Report on the SIGCOMM 2011 Conference

John W. Byers
Boston University
byers@cs.bu.edu

Jeffrey C. Mogul
HP Labs, Palo Alto
Jeff.Mogul@hp.com

(report editors)

Fadel Adib (American University of Beirut), Jay Aikat (UNC – Chapel Hill), Danai Chasaki (U. Mass. Amherst), Ming-Hung Chen (National Taiwan Univ.), Marshini Chetty (Georgia Tech.), Romain Fontugne (Graduate Univ. for Advanced Studies), Vijay Gabale (IIT Bombay), László Gyarmati (Telefonica Research), Katrina LaCurts (MIT), Qi Liao (Central Michigan Univ.), Marc Mendonca (UCSC), Trang Cao Minh (Univ. Pompeu Fabra), S. H. Shah Newaz (KAIST), Pawan Prakash (Purdue), Yan Shvartzshnaider, (NICTA/Univ. of Sydney), Praveen Yalagandula (HP Labs), Chun-Yu Yang (National Taiwan Univ.)

(note takers)

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The editors take full responsibility for this article's content. Comments can be posted through CCR Online.

ABSTRACT

This document provides reports on the presentations at the SIGCOMM 2011 Conference, the annual conference of the ACM Special Interest Group on Data Communication (SIGCOMM).

Categories and Subject Descriptors

C.2.0 [**Computer-communication networks**]: General—*Data Communications*

General Terms

Algorithms, Design, Theory

Keywords

Conference Session Reports

Introduction

The SIGCOMM 2011 Conference was held August 15–19, 2011, in Toronto. This report provides notes on the papers presented, and on the discussion/question-and-answer session after each paper.

These notes were taken by a team of volunteer scribes, and then merged and edited after the conference. Please realize that the result, while presented in the form of quotations, is *at best* a paraphrasing of what was actually said, and *in some cases may be mistaken*. Also, some quotes might be mis-attributed, and some discussion has been lost, due to the interactive nature of the question-and-answer interactions.

The papers and presentation slides are available at `http://conferences.sigcomm.org/sigcomm/2011/conf-program.php`. Papers and audio/video recordings are available through the ACM Digital Library at `http://dl.acm.org/citation.cfm?id=2018436`, under the "Table of Contents" tab. We do not include notes on Vern Paxson's keynote talk, but a recording is available at that URL.

Session 1: Security

They Can Hear Your Heartbeats: Non-Invasive Security for Implanted Medical Devices

Authors: Shyamnath Gollakota, Haitham Hassanieh (MIT); Benjamin Ransford (University of Massachusetts Amherst); Dina Katabi (MIT); Kevin Fu (University of Massachusetts Amherst)
Presenter: Shyamnath Gollakota
Notes by Romain Fontugne, Qi Liao, and Yan Shvartzshnaider

Winner of the SIGCOMM 2011 Best Paper Award

Summary:

Shyamnath Gollakota presented a system that protects wireless Implantable Medical Devices (IMDs) from malicious users. Modern wireless implants permit doctors to remotely monitor and reprogram the implants, but this exposes IMDs to two types of security attacks: (1) passive attacks (i.e. eavesdropping), which compromise patient data, and (2) active attacks, in which malicious users execute unauthorized commands (e.g., to disable therapies). Securing IMDs is not trivial, because techniques such as cryptography cannot be implemented in the millions of devices currently in use, so the goal of this work is to secure IMDs without modifying them.

The authors delegate security to an external *shield*. The shield jams all communications with the implant, preventing the execution of unauthorized commands and preserving privacy. Since the shield is the source of the jamming signal, it is the only device able to generate an antidote signal and read the data sent by the implant. The doctor uses the shield as a proxy to communicate with the IMD. The main technical innovation is a two-antenna radio design that allows a small, wearable device to do full-duplex wireless communication.

They evaluated the shield with real medical implants, surrounded by beef and bacon to simulate the human body. Attack commands with normal power can be successful at up to 14 meters from the patient, but are blocked by the shield if the adversary is as close as 20 cm. The shield blocks a high-powered adversary with 5 meters, but the shield is intrinsically limited in its ability to prevent high-powered attacks.

Discussion:

Mark Handley (UCL): Can a passive eavesdropper with multiple antennas use the difference in phase between the transmissions from the IMD and the shield to separate the signals? Answer: the

shield is worn much closer than 1/2 wavelength from the IMD, so theory says that the signals cannot be separated.

