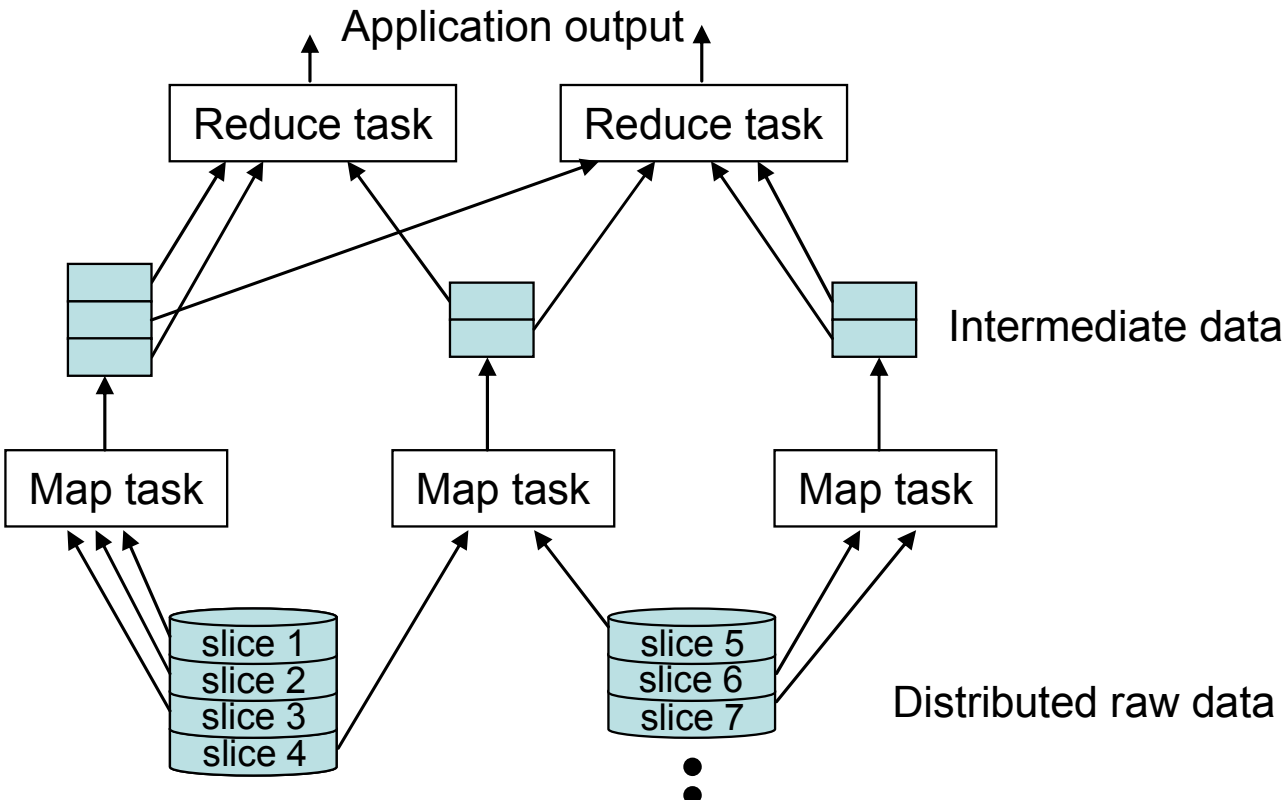


Figure 7.4: Map Reduce

processing model. Parallelism is achieved by running many instances of the *map*() function in parallel across multiple processing nodes, along with many instances of the *reduce*() function, which consume the results from the map stage. Data latencies can be kept low by dividing the stored data into slices and allocating each slice to a *map*() instance running on the local machine as far as possible. This scheme helps to reduce data transport across the network, keeping data latency low. Parallel programming is greatly simplified using this approach since the details of processor communication and system organization are typically performed by an underlying distributed file system and are hidden from the application developer. This combination of efficiency and ease of programming makes MapReduce particularly attractive for developing applications for a constantly evolving electricity grid infrastructure.

Some early steps in leveraging Map Reduce have been taken in the power grid industry. Specifically, Hadoop was used recently by Tennessee Valley Authority (TVA) to analyze 15 terabytes of PMU data collected from 103 PMUs [van Amerongen, 1988]. Their goal was to perform a forensic analysis of historical PMU data collected over the years to discover abnormal events in the power grid. The data collected was in the IEEE synchrophasor format C37.118 2005 [Min and Shengsong, 2005] that consists of frequency and three phases of voltages and currents. Signal processing algorithms such as FFT and low/high pass filters were used to discover abnormal patterns and grid instabilities. Even so, 15 TB of data is much smaller than the petabyte range that is expected as grid instrumentation scales up over the coming years.

An example data processing for the smart grid that uses stream computing for real-time applications and MapReduce for high data volume offline applications. Many data processing tasks need to access some subset of a massive data set. For example, to analyze trends in all the smart meter data from homes that experienced average daily temperature of more that 90°F during the last year, the first step is to extract the relevant records from all the recorded data (smart meter or other) for the entire year. Such data tasks are typically referred to as queries and performed using some database programming language, usually the Structured Query Language (SQL) [Chamberlin and Boyce, 1974].

Traditional database systems may not provide desired performance on these tasks as the volume of data increases toward the petabyte range. To address this challenge, the latest generation of database systems is beginning to employ hardware accelerators [Scofield et al., 2010]. In particular, dedicated field-programmable gate arrays (FPGAs) and multi-core processors are programmed to perform atomic database query tasks, like filtering unwanted records and fields or combining records from different tables.

FPGAs are particularly well suited to these acceleration tasks because of the following reasons: 1) An FPGA provides a re-programmable array of basic circuit blocks called lookup tables (LUTs). The circuit blocks and the electrical connections between them can be re-programmed to implement any logic functionality. This enables a

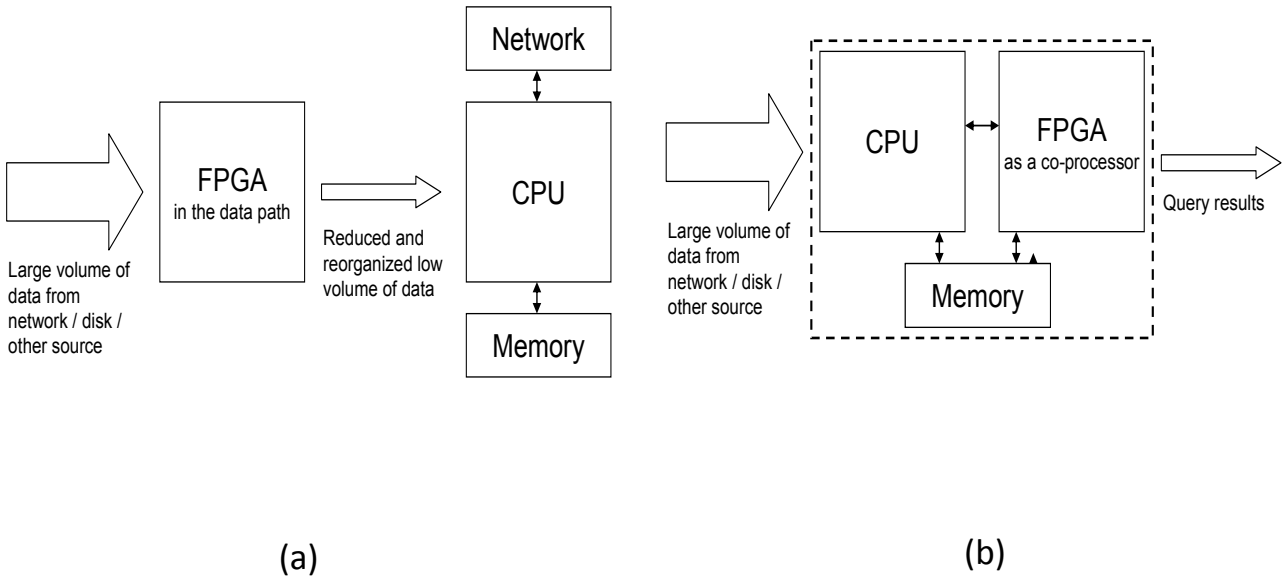


Figure 7.5: FPGA based Parallel Computing

re-optimization of the data access system if the data or application needs change over time. Such re-optimization allows the processing to occur at very high performance and very low power consumption levels. 2) A common issue when using general purpose processors for parallel processing is that the parallel jobs have to compete for shared memory resources, which can enforce slow sequential behavior instead of the desired fast, parallel behavior. FPGAs now typically contain a reasonably large number of on chip, distributed memory blocks that can be accessed in parallel. This enables many processes to operate in parallel without significant resource contention, breaking through the performance barriers imposed by sequential, von Neumann-style computation. Two ways of integrating FPGAs in the data query flow have been proposed, as shown in Figure 7.5 are:

- **FPGA in the data path**: Figure 7.5 (a) shows an outline of this architecture. High volume data coming in from the data source (disks, network, etc.) is first processed by an FPGA to reduce the amount of data significantly before sending it on to downstream general purpose processor or memory. The data reduction may be achieved by filtering out the records and fields of the data that are not required by the query being performed. The goal is to filter data in the accelerators as fast as data can be read from the disk. This removes IO bottlenecks and allows for better memory utilization. Typically thousands of filtering streams are executing in parallel on the FPGA and performance is improved by 4× - 8× and data reduction of $> 90\%$ is not uncommon [Francisco, 2009]. More advanced data processing functions may also be incorporated in the FPGA that can further increase the performance and decrease the data volume that must be processed downstream [Mueller et al., 2009].

- **FPGA as a co-processor**: The approach here is to use the customized FPGA as an assistant for the general-purpose processor (CPU). The FPGA circuitry is optimized to perform certain atomic tasks that can be combined together to form SQL queries. These atomic tasks can be SQL functions like Join, Sort and GroupBy [Chamberlin and Boyce, 1974]. As Figure 7.5(b) shows, the CPU transfers these tasks over to the FPGA while it performs other tasks required to complete queries on the data. The figure is only conceptual and the actual implementation may have further enhancements, such as multiple FPGA engines that can stream processed data between each other before sending it back to the CPU or memory, or a network of processing nodes, each of which contains a CPU, FPGA and memory within it [Scofield et al., 2010]. For a given application, these systems typically require a one-time optimization and compilation to program the FPGAs for optimum performance. One can envisage an online re-programming capability that reprograms the FPGA whenever there are changes in the application.

A similar approach that has also gained much research attention recently is to use graphics processing units (GPUs) to accelerate database query tasks [P. and Skadron, 2010]. GPUs have the advantage, like FPGAs, of supporting highly parallel processing on the chip because of highly pipelined and parallel processing elements. GPUs are able to provide higher processing power on the chip for the same power consumption in comparison to FPGA because they do not have to support the flexible field programmability of FPGAs. On the other hand FPGAs provide the flexibility of optimizing the hardware for the specific application requirements, while GPUs are not optimized for data access and processing applications. An important value that these hardware accelerators provide in the context of smart grid application is the possibility of accelerating pieces of these applications

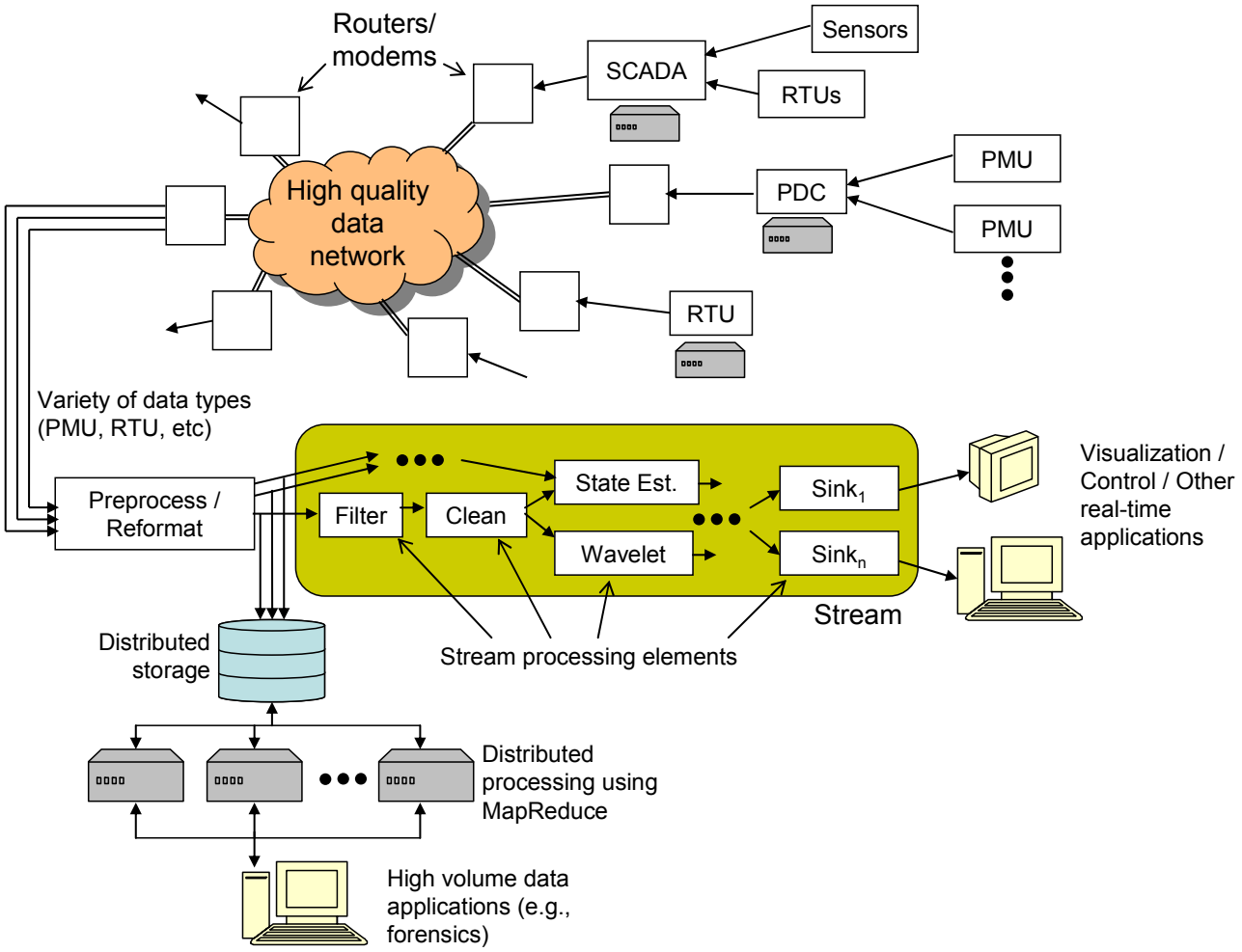


Figure 7.6: An example data processing infrastructure for the smart grid that uses stream computing for real-time applications and MapReduce for high data volume offline applications.

using FPGAs and/or GPUs. For instance, computational algorithms like wavelet transforms, nonlinear regression and simulation techniques, that form the core of many real-time situational awareness applications, can potentially be accelerated by exploiting the highly parallel computational fabric made available by GPUs and FPGAs. Similar ideas have been demonstrated in other areas of electrical and computer engineering, such as circuit simulation on GPUs [Feng and Li, 2008] and speech recognition on FPGAs [Lin et al., 2006]. Indeed, gains have been demonstrated in very recent and early research [Jalili Marandi and Dinavahi, 2010] when using GPUs to accelerate transient stability simulation of large grids. Hence, these accelerators could be exploited both in the massive data at rest systems and within a stream computing framework to accelerate processing elements in the stream.

Figure 7.6 shows an example data management infrastructure (one out of many possible) to illustrate how the discussed technologies may be deployed in a smart grid. The architecture consists of the following main parts:

- Data acquisition
- Data communication
- Data storage
- Data processing

The data acquisition sub-system consists of the sensors and monitors: the example in Figure 7.6 shows data being acquired from PMUs, sensors and remote terminal units (RTUs) among others. The data communication sub-system may consist of data concentrators (e.g., PDCs or SCADA systems) and a highly reliable and extremely fast network that accumulates data from the data acquisition sub-system and transports it to receivers such as control centers and distributed mass storage systems. The key function of such data concentrators is to accumulate sensor/monitor data coming in from a region of the grid. In the case of PDCs, they may be organized

in a flat or hierarchical level and may have local storage available to store the received data temporarily for a few days or weeks. Many parts of the acquisition and communication sub-systems may have local computation and storage to allow for localized and distributed real-time computation.

The data network as a whole has to satisfy some critical and special requirements. It should be able to chronologically synchronize data (e.g., PMU data) from different geographies, with different sampling rates, in different formats and with different incoming network latencies, based on the time stamps on the data packets. For PMU data, for instance, much of this synchronization may be done at the PDC level. Such networks will probably use some combination of internet protocols (IP) for communication, such as user datagram protocol (UDP) for data transfer and transmission control protocol (TCP) for control communications. Recent research has shown the feasibility of developing such networks using available technology, an example being the GridStat architecture described in [Bakken et al., 2007].

The data generated by the grid and by other sources relevant to the grid (such as weather data and grid asset records) will be stored in a storage system that will be distributed in geography, technology and ownership. Smart grid applications would require concurrent access to this data in many different modes, such as very frequent, relatively low volume for control applications, and infrequent and very high volume for forensic analysis. This wide variety in data generation, storage and use and the need for high reliability, concurrency and performance from the storage systems pose interesting engineering challenges. Fortunately, these challenges have been addressed by the information technology industry and distributed storage systems with high availability, reliability and performance have been developed. An example are internet scale file systems, like the Hadoop distributed file system [9], which are designed to work with distributed, parallel computation frameworks like MapReduce. Given, the relevance of MapReduce to high data volume applications for the smart grid, these are doubly attractive. One drawback of typical internet scale file systems is that they do not support non MapReduce style of applications very well. To bridge this gap, researchers have recently proposed adaptations of more traditional storage cluster based, distributed file systems such as the Parallel Virtual File System [1] to support the highly distributed and parallel processing framework of MapReduce [Tantisiriroj et al., 2008].

The final piece is the data processing sub-system. At the head of this sub-system, the data received from the network may need to be preprocessed before further downstream storage or consumption, for example, to remove the UDP headers and/or to re-organize in a different format. Figure 7.6 shows an example of how stream computing and MapReduce might fit into this sub-system, and the overall data system. In the stream example, there are multiple stream processing elements, each performing some atomic task on the data flowing into it. For instance, the *Filter* element filters through only those records and fields in the data that are relevant to the downstream application. The *Clean* element gracefully handles missing data, resulting, for instance, from errors in the data network. The *Wavelet* element performs wavelet transforms on a sliding window of the data, for real-time anomaly detection. The State Estimation element takes the same cleaned data to perform some state estimation task. Note that there may be multiple streams of data coming into the stream. For example, accurate state estimation will require not just synchrophasor data, but also other current grid status data such as relay status, data from RTUs and perhaps digital fault recorder data [Monticelli, 2000]. These other streams of data go through their own processing elements for pre-processing before being consumed by the State Estimation element. All streaming data is thus processed through the flow graph of stream processing elements and the results are finally presented by the Sink elements at the end of the stream, for use in tasks like visualization and control.

The MapReduce example shows that multiple distributed and heterogeneous processing units (computers and/or processors) are networked together to have access to the distributed historical data. Forensic and other high volume data applications running on a workstation are written in the MapReduce software framework and running. Each application runs as several parallel jobs on these processing units. The jobs are allocated to data chunks and processing units in a manner that reduces the amount of data movement across the storage network. To enable high performance, concurrent access to this high volume data by multiple applications, hardware accelerated techniques can prove very relevant.

7.2.3 Dynamics

Nonlinear differential equations are used throughout science and engineering. Singular perturbation theory examines limiting behavior of multiple time scale systems in which there is an infinite separation of fast and slow time scales. In the last few decades, power system dynamics has studied the issue of two-time scales as part of singular perturbation and fast and slow manifolds associated with transient stability equations [Gao and Strunz, 2009], [Zhang et al., 2002]. Significant Lyapunov theory has been developed around the concept of singular perturbation for the two-time scale problem. Power system dynamic behavior, though, includes more than two time scales, and emerging behavior may include several other relevant scales. [Gao and Strunz, 2009] proposes a method that enables modeling synchronous systems over diverse time scales in the range covering

[1] http://www.pvfs.org/

electromagnetic and electro-mechanical transients. The models use frequency adaptive simulation of transients where analytic signals are found to lend themselves to the shifting of the Fourier spectra. The shift frequency appears as a simulation parameter in addition to the time step size. When setting the shift to the common carrier frequency, the method emulates phasor-based simulation that is very suitable for extracting envelope information at relatively large time step sizes. [Chakraborty and Arcak, 2007] proposes a three time scale robust redesign technique, which recovers the trajectories of a nominal control design in the presence of input uncertainties by using two sets of high gain filters. The trajectories of the resulting three time scale redesigned system approach those of the nominal system when the filter gains are increased. In [Leeb and Kirtley, 1993] a method is proposed to detect multi-scale transients and events using non-intrusive load monitoring. Although interest is growing in adaptive frequency methods, most of the work reported deals with two-scale dynamics and is limited to transient stability.

Thus there is a need to study multi-scale and adaptive multi-scale dynamics methods that capture the emerging behavior in the entire temporal dimension, from the electro-mechanical to long-term dynamics ranges. These methods shall consider new power electronics and DER technologies. A highly desirable tool is dynamic security-constrained optimal power scheduling.

Wavelet transforms can be applied in the setting of spatio-temporal dynamical systems, where the evolution of a set of dynamical variables is of interest. In this setting, space is usually discrete and time is continuous, which allows for a multi-resolution decomposition [Kikuchi and Wang, 2010]. The interdependencies between scales are introduced through the action of wavelet decompositions. The multi-scale system dynamics can be modeled by explicit representation of the subsystem dynamics within local components. The system dynamics does not require that the subsystem has reached its asymptotic state. To construct a model for each scale, the individual components can be coupled together to form an ensemble of nonlinear oscillators, a method used in modeling neural systems [Selle and West, 2009].

7.2.4 Control

Multi-scale system control refers to a formal framework that uses multi-scale modeling and theory to realize desirable control and system response at all scales. Hierarchical control is gaining impetus among the smart grid and automation community [Cheng, 2009], but a comprehensive framework at all the relevant spatial levels is yet to be established. Furthermore, the system control must address control at all the relevant temporal scales. This is necessary for a variety of emergent problems such as distributed frequency regulation, demand response, ancillary services provision, disturbance detection, wholesale/retail markets, etc. The properties of self-similarity and scale-free systems may have implications for power system multi-scale control methods. Normalization, traditionally used in power system analysis, suggests scale-free characteristics. This would also pose the question of whether multi-scale hierarchical control, networked control, or a fusion of the two is possible in multi-scale electricity networks.

7.2.5 Probabilistic Assessment

Probabilistic power system performance assessment involves probabilistic models, such as composite reliability assessment of generation-transmission systems, require large computational effort to analyze a realistic power system especially non-linear models. Here, use of embarrassingly parallel algorithms such as Monte Carlo simulations and enumeration techniques enable use of Petascale and Exascale systems for fast throughput in analysis.

7.2.6 Large Scale Data Handling & Visualization

Electricity grid visualization has become increasingly important and a topic of significant research in the last decade [Overbye et al., 2003]. Visualization has been the mechanism to drastically increase situational awareness in bulk energy control centers. Systems with 2D-spatial visualization of the controlled region are the norm. Typically, the visualization requirements for these systems are at the level of 10^3 to 10^4 measurements every 5 to 10 seconds. Zooming and panning is usually too slow for real-time response in this environment. In general, interactive visualization in real-time has not been possible due to slow performance. Spatio-temporal and spatial security 3D visualization has been proposed, but it has not been deployed for either operations or planning due to performance limitations even with Graphics Processing Units (GPUs) [Guo, 2009] [Ingram et al., 2009]. For large scale smart grid deployments, there is a strong need for development of integrated multi-scale, multi-dimensional dynamic visualization tools with applications to multi-scale power system dynamics, PMU and smart meter massive data visualization, Nk contingency information visualization, and 4D dynamic navigation.

Because of the performance and design of the GPU, highly parallelized and efficient computation is possible. Several General Purpose GPU (GPGPU) libraries have been developed to provide a better match to

the computing domain. In recent years, numerous algorithms and data structures have been ported to run on GPUs. Scientific visualization generally has a spatial 2D or 3D mapping and can thus be expressed in terms of the graphics primitives of shader languages. In general, the current GPU programming model is well-suited to managing large amounts of spatial data. GPU visualization for large grids has been reported to be about 30 times faster than CPU-based visualization [Han, 2005] [Hurley, 2005] [Grishin, 2003]. GPU clusters can theoretically yield visualization speeds in the order of 103 times the current levels, opening a large number of research and application possibilities. Multi-scale, multi-dimensional visualization represents an enabling technology for the understanding and study of emerging power systems. While in the past, visualization has been seen as a presentation tool, there is growing acknowledgement that visualization tools are becoming an enabler for advanced analytics, especially in multi-scale, multi-dimensional systems.

7.3 Smart Grid in India

Every global driver for Smart Grids applies to India, but India also has additional drivers in the short term. The power system in India has roughly doubled in the last decade and similarly in the previous decade. With 215 GW of installed capacity with utilities, the Indian power system is now the fourth largest in the world, but per-capita consumption of electricity in India is only about one-fourth of the world average. This underscores the need to grow the power system at a rapid pace for the next several decades. This low consumption level is amplified by the lack of access to electricity to a significant proportion of the population. The potential demand by 2032 is estimated to be as high as 900 GW. India is also pursuing an aggressive renewable generation program. The 12^{th} Five Year Plan target for renewable energy (RE) generation is 36 GW which will increase the current 12% share of RE (excluding hydro) to 20% by end of this decade. A power system of this size growing at such pace (8-10% per year) with an increased share of renewable energy requires smarter systems to manage it efficiently and ensure its stability and reliability.

India has also recently launched a National Mission on Electric Mobility with a target of 6 million electric vehicles (4 million two-wheelers and 2 million four-wheelers) by 2020. For an efficient roll-out of the EV program, electrical distribution infrastructure upgrades and smarter systems are required which will control/limit simultaneous charging of hundreds of EVs from the same feeder. Beyond just timing the consumption of power, immediate policy level support is required to build enabling infrastructure to integrate the EVs in the electrical network so that these millions of EVs connected to the power system can be leveraged as virtual power plants (VPPs) that can store energy when there is surplus generation and support the grid during moments of deficit. Vehicle to Grid (V2G) technologies are evolving rapidly that can achieve these objectives.

The transmission and distribution losses are still very high in the Indian power system and distribution network (aggregate technical & commercial, or AT&C) loss reduction continues to be the top priority of both governments and utilities. Smart grid solutions will help monitor, measure and even control power flows in real time that can contribute to identification of losses and thereby appropriate technical and managerial actions can be taken to arrest the losses.