Jitu Padye (MSR): What scenario motivates the protection against passive attacks? Answer: We expect implants to be continuously monitored, at home or in public, not just at the doctor's office; eavesdroppers could be anywhere.

Tianji Li (UT Austin): To detect malicious commands during an active attack, the shield must detect the commands first in order to know whether they are legitimate or not, and then jam the signals, which means the IMD is also decoding the commands while the shield is decoding. Answer: The shield can respond quickly enough. Packets last ca. 30-40 msec, while the shield can respond in 100 usec.

Eric Anderson (CMU) and Robert Lychev (Georgia Tech): A one-time pad requires you to change the encryption key regularly; how did you implement this? Answer: The key is never reused, and changes for every transmission using a random-number generator.

David Clark (MIT): Because the shield is constantly listening for implant transmissions and uses cryptography, what is the battery life of the proposed device? Answer: Unlike IMDs that can work for 10 years, the shield needs to be recharged frequently.

Let the Market Drive Deployment: A Strategy for Transitioning to BGP Security

Authors: Phillipa Gill (University of Toronto); Michael Schapira (Princeton University); Sharon Goldberg (Boston University)
Presenter: Phillipa Gill
Notes by Jay Aikat, Qi Liao, Yan Shvartzshnaider
Summary:

The paper points out that Secure BGP and Secure Origin BGP (S*BGP) have been around almost two decades, but they are not widely deployed due to certain technical reasons, but mostly because there has been no economic incentive for any given ISP to deploy first. The paper presents a strategy to drive global S*BGP deployment that relies solely on each ISP's local economic incentives – i.e., an ISP's interest in attracting revenue-generating traffic to their networks. The strategy does not depend on ISPs being motivated directly by a desire to provide better security. Instead, governments or industry groups can create the necessary market dynamics.

Discussion:

Henning Schulzrinne (Columbia): Few ISPs profit from transit traffic; most ISPs get most of their revenue from peering. Have you talked with ISPs to understand if your approach corresponds to their business model? Answer: The issue is what portion of revenue comes from generating traffic, as well if there is an enough incentive within the organisation to deploy SBGP, such as government subsidies. We did not survey all ISPs, so we used a simplified business model.

Henning Schulzrinne: The current financial situation does not seem to permit government subsidies. Answer: Pres. Obama put SBGP on his national strategy. But getting early adopters is still a challenge.

Sergey Gorinsky (IMDEA): In your example with China Telecom, I am not sure that SBGP actually solves the security problem you used to motivate the work. Answer: Our work didn't concentrate on solving the security issue.

Jia Wang (AT&T): I like that you see this from an economic angle, but you don't seem to include any of the costs of deployment – they are huge. Not only the initial deployment, but also maintenance. Answer: We modelled deployment cost, but have not yet considered operational costs, but we plan to. We think everyone will end up spending more, in the end.

Jia Wang (AT&T): Did you actually talk with any ISPs about the business model? Answer: We talked with Level 3 and a few others at the NANOG conference; they were encouraging, but we need to talk with more of them.

Nikolaos Laoutaris (Telefonica): An ISP would want to load-balance across peering points; your approach seems to create hot-spots, and the ISP would have to pay to increase capacity at these peering points, without any new revenue to compensate. Answer: We didn't model multiple connections between ASs, and we might have included backup links that shouldn't have been used.

Ramesh Sitaraman (U. Mass.): Most money flows into ISPs from the first mile, where the content providers are. It would be useful to talk with content providers to see what part of that revenue would be lost if SBGP were not deployed? Answer: I'm not sure about the cost for content providers. They want their user to reach the content, not to deal with security issues.

Chris Small (Indiana U.): Have you tried to model the differences in link costs? Answer: We used a simplified model, where links are either customer-provider or peering links.

Finding Protocol Manipulation Attacks

Authors: Nupur Kothari (University of Southern California); Ratul Mahajan (Microsoft Research); Todd Millstein (UCLA); Ramesh Govindan (University of Southern California); Madan Musuvathi (Microsoft Research)
Presenter: Nupur Kothari
Notes by Qi Liao and Yan Shvartzshnaider
Summary:

This paper describes a threat model called a protocol manipulation attack, in which honest participants correctly follow a network protocol specification, but adversarial participants may deviate arbitrarily for their own benefit (for example, using TCP Daytona to achieve a faster transmission rate than compliant congestion-controlled TCP). The main contribution in the proposed approach combines program analysis techniques in novel ways. The authors use static analysis with symbolic execution, as well as dynamic analysis with concrete testing, to discover manipulation attacks. The paper describes the implementation of MAX (Manipulation Attack eXplorer), which enables systematic exploration of manipulation attacks in protocol implementations. By comparing protocol runs with and without modifications, developers can determine if manipulation attacks can succeed. The authors have the code available at `enl.usc.edu/projects/max`.

Discussion:

Marco Canini (EPFL): How do you prune the code that you deem irrelevant? Answer: By analyzing the control flow graph, and only analyzing the code that leads directly to the vulnerable statement. Follow-up: What percentage of code are we talking about? Answer: I don't have exact answers, but for example, if the vulnerable statement lies in congestion control, code in other modules can be pruned. Follow-up: How much time did it take to come up with the vulnerable statements and what is the running time? Answer: Identifying vulnerable statements depends on identifying what the important resources are, e.g., whether network usage (such as number of outstanding packets), or memory allocation, is critical. The running time depends on the protocol implementation. In our studies, it ran from minutes to days.

Praveen Yalagandula (HP Labs): Is it possible to find vulnerable statements easily in other protocols based on resources, and can you automate this effort? Answer: This is part of our future work, and using program analysis we believe it is possible. You need to know the resources that you are interested in.

Michael Sirivianos (Telefonica Research): Can you elaborate on how to find the attacks automatically with MAX, or does needing to know where to look make it a lot harder than studying known attacks in TCP? Answer: Again, we believe it is possible. For each protocol, we first figure out what are the important resources. For some it would be the network; for others memory. We will then try to find the vulnerable statements that control them. We tried to be protocol-agnostic at all times. We believe it is just a matter of exploring the other protocols and trying out different vulnerable statements.

Session 2: Novel Data Center Architectures

Augmenting Data Center Networks with Multi-gigabit Wireless Links

Authors: Daniel Halperin (University of Washington); Srikanth Kandula, Jitendra Padhye, Paramvir Bahl (Microsoft Research); David Wetherall (University of Washington)
Presenter: Daniel Halperin
Notes by Vijay Gabale and Pawan Prakash
Summary:

The goal of Flyways is to enable a data-center network with an oversubscribed core to act like a non-oversubscribed network. The work is based on measurements showing that the core is not always under heavy load, and there are relatively few hot spots in the network. These hot spots can be tackled by dynamically creating high-bandwidth links, where needed, via 60GHz wireless links.

In a data center, racks are densely packed, and distance between racks is short. This is a good match to 60Ghz wireless, which has high bandwidth, short range, and can use directional antennas to create line-of-sight links. These wireless Flyways improve performance at low overhead (material cost and installation complexity). Since the data center topology is known and stable, directionality alleviates the multipath problem.

The authors first verified the stability of link utilization in a data center environment. Then they showed how traffic demands for directional links can be calculated. They described an algorithm to configure the wireless links; the efficacy of Flyways depends on accurate information about traffic demands. For this algorithm, the authors decomposed the problem into two parts: (1) a Flyway picker, which establishes links according to demands, and (2) a Flyway validator, which ensures that the links are non-interfering. For the Flyway picker, the authors used 60GHz directional, steerable, phased-array antennas.

The authors also show how to leverage the wired backbone to coordinate the wireless transmissions; for example, to send ACKs.
Discussion:

Anja Feldman (TU-Berlin/T-Labs): Do other objects in the data center (metal, cooling liquid, etc.) interfere with the wireless connections? Answer: Since the antennas are at the top of the rack, they suffer from minimal interference. The paper shows measurements from a real data center.

Anja Feldman (TU-Berlin/T-Labs): What if computing is your rate-limiting resource, rather than data transfer? Answer: We had to approximate the traffic demand in our study.

Anja Feldman (TU-Berlin/T-Labs): What if you simply moved computation around to minimize communication costs? Answer: We did not optimize node layout. We don't know whether this would help.

Dina Papagiannaki (Telefonica): Did you compare the cost of this solution to one that provides a full-bisection-bandwidth wired network? 60 GHz radios are still very expensive. Answer: No cost comparison. But 60 GHz will commoditize (as 802.11 has).