In view of the growing importance and relevance for smart grids in India, MoP has taken early steps for the development and adoption of smart grid technologies. In 2010 MoP constituted the India Smart Grid Task Force (*ISGTF*), an inter-ministerial body under the Chairmanship of Shri Sam Pitroda, Advisor to Prime Minister of India; and the India Smart Grid Forum (*ISGF*), a PPP initiative of Ministry of Power, Government of India. Both the ISGTF and the ISGF have been functional for over a year and have laid the foundations for building smart grids in India. Some important measures taken by MoP so far are:

- Formulation of 14 smart grid pilot projects to be undertaken by distribution utilities in various states. 50% of the cost of these projects will be given as grant by GoI and the rest to be borne by respective utilities, states, or other stake-holders. These initial set of projects are expected to help with technology selection and establish the business case and regulatory environment for implementing larger smart grid projects in the future.

- Indigenous development of low cost smart meters for mass roll-out for low volume consumers. The specifications are being formulated and the strategy for implementation is nearing finalization.

- Development of the India Smart Grid Knowledge Portal, which is expected to serve as an effective collaboration and knowledge dissemination platform for all stake-holders involved in smart grid developments including consumers.

- Smart Grid Vision for India: The ISGF in consultation with ISGTF Secretariat has formulated a vision/mission document and recommended that it be made a National Smart Grid Mission.

In spite of all these efforts, there are multiple challenges that need to be addressed, before India can realize the full-vision of Smart Grid. The major challenge for implementing Smart Grid in India is availability of

funds. Huge investments are required in order to setup a link between the customers and the Smart Grid. The cost of setting up more plants can be deferred drastically. At that point of time, more emphasis will be on overall development of T&D (Transmission & Distribution) efficiency based on demand response, load control and many other Smart Grid technologies. With timely and detailed information provided by Smart Grids, customers would be encouraged to avoid over use, adopt energy-efficient building standards and invest continually in energy efficient appliances. To tackle the Smart Grid future, we need to have compelling Smart Grid consumer products, collaborative vendor partnerships and a willing investment community. The policy makers and regulators have to implement a robust incentive model frame work to attract more and more private investments keeping the rate of return, based on the output generated. Policy makers and regulators can mitigate this by seeking economies of scale and implementing advanced digital technologies.

With the transition to digital electricity infrastructure comes the challenge of communication security and data management; as digital networks are more prone to malicious attacks from software hackers, security becomes the key issue to be addressed. Smart Grid success depends on the successful handling of two major IT issues, i.e., security & integration and data handling. With an increase in computers and communication networks the threat of cyber attack has also increased invariably. Utilities can use and implement cyber security standards to reduce the vulnerability to the consumers and provide a higher reliability that their valuable information is being protected. Implementing cyber security measure through the use of standards will help reduce software and implementation cost.

As it is observed, there has been certain degree of backlash and apprehension to Smart Grid implementation in developed countries, particularly in the USA. IEEE-SA (Standards Association) is closely working with groups in India, such as the engineering community including vendors, utilities, academics to participate in the standards development and work towards implementing smart grid successfully in India. Also having the technical participants from India provide requirements ensures standards development groups understand and identify any possible gaps and address some of India's technical issues. Also because of the challenges that India has, a more robust grid will be easily accepted.

7.4 Oil & Gas: Demand for HPC

There has been an ever increasing gap between energy demand and supply of oil and gas worldwide due to challenges in improving recovery rates from existing wells. Further, the easily available conventional hydrocarbon resources are getting depleted leading to efforts for high risk deep off-shore drilling. This has led to exploration of unconventional hydrocarbon resources including Shale Oil & Gas and others. These factors worldwide have increased the dependence of the petroleum industry on research in High Performance Computing, Massive Scale Analytics and Numerical Analysis & Optimization.

Ensuring a ready and predictable supply of reasonably priced oil and gas is central to economic stability and growth, as well as national security. For the foreseeable future, oil will continue to be the lifeblood of the economy not only powering our transportation, industrial processes, and buildings, but also embedded in products as diverse as carpets, cosmetics, medicines, detergents, synthetic fibers, and all things plastic. Demand for oil in the United States averages three gallons each day for every person in the country, for a total national demand approaching 20 million barrels of oil per day. At the same time, growing world demand for oil and gas exerts pressure on supplies, which ultimately are finite. Oil and gas supplies depend not only on the volume of known reserves, but more directly on the known ability of suppliers to recover and deliver reserves in ways that are economically and competitively sustainable and winnable. Surmounting the current limits of scientific understanding, computational capability, and recovery technology presents an opportunity to dramatically expand inventory for national use.

7.4.1 Pushing Recovery Limits Using Technology Innovations

Current technology, oil prices, and extraction costs determine the point after which it becomes uneconomic to extract any more oil from a field. At the more inflexible point of irreducible saturation, it becomes physically impossible to extract more oil from a given field, regardless of the economics, due to the dispersion of oil and natural conditions. Technological advances could shift both limits. In fact, technology already has increased recovery considerably from an average of 30% several decades ago to an average of 50% today. These gains are due substantially to more precise computer modeling of underlying geological structures.

With the advent of cluster computing and the migration of high performance computing applications from centralized research centers into the oil fields, supercomputing has became a workbench tool for reservoir engineers and earth scientists. Increased computing capability has enabled the industry to develop computer models and software that help predict where oil is located and how to get better performance from oil drilling. These computer models incorporate 3-D seismic imaging data to create more accurate simulations of oil reservoirs. Seismic images show an approximation of geological features below the Earth's surface. Such visual images are

invaluable because they show where oil and gas may lie and provide a representation of surrounding geological features that must be dealt with in order to extract the reserves.

More recently, 4-D time-lapse seismic imaging technology (similar in concept to time lapse photography) has introduced an additional dimension, showing how flow patterns of oil are changing in the underground formations over time. This 4-D capability has the potential to push recovery rates higher. As an expert in supercomputing expressed it, *"...the industry has advanced from assumption to knowledge."*

7.4.2 Challenges in Improving Recovery

As oil and gas companies recover more and more oil from existing reservoirs, each additional increment becomes more difficult to extract. Remaining oil and gas increasingly is located in geologically complex structures that are more challenging to exploit. Geometrically more effort is often required to recover smaller and smaller amounts. While more sophisticated imaging technology is generating enormous volumes of data to characterize the underground or undersea geology, this information is still not comprehensive enough to permit scientists to map entire oil fields accurately from sub-micron particles to geological structures that are hundreds of meters. Mapping on this range of scale is necessary in order to sufficiently understand the geological details and make significant advances in future oil and gas recovery.

Scientists attempt to compensate for the missing data by using geological interpretations and approximations in the computational models they have developed to simulate oil reservoirs. By combining approximations with data and using the highest performance computers available to run the models, scientists hope to simulate the oil fields more accurately so that recovery can be increased. Unfortunately, despite the advances in computational capability, the approximations available now do not provide the level of certainty that is needed to ensure increased recovery. As one industry executive explained, the approximations used in today's reservoir models are not robust enough to drive the production simulations that guide where and how to drill with a high degree of confidence. The uncertainty is still large. Inadequate approximations may lead to a faulty understanding of the underlying geological structure, potentially leading to errors in drilling that can be expensive in terms of investment costs as well as environmental impact. These uncertainties regarding the geological make-up of the reservoirs and the resultant trial-and-error in the extraction strategy, though not as pervasive as before, still hinder optimal recovery. As a result, retrieval remains difficult and expensive.

For deep off-shore exploration and drilling seismic imaging and inversion play an extremely important role to mitigate operational risk. Rigorous research efforts include large scale seismic imaging projects such as the Kaleidoscope project involving collaboration between BSC, Repsol, 3DGeo and Stanford University. This research led to new computational model development and parallel optimizations for seismic imaging on massive scale supercomputers such as MareNostrum in Barcelona. This project won the Commercial Technology of the Year award at the Platts Global Energy Award in New York in 2009 and was also honored as the 2008 technology winner by IEEE spectrum. With KAUST, there is an active collaboration for high performance Seismic Imaging on Petaflop scale supercomputers such as IBM's Blue Gene/P and Blue Gene/Q.

With huge amount of data from sensors during drilling and other operations, there is an increasing need to develop insights and mitigate operational risk. Thus there is increase in use of data-intensive distributed computing for oil and gas where the combination of deep analytics on 10s to 100s of petabytes of production, seismic, electro-magnetic data, etc., in combination with reactive analytics (with response time in ms to μs) delivers a huge value in drilling, operational optimization, new explorations and other areas. This involves the design of massive scale machine learning algorithms and development of efficient reservoir models.

7.5 Seismic Imaging and Inversion

Seismic imaging and inversion help in modelling the subsurface and visualizing the source rocks that contain oil as well as regions that trap gas. Reverse-time migration is a state-of- the-art technique that reconstructs the source wavefield forward in time and the receiver wavefield backward in time. It then applies an imaging condition to extract reflectivity information out of the reconstructed wavefields. As a result, RTM can generate much improved subsurface images in areas where strong vertical velocity gradients generate turning waves or where rugose interfaces with strong velocity contrasts generate prism waves. In addition, because of its ability to image reflection events that cannot be imaged by other techniques, RTM can be used for refining a velocity model. The advantages of reverse-time migration over other depth migration techniques are that the extrapolation in time does not involve evanescent energy, and no dip limitations exist for the imaged structures. The RTM algorithm is becoming more and more attractive to the industry because of its robustness in imaging complex geology, e.g., sub-salt.

Processing seismic data obtained over 4-6 months of seismic survey, can be very compute and data intensive. For instance, isotropic imaging of a typical 10 TB data set using a $512 \times 512 \times 512$ model requires 112 days for one imaging run using 13.6 TFlops of compute power with 10 TB of data input and 32 GB of data output. As

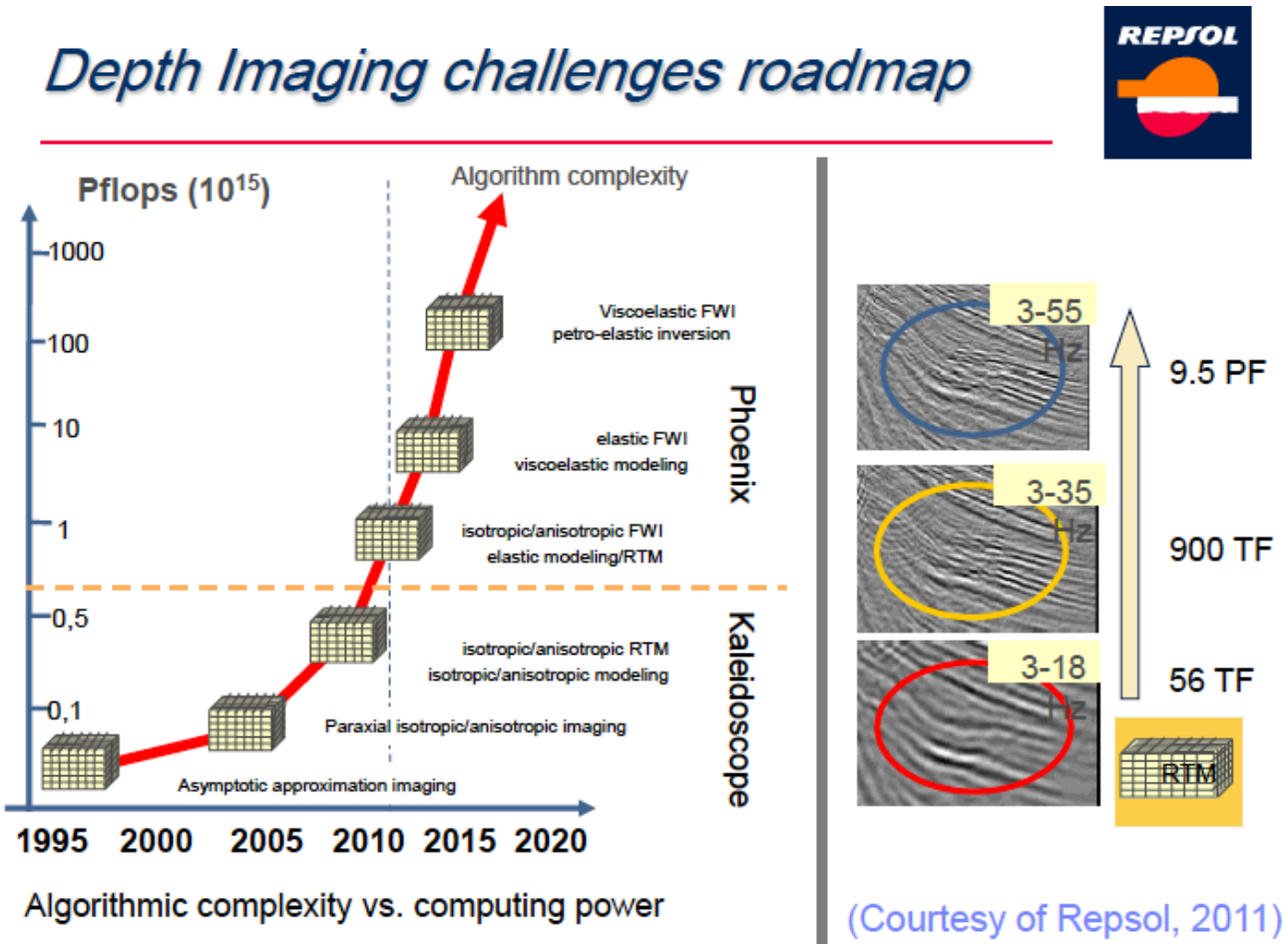


Figure 7.7: Depth Imaging Challenges With Increasing Compute Power In Future

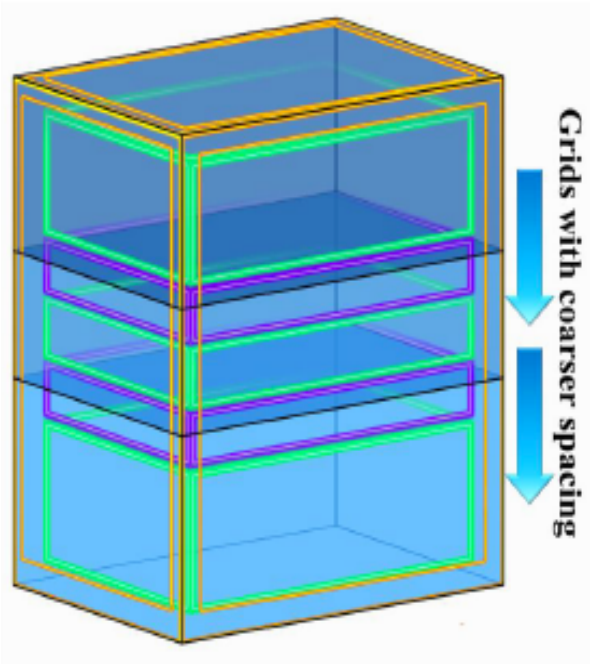
the performance of the underlying parallel architecture improves, one can see improvements (as given below) in the (a) resolution of the model, (b) the size of data that can be handled, (c) the quality (and hence accuracy of model) of the algorithm that can be used as well as (d) the time to completion:

- 80 TFlops: Model Size - 1024^3. Data Size - 30 TB, Algorithm - TTI, Time to Solution - 5 days
- 1 ExaFlops: Model Size - 1024^3. Data Size - 90 TB. Algorithm - FWI. Time to Solution - 12 days
- 10 ExaFlops: Model Size - 2048^3. Data Size - 270 TB. Algorithm - ViscoElastic. Time to Solution - 14 days

Figure 7.7 illustrates the capability improvements in seismic imaging and inversion with increasing compute power. TTI (Tilted Transverse Isotropic) RTM wave equations, such as those developed by [Fletcher et al., 2009], are well-known as an improvement over isotropic representations of wave propagation. When discretized they enable the correct imaging of complex geology when running Reverse Time Migration (RTM) [Baysal et al., 1983] [Jones et al., 2007] and are an important basis for seismic modeling. Although using the finite difference method to code discretized approximations for these coupled second-order PDEs is relatively straightforward, TTI wave propagation is highly compute intensive and optimizing for peak performance is, in general, a formidable problem.

Further, the necessary computation far exceeds that of conventional one-way wave-equation migration (WEM) and requires a large amount of core memory. Because of these requirements, RTM is considered too expensive for routine production projects with large volumes. GPGPU (General Purpose Graphics Processing Unit) has emerged as a key many-core architecture for High Performance Computing and is now being heavily used for scientific computing and visualization. Many recent research efforts have addressed the computational challenge for RTM by leveraging GPGPUs, including [Cabezas et al., 2009] There also has been research using multi-core clusters [Perrone et al., 2011]. Much of the work either considers only isotropic RTM, or deals solely with optimization and parallelization of the stencil operations when coding for GPU [Mickevicius, 2009].

Following Fletcher et al. [Fletcher et al., 2009], one can model acoustic propagation in TTI media using the coupled PDE system. The complete finite difference modelling algorithm consists of solving the above coupled



P-wave and Q-wave coupled equations for Seismic Imaging

$$\frac{\partial^2 p}{\partial^2 t} = v_{px}^2 H_2 p + \alpha v_{pz}^2 H_1 q + v_{sz}^2 H_1 (p - \alpha q)$$

$$\frac{\partial^2 q}{\partial^2 t} = \frac{v_{pn}^2}{\alpha} H_2 p + \alpha v_{pz}^2 H_1 q - v_{sz}^2 H_2 (\frac{1}{\alpha} p - q)$$

where, H_1 and H_2 are functions of the second-order partial derivatives and dip/azimuth angles. given by the equations below.

$$H_1 = sin^2\theta cos^2\phi \frac{\partial^2}{\partial x^2} + sin^2\theta \sin^2\phi \frac{\partial}{\partial y^2} + cos^2\theta \frac{\partial^2}{\partial z^2} + sin^2\theta \sin 2\phi \frac{\partial^2}{\partial x \partial y} + sin2\theta sin\phi \frac{\partial^2}{\partial y \partial z} + sin2\theta cos\phi \frac{\partial^2}{\partial x \partial z}$$

$$H_2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - H_1$$

Figure 7.8: Seismic Imaging Grid With Increasing Coarseness & Fletcher's Equations for TTI / RTM Wave Propagation

PDE system for both the source and receiver wavefields for a large number of iterations. The coupled equations for P-wave and Q-wave propagation (TTI/RTM model) are given in Figure 7.8. As a starting point for this work, a highly optimized CPU implementation of TTI RTM was used which achieved a cycles/instruction rate of $0.6 - 1.0$ in the compute intensive kernels [Narang et al., 2013]. In addition, for increased throughput performance, the complete 3D volume is split into grid zones along the Z dimension. As velocity is generally higher at greater depth the grid zones which represent deeper sections of the 3D volume may have larger grid spacing along the Z dimension to reduce computational cost with minimal reduction in accuracy. For the same reason, long spatial finite difference operators are used, their total lengths are between 21 and 25 points.

For each time-step of the source and receiver computation the kernels called are (Figure 7.8 [Narang et al., 2013]):

- **Z Kernel Internal Regions**: In this kernel, the partial 1^{st} and 2^{nd} order spatial derivatives of the P-wave and Q-wave w.r.t. Z dimension are computed. (Purple)
- **Z Kernel Overlap Regions**: To maintain continuity of wave propagation across multiple grid zones, an adapted Z kernel for "Overlap Regions" is used at each of the boundaries between them. (Green)
- **XY Kernel**: In this kernel, the partial 2^{nd} order spatial derivatives of the P-wave and Q-wave are computed and these are used to update the functions *H1p, H1q, H2p* and *H2q*. The update to the new values of P and Q-waves at time $t+1$ is performed using the values of *P(t), P(t - 1), Q(t)* and *Q(t - 1), H1p, H1q, H2p* and *H2q* along with multiple velocity and dip/azimuth angle parameters. (Purple and Green)

After the source computation, the values of P-wave at each point in the 3D volume and at every k^{th} time step is stored in memory (or disk as appropriate). During the receiver computation, the image correlation at every k^{th} time step is done by reading in the value of the P-wave from the source computation and multiplying it by the corresponding values from the receiver computation at the respective spatial points. The final image is given by summing these correlated volumes over time steps. Figure 7.9 illustrates the source computation and receiver computation loops used in the RTM Algorithm.

To maximize parallel execution, the algorithm utilizes both the available CPU and GPU hardware. Source and receiver computation, as the most compute intensive, are run on the GPU (Figure 7.10) [Narang et al., 2013]. While receiver computation is running, correlation to form the image is run on the CPU (Figure 7.10). This concurrency is achieved by transferring the values of the P-wave from GPU to CPU memory after each set of k time steps and performing the correlation with the source wavefield values there. All kernels can be optimized for higher performance on the nVidia Tesla GPU architectures (Fermi 2090, Kepler).

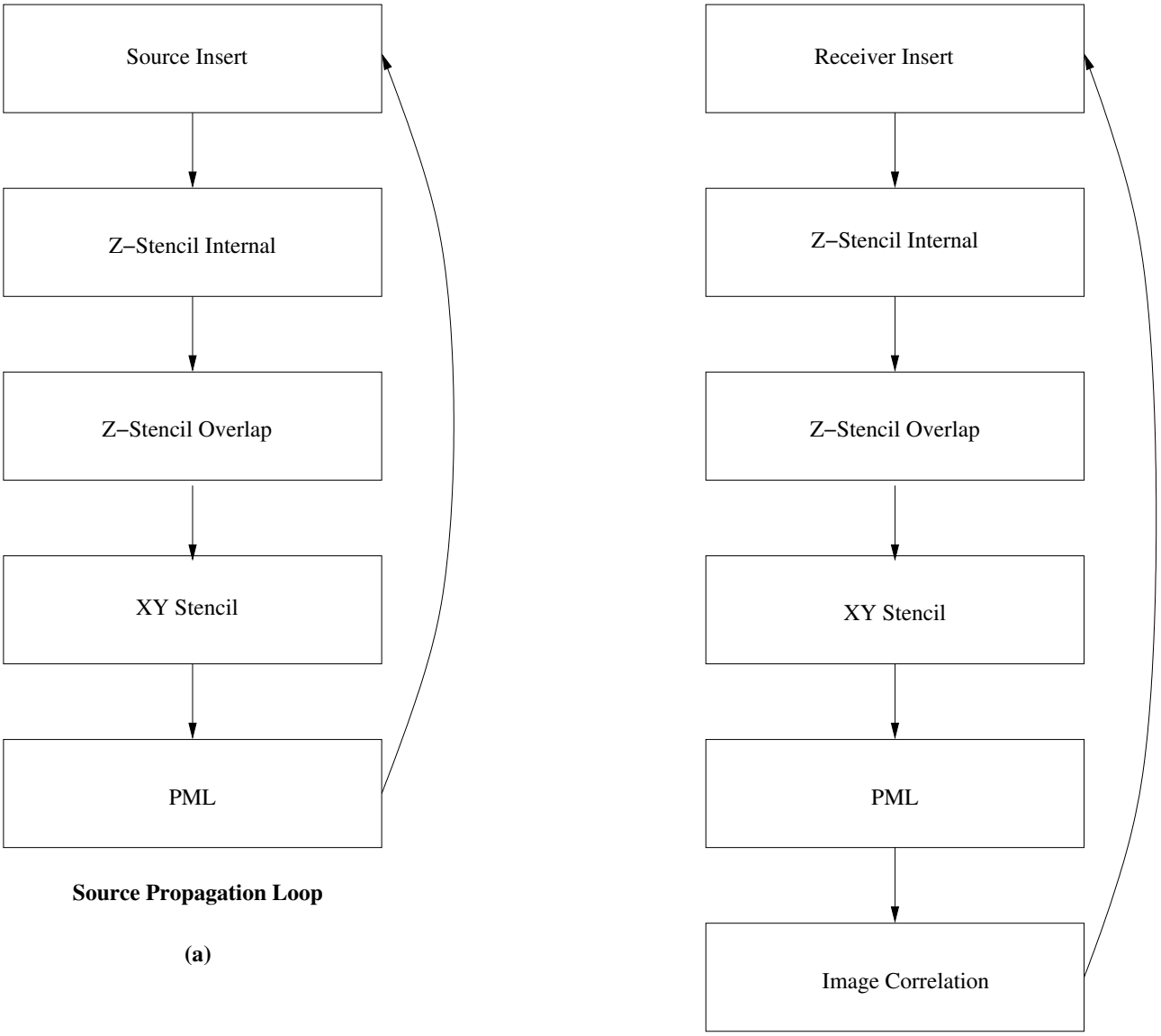


Figure 7.9: Source & Receiver Algorithm

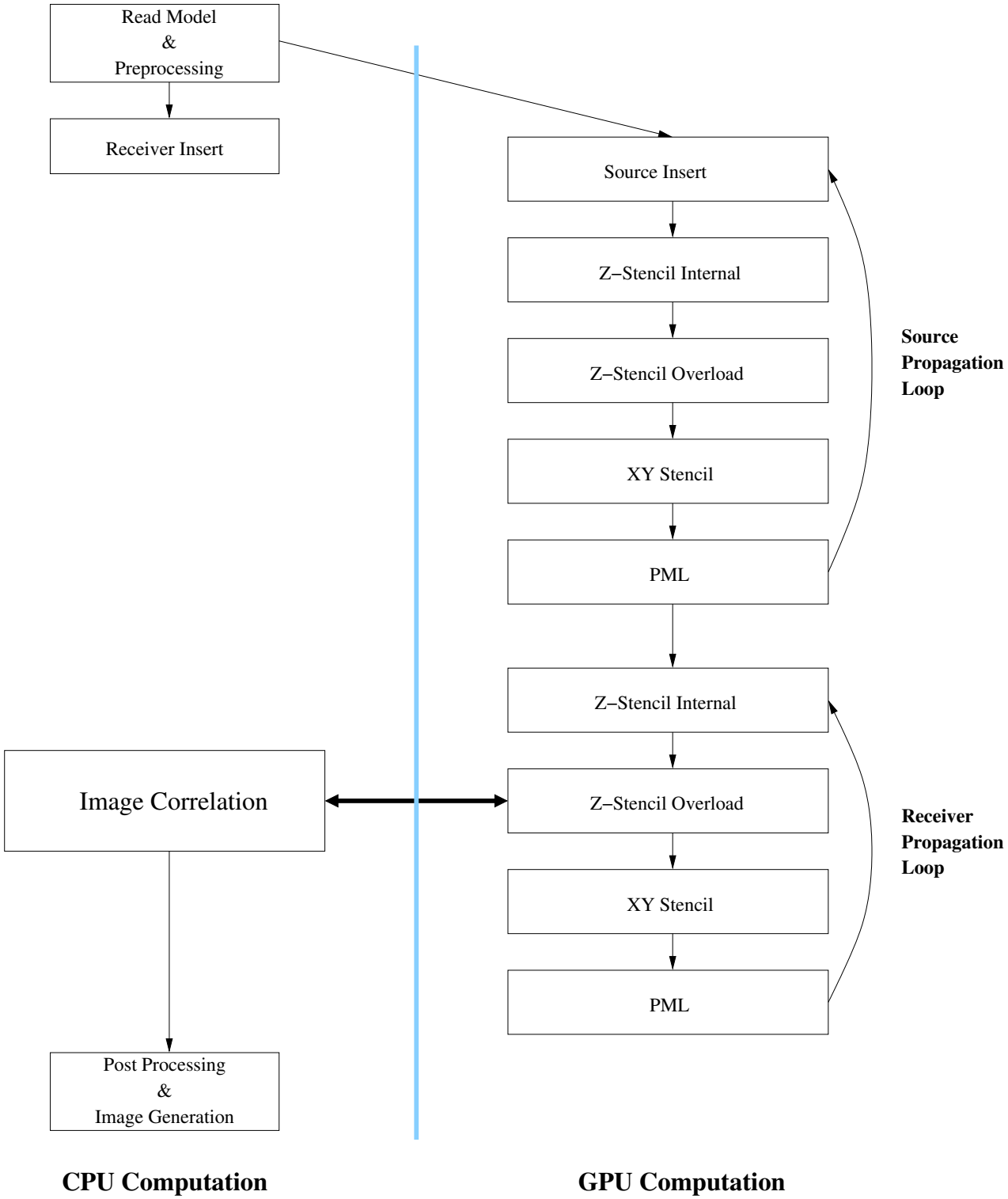


Figure 7.10: CPU-GPU Hybrid Design

Table 2: Performance Gain of Hybrid (CPU+GPU) over CPU for Source Computation Loop

Kernel Name	Src-CPU	Src-Hybrid	SpeedUp
Source Insert	96.6s	198.6s	N/A
Z Internal	582s	120s	4.84×
Z Overlap	724.2s	220s	3.3×
XY	3402.6s	660s	5.15×
PML	1450.2s	198s	7.32×
Total Time	6255s	1397s	4.48×

Table 3: Performance Gain of Hybrid (CPU+GPU) over CPU for Receiver Computation Loop

Kernel Name	Rec-CPU	Rec-Hybrid	SpeedUp
Receiver Insert	222.6s	22.8s	9.7×
Z Internal	577s	136s	4.24×
Z Overlap	766s	327.6s	2.34×
XY	3526.2s	720s	4.9×
PML	1579s	258s	6.12×
Image Correlation	1558s	0s	∞
Total Time	8228s	1465s	5.62×

Figure 7.11: Speed-Up Using GPU based Hybrid Design

The parallel TTI RTM algorithm was compiled with CUDA 4.1 & icc, and run on a 12 Core Intel CPU - Dual socket X5650 (Westmere) - with 2 Tesla M2090 (Fermi) GPUs. [Narang et al., 2013] used a large testcase, in which across all the zones the underlying 3D data volume had dimensions of 288 X 335 X 449 with 20 points extended from each face for boundary condition evaluation by the PML kernel. Figure 7.11 demonstrates the speedup obtained for each kernel. For the hybrid CPU+GPU system based run, the four kernels were optimized and run on the GPUs, while the image correlation was run on the CPU in parallel with the receiver loop iterations. The comparison is presented against a highly optimized CPU code with detailed cache optimizations and SSE SIMD instructions. The GPU+CPU code was shown to be stable and the output image quality from the GPU+CPU code was matched with the CPU code output for this large test case. The GPU+CPU code was also run for various smaller test cases with lesser dimensions and shown to be stable and generated high quality images. Assuming around 800s for the remaining sequential code in the seismic imaging flow, the end-to-end performance gain of CPU+GPU is around 4.2× over optimized CPU only implementation [Narang et al., 2013]. Further, the performance gain for only the computational bottleneck kernels is around 5×.

7.6 Basin Modeling & Simulation

Geological processes are highly complex and fully coupled. Sedimentary basins can be considered as thermo-chemical reactors in which pressure and temperature changes induce numerous mechanical and chemical transformations to produce most of the hydrocarbon, mineral and drinking water resources upon which modern civilization depends. Sedimentary basins are formed during tectonic episodes that create subsidence of the crust and lithosphere which generates the space that allows the accumulation of sediments, and also causes changes in the heat flow into the basin. As the sediments are deposited in the accumulation space, additional space is created due to their load and the associated isostatic compensation mechanisms such as flexure. The newly deposited sediments normally have high porosity with water filling the porous space. Petroleum System Modeling (PSM) consists of the analysis of coupled processes which lead to large petroleum accumulations. Basin models play a central role to integrate coupled processes in PSM, since they are a key tool that can provide insights on the timing of hydrocarbon generation, migration, and accumulation.

Basin models are numerical tools for simulating integrated geological processes during the evolution of

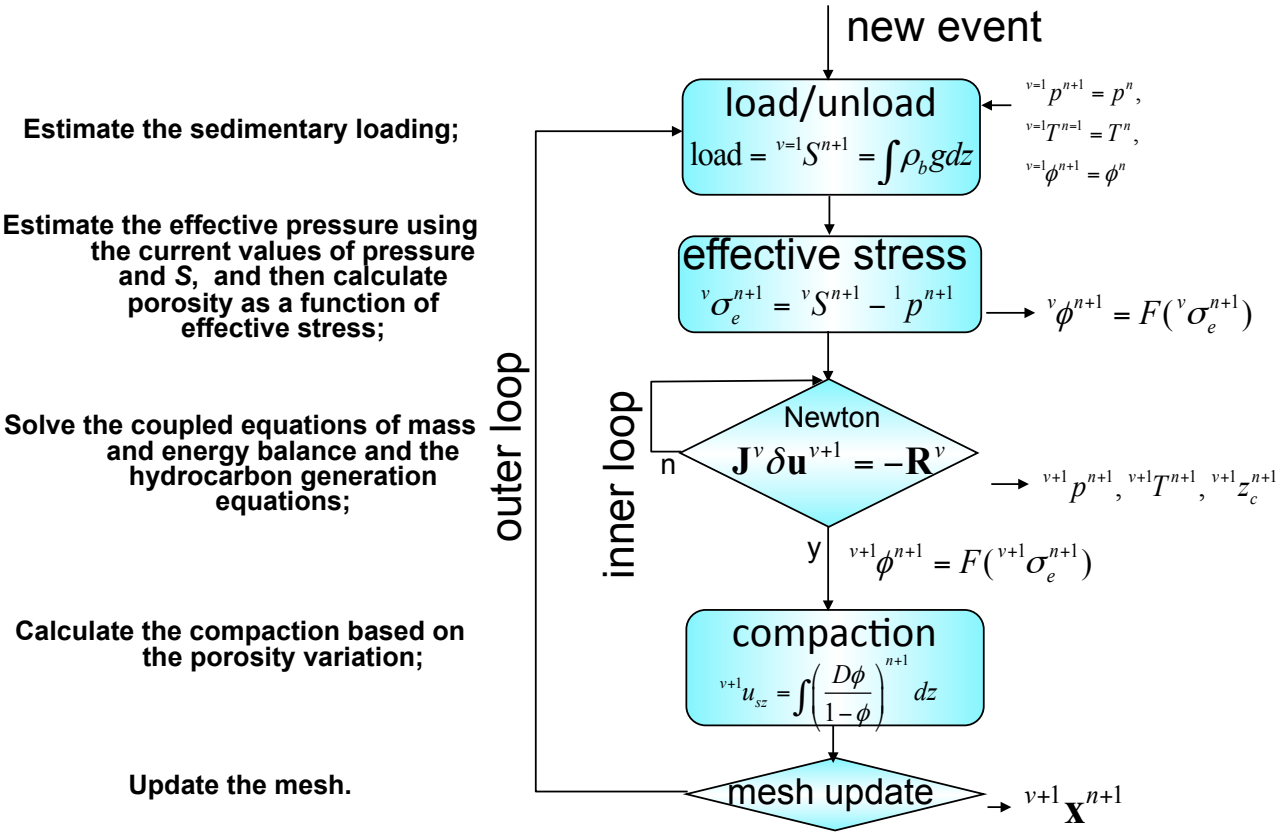


Figure 7.12: Basin Modeling Flow

sedimentary basins, including sediment deposition, compaction, tectonic deformation, heat transport, and hydrocarbon generation, migration and accumulation. The main objective of basin models is to reduce the exploration risk by integrating multidisciplinary methods and modeling the evolution of the petroleum system. These models assist explorationists to assess the hydrocarbon potential in sedimentary basins. This assessment is mainly obtained by using numerical simulations to understand the timing and relationship of the multiple processes which are typically controlled by the temperature and pressure history of the sediments as well as the hydrocarbon generation and multi-phase flow in the deforming porous media. Multi-phase basin modeling can be much more complex than reservoir simulation due to the added complexity associated with compaction and the evolving geometries due to fault and salt motion.

Modeling the heat and fluid transport in sedimentary basins requires a simulation domain, a set of flow and transport parameters, and boundary conditions. The simulation domain construction requires a geological model and a numerical mesh to represent the model. To build the geological model of a given basin, large amounts of data and pre-analyses are necessary. The main input for the model are: surfaces describing the geometry and stratigraphy of the sedimentary layers, geometrical reconstruction of sediment motion (backstripping, lateral reconstruction), thermal and mechanical rock properties (e.g., thermal conductivity, absolute permeability, compressibility), state fluid properties (density, viscosity), rock-fluid properties (relative permeabilities), organic matter properties (source rock, generation potential, kinetic parameters), paleo-bathymetry and paleo-heat flow maps, thermal and pressure paleo-indicators (vitrinite, fluid inclusions).

The layers in the model are subdivided into grid cells within which properties are uniform. Computer programs simulate physical processes that act on each cell, starting with initial conditions and progressing by a selected time increment to the present. Model outputs, such as porosity, temperature, pressure, vitrinite reflectance, accumulation volume or fluid composition, can be compared with independent calibration information, and the model can be adjusted to improve the match. Basin and petroleum system modeling is an iterative process with many interrelated steps, each of which is a scientific discipline in itself. Figure 7.12 below ([Mello et al., 2009]) illustrates the typical steps in basin modelling. These include estimation of sedimentary loading, estimating the effective pressure using the current values of pressure and S and then calculating the porosity as a function of effective stress, solving coupled equations of mass and energy balance and hydrocarbon equations, calculating the compacted based on porosity variation, and finally, updating the mesh Figure 7.13.

For real-data consisting of large grids, the basin and petroleum system modelling iterations can be highly computationally expensive [Mello et al., 2009]. For instance, one complete simulation for a 200 km $\times$ 800 km $\times$ 10 km basin at 1km $\times$ 1km $\times$ 1km resolution would require a full day at 20 TFlops with 10 GB of data input

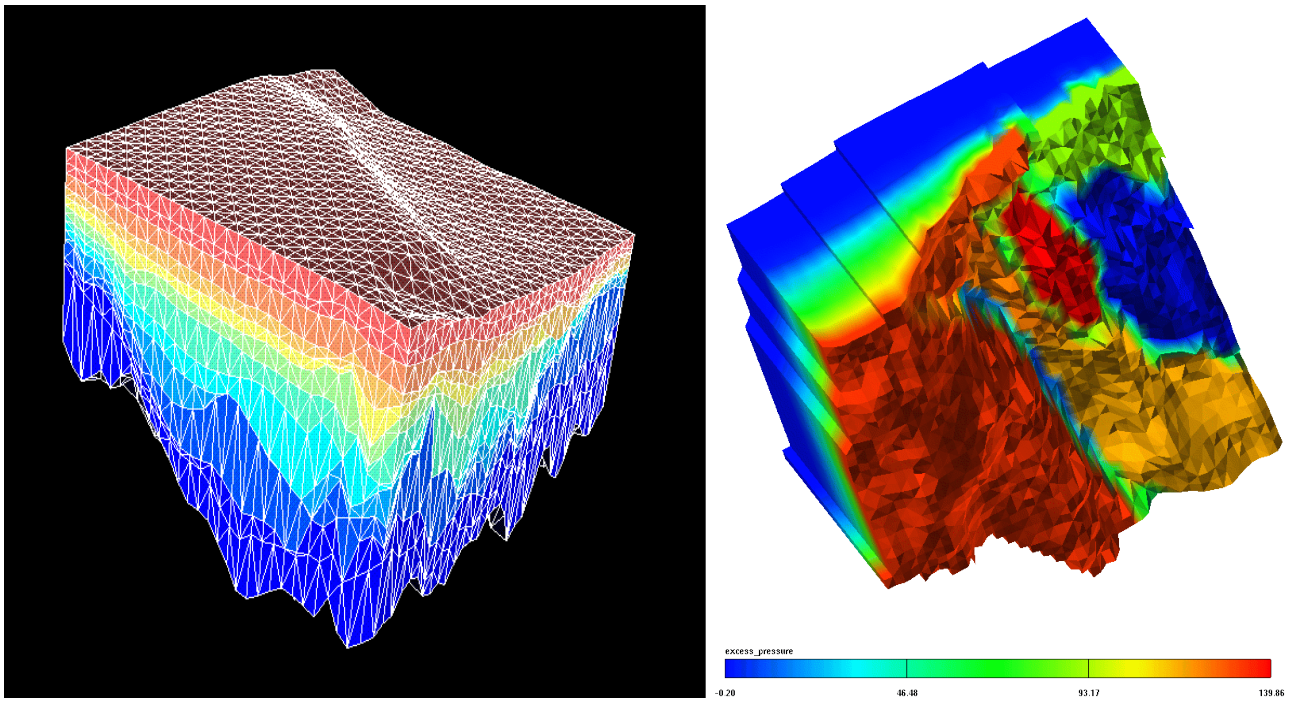


Figure 7.13: Basin Modeling Simulation: Tetrahedral Meshing & Temperature Distribution

and 400 GB of data output. As the computing power is increased, one can observe the following improvement in resolution and processing time:

- 60 TFlops: Resolution of 250 m × 250 m × 250 m, Completion Time: 12 hrs, Data scale per simulation: 64 teraytes
- 2 ExaFlops: Resolution of 100 m × 100 m × 100 m, Completion Time: 6 hrs, Data scale per simulation: 2 petaytes
- 20 ExaFlops: Resolution of 20 m × 20 m × 20 m, Completion Time: 6 hrs, Data scale per simulation: 20 exabytes!

Figure 7.14 illustrates mesh partitioning for 3D parallel basin modelling along with results [Mello et al., 2009] using 1024 nodes on Blue Gene/P. By partitioning the mesh across 1024 nodes of Blue Gene/P, each node (with 4 cores and 4 GB memory) can independently perform iterations for basin simulation on a part of the full basin, while performing fine-grain and coarse-grain synchronization across the nodes in the system to ensure correctness and maintain accuracy.

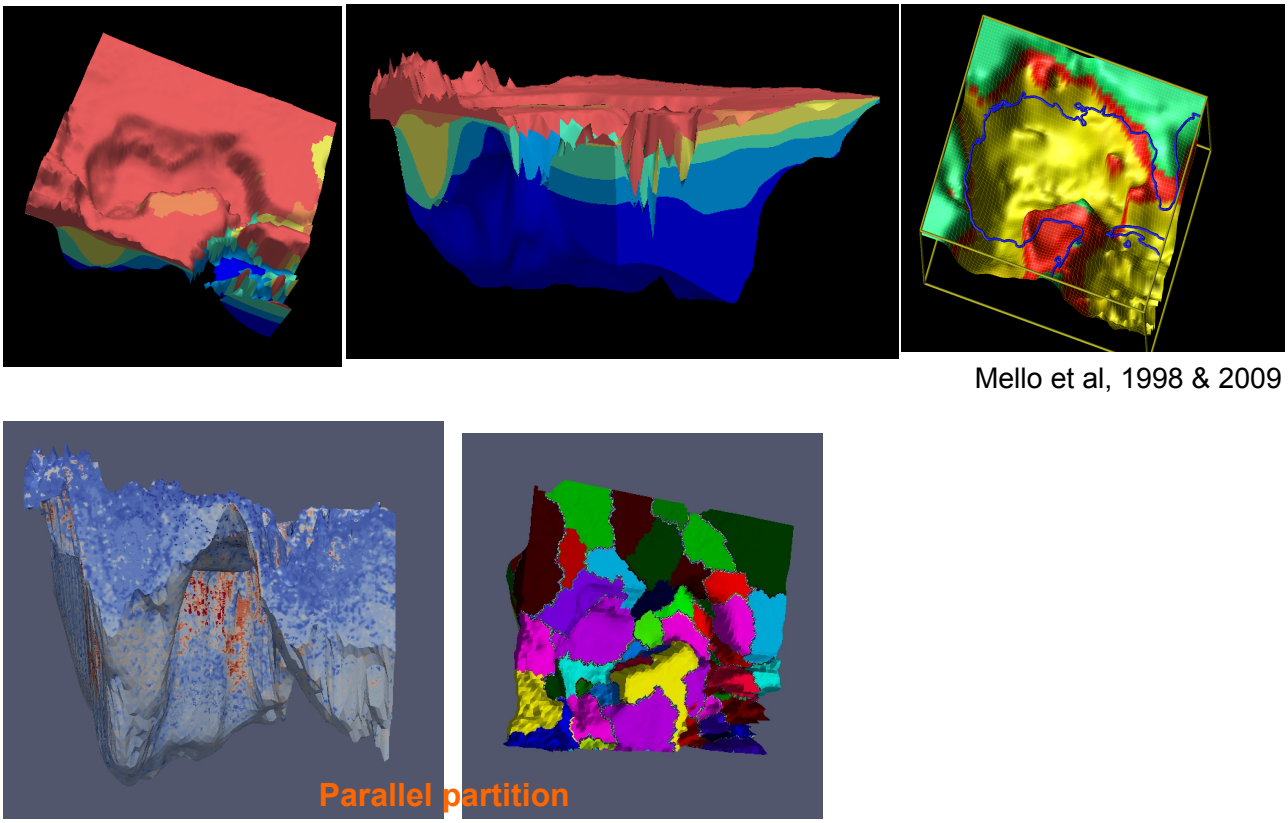


Figure 7.14: Parallel Basin Modeling Simulation: Mesh Partitioning & Simulation Results on 1024 nodes of BG/P

Bibliography

[Alsac and Stott, 1974] Alsac, O. and Stott, B. (1974). Optimal load flow with steady state security. *IEEE Transactions on PAS-93*, 3:745 – 751.

[Attinger and Koumoutsakos, 2004] Attinger, S. and Koumoutsakos, P. (2004). *Multiscale Modeling and Simulation*. Springer.

[Bakken et al., 2007] Bakken, D., Hauser, C., Gjermundrød, H., and Bose, A. (2007). Towards more flexible and robust data delivery for monitoring and control of the electric power grid. Technical Report EECS GS 009, Elec. Engg. and Comp. Sc., Washington State University.

[Bakkum and Skadron, 2010] Bakkum, P. and Skadron, K. (2010). Accelerating SQL database operations on a GPU with CUDA. In 3^{rd} *Workshop on General-Purpose Computation on Graphics Processing Units*.

[Bank et al., 2009] Bank, J., Omitamou, O., and Liu, Y. (2009). Visualization and classificationn of power system frequency data streams. In *IEEE Conference on Data Mining Workshops*, pages 650–655.

[Baysal et al., 1983] Baysal, E., Koslo, D., and Sherwood, J. (1983). Reverse time migration. *Geophysics*, 48:1514 – 1524.

[Cabezas et al., 2009] Cabezas, J., Araya-Polo, M., Gelado, I., Navarro, N., Morancho, E., and Cela, J. (2009). High performance reverse time migration on gpu. In *Intl. Conference of the Chilean Computer Science Society, SCCC*, pages 77 – 86.

[Capitanescu and Wehenkel, 2007] Capitanescu, F. and Wehenkel, L. (2007). Improving the statement of the corrective security-constrained optimal power flow problem. *IEEE Transactions on Power Systems*, 22:887 – 889.

[Chakraborty and Arcak, 2007] Chakraborty, A. and Arcak, M. (2007). Three timescale redesign for robust stabilization and performance recovery of nonlinear systems with input uncertainties. In *46th IEEE Conference on Decision and Control*, pages 3484 – 3489.

[Chamberlin and Boyce, 1974] Chamberlin, D. and Boyce, R. (1974). SEQUEL: A structured English query language. In *ACM SIGFIDET Workshop on Data Description, Access and Control*.

[Cheng, 2009] Cheng, M. e. a. (2009). Hierarchical Utilization Control for Real Time and Resilient Power Grid. In *21st Euromicro Conference on Real Time Systems (ECRTS)*.

[Dean and Ghemawat, 2010] Dean, J. and Ghemawat, S. (2010). MapReduce: A flexible data processing tool. *Communications of the ACM*, 53(1).

[DeMarco, 2010] DeMarco, C. (2010). Situational awareness: singular value methods for PMU data interpretation. In *PSERC Seminar*.

[Eto and R.J., 2011] Eto, J. and R.J., T. (2011). Doe workshop on computational needs for next generation electricity grid.

[Feng and Li, 2008] Feng, Z. and Li, P. (2008). Multigrid on GPU: tackling power grid analysis on parallel SIMT platforms. In *Proc. of IEEE/ACM Intl. Conf. on Computer−Aided Design*.

[FERC, 2008] FERC (2008). Assessment of demand response and advanced metering. Technical report, Federal Energy Regulatory Commission.

[Fletcher et al., 2009] Fletcher, R., Du, X., and Fowler, P. (2009). Reverse time migration in tilted traversely isotropic (tti) media. *GEOPHYSICS*, 74(6):179 – 187.

[Francisco, 2009] Francisco, P. (2009). The Netezza data appliance architecture: a platform for high performance data warehousing and analytics. `http://www.ibmbigdatahub.com/sites/default/files/document/redguide_2011.pdf`.

[Gao and Strunz, 2009] Gao, F. and Strunz, K. (2009). Frequency Adaptive Power System Modeling for Multiscale Simulation of Transients. *IEEE Transactions On Power Systems*, 24:561 – 571.

[Gedik et al., 2008] Gedik, B., Andrade, H., Wu, K., Yu, P., and Doo, M. (2008). SPADE: The System S declarative stream processing engine. In *SIGMOD*, pages 1123 – 1133.

[Glover et al., 2008] Glover, J., Sarma, M., and Overbye, T. (2008). *Power Systems Analysis and Design*. Thomson Learning, Toronto.

[Grijalva et al., 2007] Grijalva, S., Dahman, S., Patten, K., and Visnesky, A. (2007). Large Scale Integration of Wind Generation Including Network Temporal Security Analysis. *IEEE Transactions on Energy Conversion*, 22.

[Grishin, 2003] Grishin, V. (2003). Pictorial Analysis: A Multi resolution Data Visualization Approach for Monitoring and Diagnosis of Complex Systems. *Information Sciences*, 152:1 – 24.

[Guo, 2009] Guo, D. (2009). Mapping and Multivariate Visualization of Large Spatial Interaction Data. *IEEE Transactions on Visualization and Computer Graphics*, 15:1041 – 1048.

[Han, 2005] Han, J. e. a. (2005). Stream cube: Architecture for multi−dimensional analysis of data streams. *Distributed and Parallel Databases*, 18:173 – 197.

[Huan and Orban, 2011] Huan, L. and Orban, D. (2011). A MapReduce Implementation On Top of a Cloud Operating System. In *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pages 464 – 474.

[Hurley, 2005] Hurley, C. (2005). Clustering Visualizations of Multidimensional Data. *Journal of Computational and Graphical Statistics*, 13:788 – 806.

[Ingram et al., 2009] Ingram, S., Munzner, T., and Olano, M. (2009). Glimmer: Multilevel MDS on the GPU. *IEEE Transactions on Visualization and Computer Graphics*, 15.

[Jalili Marandi and Dinavahi, 2010] Jalili Marandi, V. and Dinavahi, V. (2010). SIMD based large scale transient stability simulation on the graphics processing unit. *IEEE Transactions on Power Systems*, 25.

[Jones et al., 2007] Jones, I., Goodwin, M., Berranger, H., Zhou, H., and Farmer, P. (2007). Application of anisotropic 3D reverse time migration to complex north sea imaging. *In 77th Annual International Meeting, SEG, Expanded Abstracts*, pages 2140 – 2143.

[Kikuchi and Wang, 2010] Kikuchi, K. and Wang, B. (2010). Spatiotemporal Wavelet Transform and the Multiscale Behavior of the Madden Julian Oscillation. *Journal of Climate*, 23.

[Leeb and Kirtley, 1993] Leeb, S. and Kirtley, J. (1993). Multiscale Transient Event Detector for Nonintrusive Load Monitoring. *IECON Proceedings (Industrial Electronics Conference)*, 1:354 – 359.

[Li et al., 2010] Li, H., Treinish, L., and Hosking, J. (2010). A statistical model for risk management of electric outage forecasts. *IBM Journal of Res. & Dev.*, 54(3):8:1–8:11.

[Li, 2008] Li, Y. (2008). *Decision making under uncertainty in power system using Benders decomposition*. PhD thesis, Iowa State University, Iowa.

[Li and McCalley, 1994] Li, Y. and McCalley, J. (1994). Asynchronous programming model for the concurrent solution of the security constrained optimal peer flow problem. *IEEE Transactions on Power Systems*, 9:2021 – 2027.

[Li and McCalley, 2009] Li, Y. and McCalley, J. (2009). Decomposed SCOPF for improving efficiency. *IEEE Transactions on Power Systems*, 24:494 – 495.

[Lin et al., 2006] Lin, E., Yu, K., Rutenbar, R., and Chen, T. (2006). Moving speech recognition from software to silicon: the In Silico Vox project. In *Proc. Intl. Conf. on Spoken Language Processing (InterSpeech2006)*.

[Liu et al., 2010] Liu, C., Shahidehpour, M., and Wu, L. (2010). Extended Benders Decomposition for Two Stage SCUC. *IEEE Transactions on Power Systems*, 25:1192 – 1194.

[Mahendra Patel, 2010] Mahendra Patel, e. a. (2010). Real-time application of synchophasors for improving reliability. Technical report, North American Electric Reliability Corporation.

[Mello et al., 2009] Mello, U. T., Rodrigues, J. R. P., and Rossa, A. L. (2009). A control-volume finite-element method for three-dimensional multiphase basin modeling. *Marine and Petroleum Geology*.

[Mickevicius, 2009] Mickevicius, P. (2009). 3D finite difference computation in GPUs using CUDA. In *GPGPU-2, 2^{nd} Workshop on General Purpose Processing on Graphics Processing Units*, pages 79 – 84.

[Min and Shengsong, 2005] Min, W. and Shengsong, L. (2005). A trust region interior?point algorithm for optimal power flow problems. *International Journal of Electrical Power & Energy Systems*, 27:293 – 300.

[Monedero et al., 2007] Monedero, I., Leon, C., Ropero, J., Garcia, A., Elena, J., and Montano, J. (2007). Classification of electrical disturbances in real time using neural networks. *IEEE Transactions on Power Delivery*, 2:1288 – 1295.

[Monticelli, 2000] Monticelli, A. (2000). Electric power system state estimation. *Proc. of the IEEE*, 88.

[Monticelli et al., 1987] Monticelli, A., Pereira, M., and Granville, S. (1987). Security-constrained optimal power flow with post-contingency corrective scheduling. *IEEE Transactions on Power Systems*, 2:175–180.

[Mueller et al., 2009] Mueller, R., Teubner, J., and Alonso, G. (2009). Streams on wires: a query compiler for FPGAs. In *ntl. Conf. on Very Large Data Bases*.

[Muralidharam et al., 2008] Muralidharam, K., Mishra, S., Frantziskonis, G., Deymier, P., Nukala, P., Simunovic, S., and Pannala, S. (2008). Dynamic compound wavelet matrix for multiphysics and multiscale problems. *Physics Review E.*, 77:026714.

[Narang et al., 2013] Narang, A., Wade, D., Kumar, S., Soman, J., Perrone, M., Bendiksen, K., Slĺtten, V., and Rabben, T. (2013). Maximizing TTI RTM Throughput for CPU+GPU. In *75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013*.

[NASPI, 2009] NASPI (2009). SynchroPhasor Technology Roadmap.

[Neumeyer et al., 2010] Neumeyer, L., Robbins, B., Nair, A., and Kesari, A. (2010). S4: Distributed stream computing platform. In *International Workshop on Knowledge Discovery Using Cloud and Distributed Computing Platforms, KDCloud*.

[Niimura, 2006] Niimura, T. (2006). Forecasting techniques for deregulated electricity market prices - extended survey. *Power Systems Conf. and Expo.*

[Overbye et al., 2003] Overbye, T., Klump, R., and Weber, J. (2003). Interactive 3D Visualization of Power System Information. *Electric Power Components and Systems*, 31:1205 – 1215.

[P. and Skadron, 2010] P., B. and Skadron, K. (2010). Accelerating SQL database operations on a GPU with CUDA. In *3rd Workshop on General Purpose Computation on Graphics Processing Units*.

[Padhy, 2004] Padhy, N. (2004). Unit Commitment - A bibliographical survey. *IEEE Transactions on Power Systems*, 19:11196–2005.

[Perrone et al., 2011] Perrone, M., Liu, L., Lu, L., Fedova, I., and Semenikhin, A. (2011). Fast scalable reverse time migration seismic imaging on Blue Gene/P.

[Qiu et al., 2005] Qiu, W., Flueck, A., and Tu, F. (2005). A new parallel algorithm for security-constrained optimal power flow. In *IEEE Power Engineering Society General Meeting*, pages 2422 – 2428.

[Scofield et al., 2010] Scofield, T., Delmerico, H., Chaudhary, V., and Valente, G. (2010). XtremeData dbX: An FPGA based data warehousing appliance. *Computing in Science and Engineering*, pages 66 – 73.

[Selle and West, 2009] Selle, C. and West, M. (2009). Multiscale Networks for Distributed Consensus Algorithms. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4753 – 4758.

[Sheble and Fahd, 1994] Sheble, G. and Fahd, G. (1994). Unit Commitment - Literature Synopsis. *IEEE Transactions on Power Systems*, 9:128–135.

[Tantisiriroj et al., 2008] Tantisiriroj, W., Patil, S., and Gibson, G. (2008). Crossing the chasm: sneaking a parallel file system into Hadoop. In *SC08 Petascale Data Storage Workshop*.

[Torres-Hernandez and Velez-Reyes, 2008] Torres-Hernandez, M. and Velez-Reyes, M. (2008). Hierarchical control of Hybrid Power Systems. In *11th IEEE International Power Electronics Congress (CIEP)*, pages 169–176.

[van Amerongen, 1988] van Amerongen, R. (1988). Optimal power flow solved with sequential reduced quadratic programming. *Electrical Engineering*, 71:213 – 219.

[White, 2009] White, T. (2009). *Hadoop, The Definitive Guide*. O'Reilly Press.

[Wood and Wollenburg, 1996] Wood, A. and Wollenburg, B. (1996). *Power Generation Operation and Control*. John Wiley and Sons, New York.

[Zhang et al., 2002] Zhang, L., Pan, Q., Bao, P., and Zhang, H. (2002). The Discrete Kalman Filtering of a Class of Dynamic Multiscale Systems. *IEEE Transaction on Circuits and Systems II: Analog and Diginal Signal processing*, 49.

Chapter 8

Blockchain: Revolution in TRUST

RK Shyamasundar and Vishwas T Patil
IIT Bombay

8.1 Introduction

TRUST is the cornerstone of our relationships; whether in business, society, or with the institutions that govern us. As Internet has extended the sphere of our ability to do business and conduct personal interactions across the world, its trustworthiness has come under stress in the past two decades. Our collective journey through the Internet seems inalienable now as almost all the services that we use today rely on it. Anyone who controls the Internet or the information flowing through it, controls and manipulates access to online services. It had been a continuous quest to bring trustworthiness to the Internet and the services it enables. Invention of blockchain seems to have quenched this quest.

Since Gutenberg invented the modern printing press more than 500 years ago, making books and scientific tomes affordable and widely available to the masses, no other new invention empowered individuals and transformed access to information as profoundly as Google. Access to information combined with global supply and demand is reshaping established conventions and destroying old definitions. Technology does not create prosperity any more than it destroys privacy. In this digital era, technology is at the heart of just about everything – the good and the bad. The explosion in online communications for social interactions, online financial transactions and commerce has lead to enormous opportunities for cyber crime. While there had been serious efforts to solve Internet's issues of security and privacy via cryptographic technology, there had always been information leaks due to the underlying trusted third parties (TTP). TTPs have evolved as facilitators of e-transactions. They act as trusted intermediaries between two transacting parties. For example; banks, DNS servers, search engines, news outlets, government registries for land or voters, et al – some interchangeable, others enforced. TTPs wield enormous power over our transactions. This centralized power constitutes knowledge of information about our transactions, their inclusion and exclusion from ledgers or just plain rent-seeking. Even paying online with credit cards reveals too much personal details with the added issue of high transaction cost. TTPs can quickly turn from facilitators to controllers, at times, with the blessings of the regulators that are statutorily entrusted by the people through their elected governments. The role TTPs play is undeniably important. However, it is equally important to be able to verify the trust enshrined in them is not being abused. So far, we have taken refuge in legislation, to do the verification, and we have failed miserably time and again as we have tried to solve a technological challenge through non-technological tool, which itself is subject to misuse, favoritism, and malfeasance. Thus, one of the challenging engineering quest had been to build a *Trust Protocol*[1]*, that would naturally blend with the trust as conceived in the society. The underlying protocol of Bitcoins referred to as Blockchains is expected to disruptively revolutionize the notion of Trust among citizens and governments with respect to currency, societal transactions, finance, asset management, etc.*

Blockchain – *a type of database that is immutable, auditable, and distributed* – is expected to overcome persistent structural and systemic obstacles confronting people with limited means in getting societal benefits by bringing in transparency to actions by the stakeholders (that includes, among others, government as well) of asset and financial management – thus leading to overcoming excessive bureaucracy, cultural snobbery and corruption. In this exposition, we shall address the possible impact of these concepts in various social sectors for realizing trust and transparency, particularly in the Indian context.

[1] Or as Nick Szabo termed it *The God Protocol*.

8.2 Money, Currency and the History of Trust

The notion of "trust" as conceived in the society demands a glance at the history of money and currency. Historically, those who had the strongboxes and those who had strong moral fiber emerged as the custodians of people's money and other forms of wealth [Pitroda and Desai, 2010]. Wealth requires protection. In the era of Kings, who had armies to protect the wealth, people used to store their produce/wealth against receipts issued by the accountants of the kingdom. These receipts were used as a medium of exchange for trade. Scarce metals like gold and silver[2] were also used as a medium of exchange in the form of coins having additional benefits of divisibility, unit of value, fungibility, and universal store of value (being acceptable across kingdoms). However the precious metal coins suffered by debasement. With the invention of printing, paper notes were introduced as currency. Classically, sovereign states appoint central banks to perform this task. In 1609, the Bank of Amsterdam was guaranteed by the City of Amsterdam (major commercial center at that time) and was tasked with bringing order and efficiency to the wide range of coinage in circulation in Amsterdam. The Bank accepted local, foreign and debased coins, valued them according to common standards, and then gave credit in an account with a common value, bank *currency*, for which it issued receipts (and charged a small administrative fee.) This standardization of values significantly diminished the incentives to debase money (and the profitability of doing so) and was an important step in making money more efficient. The act of state becoming a guarantor of a bank for the protection of wealth is an act of transitive flow of trust from state to bank (a public or private entity.) The Bank of Amsterdam, initially operated solely as a depository institution, on a 100% reserve basis. In other words, none of the precious metals on deposit were loaned out to other parties. Each receipt issued by the Bank of Amsterdam had equivalent amount of metal deposits in its vault; thus maintaining a full convertibility of receipts into metals and vice versa. However, the Bank of Amsterdam started lending money to the Dutch East India Company, initially on a short-term basis, out of the deposits of others and this activity is known today as *fractional reserve banking.* This was one of the earliest steps toward modern fiat currency, generating notes that were only fractionally backed by metal deposits.

In 1694, the Bank of England was founded as a private bank, incorporated to allow William III to borrow 1.2M Sterling that the city goldsmiths could not support. In exchange for the share rights offer of 1.2M Sterling (that was then lent to the government), the bank gained the right to issue notes, including against the government bonds it had received. This was an important right and another step towards modern fiat currency. In time, through a succession of Acts restricting its competitors, the Bank of England came to monopolize bank note issuance in England and Wales, and effectively became the Central Bank of UK. Pound Sterling became the world reserve currency during the period of British East India Company dominating the world trade.

By the 20^{th} century, the US dollar (USD) had replaced the pound Sterling (GBP) as the most important reserve currency in the world and, as a consequence, the Federal Reserve became the key Central Bank in the world. Like the GBP, the USD exhibited a long history of fluctuating through periods of convertibility and non-convertibility to metals throughout its history. The US adopted the gold standard in 1879. Having a currency backed by an actual precious metal helped lend credibility to the governments that issue it. *It facilitated the trust these institutions needed to make their financial system work.*

In 1933, President Roosevelt and Congress began taking the US off the gold standard with a resolution[3] nullifying the right of citizens to demand payment in gold for their currencies. People were also required to deliver all gold coins, gold bullion, and gold certificates owned by them to the Federal Reserve at a pre-set price of USD 35. By hoarding all of the gold and controlling its price, the Federal Government effectively controlled how much money was in circulation. The irony of the situation is that abandoning the gold standard (people's trust in fully convertible currency) was done to build confidence in the economic system. This allowed introduction of Keynesian model of stimulating economy in recession through state spending.

In 1971, President Nixon announced that the US was no longer in the business of converting dollars to gold at the fixed value of USD 35 per ounce, and thus the gold standard was abandoned completely. **With the absence of a gold-backed dollar, US citizens inherited a fiat currency system backed by nothing but the trust in the government.**

Today, the USD is a 100% fiat currency with no redeemability into any commodity assets, managed by the Federal Reserve. Almost all national currencies that exist today are fiat currencies managed by their respective central banks. U.S. law allows foreign central banks and several international organizations to maintain dollar-denominated deposit accounts at the Federal Reserve. The Federal Reserve is the fiscal agent of the U.S. Treasury. Major outlays of the Treasury are paid from the Treasury's general account at the Federal Reserve. Similar relationships exist between national treasuries (i.e., the governments) and national central banks across the globe.

Thus, the societal TRUST has moved from gold deposits to fiat currency system in each country and also among the countries (that are often linked through the US dollar). In other words, for a functional financial system to work, citizens need to keep TRUST in it, mediated/guaranteed by its

[2]Money must be a store of value and maintain its purchasing power over long periods of time.

[3]an authoritative order or official decree – known as fiat.

(elected[4]) government. A point to be seriously noted is that citizens may lose trust due to excessive bureaucracy, cultural snobbery and corruption. Thus, if citizens don't trust a government to represent their interests, they won't trust its currency–or better put, they won't trust the monetary system around which their economy is organized. So when given a chance, they will sell that currency and flee it for something they regard as more trustworthy, whether it is the US dollar, gold, or some other safe haven [Vigna and Casey, 2016]. The question is where does the Trust flee? **Trust needs an anchor.** And a government's fiat as a foundation for anchoring trust is as credible as the government. As the proverb goes: ***trust, but verify;*** *the promise of fiat cannot be verified in present but only in future, since fiat is a promissory note on future good and it is backed by the strength and stability of a geopolitical system, legal system, and the economy.*

8.3 TRUST in the Internet Era

With the increase in online transactions, and e-commerce, there is naturally a significant increase in privacy leaks and financial fraud – mostly due to the negligence/malfeasance of the TTPs. Thus, started a huge effort on arriving at a cohesive trust protocol to overcome these issues. *One of the main impediments for electronic cash* a la *currency was* double-spending *that reflects the capability of spending the spent cash again and again – arising due to copying coming for free in the digital world.* In 1993, a brilliant, secure, anonymous payment system over the Internet, called ecash [Chaum et al., 1988] by Chaum, Fiat, and Naor, was invented mimicking the societal traits and solving the problem of double-spending in digital currency. As perhaps the e-commerce volume had not yet reached its tipping point, the scheme did not go far. Also, centralization of trust became a contentious issue with the Cypherpunks[5], since to check the double-spending efforts of the ecash in circulation a central trusted server was required. ecash solved the problem of double-spending and brought anonymity to buyers from the merchants but the central server verifying the double-spend efforts would know behavior of its ecash clients. Cypherpunks wouldn't settle for this drawback. And thus, the quest of a universal, decentralized trust protocol continued.

In the meantime, the relevance of the quest for universal trust protocol seemed urgent in light of the following events [Vigna and Casey, 2016]:

1. The remittance of money had increased enormously (the transaction cost and the settlement time remaining unreasonably quite high.)

2. The privacy issues involved and the (hidden!) cost of transactions of credit cards had increased significantly (realized both in the developed and the developing world.)

3. While the overall literacy in the world increased, a vast majority of the population in poor countries and a large fraction in middle income countries did not have bank accounts. The important point to note is that the reason for not having bank accounts was not education or literacy but due to persistent structural and systemic obstacles [of India, 2015, Force, 2016][6] confronting people with limited means; in other words, it was due to undeveloped systems of documentation and property titling, excessive bureaucracy, cultural snobbery and corruption.

4. In past decades, hyper-inflation had been experienced in countries like Zimbabwe, Venezuela, Greece, etc. as an outcome of systemic deficiencies in their respective financial systems. This shows that the elected governments are susceptible to tread a financially disastrous path at the expense of populist decisions to get re-elected. It is understandable that no government would like to undertake arduous path of fiscal prudence to rectify the inherited financial mess, instead they tend to pass it on to the next government, thus increasing the severity of the economic consequences to all. Hyper-inflation, once set in motion, can gradually debase[7] the fiat currency of respective state.

5. In 2008, global financial system collapsed. It was an epic outcome of lack of transparency in evaluation of toxic assets with banks, failure of (or abuse by) regulators (entrusted legal entities) to identify discrepancies in audits. In hindsight, it appeared to be a collusion between auditors and regulators.

[4]Election is a process of *entrusting* a set of people, for a stipulated period of time, to carry out an agenda that is agreed upon by the majority of the people – irrevocable transfer of trust from the people to *the elected.*

[5]An informal group, since late 1980s, aimed to achieve privacy and security through proactive use of cryptography. PGP (by Phil Zimmermann) was one of the first notable tools from this movement. David Chaum was also part of this movement.

[6]Excessive KYC requirements can hinder financial inclusion as providers might find it too onerous to deal with the poor. The Goal: Design KYC rules that are adequate to the task of maintaining financial integrity, yet do not create unnecessary barriers to financial inclusion. OR, get rid of KYC altogether? We shall see one such possibility in later part of the report.

[7]In 1609, Bank of Amsterdam had done away with the problem of debasement of precious metals by introducing paper currency. Inflation (beyond a moderate level, usually above 2%) is a way of debasing a currency by its issuer! So, how do we control such a manipulation of currency? In other words, such a control is a desired property of a digital currency.

Around this time, arrived a new decentralized protocol for peer-to-peer digital currency system, using standard cryptographic functions, called *bitcoin* [Nakamoto, 2008] by a pseudonymous person or a group of people under the name Satoshi Nakamoto. This digital currency, due to the use of cryptographic functions, also referred as cryptocurrency, is different from the fiat currencies as it is neither created by any country nor controlled by any country but governed by cryptographic algorithms. Bitcoin established a protocol involving distributed computations by the disparate stakeholders that collectively ensure integrity of the data exchanged among billions of subjects without involving a trusted third party. The data created is essentially a distributed ledger denoting the actions by the stakeholders. This collective data about transactions among subjects, generated periodically as blocks, is referred to as "blockchains". Note that blockchain is cryptographically protected, and resides on distributed network and not on some central database that is under the purview or control of some organization like central bank and hence public! Each stakeholder can see every transaction (transfer of currency from one subject to the other) on the network and terms it to be valid only if it is not double-spend. Thus an immutable, append-only, global database is generated and maintained by the subjects without allowing any single stakeholder to manipulate the entries in the database, also called as ledger. Therefore, blockchain can be termed as a special type of database in which entries only can be appended and old entries in the database cannot be updated. Thus giving its transactions immutability, integrity, transparency.

In the digital era, all transactions are recorded in ledgers, i.e., in the databases of transacting peers and a copy in the ledger of the TTP facilitating these transactions. Integrity of these ledgers is of prime importance. Transacting peers implicitly trust a third-party who is tasked with maintenance of the transaction ledgers. To understand the importance of provable guarantees on immutability of transaction data in ledgers we need to first understand the shortcomings of digitally-signed entries in ledgers prevalent in pre-bitcoin internet era.

8.3.1 Triple-entry accounting & digital ledgers

Databases play an important role in Internet era accounting. They replaced traditional paper based ledgers with double-entry[8] accounting (a *balance sheet equation* matching the two columns of *assets* and *liabilities* – a correct entry must refer to its counterparty) that helps in identifying unintentional human errors in the ledgers and correct them. In paper based ledgers an attempt to fudge the ledger leaves a physical trail of evidence which later could help in investigation of source of malfeasance. By intrinsic nature of digital records, it is not possible to rely on physical evidence of tampering. That is, there is a need for an out-of-the-ledger system to be deployed again in digital form – for which again the same issue applies. The challenge is to terminate the recursion at a time acceptable to the stakeholders. This is resolved through notion of signed-receipts, which is captured below.

Double-entry accounting using ledgers is prevalent in all organization including governments as they give an auditable state of movement of assets of an organization. Similarly, inter-connected double-entry ledgers give a state of movement of assets across organizations. Unlike the physical ledgers, digital ledgers are remotely accessible and thus can be altered by an attacker without leaving a physical trail. Therefore, while transitioning from physical ledgers to digital ledgers, integrity of ledgers was an important requirement. Double-entry bookkeeping provides evidence of intent and origin, leading to strategies for dealing with errors of accident and fraud. The invention of the signed-receipt in the field of financial cryptography brought in these above-mentioned benefits of double-entry bookkeeping to digital ledgers. Signed receipts are the digitally signed proofs of transactions – *at a given point in time, this information was seen and marked by the signing computer.* Digital signatures introduced a new way to create reliable and trustworthy entries, which can be constructed into accounting systems. There are three parties to such transactions: sender of a value, receiver of the value, and the contract manager of this transfer – receipt issuer; a trusted party. For example, when Alice wishes to transfer value to Bob in some unit or contract managed by Ivan, she writes out the payment instruction and signs it digitally, much like a cheque is dealt with in the physical world. She sends this to the server, Ivan, and Ivan presumably agrees and does the transfer in his internal set of ledgers. He then issues a receipt and signs it with his signing key. As an important part of the protocol, Ivan then reliably delivers the signed receipt to both Alice and Bob, and they can update their internal ledgers accordingly. This results in three active agents who are charged with securing the signed entry as their most important record of transaction. In evidentiary terms, the signed-receipt is more powerful than double-entry records due to the technical qualities of its signature. Triple-entry accounting is a logical arrangement of three-by-three entries, which is a meld of signed-receipts (providing evidentiary power) with double-entry accounting (providing convenience as well as the power to cross-check records locally.)

Triple-entry accounting was one of the fundamental contributions of financial cryptography that paved way for

[8]More than 500 years ago a new accounting technique, later known as double-entry bookkeeping, emerged in northern Italy. It was a big step in the development of the modern company and economy. Werner Sombart, a German sociologist who died in 1941, argued that double-entry bookkeeping marked the birth of capitalism. It allowed people other than the owner of a business to keep track of its finances [Economist, 2017].

modern digital ledgers that not only provide ACID (atomic, consistent, isolated, and durable) properties to the transactions but also the evidentiary property through signed-receipts.

8.3.2 Perils of Centralization of Trust

In this era of globalization business processes, workflows, supply-chains generally span across many organizations. Therefore, ledger of an organization gets interfaced with the ledgers of its collaborators. For example, a purchase transaction on Amazon's online store not only leads to an entry in Amazon's ledger but also in the ledgers of Amazon's sellers, couriers and also in respective bank ledgers of the buyer and seller. A curious look around us will lead us to realize that everything around us is recorded in ledgers somewhere down the line, across organizations/nations/continents: the phone calls, travel commutes, expenses, property titles, share markets, remittances, et al. – almost everything spanning from personal finance to businesses! **Whoever controls a ledger[9], wields an enormous power (financial, political) over the subjects of the ledger.** Digital ledgers with triple-entry feature, which are ubiquitous in our current digital economy are deficient in following aspects:

1. **efficiency:** in a distributed setup, to preserve the atomicity of a transaction, each entity needs to wait for a signed-receipt before updating the local ledger. Usually, a highly-available, trusted third-party assists the transacting peers to settle transaction efficiently. This leads to a hierarchy where a TTP is at the top and has a view of all transactions being settled through it, which leads to generation of meta-data (data about data) that again forms a new proprietary ledger owned by the TTP!

2. **cost:** TTPs facilitating online transactions do charge a fee. The problem arises when a TTP achieves a dominant market position (e.g., Western Union, Visa, Uber), the cost of facilitation appears exorbitant. The facilitation is not necessarily be always charged in legal currency, it could be recovered [Patil and Shyamasundar, 2017] by aggregating transaction's meta information and using such information to earn legal currency (e.g., OpenDNS, JustDial.) Despite charging a fees on transaction settlement, there is nothing that stops TTP from monetizing the transactions' meta information. Another input to the transaction cost is the cost of dispute resolution.

3. **transparency:** TTPs tasked with managing a centralized ledger, against which the state/existence of transactions can be checked, derive an implicit trust of relying parties. Thus, TTPs derive an enormous power over sanctity of past transaction and inclusion/exclusion of on-going/future transactions. In a distributed setup of inter-connected ledgers, a deliberate or accidental modification or suppression of transaction adversely impacts entries in connected ledgers (e.g., propagation of toxic loans in 2008 US mortgage crisis.) Transparency is a trust enhancing property and improves accountability.

4. **control:** being the middle-man for transactions TTPs have quasi-control over whose transactions can go through their system (e.g., 2010 financial blockade against WikiLeaks) and at what fee. Furthermore, as our digital identities have become our primary identities, TTPs can accidentally or maliciously may annihilate an individual's digital presence. The serious impact of control is on the personal data front. In the data-driven economy, end users interact with online services that are run by algorithms, which in turn make decisions based on the supplied user data. An error, omission of user data affects the algorithms behavior. Though users are coerced/compelled to share personal data, their interaction with services generate meta-data, which is generated collectively but aggregated and controlled by the service provider without any curative interface for users. This has created a huge information inequality in the ecosystem. This stifles competition among incumbent service providers and puts high barrier for the new ones.

These problems stem from our reliance on centralized, trusted third-parties; such as banks, clearinghouses, telcos, credit-rating agencies, government departments and many other big players of our digital economy like Google, Amazon, Facebook that collect and control our personal data under the pretext of personalization and convenience. Computational and communication advances are enhancing the speed of transactions and reducing costs of transactions. But, on the fronts of transparency, fraud-prevention, and control over the data we have not seen much advancement. The reasons are two-fold: i) in the digital economy, data and meta-data is equivalent to gold. It is a compelling differentiator and there is an on-going rush to hoard and control as much data as possible. ii) lack of a global platform to orchestrate data life-cycle management.

In a stark comparison with old economy, where trust was under strain due to central banks and governments, new digital economy further aggravated the strain on trust due to the necessity of trusted-third-parties. Trust continued to erode from public sphere in light of large-scale data breaches and a continued intrusion of businesses in personal sphere. At times the regulators appeared to be in collusion with the businesses. In the

[9]There are ledgers about ledgers that are usually maintained by entities that are positioned at the top of our communication infrastructure – ISPs, telcos, Governments, PKIs, DNSs – do collect data about data called meta-data, which constitutes the ledgers about ledgers.

meantime, Cypherpunks continued their journey beyond ecash to build an electronic payment system based on cryptographic proof instead of trust, allowing any two willing parties to transact directly with each other without the need for a trusted third party.

8.4 Bitcoin: Currency without Fiat

In response to the above mentioned serious impediments to achieve verifiable trust, an experimental, decentralized, P2P platform called blockchain was invented by Satoshi Nakamoto in 2008 (of course, perhaps worked over the years.) The platform was specifically designed for keeping track of cryptocurrency called bitcoin, which is generated by the platform itself. It is a self-breeding platform that generates its own currency to keep itself running. The currency is issued transparently and continuously accounted for. That is: new bitcoins are generated to represent some work done by someone in the network and are rewarded to the worker by making an entry in the ledger of the network. The worker is allowed to spend this reward at will, provided such a will to spend is broadcasted to the network and recorded in the ledger of the network. Similarly, a recipient of bitcoins from a worker is allowed to spend them at will, provided such a will to spend is again broadcasted to the network and recorded in the ledger of the network. And this goes on. Every node in the network can read the ledger and thus can precisely knows the owners of bitcoins at any given time. Therefore, there is transparency and freedom to verify each bitcoins origin and its traversal to current owner. Who (among the many) is authorized to write to the ledger is the challenge that Satoshi solved. Remember that whoever controls a ledger wields enormous power over the subjects relying on the ledger.

In a nutshell, Bitcoin is a global ledger of values that is collectively owned and governed by rules that cannot be amended without a global consensus. The Bitcoin system [Nakamoto, 2008] consists of two intertwined components:

- **blockchain:** the protocol to maintain the global ledger, and
- **bitcoin:** the currency to incentivise the maintenance.

Since the ledger is maintained collectively there is no dependence on a TTP – thus not inheriting the perils of relying on a TTP. Being a global, collectively maintained ledger, everybody can read & validate the transactions in the ledger. The issuance of currency is done as per the ledger maintenance work. Each peer in the Bitcoin network can read the ledger and be assured of ownership of currency at any point in time – thus transparent and publicly auditable.

> *The root problem with conventional currency is all the trust that is required to make it work. The central bank must be trusted not to debase the currency, but the history of fiat currencies is full of breaches of that trust. Banks must be trusted to hold our money and transfer it electronically, but they lend it out in waves of credit bubbles with barely a fraction in reserve. We have to trust them with our privacy, trust them not to let identity thieves drain our accounts.* – *Satoshi Nakamoto*

Through Bitcoin, Satoshi showed an alternative currency system without a central trusted party who plausibly can debase, control, kill a currency. The anchor for trust is rooted in cryptographic proofs rather than in governments' fiat. Therefore, very quickly this currency received a global appeal and acceptance.

Satoshi managed to engineer the concepts of economics like scarcity, supply, demand, unit of work, incentive into computer science. At the core of all these concepts is unit-of-work: the universally acceptable and verifiable method to quantify a unit of work using computers. Borrowing from the works of Dwork and Naor [Dwork and Naor, 1993], and Adam Backs [Backs, 2002], Satoshi resorted to SHA256 cryptographic hash function[10], which is universally available on all computing devices, to define unit-of-work. SHA256 function takes an input string and produces a 256-bit long output. Therefore, given an input string, all computers in the world will produce the same 256-bit long output using SHA256 function. Successive invocations of this function constitutes amount of work. And to define work itself, Satoshi resorted to a simple condition of having first n bits of the output string to be zero, where n is a measure of determining hardness/difficulty of the work. In a given time period, a computer that can invoke SHA256 more number of times than another computer has higher chance of completing the work. How this definition of work is used to build a decentralized, global, ledger management system and a currency to incentivize ledger management, we should understand the notion of proof-of-work. Proof-of-work is a method to tie an entity to its successful completion of work before the others and therefore claiming a reward from the system for successfully completing the work. The work is: to extend the ledger with previously unrecorded transactions. The extension is periodic and constructed as a block consisting of a subset of valid transactions during that period. A block is a unit of successfully completed work – therefore aptly named as *blockchain.*

[10]It is a mathematical algorithm that maps data of arbitrary size to a bit string of a fixed size; 256 in the case of SHA256.

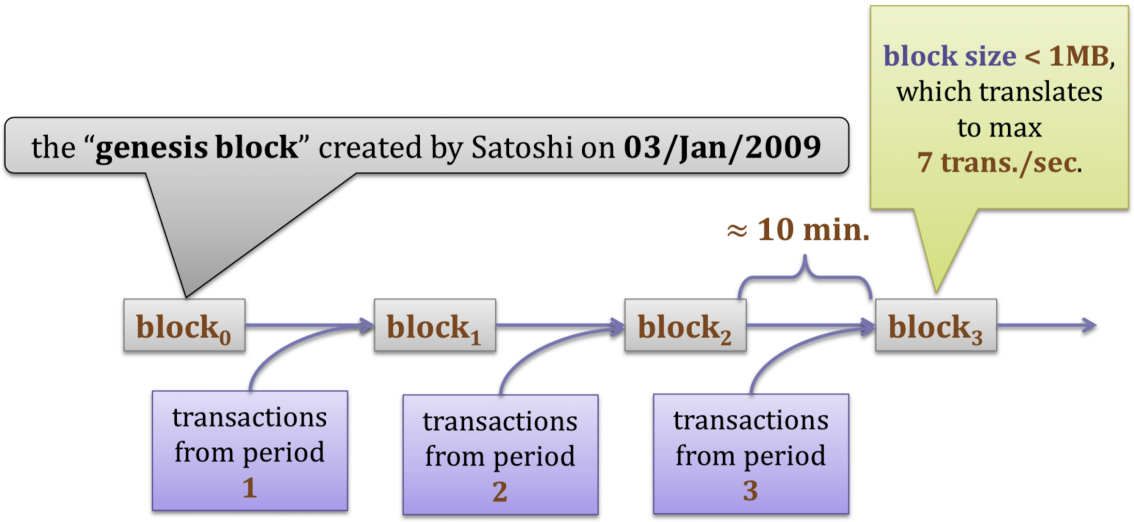


Figure 8.1: Genesis block as the first block of blockchain (Credit: Stefan Dziembowski)

8.4.1 Proof-of-Work

It is an algorithmic crux of the Bitcoin system, where nodes are incentivized to do work. The first node that publishes a proof of doing a pre-defined work is rewarded by a fixed number of bitcoins. All other nodes give up that work upon verifying the correctness of winning node's proof. If the proof is correct a new round to do new work starts. Nodes put efforts to complete the new work before others in order to claim the associated reward upon submitting a proof of work completion. The algorithm adjusts hardness of the work such that, on an average, for every 10 minutes, a unit of work is done by someone in the network.

To collect rewards from the system, the nodes need to be identified, which is done by allowing the nodes to independently generate a cryptographic key-pair where the public-key part of the key-pair is node's identifier and the associated private-key is the guardian of the rewarded bitcoins. Rewards (bitcoins) are issued to a public-key iff the corresponding node submits a valid proof-of-work computed before others. A node that receives bitcoins as a reward is free to transfer these bitcoins (as payment/gift/donation) to others. To transfer a bitcoin, a node composes a transaction that consists of information about how it has received the bitcoin and to whom it wants to transfer that bitcoin. Similar such transactions constructed by other nodes are observed by all nodes and are used as input to generate proof-of-work in the hope of receiving reward from the system. Thus, the consecutive submissions of proofs-of-work that are peer validated and universally accepted, produce a series of blocks (therefore collectively called blockchain, depicted in Figure 8.1.) Each block represents a proof-of-work and its reward is assigned to the respective worker's (winner's) public key. A block contains transactions that were submitted for confirmation before the creation time of that block. Blocks being of a fixed size, i.e., less than 1MB, it is not guaranteed that all the unconfirmed transactions floating around in the Bitcoin network will be accommodated in current block. Unaccommodated (unconfirmed) transactions may get accommodated in subsequent rounds of proof-of-work. Transactions may optionally offer a fees as a premium so that its inclusion in current block can be prioritized. The creator of a block collects transaction fees on top of the guaranteed reward.

8.4.2 Programming the Concepts of Economics

Through proof-of-work we saw how Satoshi succeeded in defining a universally acceptable unit of work and a mechanism to irrevocably tie the proof to a public-key. In the following we shall see how ingeniously Satoshi encapsulated the other concepts of economics using computational engineering.

Engineering scarcity, supply, and demand: In order to induce value in something, it has to be scarce and known to be limited in supply. Satoshi fixed the total number of bitcoins, to be issued by the Bitcoin network, to 21 million. New bitcoins come to existence approximately every 10 minutes upon a successful proof-of-work round (in other words, upon creation of a new block). For the first 210,000 blocks the reward was 50 bitcoins/block. For the next 210,000 blocks it halved to 25 bitcoins/block. At present (November 2017) it is 12.5 bitcoins/block. Alternatively,

$$210,000 * (50 + 25 + 12.5 + 6.25 + ...) \rightarrow 21,000,000$$

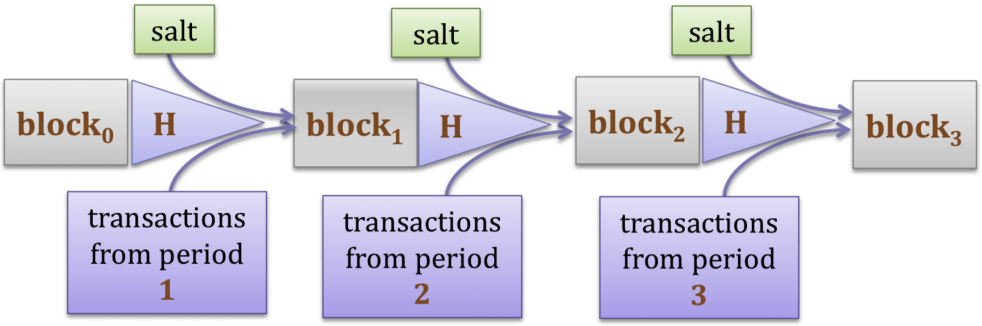


Figure 8.2: PoW Construction, H is SHA256 (Credit: Stefan Dziembowski)

where the last block to be mined is expected in year 2140. Thus, there is a constant but reducing supply of bitcoins from the network. So far, 80% of the total bitcoins have been mined and each trading at USD 11,000 (November 2017) on global bitcoin exchanges. Whereas, it was trading at USD 376 and USD 742 in November 2015 and November 2016 respectively; this highlights the increase in demand of bitcoin. Each bitcoin is divisible up to 10^8 Satoshis – the indivisible unit of bitcoin currency.

Engineering fairness, integrity, and incentive (through proof-of-work): The indivisible unit of work in Bitcoin system is SHA256 – a cryptographic hash function. By definition, a cryptographic hash function maps an arbitrary sized input to a fixed size output such that it is infeasible to determine the input from a given output string. In other words, there is no efficient way to determine an input value mapping to a specific output value. Therefore, the only way to find an input value leading to a specific output value is by repeatedly trying out different input values. The time required to find an input value leading to a specific output value using SHA256 function is directly proportional to the number of invocations of SHA256 function with different inputs. These properties are exploited to construct the proof-of-work algorithm, where the input string consists of 3 elements (2 fixed and 1 random), which are: i) Merle-root of unconfirmed, valid transactions viewed in the network, ii) hash of most recent block in the blockchain, and iii) a random value (aka salt/nonce), producing an output string of length 256-bits. Proof-of-work algorithm demands the output string conform to a pattern in which first n bits of the 256-bit string are zeros. The algorithm invokes SHA256 function recursively until an acceptable target string is not obtained. This is depicted in the equation below and in Figure 8.2.

$$H(salt, H(block_i), transactions) = target$$
$$\text{such that } target \text{ starts with } n \textbf{ zeros}$$

where n is the hardness parameter for proof-of-work algorithm for a period of time, which is approximately 2 weeks. Hardness parameter is periodically adjusted because the computing power of nodes[11] changes. The hardness adjustment is automatic, and depends on how much time it took to generate last 2016 blocks (i.e., 2016×10 mins $= 14$ days.) If the previous 2016 blocks took more than two weeks to find their respective targets, the hardness is reduced. If they took less than two weeks, the hardness is increased.

Fairness: The probability of finding an acceptable target value for a successful proof-of-work is directly proportional to the disposable hash power of a miner. Miners sell their bitcoins in order to increase their hash power so that their probability to find proof-of-work increases. The system takes care of potential spike in computational power in the network by adjusting the hardness value (depicted in Figure 8.3) in finding a target value.

Integrity: The cost of changing the contents of old blocks is compounded by each new block that gets added to the chain. When a new block is made, it contains the hash of the one before it. Any changes in old blocks will result in invalid hashes for all subsequent blocks. Therefore, it is impossible to insert bogus modifications into a previous block without having to repeat all the work that was performed after that block.

Incentive: Proof-of-work produces a block containing a special transaction (coinbase) that transfers the reward to the miner. Reward provides incentives to be a miner. It also makes the miners interested in broadcasting new block as soon as possible. On top of this, a specification in Bitcoin states that "from two blocks of equal length mine on the first one that you received", which brings sense of urgency in broadcasting successful proof-of-work to the network at the earliest.

[11]Nodes having relatively large dedicated hash power are called bitcoin miners. Analogy of miners for nodes is derived from gold miners who voluntarily spend their efforts to find gold in mines with the hope of finding gold. Finding gold is rewarding and vice versa is penalizing.

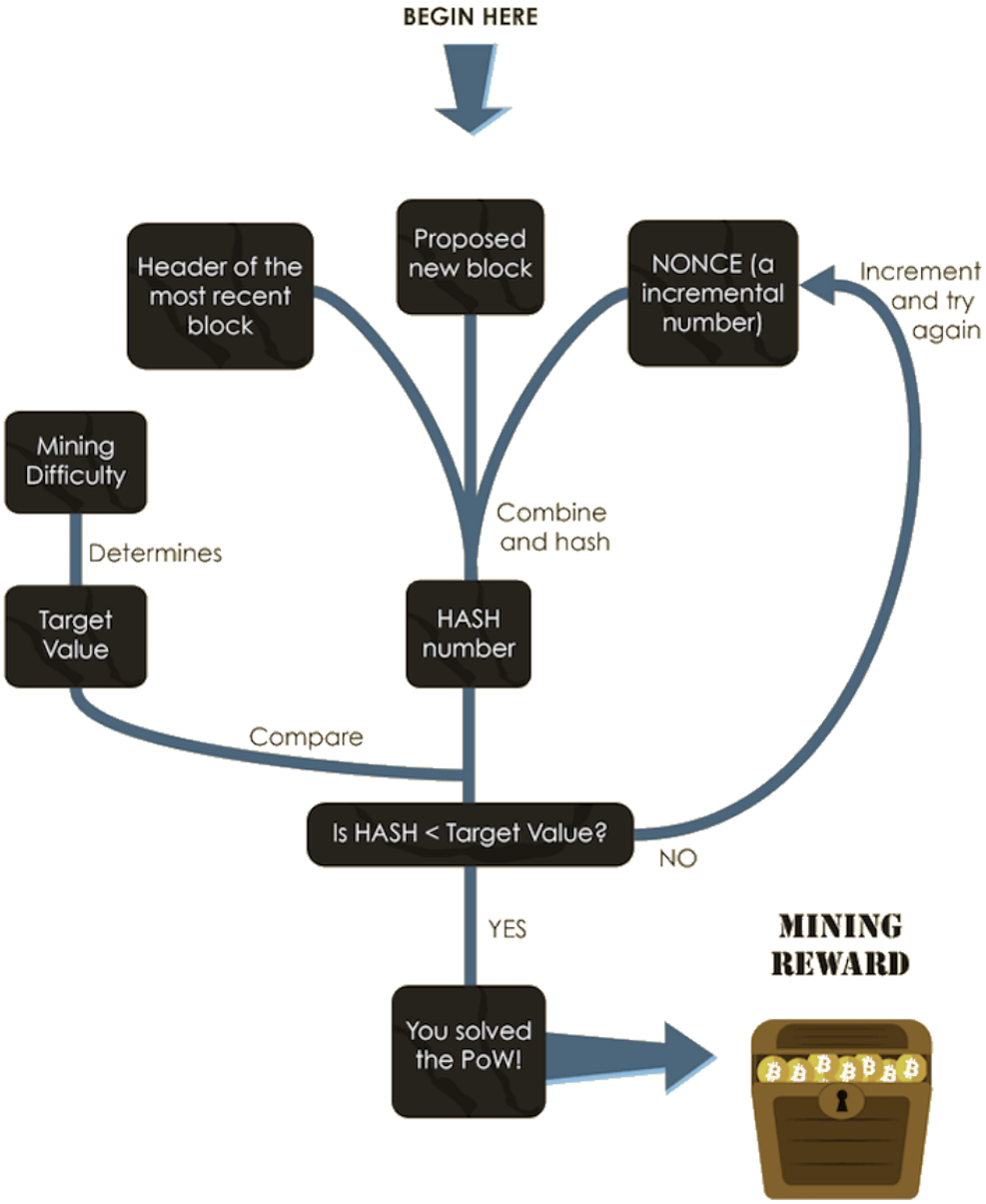


Figure 8.3: Hardness determines the target value for proof-of-work in a given period (Credit: Patricia Estevão)

Engineering financial inclusion (through pseudonymity): Owners of bitcoins are identified by their public keys. Public keys are a very peculiar type of identifiers. They can be generated by anyone and are always associated with a corresponding private key which is generated simultaneously. As an analogy, if public key is considered as login, private key is its permanent password – and it can be used without providing it to the verifier. This is in stark contrast with other identifiers like email, bank account, mobile number, because these type of identifiers are generated by and assigned to subjects. And a copy of passwords associated with these identifiers is stored with its issuer in order to perform authentication. Therefore, subjects using such identifiers can be identified and revoked (excluded from respective systems) by identifier issuer as the issuer owns the identifiers in its namespace. Whereas, a public key is an identifier generated and issued by a subject to self along with its corresponding private key. This helps a subject to remain pseudonomous as long as it desires.

Engineering transparency and auditability (therefore accountability): The chain of blocks at any given time provides the list of verified transactions accepted by the network till that time. In its simplest configuration, the protocol allows any peer to scan through the verified transactions. The participants can trust the integrity of the networked verified transactions because it is computationally infeasible for an adversary to change any network verified transaction. This is so because changing any transaction in a block will change the block's output hash, which will impact the proof-of-work value of the next block in the chain. Note that each block's creation requires previous block's hash value as an input to obtain proof-of-work. Furthermore, availability of all the previous transactions to each participant of the network brings non-repudiation to transactions and transparency in the network.

Bitcoin is the first real-world application of blockchain protocol where proof-of-work is used as a type of consensus algorithm. It is a trusted, self-regulating, transparent application of global transfer of money where the transactions listed in the chain of blocks are equivalent to the ledger entries of any traditional bank. Today's value-transfer systems rely on central ledgers. Banks, governments, telcos and other firms have a big computer that keeps track of who owns what. And when one wants to make a payment, the central ledger is updated. Bitcoin does the ledger updates in a completely different way. It does not have a centrally controlled ledger. Instead, everybody who runs the (full) software has their own copy of the ledger. Hundreds of thousands of people have a full copy of the ledger. This means no single person/entity can deny availability of the ledger and entries in it, confiscate the value/asset marked against an identity, or charge an unfair fee for transactions to go through. And the genius of Bitcoin was to figure out a way to encourage people to maintain these ledgers and to do so honestly and with no trusted third parties.

This new way of deriving trust and transparency in a distributed environment like Internet has tremendous potential to re-engineer all the prevalent systems and applications that are under stress due to lack of trust and transparency. Bitcoin is one such attempt to put forward an alternative financial system where trust is anchored in cryptographic algorithms (that are verifiable in present time) instead of fiat (which cannot be verified until tested in future time) of a government. The technology worked on the principle that, at its foundation, money is just an accounting tool – a method for abstracting value, assigning ownership, and providing a means for transacting. It turns out that such a system may be useful for much more than just money.

8.4.3 Beyond Bitcoin

By forcing miners to provide costly proofs and then repaying them for their work, Satoshi created the first viable peer-to-peer digital currency. But he also solved a more general problem that had vexed computer scientists for decades – consensus. Consensus in distributed systems has been rigorously studied in Computer Science for past few decades as Byzantine Generals Problem or Chinese Generals Problem, in which two generals have to come to a common agreement on whether to attack or retreat, but can communicate only by sending messengers who might never arrive.

> Reliable computer systems must handle malfunctioning components that give conflicting information to different parts of the system. This situation can be expressed abstractly in terms of a group of generals of the Byzantine army camped with their troops around an enemy city. Communicating only by messenger, the generals must agree upon a common battle plan. However, one or more of them may be traitors who will try to confuse the others. The problem is to find an algorithm to ensure that the loyal generals will reach agreement. It is shown that, using only oral messages, this problem is solvable if and only if more than two-thirds of the generals are loyal; so a single traitor can confound two loyal generals. With unforgeable written messages, the problem is solvable for any number of generals and possible traitors.
>
> Achieving reliability in the face of arbitrary malfunctioning is a difficult problem, and its solution seems to be inherently expensive. The only way to reduce the cost is to make assumptions about the type of failure that may occur. For example, it is often assumed that a computer may fail to

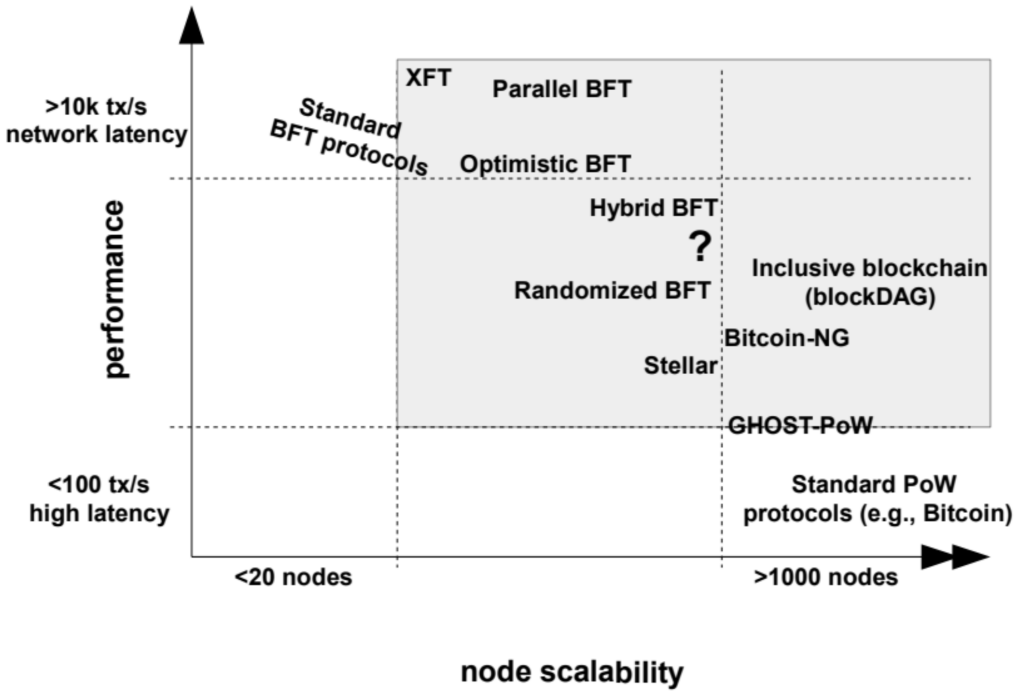


Figure 8.4: Illustration of performance and scalability of different families of PoW and BFT protocols as discussed in [Vukolić, 2016].

> respond but will never respond incorrectly. However, when extremely high reliability is required, such assumptions cannot be made, and the full expense of a Byzantine Generals solution is required [Lamport et al., 1982].

Bitcoin system through its proof-of-work algorithm solved this long standing problem of consensus in distributed system. Bitcoin is at its core, a technology that enables a series of achievements that were not possible before, and not just a global cryptocurrency. Decentralized consensus can create more robust systems in a multitude of ownership or attestation related roles. Currency should be considered as the first application of this technology. Since BGP is a general problem in distributed systems, the same concept can be employed for other purposes. Motivated by Bitcoin, there is a flurry of projects tweaking components of the system and solving different problem of practical importance, which were not addressable before due to lack of a practical consensus method. We shall not get into the details of the tweaks but broadly categorize them into two verticals: permissioned and permissionless – both as types of trust management system with an increasing degree of underlying trust. Some variants built for the want of speed of transaction at the cost of trust, some built for the want of capturing value representation other than currency like land records. This whole family of variants is conveniently called as blockchains, each differing from the other based on the underlying consensus mechanism. There is a subset of variants that use BFT (Byzantine Fault Tolerance) algorithm to construct their consensus algorithm. We briefly mention the prominent ones below.

So far, Bitcoin is the most successful deployment of blockchain protocol with proof-of-work (PoW) as its consensus algorithm. Similar to bitcoin, several alternative cryptocurrencies (altcoins.com) were deployed with slight improvements in objectives. Projects like Stellar (stellar.org), Ripple (ripple.com) are using concept of blockchain to perform global inter-bank settlements with their own private cryptocurrencies; lumens and XRP respectively. An interesting proposal of programmable (Turing-complete unlike Bitcoin, which has limited set of operations) blockchain called Ethereum [Buterin, 2013, Wood, 2014] was floated in year 2013, which is gathering momentum recently in business domain [ConsenSys (consensys.net), Corda (corda.net), Augur (augur.net), et al.] Ethereum is inspired by Bitcoin and presented an alternative consensus forming algorithm called proof-of-stake (PoS) to assuage the concerns of power consumption and latency in verification of transactions in Bitcoin. A prominent permissioned variant called Hyperledger Fabric (`https://www.hyperledger.org/`hyperledger.org) is an open-source project championed by IBM et al that uses PBFT (practical Byzantine fault tolerance) [Castro and Liskov, 1999] as its consensus algorithm. Several other noteworthy consensus algorithms in this space are: Paxos/RAFT [Ongaro and Ousterhout, 2014], Hashgraph [Baird, 2016], Algorand [Micali, 2016], PoET (Proof of Elapsed Time), Blockstack [Ali et al., 2016]. Figure 8.4 illustrates a comparison between Proof-of-Work and BFT (Byzantine fault tolerance) types of consensus algorithms for performance and scalability. And, in Table 8.1 their high level feature-wise comparison is presented.

	PoW consensus	BFT consensus
Node identity management	**open, entirely decentralized**	permissioned, nodes need to know IDs of all other nodes
Consensus finality	no	**yes**
Scalability (# of nodes)	**excellent (thousands of nodes)**	limited, not well explored
Scalability (# of clients)	**excellent (thousands of clients)**	**excellent (thousands of clients)**
Performance (throughput)	limited (due to possibility of chain forks)	**excellent (tens of thousands tx/sec)**
Performance (latency)	high (due to multi-block confirmations)	excellent (matches network latency)
Computational requirement	high	**moderate**
Network synchrony assumptions	physical clock timestamps (e.g., for block validity)	**none for consensus safety** (synchrony needed for liveness)
Correctness proofs	no	**yes**

Table 8.1: High-level comparison between PoW and BFT blockchain consensus families for a set of important blockchain properties. Entries in bold suggest desirable features and highlight advantages of one consensus family over the other. [Vukolić, 2016]

Bitcoin ushered a completely revolutionary protocol through blockchain. It is revolutionary because it showed a way to handle trust without TTPs and suddenly there is an invigorating stock-taking of all the relationships involving TRUST (in business, society, or with the institutions that govern us). Old ways of doing transactions are being re-engineered and a completely new set of applications are engineered – with blockchain at their core as-a-service that offers trust, similar to cloud service that offers on-demand compute, storage, network. In the following we take an abstract view of blockchain and treat it as a machine that provides trust as-a-service!

8.5 Blockchain: The Trust Machine

Conceptually, a blockchain as a machine;

1. stores data (in a shared, distributed ledger),
2. performs some computation (read data from ledger, append data to ledger),
3. reach consensus about both (through algorithms like PoW), and
4. at each epoch changes its internal state to new.

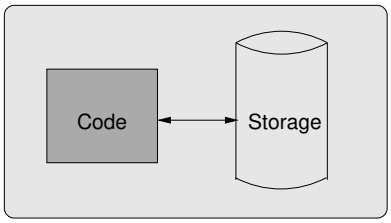


Figure 8.5: Bitcoin code interacting with immutable storage

Figure 8.5 depicts the Bitcoin blockchain protocol as a simple state machine, where "code" starts with a state fetched from the "storage" and the new state is written back (appended) to the "storage." Figure 8.6

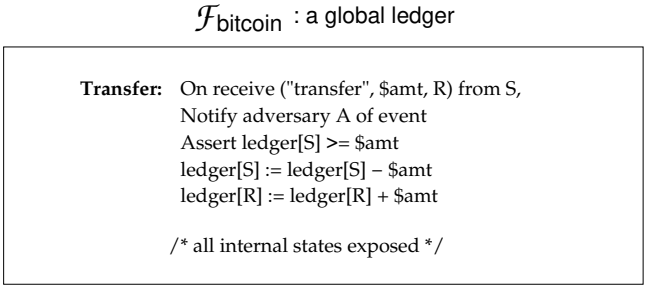


Figure 8.6: Change of state upon invocation of Transfer function

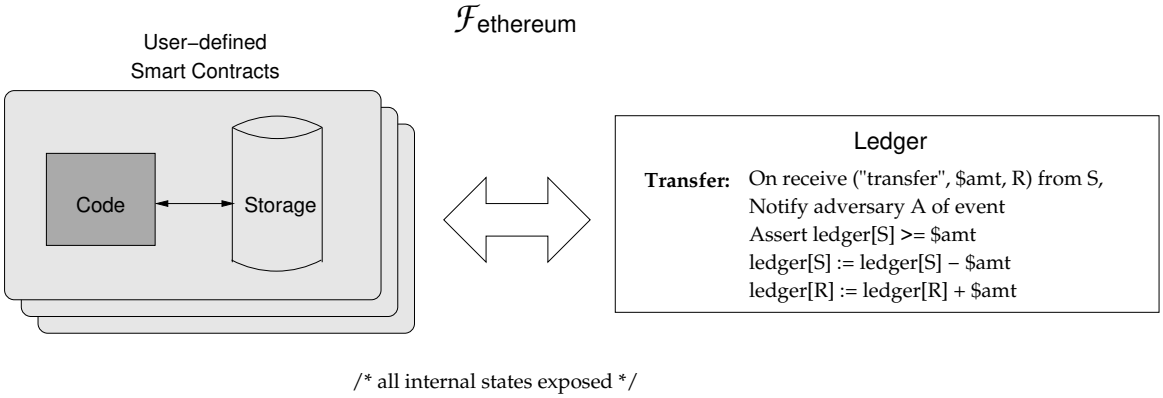


Figure 8.7: Ethereum Trust Machine: local and global state interaction

shows a simplified version of the bitcoin script responsible for transfer of value from sender "S" to receiver "R". In an abstract way, blockchain is trusted for correctness but not for privacy since it exposes internal state to everyone, at least in its primitive form.

Bitcoin can be said as a special purpose program running on blockchain. Functionally it serves only one purpose – transfer of value. The underlying technology blockchain ensures users' trust in the system by thwarting double-spending attempts by malicious users. Bitcoin and blockchain are inseparable. Bitcoin runs on blockchain and blockchain requires bitcoins, as a currency, to incentivize PoW. The elegance of the protocol is in delivering trust without a TTP. It is a self-sustained, self-regulating, transparent "trust machine." Anyone can rely on it but for only one functionality, that is, transfer of value.

In year 2013, Ethereum [Buterin, 2013] was proposed as a new general purpose blockchain that promised more than "transfer of value". It proposed a Turing-complete language to write code that not only does "transfer of value" but also any functionality that can be digitally controlled/interfaced (e.g., transfer of shares, real-estate, etc.)

To put Bitcoin and Ethereum in perspective; Bitcoin is a special-purpose blockchain (like a stand alone Calculator) whereas Ethereum is a general-purpose blockchain (like Android - on which Calculator is an app along with many other apps). Ethereum uses Proof-of-Stake as its consensus algorithm, which is bootstrapped from Proof-of-Work initially. Ether is the currency on Ethereum platform that can be used to buy "stake". Stake provides proportionate voting (consensus) rights. Gas is another concept introduced in Ethereum. A predefined amount of gas is required to execute a smart contract, which is nothing but a program having its own code and storage, that is, its own state. Gas measures how much "work" an action or set of actions takes to perform. Every operation that can be performed by a transaction or contract on the Ethereum platform costs a certain number of gas, with operations that require more computational resources costing more gas than operations that require few computational resources. The reason gas is important is that it helps to ensure an appropriate fee is being paid by transactions submitted to the network. By requiring that a transaction pay for each operation it performs (or causes a contract to perform), we ensure that network doesn't become bogged down with performing a lot of intensive work that isn't valuable to anyone. This is a different strategy than the Bitcoin transaction fee, which is based only on the size in kilobytes of a transaction. Since Ethereum allows arbitrarily complex computer code to be run, a short length of code can actually result in a lot of computational work being done. So it's important to measure the work done directly instead of just choosing a fee based on the length of a transaction or contract.

Figure 8.7 shows a simplified notion of two states in Ethereum "trust machine". Smart contracts have their local state, which is also recorded in the underlying blockchain and the system as a whole has a global state on

which all other smart contracts rely upon.

8.5.1 Smart Contracts – the code on the Machine

A Smart Contract is a contractual agreement that is implemented using software. Unlike a traditional contract where parties may seek remedial action through the legal system, a smart contract is self-enforced (possibly also self-executed), depending on whether specific conditions, that are monitored through software, are met. Smart contracts may provide several benefits, for instance:

- automatically enforce power equality of all parties involved,
- protect an individual's rights by enforcing reasonable expectations for the signee,
- eliminate the possibility of any signatory defaulting on their obligations.

Most financial instruments are essentially a contract depending on the issuer and the set of rules or dependencies set by them. In regulated markets, the relevant security and exchange authorities monitor the compliance of the issuer and user of the contract/instrument to the rules-set. What if we could replace these with cryptographic guarantees? Oracles [Buterin, 2013], in this case, can act as the authority that determines compliance and adherence to the rules set – done objectively, transparently and without trust between contractual parties.

Like Bitcoin, Ethereum uses a blockchain that has its own currency, called ethers. Unlike Bitcoin, Ethereum uses transactions that are miniprograms, called smart contracts, that can be written with an unlimited amount of complexity. Users can then interact with programs by sending them transactions loaded with instructions, which miners then process. In practice, this means that anyone can embed a software program into a transaction and know that it will remain there, unaltered and accessible for the life span of the blockchain.

In other words, a smart contract is an event driven program, with state, which runs on a distributed, shared ledger and which can take custody of assets on that ledger [Weber et al., 2016]. An abstract smart contract model under Ethereum has:

1. Shared public ledger
2. Replicated states (smart contracts)
3. Cryptocurrency as reward for contract execution
4. Contracts that involve financial gains or losses
5. Event driven execution flow
6. Consensus (smart contract state change and recording in global ledger)
7. Participants are not trusted (can read contract before execution)
8. Inter-dependent contracts communicating via the global ledger

Business processes vying for efficiency, transparency, reliability of actions and deliverables upon fulfilling a task are aggressively exploring this space. In our globalized economy almost all workflows span across boundaries and disparate collaborating organizations. The whole workflow loses its efficiency if any of the participating entity acts maliciously or does not perform as expected. It causes litigation and have cascading effects on other organizations. The logic of existing business processes/workflows can be captured and automated through smart contracts and the underlying "trust machine" keeps track of state changes for continuous auditable visibility of the workflow for all users of the machine.

Smart contracts promise to change the economy more than any other feature of the blockchain. They could take over most routine business processes. Some companies could be no more than a bundle of smart contracts, forming true virtual firms that live only on a blockchain [Economist, 2017]. DAO (decentralized autonomous organization) is an example of formation of such virtual venture-capital fund where stakes in the firm can be purchased using ether – cryptocurrency of Ethereum platform. ICOs (initial coin offerings) is yet another simpler version of such structures of automated crowd-funding for startups whose functionality is publicized as a white paper or prospectus for investors in the form of smart contracts. Investors can then send ether to the smart contract, which automatically creates "tokens" that can be traded like shares.

DAOs and ICOs are a type of permissioned or private blockchains that can be realized on permissionless Ethereum platform. There is another form of private or permissioned blockchain that can be realized using Ethereum source code in which the genesis block (zeroth block) of the realized blockchain is shared among select group of participants.

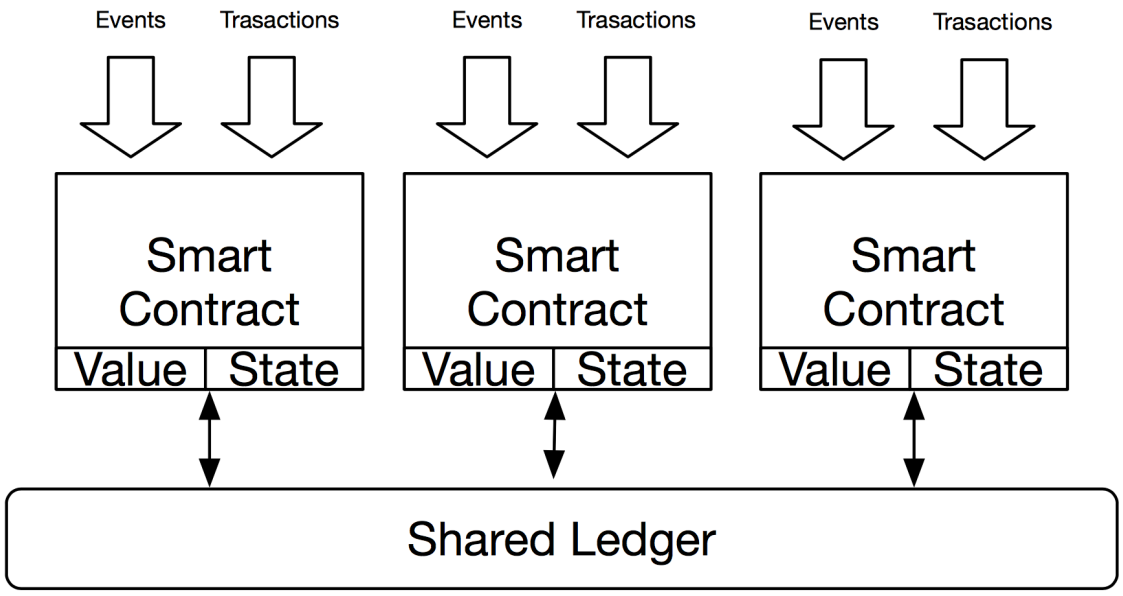


Figure 8.8: Ethereum Inter-Contract Communication

8.5.2 Triggers & Signals – the interrupts to the Machine

Smart contracts are capable of taking inputs from external sources. This makes them extremely useful in addressing and integrating external data sets and proprietary business interfaces that cannot be readily ported to the trust machine either due to legacy issues or privacy concerns. Programmers have to be aware of the fact that each action listed in the code of a smart contract has an associated execution cost. If all of the business logic is as it is imported into the smart contract the gas cost of running the contract increases. Smart contracts should be used as special code snippets of business logic that are critical in communicating state change in the workflow to all other participants in a verifiable, non-repudiable fashion. Non-critical part of the business logic should be off-loaded from the blockchain to reduce the cost of running a smart contract on the "trust machine".

The shared global ledger among the participants acts as a shared communication bus from/to which each participant receives/sends triggers to others via recording a local state change. Figure 8.8 depicts the inter-contract communication using shared ledger. Special smart contracts can be written that specifically act as triggers to other contracts by capturing events in the environment in which they are deployed [Weber et al., 2016]. For example, a stock price tracking contract can trigger a sell/buy contract automatically. In [Azaria et al., 2016], smart contracts on blockchain are used to specify access control policies for medical records of patients. Actual medical records are stored in an encrypted fashion off-blockchain to reduce cost, latency and for privacy preservation. Whereas who can access the data and the keys to decipher are delivered via blockchain as "signals" to the legacy database systems holding actual medical records.

Ability of smart contracts to integrate traditional IT system interfaces into the "trust machine" has brought benefits of automation, efficiency, integrity, continuous auditability, transparency, optimization, etc. to traditional IT systems. IoTs can pave way for similar impact to non-IT systems like physical assets (cars, houses) by facilitating the trigger & signaling interface to the "trust machine".

8.5.3 IoT – the peripherals of the Machine

The advances in networking protocols and miniature, power-efficient computational chips have made our surrounding intelligent and interactive through IoTs. Current deployments (c.f. Figure 8.9) are cloud-centric [Pureswaran and Brody, 2015] and derive their intelligence from cloud, which is privacy invasive. This is largely because of lack of alternatives to deploy and manage IoTs in a naturally ambient fashion. The potential of disruption because of this technology alone is summarized in Figure 8.10 by IBM [Pureswaran and Brody, 2015].

The sort of programmability Ethereum offers does not just allow people's property to be tracked and registered. It allows it to be used in new sorts of ways. Imagine a digitized car-key (password that is needed to start the engine) embedded in the Ethereum blockchain could be sold or rented out in all manner of rule-based ways – enabling new P2P schemes for renting or sharing cars (bypassing TTPs like Avis, Hertz.) Imagining further, smart contract enabled self-driving cars can be self-owning [Economist, 2015a]. Such vehicles could stash away some of the digital money they earn from renting out their keys to pay for fuel, repairs and parking spaces; all

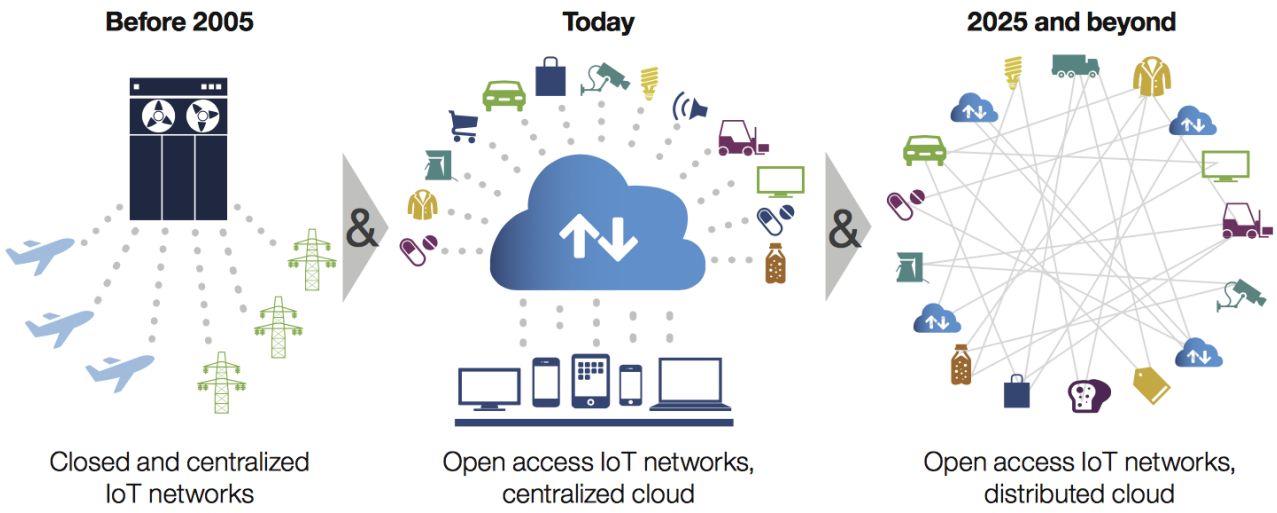


Figure 8.9: IoT progressing towards decentralization [Pureswaran and Brody, 2015]

Vectors of disruption	Liquification of the physical world
Unlock excess capacity of physical assets	Instantly search, use and pay for available physical assets
Create liquid, transparent marketplaces	Real-time matching of supply and demand for physical goods and services
Enable radical re-pricing of credit and risk	Digitally manage risk and assess credit, virtually repossess and reduce moral hazard
Improve operational efficiency	Allow unsupervised usage of systems and devices, reduce transaction and marketing costs
Digitally integrate value chains	Enable business partners to optimize in real-time, crowdsource and collaborate

Figure 8.10: Five vectors of disruption: How the IoT will increase our leverage of physical assets [Pureswaran and Brody, 2015]

according to preprogrammed rules as smart contracts, where IoTs are acting as peripheral devices (interfaces to) connected to the "trust machine".

With respect to the future of IoT, as highlighted in [Pureswaran and Brody, 2015], blockchain is a suitable platform for facilitating transaction processing and coordination among interacting devices. Each managing its own role and behavior, resulting in an "Internet of Decentralized, Autonomous Things" – and thus the democratization of the digital world and cutting the cloud's disproportionate control over ubiquitous, autonomic computing. As a consequence, plausibly reigning back privacy; since IoTs are going to be the most ingrained computational sensors in our immediate surroundings.

8.6 Applications of the Trust Machine

Blockchains are clunky databases, so why would you want to use one? Traditional systems have inherent flaws that make them easy targets for corruption of data either by technical error or by human intention. When financial firms do business with each other, the hard work of synchronizing their internal ledgers can take several days, which ties up capital and increases risk. All sorts of companies and public bodies suffer from hard-to-maintain and often incompatible databases and the high transaction costs of getting them to talk to each other. Distributed ledgers that settle transactions in minutes or seconds could go a long way to solving such problems and fulfilling the greater promise of digitization and automation with trust and transparency.

A list of efforts to solve business and social use cases is enumerated at: http://dcg.co/portfolio/. These efforts give a taste of what will be possible. Table 8.2 gives a domain-wise list of applications where the "trust machine" has a promise to play revolutionary role.

8.6.1 Blockchain Applicability Test

Can't computers already execute transactions based upon pre-programmed conditions? Indeed they can; however, several intermediaries are often needed to verify and validate the details of the transaction. If the intermediaries fall under a single administrative domain, they derive the same source of trust/allegiance. If the intermediaries are from disparate administrative domains and are susceptible to external breach, influence, malice, laxity, etc. then blockchain brings all domains to a common immutable ledger. Andreas M. Antonopoulos, the author of book "Mastering Bitcoin" [Antonopoulos, 2014], has coined a simple test to identify whether a use case is really a blockchain use case or pure classical database use case. He states:

> If you replace the word Blockchain by Database and the implementation deliverables remain the same, then the implementation does not require a Blockchain.

Blockchain practitioners should use this test, further elaborated in Figure 8.11, as a guiding principle, while evaluating blockchain as a solution to the problem at hand, apart from the judicious consideration of other aspects like efficiency, integrity, non-repudiation, and potential for collusion in the proposed solution.

8.6.2 Challenges in Deploying the Blockchains

As blockchain applications have evolved from potential to actual use cases, we can see that particular use cases will raise specific governance questions best answered at the level of each use case (e.g. payments, contracts, securities clearance, insurance, etc.) There will not be a single blockchain but many, some of which may serve specific industries or geographies.

1. Interoperability: At the highest level, we need to focus on interoperability. Commercial blockchain applications are taking off, and governance will be critical to their success. For example, Ripple's global payments steering group, a blockchain bankers network with defined rules and governance, has been a major step forward in terms of adoption and industry acceptance. In case of organizations from different functional domains where their collaboration is ad-hoc, a token based approach to interoperability will help [Tapscott and Tapscott, 2017].

2. Privacy: Blockchains are open ledgers where all past transactions are recorded thus posing a dilemma of constructing transactions in either a transparent way or obfuscated way. In case of smart contracts closely resembling an organizations business process flow and logic, which at times is a trade secret, the issue of privacy becomes a serious challenge. Privacy and transparency run orthogonal to each other. Deriving trust while balancing privacy and transparency will be a challenge worth addressing.

3. Regulation: Drawing up regulations for blockchains at this early stage would be a mistake: the history of peer-to-peer technology suggests that it is likely to be several years before the technology's full potential becomes clear. In the meantime regulators should stay their hands, or find ways to accommodate new

Domain/Class	Examples
General	Escrow transactions, bonded contracts, third-party arbitration, multiparty signature transactions, messaging (Whisper), carbon credit, personal data ecosystem
Financial transactions	Remittance, trade settlement, stock, KYC/AML, private equity, crowdfunding, micro-lending, P2P lending, bonds, mutual funds, derivatives, prediction market, annuities, pensions, insurance
Businesses	transparent and efficient workflow composition, trade settlement, shareholder agreements, continuous compliance and audit, efficient & deterministic composition of services and business processes
Governance	Tendering, auctions, judiciary, regulation, agile taxation (GST), national digital currency [Economist, 2016b], accountability and transparency (RTI), platform for citizen engagement
Public services	Smart-grid metering, traffic congestion management, direct benefit transfer, dynamic pricing of services
Agriculture	livestock digitization for collateral, organic food provenance, supply chain formation, community-driven shared resources (equipments, warehouses), crop insurance, targeted subsidy disbursement, soil & crop management
Public records	Land and property titles [Economist, 2015b], vehicle registrations, business licenses, marriage certificates, death certificates
Semi-public records	Degree, vocational certifications, learning outcomes, grades, HR records (salary, performance reviews, accomplishment), healthcare (performance tracking of doctors)
Private records	IOUs, loans, contracts, bets, wills, trusts, escrows, tax returns, credit score, medical records
Identification	Driver's licenses, identity cards, passports, voter registrations, federated authentication platform (Aadhaar V2)
Attestation	Proof of insurance, proof of ownership, notarization
Physical asset keys	Home (Airbnb), hotel rooms, rental cars, automobile repair access
Intangible assets	Patents, trademarks, copyrights, reservations, domain names

Table 8.2: Blockchain (the Trust Machine) Applications

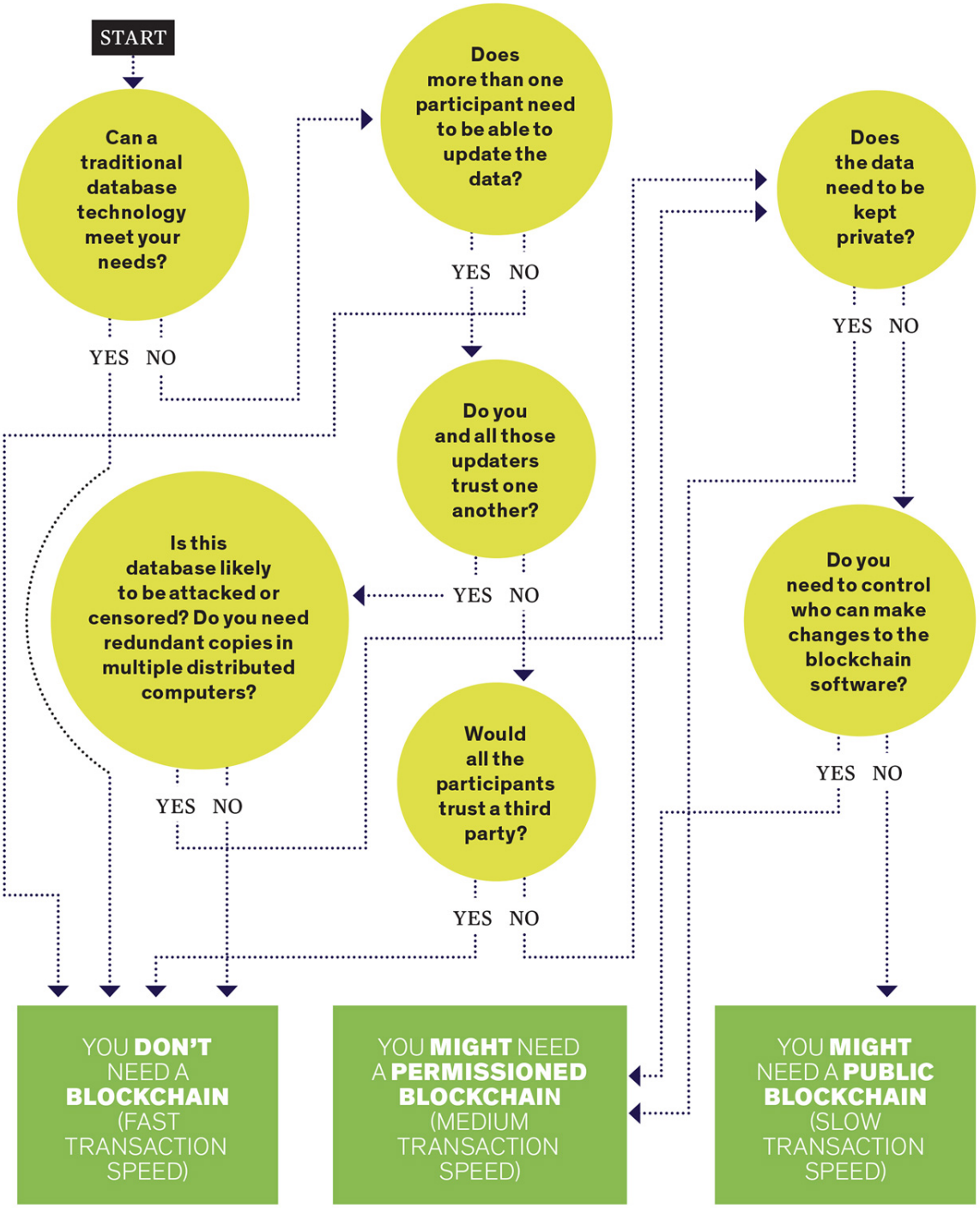


Figure 8.11: Blockchain Applicability Test [Peck, 2017]

approaches within existing frameworks, rather than risk stifling a fast-evolving idea with overly prescriptive rules [Economist, 2015c].

4. Testing: These technologies introduce a novel programming framework and execution environment, which are not satisfactory understood at the moment and have faced some major glitches in their nascent lifespan [Economist, 2016a, Atzei et al., 2017]. Multidisciplinary and multifactorial aspects affect correctness, safety, privacy, authentication, efficiency, sustainability, resilience and trust in smart contracts. Existing frameworks, which are competing for their market share, adopt different solutions to issues like the above ones. Merits of proposed solutions are still to be fully evaluated and compared by means of systematic scientific investigation, and further research is needed towards laying the foundations of Trusted Smart Contracts (http://fc17.ifca.ai/wtsc/).

5. Scalability: Industries also differ in their need for speed. For the bitcoin blockchain network, the process of clearing and settling transactions takes about 10 minutes, which is far faster end to end than most payment mechanisms today. But clearing transactions at the point of sale instantaneously is not the issue; the real problem is that 10 minutes is simply too long for the IoT where devices need to interact continuously. Former core developer Gavin Andresen said solving for a trillion connected objects is a different design space from bitcoin, a space where low latency is more critical and fraud is less of an issue or where parties could establish an acceptable level of trust without the bitcoin network [Tapscott and Tapscott, 2016].

6. Standardization: Like Internet, blockchain is being treated as a global resource. There are already efforts underway to steward this resource for standardization on the lines of what IETF/ICANN does for the Internet. Without standardization and stewardship invisible powers could emerge.

7. Digitization of Resources: Full potential of this technology cannot be reaped unless the resources around us can interact with the digital world. In many of the developing and under-developed countries, where this technology will have the highest impact, yet do not have governmental records in digital form. Without this availability of resources in digitized form it will be extremely difficult to realize the full potential of this technology.

8. User Interface: User interface will remain as Achilles heel given the fact that even a sophisticated user finds using crypto-wallets as a daunting task.

8.7 Blockchain in Indian Context

Investments in blockchain start-ups are similar in scale to that happened for dot-coms in the 1990s. While the invention was for creating a currency, there has been a widespread belief that the underlying trust protocol lends itself for reconfiguring our institutions and economy. Though there are certainly great challenges in creating such a future for which some of the emerged principles over this short period are: (i) networked integrity, (ii) distributed power (by consensus), (iii) value as incentive, (iv) security-by-design, (v) pseudonymity, (vi) preservation of rights, and (vii) democratic platform for inclusion with efficiency and transparency. Breakthroughs in these will lead to a great impact on building viable democratic societal applications, and a smart economy.

8.7.1 Sector-wise Potential

Some of the sectors that will have a positive impact of using such a technology are briefed below:

1. Policy: Management thinker Peter Drucker is often quoted as saying that "you cannot manage what you cannot measure." Drucker means that you can't know whether or not you are successful unless success is defined and tracked. What best can give a platform other than blockchain to define KPIs and triggers/conditions to track their progress in real-time?

2. Judiciary: A growing pool of empirical studies suggests that slow court systems discourage the growth of new businesses. With 2.8 crore of pending cases, blockchain's smart contract technology can be used to resolve the cases involving economic contract breaches, as a first step to experiment with. With the advances in AI (machine learning) and NLP technologies, effort can be made to resolve cases that have a clear precedent to rely on.

3. National Identity Platform: Identity is a critical part of a modern, advanced nation. Identity plays vital role in correct identification of individuals for various purposes: for economic, public service delivery, etc. Duplication of identities without holistic view gives rise to leakages as each department/institution maintains its own database – resulting in parallel expenses for same goal. Malfeasance to such databases

create situations where genuine individuals are excluded from being identified. Issues arising out of privacy violations generate resistance to evolution of a cohesive platform. Blockchain can provide a cohesive registry of identities and associated attributes that can be accessed by authorized entities under well-defined circumstances and contexts with appropriate authentication loop involving the subject being identified. Smart contracts can help improving Aadhaar framework into an intelligent, privacy-preserving national identity platform. Such a system will save cost of doing KYC for financial institutions and provide uniform view and control over data to end users.

4. Public Distribution System & DBT: Blockchain can reduce the number nodes through which a benefit/value traverses from issuer to receiver to zero. Thus the traditional intermediate nodes in value transfer to beneficiary will have the role of actuators only in which they have to just verify the validity of eligibility conditions for a beneficiary. Eligibility of a beneficiary can be evaluated in real-time instead of current periodic evaluation. Having a inter-connected national identity platform will greatly help in accurate evaluation of any beneficiary. Blockchain plays a role of a universal, all-knowing database to which any authorized entity can make a query.

5. Governance & Service Delivery: In a democratic country like India, health of the democracy is dependent on the active participation of its populace. Government spends huge amount of money on public welfare projects where the project executioner and the project auditor are exclusive of populace supposed to be the beneficiaries of or affected by the project. We can borrow the idea championed by ixo Foundation (http://ixo.foundation) to use blockchain for measuring impact of UN SDG (sustainable development goals) by making the populace as an auditor of the projects being implemented. Upon completion of execution of a project the affected people vote or provide feedback about the quality and degree of completion of the project. Thus making it difficult for the project executioner to influence or bribe the auditor.

6. Energy: With a huge potential of roof-top solar power generation, the national power grid will have to be equipped with an ability to dynamically adjust its transmission and distribution capacities. Reporting inaccurate data by error or malice can have cascading effect on the grid's stability. It will be of paramount importance to bring unified view across the grid for import/export of electricity. IoT-enabled controllers & meters with blockchain as an underlying data reporting, billing system will be a natural fit.

7. Agriculture: In a country like ours where a large population is engaged in agriculture, any gains in matching the produce with the best market will benefit the farmers. APEDA actively assists farmers to sell their produce in foreign markets by certifying the produce. Authenticity of these certificates and time taken to issue them is critical for perishable items. Integration of blockchain in supply chain consisting of certifiers like APEDA, cold-storage chains, port authorities, shipping lines will be a game changer. Another great benefit of blockchain in this sector would be a system that digitizes immovable assets and livestock of marginal farmers who have Jan Dhan Account but no credit profile thus excluded from formal financial services. Representation of livestock, ancestral property in shared custody of undivided joint family onto a blockchain will build their credit profile for NBFCs.

This technology has great potential to transform almost all sectors fundamentally. With a proper action plan and strategy, government can nurture and promote this technology by becoming its promoter and user.

8.7.2 Design & Deployment Considerations

While constructing a "trust machine" for a national (governamental) initiative a few subtle decisions need to be made in line with the spirit of Bitcoin highlighted below:

1. Permissioned vs Permissionless blockchain: It is going to be a great conundrum because by relevance it has to be a permissioned blockchain at global level, whereas it has to be a permissionless blockchain at national level. Identity-cum-authentication will help segregating the users interacting with the national level blockchain. Luckily, India has a national level identification mechanism for its citizens and businesses. It will be an interesting proposition to build such a blockchain also because businesses operating out of India will have their workflow spanned across the world. How do we provide the interoperability will be an important design criteria.

2. Evoking the Trust: Being a national blockchain, either backed by the Indian government or by a consortia of public-private partnership, the obvious fact will be the ownership of the setup. Blockchain is a P2P system in its original form with no entry or exit barrier for the nodes and no ownership of the whole. Whereas, having a owner of the permissioned setup does not bode well for evoking public trust into the system. Pragmatic approach like setting up an independent statutory body similar to Election Commission of India will assuage the concern.

3. Choice of the consensus algorithm: PoW is the only proven practical consensus algorithm that scales for a large number of nodes, as seen in case of Bitcoin. PoS of Ethereum is bootstrapped from PoW in the beginning to denote generated currency units as "Stake" or Ether. Choosing PoW type of consensus algorithm has to be extremely careful while making the choice of one way hash function to perform actual PoW construction. Bitcoin miners have reached to such gigantic levels of hashing power that the biggest miner on Bitcoin can easily overwhelm combined power of all supercomputers in the world put together. A different hash function has to be chosen while keeping in mind the sophistication of existing Bitcoin miners for SHA1 family of hash functions. PoS without PoW for bootstrapping could be a good option since the national blockchain will have option of using Aadhaar-unique-IDs to offer a pre-determined stake to each individual a priori.

8.8 Takeaways

Bitcoin is the first application of a technology that paves the way forward, revealing an opportunity for innovation that was not apparent before. Bitcoin is wholly open source (in important trust evoking decision), so every element of it can be tweaked, modified, altered and tested for potentially improved iterations, just like evolution.

1. Blockchain is an idea of making trust a matter of coding, rather than of democratic politics, legitimacy and accountability. The blockchain lets people, who have no particular confidence in each other, collaborate without having to go through a neutral central authority. Simply put, it is a machine for creating trust. In essence it is a shared, trusted, public ledger that everyone can inspect, but which no single user controls.

2. Ledgers that no longer need to be maintained by a company or a government may in time spur new changes in how companies and governments work, in what is expected of them and in what can be done without them.

3. A realization that systems without centralized record-keeping can be just as trustworthy as those that have them may bring radical change.

4. People and institutions today can solve hard problems and change the world for the better when they have a reliable framework to build on.

5. Systems that are honest free up dead capital. The transparency provided by blockchain can help eliminate forgery and provide efficient service delivery.

6. Blockchain is an important technology of Internet Era and has global appeal. Any nation embracing this technology (e.g., Estonia, Singapore, Japan) will have a competitive advantage over the laggards. Industry (through innovation) as well as government (through calculated policy oversight, being promoter of common standards for interoperability) have a responsibility to invest in this potentially revolutionary technology for trust management in our digital economy.

One reason why this technology works is that it has socially engineered the game mechanics based on one assumption, that there are more good people than bad people. This is the underlying hope on which blockchain resides. – Pindar Wong, VeriFi.

Bibliography

[Ali et al., 2016] Ali, M., Nelson, J., Shea, R., and Freedman, M. J. (2016). Blockstack: A global naming and storage system secured by blockchains. In *Proceedings of the 2016 USENIX Conference on Usenix Annual Technical Conference*, USENIX ATC '16, pages 181–194. USENIX Association.

[Antonopoulos, 2014] Antonopoulos, A. M. (2014). *Mastering Bitcoin: Unlocking Digital Crypto-Currencies*. O'Reilly Media, Inc., 1^{st} edition.

[Atzei et al., 2017] Atzei, N., Bartoletti, M., and Cimoli, T. (2017). A survey of attacks on ethereum smart contracts sok. In *Proceedings of the 6th International Conference on Principles of Security and Trust - Volume 10204*, pages 164–186, New York, NY, USA. Springer-Verlag New York, Inc.

[Azaria et al., 2016] Azaria, A., Ekblaw, A., Vieira, T., and Lippman, A. (2016). Medrec: Using blockchain for medical data access and permission management. In *2016 2nd International Conference on Open and Big Data (OBD)*, pages 25–30.

[Backs, 2002] Backs, A. (August 2002). Hashcash - A Denial of Service Counter-Measure. `http://www.hashcash.org/papers/hashcash.pdf`. Technical Report.

[Baird, 2016] Baird, L. (2016). The swirlds hashgraph consensus algorithm: Fair, fast, byzantine fault tolerance. `http://www.swirlds.com/downloads/SWIRLDS-TR-2016-01.pdf`. SWIRLDS Tech Report.

[Buterin, 2013] Buterin, V. (2013). Ethereum: A next-generation smart contract and decentralized application platform. `https://github.com/ethereum/wiki/wiki/White-Paper`.

[Castro and Liskov, 1999] Castro, M. and Liskov, B. (1999). Practical byzantine fault tolerance. In *Proceedings of the Third Symposium on Operating Systems Design and Implementation*, OSDI '99, pages 173–186. USENIX Association.

[Chaum et al., 1988] Chaum, D., Fiat, A., and Naor, M. (1988). Untraceable electronic cash. In *Advances in Cryptology - CRYPTO '88, 8th Annual International Cryptology Conference, Santa Barbara, California, USA, August 21-25, 1988, Proceedings*, pages 319–327.

[Dwork and Naor, 1993] Dwork, C. and Naor, M. (1993). *Pricing via Processing or Combatting Junk Mail*, pages 139–147. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Economist, 2016a] Economist, T. (Jul 2016a). Not-so-clever contracts. `online`. The Economist.

[Economist, 2017] Economist, T. (Jul 2017). Disrupting the trust business. `online`. The Economist.

[Economist, 2016b] Economist, T. (Mar 2016b). Redistributed ledger. `online`. The Economist.

[Economist, 2015a] Economist, T. (Oct 2015a). The great chain of being sure about things. `online`. The Economist.

[Economist, 2015b] Economist, T. (Oct 2015b). The great chain of being sure about things. `online`. The Economist.

[Economist, 2015c] Economist, T. (Oct 2015c). The trust machine. `online`. The Economist.

[Force, 2016] Force, A. C. T. (April 2016). Financial regulations for improving financial inclusion. `https://www.cgdev.org/sites/default/files/financial-access-task-force-brief_0.pdf`.

[Lamport et al., 1982] Lamport, L., Shostak, R., and Pease, M. (1982). The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401.

[Micali, 2016] Micali, S. (2016). ALGORAND: the efficient and democratic ledger. *CoRR*, abs/1607.01341.

[Nakamoto, 2008] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. `http://bitcoin.org/bitcoin.pdf`.

[of India, 2015] of India, R. B. (2015). Financial inclusion in india – an assessment. `https://rbidocs.rbi.org.in/rdocs/Speeches/PDFs/MFI101213FS.pdf`.

[Ongaro and Ousterhout, 2014] Ongaro, D. and Ousterhout, J. (2014). In search of an understandable consensus algorithm. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference*, USENIX ATC'14, pages 305–320. USENIX Association.

[Patil and Shyamasundar, 2017] Patil, V. T. and Shyamasundar, R. K. (2017). Privacy as a currency: Unregulated? In *Proceedings of the 14th International Conference on Security and Cryptography: SECRYPT*, pages 586–595. INSTICC, SciTePress.

[Peck, 2017] Peck, M. E. (2017). Blockchain world - do you need a blockchain? this chart will tell you if the technology can solve your problem. *IEEE Spectrum*, 54(10):38–60.

[Pitroda and Desai, 2010] Pitroda, S. and Desai, M. (2010). *The March of Mobile Money: The Future of Lifestyle Management*. HarperCollins Publisher, Collins Business.

[Pureswaran and Brody, 2015] Pureswaran, V. and Brody, P. (2015). Device democracy: Saving the future of the internet of things. `http://www-935.ibm.com/services/multimedia/GBE03620USEN.pdf`. IBM Institute for Business Value.

[Snow et al., 2014] Snow, P., Deery, B., Lu, J., Johnston, D., and Kirby, P. (Nov 2014). Business processes secured by immutable audit trails on the blockchain. `online`. Factom White Paper.

[Swan, 2015] Swan, M. (2015). *Blockchain: Blueprint for a New Economy*. O'Reilly Media, Inc., 1^{st} edition.

[Tapscott and Tapscott, 2016] Tapscott, D. and Tapscott, A. (2016). *Blockchain Revolution: How the Technology Behind Bitcoin Is Changing Money, Business, and the World*. Portfolio Penguin.

[Tapscott and Tapscott, 2017] Tapscott, D. and Tapscott, A. (June 2017). Realizing the potential of blockchain: A multistakeholder approach to the stewardship of blockchain and cryptocurrencies. `online`. World Economic Forum.

[Vigna and Casey, 2016] Vigna, P. and Casey, M. J. (2016). *The Age of Cryptocurrency: How Bitcoin and the Blockchain Are Challenging the Global Economic Order*. St. Martin's Press.

[Vukolić, 2016] Vukolić, M. (2016). *The Quest for Scalable Blockchain Fabric: Proof-of-Work vs. BFT Replication*, pages 112–125. Springer International Publishing, Cham.

[Weber et al., 2016] Weber, I., Xu, X., Riveret, R., Governatori, G., Ponomarev, A., and Mendling, J. (2016). *Untrusted Business Process Monitoring and Execution Using Blockchain*, pages 329–347. Springer International Publishing, Cham.

[Wood, 2014] Wood, G. (2014). Ethereum: a secure decentralised generalised transaction ledger. `http://gavwood.com/paper.pdf`.

Chapter 9

Computational Thinking

RK Shyamasundar
IIT Bombay

Computer Science, Engineering and Technology have not only made an amazing progress in the past few decades but also have made a huge impact on science, and society. As early as 1987, Nobel Laureate Ken Wilson said, "Computation has become the third leg of science". From the time Computer Science was regarded as a provider for tool support, it has reached a point of being declared by the scientific community that a time has come when it has become very essential to integrate Computer Science concepts, tools and theorems into the very fabric of science for scalable discoveries. Even though on the face of it, the change may seem subtle, it should be seen as fundamental to science and the way science is practiced. This reflects the foundations of a new revolution in science. Such an evolution has been nicely captured through the phrase "Computational Thinking" [Wing, 2006] by Jeannette Wing of CMU to refer to a universally applicable attitude and skill set everyone, not just computer scientists, would be eager to learn and use. Computational thinking refers to approaches to problem solving methodologies, design of systems as well as an understanding of human behavior drawing concepts from Computer Science.

One of the basic concepts of Computer Science in understanding and conquering complex phenomena is the concept of "Abstraction". The distinguished computer scientist CAR Hoare says:

> *In the development of the understanding of complex phenomena, the most powerful tool available to the human intellect is Abstraction.*

A gentle introduction to this notion is discussed below:

Suppose we were asked the question whether 376 * 485 * 253 * 252 * 783 * 938 is an even number. One way of finding out is to start multiplying all the six numbers and see whether the number is even. Another way of looking at is to use abstraction whether the number is even or odd, and compare the two results for revalidation. Such an abstraction is depicted in Figure 9.1 shown below: In this Figure, the top layer denotes a

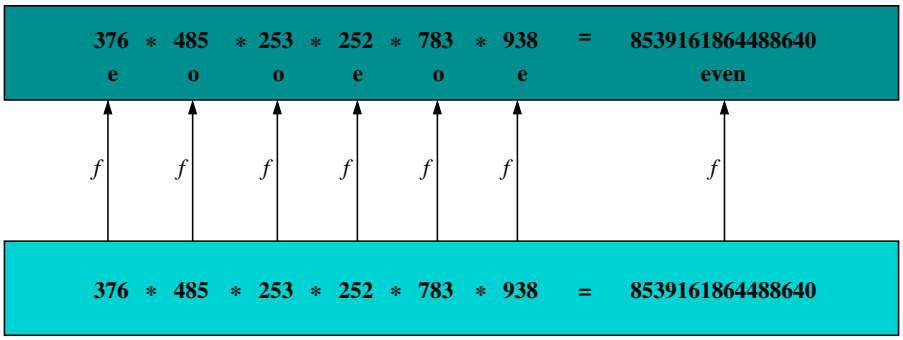


Figure 9.1: Example of applying abstraction to decide parity of a number

concrete abstraction of the problem and the bottom one is the concrete result or implementation. It so happens, the above abstraction was good enough to realize correctly what one was looking for without fully calculating the expression. This is due to the fact that we have the rule: $even * odd = even$. However, quite often the abstraction used by the user may not suffice even though the mapping may be correct; in other words, it can be interpreted to mean that the mapping was not precise enough to capture the intended property. Further, the process need not be a one step process; there could be multiple layers one refining the other. In general, we

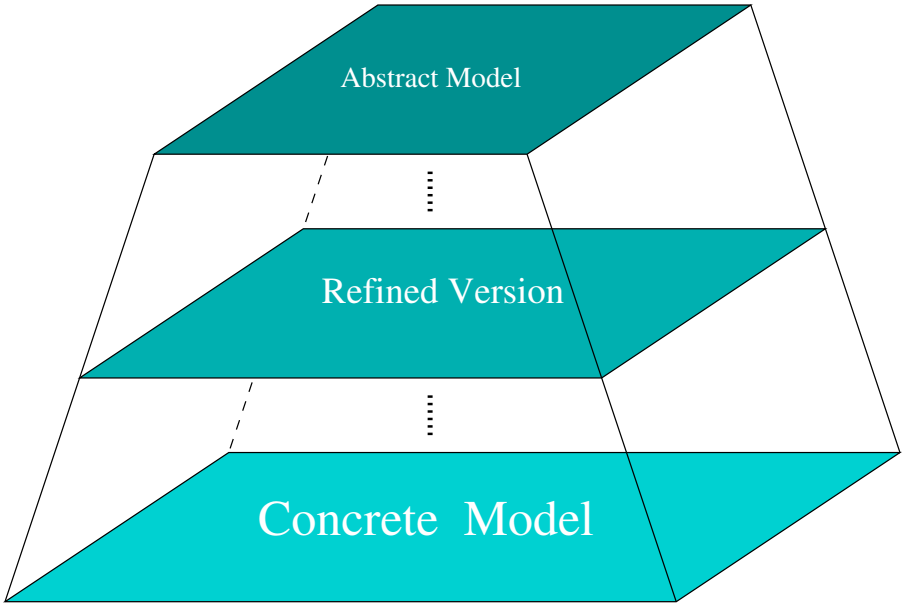


Figure 9.2: Layers of refinement

could have several layers like the one shown in Figure 9.2: The notion of "refinement" is one of the well-studied aspects in Computer Science and the formal setting of Abstract Interpretation provides a clean model for the same. In fact, it provides a basis of correctness of implementation from specifications. For instance, a common example used in Computer Science is that of abstract data types used widely in programming wherein the operations of *push* and *pop* are defined on a given stack as an artifact in a programming language. Using a formal abstraction, one can realize the artifact of stacks and establish the correctness of the operations with a possibility of not even interpreting the data and thus get artifacts like stack of integers, stack of reals, etc. We shall not go into further formal aspects of these. We just quote from another distinguished computer scientist, Edsger W. Dijkstra:

> *The purpose of abstraction is not to be vague but to create a new semantic level in which one can be absolutely precise.*

In short, abstractions as used in Computer Science are more general than that used in Mathematics and Physical Sciences as they need not have the elegant properties of the structures known. In a sense, the abstractions in Computer Science denote architecture for solving the problem leading possibly to several layers, some of which could be mechanizable. The beauty of the abstraction process is that it introduces layers and at any point, one need to look usually at two layers – the layer of interest and the layer below or above and thus, making the reasoning simple confining to the requisite information and the implementation well structured. Thus, a complex task could be mastered with good abstractions. Well-defined interfaces between layers enable us to build large complex systems. Thus, the process of abstraction enables the user to ignore the details one is not concerned with and enables the implementer to build correct systems or correct applications.

Another point to be noted is that the layers of abstraction highlighted could be viewed as an object, an artifact or even a specification. For instance, one could consider axiomatic specification and denotational semantics of programming languages as two layers. Now considering these two as the layers, one could either interpret the complementarity of the definitions or the consistency of the definitions of the underlying programming language. This generality has given the power underlying *multi-scalar approach* – a paradigm used widely in science and engineering simulations. We shall discuss these aspects later.

The crux of computational thinking lies in defining well-defined structured abstractions and a succinct elucidation of relationships between various layers. In other words, abstractions are nothing but the mental tools of computing. Thus, a proper formal modeling of the abstraction leads to a computable abstraction. Perhaps, the computing engine could be a real computer or even a human being. In case we can combine the two i.e., a computer and a human, we will really get a capability that cannot be handled by any one of these entities and thus, overcome the limits of either. It is to be noted that *Computational Thinking* is not necessarily tied to a computing engines and thus, explores architectures to solve problems and when one needs mechanization it is possible to resort to different mechanizable engines including a computer.

As Jeannette Wing puts it:

> *Computing is the automation of our abstractions.*

CT in Sciences, Math, and Engineering	Computational Paradigm
Biology	∗ Shotgun algorithm expedites sequencing of human genome ∗ DNA sequences are strings in a language ∗ Protein structures can be modeled as knots ∗ Protein kinetics can be modeled as computational processes ∗ Cells as a self-regulatory system are like electronic circuits
Brain Science	∗ Modeling the brain as a computer ∗ Vision as a feedback loop ∗ Analyzing fMRI data with machine learning
Chemistry	∗ Atomistic calculations are used to explore chemical phenomena ∗ Optimization and searching algorithms identify best chemicals for improving reaction conditions to improve yields
Geology	∗ Modeling the earth's surface to the Sun, from the inner core to the surface ∗ Abstraction boundaries and hierarchies of complexity model the earth and our atmosphere
Astronomy	∗ Sloan Digital Sky Server brings a telescope to every child ∗ KD-trees help astronomers analyze very large multi-dimensional datasets
Mathematics	∗ Discovering E8 Lie Group: 18 mathematicians, 4 years and 77 hours of supercomputer time (200 billion numbers) ∗ Profound implications for physics (string theory) ∗ Four-color theorem proof
Engineering	∗ Calculating higher order terms implies more precision, which implies reducing weight, waste, costs in fabrication ∗ Boeing 777 tested via computer simulation alone, not in a wind tunnel
Social Sciences (Society)	∗ Social networks explain phenomena such as MySpace, YouTube ∗ Statistical machine learning is used for recommendation and reputation services, e.g., Netflix, affinity card
Medicine (Society)	∗ Robotic surgery ∗ Electronic health records require privacy technologies ∗ Scientific visualization enables virtual colonoscopy
Arts, Films, Games (Society)	∗ Arts (e.g., Robotticelli) ∗ Movies: − Dreamworks uses HP data center to render Shrek and Madagascar − Lucas Films uses 2000-node data center to produce Pirates of the Caribbean
Sports (Society)	∗ Synergy Sports analyzes digital videos NBA games

Table 9.1: *Computational Thinking* (CT) paradigms in various areas

This gives the power and ability to scale. Thus, computational thinking is nothing but:

- Choosing the right abstractions, etc.
- Choosing the right *computer* for the task.

Some of the *Computation Thinking* paradigms as captured in various areas of science, engineering, mathematics and societal applications are given in Table 9.1 [Wing, 2008]. Scaling up of education is being attempted through MOOC using the ICT.

Skill Development During High School Education

Computational thinking has been suggested as an analytical thinking skill that draws on concepts from computer science but is a fundamental skill used by, and useful for, all people.

These powerful ideas and processes have begun to have significant influence in multiple fields, including biology, journalism, finance, and archaeology, making it important to include computational thinking as a priority for K–12 education in North America. The National Research Council (NRC) highlighted the importance of exposing students to computational thinking notions early in their school years and helping them to understand when and how to apply these essential skills.

Barr and Stephenson argued that, given that students will go into a workforce heavily influenced by computing, it is important for them to begin to work with computational thinking ideas and tools in grades K–12 Specifically, they discussed the need to highlight "algorithmic problem solving practices and applications of computing across disciplines, and help integrate the application of computational methods and tools across diverse areas of learning."

Recent educational reform movements (such as the Next Generation Science Standards and the Common Core) have also focused on computational thinking as a key skill for K–12 students. For example, the Next Generation Science Standards (NGSS) have identified computational thinking as key scientific and engineering practices that must be understood and applied in learning about the sciences. Computational theories, information technologies, and algorithms played a key role in science and engineering in the 20^{th} century; hence, NGSS suggested allowing students to explore datasets using computational and mathematical tools.

From: Aman Yadav, Chris Stephenson and Hai Hong, Computational Thinking for Teacher Education, Comm., ACM, April 2017, Vol. 60, No. 4, pp. 55-62.

To summarize: Computational thinking is revolutionizing a wide spectrum of areas like computational statistics, computational learning, computational chemistry, computational fluid dynamics, computational biology, computational genomics, computational micro-economics, humanities, etc. In a sense, it has created a remarkable intellectual revolution around us. In fact, Computational Thinking will lead to scaling up of education through the effective use of ICT. The nice point to note is that the advancement of computing engines has enabled us to ask different kinds of questions and get different kinds of answers through a large data collection and navigation – often called Big Data. The depth and breadth of Computational Thinking can be see in the remarks of Moshe Vardi who says:

> *Computational Thinking thoroughly pervades both legs of science – theory and experiment. It is the universal enabler of science, supporting both theory and experimentation.*

There have been several discussions on Computational Thinking to highlight the breadth of its impact on every kind of thought [Marcia Linn, 2011]. However, the main observation is that the same key themes keep emerging from each discipline: the ability that computation provides to investigate new kinds of questions; the infiltration of computational concepts into other disciplines' theories; and that computation's influence is seldom what you might initially predict, but is often both more subtle and more profound. The drivers of computing can be captured in Figure 9.3.
To push forward such a revolution, the challenge is to:

> *Recast science education through integrating computational thinking with the sciences – leading to science based innovation that would have impact on society, economics as well as science discoveries.*

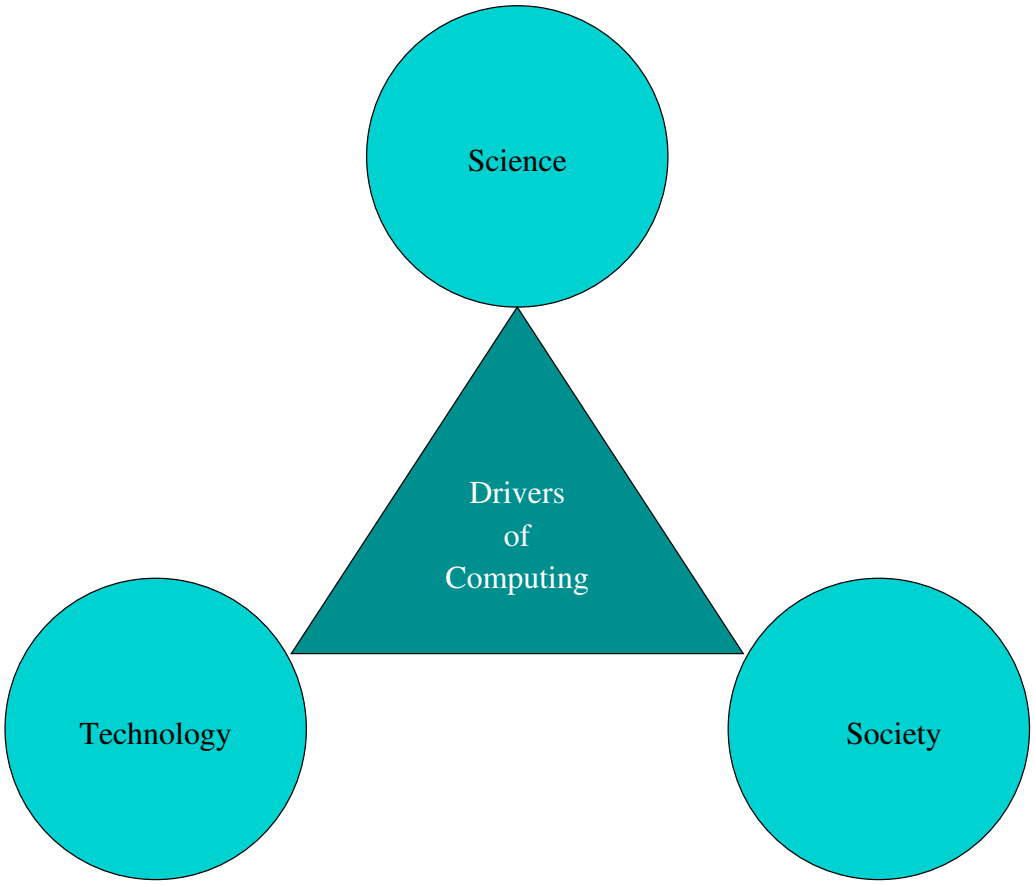


Figure 9.3: Drivers of Computing

Bibliography

[Hopcroft et al., 2011] Hopcroft, J. E., Soundarajan, S., and Wang, L. (2011). The Future of Computer Science. *Int. J. Software and Informatics*, 5(4):549–565.

[Marcia Linn, 2011] Marcia Linn, e. a. (2011). Report of a Workshop on the Pedagogical Aspects of Computational Thinking. `http://people.cs.vt.edu/~kafura/CS6604/Papers/NRC-Pegagogy-CT.pdf`.

[Wing, 2006] Wing, J. M. (2006). Computational Thinking. *Commun. ACM*, 49(3):33–35.

[Wing, 2008] Wing, J. M. (2008). Computational Thinking and Thinking About Computing. Carnegie Mellon University.

Chapter 10

Future of Computing Science

RK Shyamasundar
IIT Bombay

If people do not believe that mathematics is simple,
It is only because they do not realize how complicated life is
John von Neumann, 1947.

Computing has made a huge impact on science, society and even national security. We have reached a point, wherein significant progress either in science or society is dependent on the computing power. To mention a few of the areas where *computing* has made a huge impact is: smart materials, understanding structures, earthquake engineering, epidemiology, genomics, molecular modeling, chemistry astronomy, biology, e-commerce, e-governance, health-care, disaster management, national security, and public infrastructures.

Computer science is undergoing a fundamental change and is reshaping our understanding of the world. An important aspect of this change is the theory and applications dealing with the gathering and analyzing of large real-world data sets. In this section, we introduce four research projects in which processing and interpreting large data sets is a central focus. Innovative ways of analyzing such data sets allow us to extract useful information that we would never have obtained from small or synthetic data sets, thus providing us with new insights into the real world.

Modern computer science is undergoing a fundamental change. In the early years of the field, computer scientists were primarily concerned with the size, efficiency and reliability of computers. They attempted to increase the computational speed as well as reduce the physical size of computers, to make them more practical and useful. The research mainly dealt with hardware, programming languages, compilers, operating systems and databases. Meanwhile, theoretical computer science developed an underlying mathematical foundation to support this research, which in turn, led to the creation of automata theory, formal languages, computability, and algorithm analysis. Through the efforts of these researchers, computers have shrunk from the size of a room to that of a dime, nearly every modern household has access to the Internet and communications across the globe are virtually instantaneous.

Computers can be found everywhere, from satellites hundreds of miles above us to pacemakers inside beating human hearts. The prevalence of computers, together with communication devices and data storage devices, has made vast quantities of data accessible. This data incorporates important information that reveals a closer approximation of the real world and is fundamentally different from what can be extracted from individual entities. Rather than analyzing and interpreting individual messages, we are more interested in understanding the complete set of information from a collective perspective. However, these large-scale data sets are usually far greater than the data sets that can be processed by traditional means. Thus, future computer science research and applications will be less concerned with how to make computers work and more focused on the processing and analysis of such large amounts of data.

Various strategies and specific takeaways have been discussed in various chapters in articulating a Roadmap of investment on ICT.

Looking at the impact computing has been making along with possible disruptions in science, technology and society, the following simple takeaways as captured in the book from Microsoft entitled "The Fourth Paradigm: Data Intensive Science Discovery" provide simple advice that need to be adhered by one and all to carry the benefits of computing to mankind:

1. If you are a scientist, talk to a computer scientist about your challenges, and vice versa.

2. If you are a student, take classes in both science and computer science.

3. If you are a teacher, mentor, or parent, encourage those in your care toward interdisciplinary study in addition to giving them the option to specialize.

Appendix

In this Appendix, we shall briefly provide the recommendations made by several experts at a conference held at TIFR in commemoration of Homi Bhabha. This has been included in this book since the recommendations still look relevant today.

Chapter 11

CSDI Panel Recommendations: Computing for Science Discovery and Innovations – A Roadmap

RK Shyamasundar and MA Pai
IIT Bombay and UIUC

Background

Many in the media and general public attribute the IT revolution in India to the process of liberalization in the 1990's. Those familiar with the Science and Technology (S&T) scene in India after independence, however, have a different take on this issue, as well as on other issues such as the green, white, and the telecom revolutions that took place in the first three to four decades after independence. These contributed significantly to the success of the reforms, which took place in the 90's. The IT revolution began in the late 50's in India when Homi Bhabha saw the need to develop know-how about computers and its scientific usage in the country. As a tribute to this visionary on his birth-centenary year, a book titled *"Homi Bhabha and the Computer Revolution"*, edited by RK Shyamasundar (TIFR) and MA Pai (UIUC), was released by the Honorable Chief Minister of Maharashtra, Shri Prithviraj Chavan on February 18, 2011 at the Homi Bhabha Auditorium, TIFR, Mumbai.

Continuing the tradition initiated by Homi Bhabha, TIFR organized the above conference on February 18-19, 2011 at TIFR, Mumbai, to carve out a road-map of research and development in Computer Science, Communication and IT for science and societal applications. The conference was organized by Dr. Sam Pitroda (Adviser to PM, Public Information Infrastructure and Innovations), Prof. RK Shyamasundar (TIFR) and Prof. MA Pai (UIUC). The conference had very distinguished invited talks, panels on R&D activities in ICT and the required eco-systems for fast growth. The conference had a distinguished participation from leading universities, R&D institutions, industry, and the Government. A summary of the key participants is given as an annexure. This report contains the broad recommendations of the Conference, recommendations of the individual panels and white papers/discussions deliberated during the panels.

We do hope the recommendations and summary of the deliberations of the conference will be of help to S&T policy makers, academic researchers, educationists, industrial researchers, etc., and will help in carving out a road-map to realize the potential of Computing and ICT for science discovery, and growth of the society.

RK Shyamasundar, *TIFR*
MA Pai, *UIUC*
August 2011.

Executive Summary

Overall Recommendations of the Panels

Panel Co-ordinators: RK Shyamasundar (*TIFR*), A Paulraj (*Stanford*), N Viswanadham (*ISB*), Subhasis Chaudhuri (*IIT Bombay*), V Rajaraman (*IISc*), and MA Pai (*UIUC*).

Panel 1. **Building a Strong R&D Ecosystem**

- Frontier ICT areas such as (i) design of algorithms for multi-core processors, (ii) building large scale information systems including cyber physical systems, (iii) science and technology of cyber security and privacy, secure trustworthy systems, (iv) learning in large data sets and social networks, (v) image understanding, computational biology and cognitive computing and (vi) applications to service engineering and (vii) ICT for public infrastructures like transport, power, etc. Build and nurture groups in sizes reasonably above respective threshold sizes.
  - Build a strong graduate programme that will lead to a sizable pool of quality Ph.D.s
- Generous R&D grants (in 100s of Crores) to help companies that show world class product traction. This will enable them to compete on a level ground in international markets against international competition.
- Initiate visionary projects across lead institutions and researchers and industries (with plausible international participation) in lead frontier areas that will lead to building systems that are expected to have a huge impact on science discovery, economics and societal benefits keeping in view:
  - An expert (possibly international experts as well) team need to continuously monitor and advice to realize projected goals (both for open-ended projects and goal-oriented projects).
  - Evaluation structure of ESPRIT projects could be adapted and also the projects need to be bounded in time even if the vision attains the level of a center of excellence (model on the lines of fifth generation project is one such example).
  - Technology transfer should be handled afresh with its own evaluation and structure.
- Establish a Strategic Review Center for Impact of ICT on Science and Society. It should use the expertise in the country to arrive at white papers of relevance.

Panel 2. **Industry-Academia Centers for Excellence:** Under a public-private partnership model establish centers of excellence in institutes/universities of excellence with R&D targets for futuristic demands (take into account the regional needs as well).

Panel 3. **Innovations:** Build a strong Intellectual Property (IP) pipeline and ecosystem by;

- encouraging inventors and supporting transfer of technology through proactive evaluations of IP, and
- soliciting Requests for Innovations and attracting talent

Panel 4. **Attracting Venture Capital:** Create incentives through preferential market access for companies (Indian and perhaps even foreign) that develop world class technology and do high end value addition within the country. Such a policy will attract the much-needed venture capital to fund innovations of local companies.

Panel 5. **Education:**

- Building a strong Graduate Programme: quality faculty is the crux
- Training Faculty Centers: rigorous training to teachers (teaching + evaluation); can be extended to training school-teachers teaching ICT
  - Lead to quality teachers for UG, PG and Graduate education (cannot be done through symbolic qualification prescriptions);
  - Strict evaluation of teachers, Ph.D. guides, and professors at a national level.
  - Scalable e-learning for enhancing the deliverability (teaching content + Lab)
- Introduce new Certifications, Masters Courses based on the need (regional and national).
  - Example: ICT Certifications consisting applications of ICT in infrastructures like power, transport, etc.

Panel 6: **ICT in Power Infrastructure:**

- Transmission and Distribution
- Education and Training keeping in view smart-grids, smart-cities etc.

Panel 1

Computer Science Research: Basic Research, Embedded Systems, and Building Scalable Systems

---

Panel: P Anandan (*Microsoft Research*), Manish Gupta (*IBM Research*), Ravishankar Iyer (*UIUC*), Heiko Mantel (*Darmstadt*), SV Raghavan (*GoI*), S Ramesh (*GM R&D*), RK Shyamasundar (*TIFR*)

(a) Build a strong R&D ecosystems in Computer Science that caters to frontier areas such as: (i) Design of algorithms for multi-core processors, (ii) Large-scale information systems and cyber physical systems (iii) Cyber security and privacy, secure trustworthy systems, (iv) Learning in large data sets and social networks, (v) Image understanding, Computational biology, and Cognitive computing and (vi) applications to service engineering, and (vii) ICT for public infrastructures like transport, power, etc. It is important to nurture groups that have a size above a threshold.

(b) Encourage growing the theory to serve information technology drivers for the next few decades. This should be the basis for building a strong graduate programme that encompasses Computer Science, Communications and multiple other disciplines ranging from engineering to design. This will lead to building a strong human resource both in quality and quantity. This will create the much-needed Ph.D. pool for India

- Sub-standard or mediocre Ph.D.s will cause havoc. Scientific evaluation methodologies should be created to assess the capabilities of guides to Masters and Ph.D. programmes.
- The effort will lead to a sound curriculum that can be adapted from various perspectives and build an enviable teachers in ICT.

(c) There is a need to build and strengthen the System Building area that will impact Science Discovery, Innovations and Society. This will require a sustained encouragement and support to building scalable computing systems that can leap into Exascale computing capabilities that will be the backbone for Science Discovery, Innovations, and societal applications like cyber-physical systems, e-governance systems.

- Initiate large visionary projects in Theory and Applications across institutions and industries in India with well-defined goals that need to be closely monitored and assessed by real-experts in respective areas (could as well include international experts). An expert continuance evaluation-cum-guidance is needed for good results.
  - One can keep in mind the evaluation structure of ESPRIT projects in Europe and also the bounded time for visionary projects – it is possible that failures may lead to better insights. But it is important to close down the projects in definite time framework. The engineering of a successful system should be handled with its own evaluation and structure.
- Note that, a few of the above open ended projects while a large number should be goal oriented.
- Several areas mentioned in (a) need scalability of systems to realize newer heights whether for science or for information systems.

Panel 2

Telecom and Networking Panel: Telecom and Networking Equipment Industry in India – Rising to World Class

Panel: Dr. Girija Narlikar (*AT&T Bell Labs, Mumbai*), Prof. A Paulraj (*Stanford*), Prof. Bhaskar Ramamurthy (*IIT Chennai*) , Dr. Kumar N Sivarajan (*Tejas Networks*), Dr. Rahul Vaze (*TIFR, Mumbai*)

Need for an Indian Telecom and Networking Equipment Industry

India is world's fastest growing telecom market with a subscriber base of 750 million. However, all of the telecom and networking equipment (Rs. 135,000 Crores in 2009-10) is imported with minimal, if any, value addition in India. This is inevitable since India lacks a competitive telecom and networking equipment (TNE) industry.

There are many compelling reasons for building an Indian TNE industry. The first group is economic. A world-class TNE industry in India can increase internal value addition significantly, and help achieve a better balance of trade in this high-tech sector. This will also create Lakhs of high-end jobs and attract large foreign direct investments. Further, such an industry can serve Indian military communication needs, which also relies on massive imports today. Additionally, improving national productivity requires penetration of custom telecom, networking, and computing technologies into many vertical industries, and this is best accelerated if we have a local TNE capability. Finally, half of the GDP growth in developed countries comes from high technology industry, and India too will have to find sustained growth from such segments soon.

A second group of motivations comes from national security. Since telecom infrastructure underpins so much of the nation's critical infrastructure, its disruption can have severe economic impact. An indigenous TNE capability can significantly reduce the vulnerability associated with imports. Likewise, local TNE can support India's defense and government sector where there are inevitable security concerns with imported equipment.

Therefore, it is imperative for India to develop a world-class TNE industry which can be both a major force in the huge domestic market and a significant global player.

Developing a TNE Industry

The value addition in TNE industry is captured in two layers. First – Semiconductor design (known as fabless semiconductor) and related semiconductor fabrication, and Second – System design, integration, and related manufacturing/assembly. Semiconductor fabrication contributes about 30% of the total equipment costs and this share is growing. The creation of semiconductor fabrication facility is a major investment and needs to be discussed separately and is outside the scope of the panel. Some assembly of phones and equipment is being done in India, but this adds only about 5-7% of the total TNE value addition. There still remains about 60-65% of value addition in fabless semiconductor and system integration/design. The system companies directly service the needs of the telecom operators using the products/IP developed by semiconductor companies. Therefore, our immediate focus must be on developing fabless semiconductor and system companies.

India has some of the important assets necessary to grow fabless semiconductor and system companies. First, almost all the R&D skills required are available in abundance today thanks to the on-the-job learning in Indian Eng. Service (IT) sector, and R&D subsidiaries of MNCs. In fact, India has become a major global outsourcing destination for algorithms development, fabless semiconductor design and advanced software. These are the very same skills needed to develop leading edge TNE products. e.g., Beceem, now part of Broadcom, the worldwide leader for 4G wireless semiconductor did most of its R&D in India. Complex system integration skills are also available as demonstrated by Tejas Networks. India's universities, including IISc and IITs, produce a constant supply of high-quality students that can feed the TNE industry. In addition to R&D capability, India has a huge and dynamic telecom services market. This is fertile ground for growing a high volume industry.

However, other elements that are important to create a world-class TNE industry are absent or weak.

India lacks venture capital required to nurture TNE technology companies which are necessarily very high risk. Venture capital can be attracted to invest in Indian companies but will need significant incentives in terms of market access (see below) to reduce perceived risk. Notably, both Beceem and Tejas (mentioned earlier), were backed by US venture capital. A single TNE company needs about Rs. 300-600 Crores to reach sustainability. Also, many start-ups fail, making venture funding risky.

Next, the research sector in India – universities, government research laboratories and engineering service companies create very little valuable and patent protected IPR, which traditionally forms the basis for venture backed technology companies. This is born out of lack of appreciation of the value of such IPR, but mostly due to the absence of TNE start-up ecosystem.

Third, India should build a Telecom Standards Development Organization on the lines of ETSI or ANSI that can work with international standards to the benefit of Indian equipment companies and operators.

Finally, but most importantly, India needs a strong government policy support for:

(a) Creating incentives through preferential market access for companies (Indian and perhaps even foreign) that develop world-class technology and do high-end value addition within the country. Such a policy will attract the much-needed venture capital to fund local companies.

(b) Generous R&D grants (in 100s of Crores) to help companies that show world class product traction. This will enable them to compete on a level ground in international markets against competition like Huawei and ZTE, whose R&D is heavily subsidized by the Govt. of PRC.

Without such Govt. support, India will continue to depend entirely on imports, which will grow to 300,000 Crores annually in a few years, with all its associated problems and dangers.

Conclusion

It is time to harness the imagination and willpower necessary to help India reach its rightful position as an equal player with other world leaders like China, US, Japan, Korea, and Europe in TNE sector. The main missing enabler is risk-taking venture capital. And the key to attract such venture capital is: a sustained and carefully crafted Govt. policy of providing market access and R&D support to companies that add significant value in TNE sector.

Dr. Homi Bhabha will no doubt be a forceful advocate for this worthy cause had he been here with us today.

Panel 3

Information Communication Technology (ICT) Solutions for Social, Environmental and Health-care

---

Panel: Prof. Narendra Ahuja (*UIUC*), Dr. Sudhir Dixit (*Director, Hewlett-Packard India Laboratory, Bangalore*), Ashok Misra (*Chairman-India & Head of Global Alliances Intellectual Ventures, Bangalore*), Dr. Basant Rajan (*CEO, Coriolis Technologies*), N Viswanadham (*ISB, Hyderabad*)

(a) **Innovation Center for New Products:** As a nation, we need to run something akin to a *new products initiative* in the area of ICT, which will be able to objectively prune proposals, fund PoCs (Proof of Concepts), and monitor execution through a joint Govt., Corporate and University participation; until such time that we do not have a viable ecosystem for indigenous product development in this field.

(b) **Man-Machine user Interface:** ICT adoption, through simplifying man-machine interaction and immersive experience, will bring out applications of IT for inclusive growth. Such a growth will lead to ICT acceptability not only from the educated class but also from illiterate and uneducated masses.

(c) **Inter-disciplinary Innovation Centers:** India has an opportunity to be the hub for development of inventions and high value services through ICT but the creation of specialized knowledge is often the rate-limiting step and this is where innovators play a role. Thus, investing in inter-disciplinary innovation centers with inputs from electronics, communication, medical sciences, etc. will lead to opportunities in building up a huge IP bank that has positive effect on economy and society.

(d) **Logistics and Supply Chain:** Organize logistics and supply chain that can govern various schemes. This should cover future and existing schemes like the public distribution scheme, the mid-day meal program for school children, the National Rural Employment Guarantee Scheme (NREGS), etc. Such a governance will lead to put together disparate efforts and organize them as an effective food security supply-chain that will have a high impact food security solution, which serves millions of below poverty line people and also in the process generates millions of jobs.

(e) **Developing a Culture:** Invest in the development of right culture to accelerate innovation in ICT field. And a push to solve problems that adversely affect many, which in turn can draw on those affected for effective solutions as well.

Panel 4

Bioinformatics, Medical Informatics and Imaging, and Cognitive Computing

---

Panel: Somnath Biswas (*IIT Kanpur*), Vijay Chandru (*Strand Life Sciences, Bangalore*), Subhasis Chaudhuri (*IIT Bombay*), Daniel Cremers (*TUM, Germany*), Manoj Gopalkrishnan (*TIFR*), Bud Mishra (*Courant Institute, NYU*)

(a) Establish quality research groups which are essentially inter-disciplinary cutting across Bioinformatics, Medical Informatics & Imaging, Cognitive Computing, etc.

(b) The educational institutions must expand to meet the challenge of providing quality manpower in this area.

(c) Both the educational institutions and the industrial R&D sector must join hands to solve some of these pressing issues in structural biology and cognitive computing.

Panel 5

Rethinking Teaching of CS, IT and Computational Science (A viable ecosystem for Higher Education and Research)

Panel: Narendra Ahuja (*UIUC*), Prahladh Harsha (*TIFR*), Anil Kakodkar (*BARC*), V Rajaraman (*IISc*), Rajeev Sangal (*IIIT, Hyderabad*)

(a) **Curriculum Drivers in CS and IT:** There has been a dramatic change in the drivers of Computer Science due to: (i) ubiquity of computers, (ii) networked world with devices, sensors, computers, actuators, (iii) integration of computing and communications, and computation becoming cheaper compared to communication, (iv) availability of data in digital form, etc. Further, CS and IT have become part and parcel of several application domains of Sciences and Engineering. Thus, CS and IT curriculum needs to be appropriately integrated from school level to graduate programs.

1. **Computer Science:** The main challenge of Computer Science is to grow the theory to serve Information Technology drivers for the next few decades.
2. **Computational Science:** Teaching in computational science should be at a level that removes any mystery surrounding the advanced tools that we are talking about and students are able to develop a step-by-step understanding of the subjects concerned, including the tools that are used to explain the concepts involved.
3. **Exciting the Young:** We need to teach real Computer Science to high-school teachers. There is a strong need to popularize, among ourselves, and especially among high-school students and college students that Computer Science is an exciting science and in today's world – *indispensable*.

(b) **Strong Graduate Programmes in Computer Science, Computational Science and ICT Engineering**

1. Need to have strong graduate programmes.
2. Provide incentives for the young to do research, teach, and innovate.

(c) **E-Learning Systems:** Unified platforms of content and laboratories that are scalable, cost effective, and with a high utilization factor.

1. Could be used as an additional source, or as a continuing programme or independently.

Panel 6

ICT in Power Infrastructure

Panel: Jay Giri (*Alstom*), Prashant Gopalkrishnan (*Kalki Tech*), MA Pai (*UIUC*), CN Raghupati (*Infosys*), S Roy (*TCS*), Sushil Soonee (*POSOCO*)

The panel agreed that that there is now a rapid convergence of ICT and power infrastructure both at the transmission and distribution end. These recommendations are specific in this regard.

(a) On the transmission side while the network is state-of-the-art, the problem of massive data handling in implementation of Wide Area Measurement System (WAMS) as well as System Integrity Protective System (SIPS) requires that expertise be built within the country instead of relying on foreign vendors. Both WAMS and SIPS require a good understanding of communication and computer domains. In the West this expertise is provided by consulting firms. India must adopt that model by encouraging the private sector to get involved. A PPP model is also possible.

(b) Urgent R&D efforts should be initiated in the area of cyber security of power-grid and power-plants to prevent hackers from disrupting the grid infrastructure. This requires a high level of Computer Science expertise and the IT companies must take the lead.

(c) At the distribution end, a lot remains to be done to cut the AT&C (Aggregate Technical & Commercial) losses from 33% to 15% in a span of 5 years. A conservative calculation shows that this is equivalent to electricity generation in the tune of 16,000 to 20,000MW. This is a huge amount of loss. We can remedy this through:

1. Successful implementation of the R-APDRP scheme, SS automation, AMI, proper billing systems, etc. Also, procure simulation tools available today for the EHV and distribution systems that will allow one to demonstrate these *loss location* solutions in an off-line manner.
2. Redesign of the distribution system so that 11KV feeders run close to points of consumption and then step down the voltage. A study team from an IIT or IISc must look into this and submit a report for the DISCOMS in a time bound manner.

(d) Contrary to popular belief the AT&C losses are *both* due to (i) theft, unmetered load, billing issues, as well as, (ii) poor design of the distribution system.

(e) While, item (i) can be addressed through strong IT application as is being done under R-APDRP scheme; item (ii) is purely a power system problem and has to be handled by the planners.

1. The entire panel agreed that training of both transmission and distribution people is critical. This has to be taken on a "mission mode" basis. An idea of a one year certificate course in ICT and Power Infrastructure for engineers from *all* backgrounds is necessary to work in ICT related areas of the power sector. This can be modeled after the the BARC training school model done 5 decades ago. With a vast pool of power engineers, India can be a global leader in power sector just as in IT.