Hari Balakrishnan (MIT): Did you characterize traffic patterns for which the system will not work as well? And what is the best you could hope to do? Answer: It won't work well if you need full bisection bandwidth, or if the workload is not predictable over reasonable intervals. Regarding the best you could hope to do: we are mainly limited by the bandwidth of the unlicensed band.

[Didn't give name]: Please compare with optical work (Helios, Glimmerglass). Answer: While I was an intern, I tried to get a Glimmerglass switch to work, but ran into technical difficulties. These switches have been around for a decade, and their cost has not declined. 60 GHz parts are CMOS and should decline in cost.

Eric Anderson (CMU): What makes this a good application for wireless, vs. a fully wired topology with less than full bisection bandwidth? Answer: You can put the wireless bandwidth where it is needed, on demand.

Better Never than Late: Meeting Deadlines in Datacenter Networks

Authors: Christo Wilson (UCSB); Hitesh Ballani, Thomas Karagiannis, Ant Rowstron (Microsoft Research)
Presenter: Thomas Karagiannis
Winner of the SIGCOMM '11 Honorable Mention Award
Notes by Pawan Prakash
Summary:

The basic premise of this work is that application SLAs cascade into the need to enforce deadlines for communication between components of an application. A flow is useful if and only if it satisfies its deadline. Today's protocols, however, are deadline-agnostic and strive for fairness, which may not be the most optimal strategy given a deadline-dependent workload. The authors propose a deadline-driven delivery protocol (D^3), in which the flows are prioritized based on their deadlines. This technique improves the quality of responses and also saves resources.

If information on a flow's size and deadline is available beforehand, the system calculates a desired rate for the flow. The routers on the path of the given flow then promise to provide that rate, which increases the chances that the flow will finish before its deadline. A flow that has missed its deadline is "quenched" in the network.

In their evaluation, the authors demonstrated that D^3 can support roughly twice as many workers, while satisfying application deadlines.
Discussion:

Marco Canini (EPFL): How realistic is it to require the application to know its flow size and deadline when it opens the connection? And can D^3 be used with legacy applications that do not provide this information? Answer: This information more or less exists today, but we do need the mechanism to provide it to the protocol. In our evaluation, we included background flows without deadlines, but we give priority to those that do.

Nina Taft (Technicolor): How do you actually get the deadlines – this seems hard to do? Answer: In today's data center, the SLAs already mostly are known, but they aren't enforced.

Dina Papagiannaki (Telefonica): Are you just quenching specific flows, or are you stopping entire jobs? You could still end up with the entire job failing its SLA. Answer: You could think of more clever implementations than our simple one. It depends on what the operator wants to do.

Anja Feldman (TU-Berlin/T-Labs): How would this work with cloud-based VMs which implement their own TCP stack? Answer: We assume that the entire data-center is controlled by a single entity.

Anja Feldman (TU-Berlin/T-Labs): Data-center RTTs will be very small; your calculation seems like it might take longer. Answer: we have observed RTTs on the order of 100 microseconds.

Amin Tootoonchian (U. Toronto): You are requiring the network control plane to do work on each flow. A burst of flow arrivals could use up a lot of CPU within the network switches. Answer: The overhead, even in a user-mode implementation, was less than 1 microsecond/packet; there is no per-flow state in the routers. We can handle bursts of thousands of flows.

Amin Tootoonchian (U. Toronto): What is the signalling overhead? Answer: ca. 20 bytes for the rate request.

Srikanth Kandula (Microsoft): It seems like shorter flows are the ones that care about deadlines, and your control loop cycle is about 1 RTT. Does your ability to meet deadlines negatively correlate with flow length? Answer: It depends on flow length and how tight is the deadline.

NetLord: A Scalable Multi-Tenant Network Architecture for Virtualized Datacenters

Authors: Jayaram Mudigonda, Praveen Yalagandula, Jeffrey C. Mogul (HP Labs); Bryan Stiekes, Yanick Pouffary (HP)
Presenter: Jayaram Mudigonda
Notes by Pawan Prakash
Summary:

The authors try to help providers of "Infrastructure-as-a-Service," who need data center networks that support multi-tenancy, scale, and ease of operation, at low cost. NetLord is a multi-tenant network architecture that provides a virtualized layer 2 and layer 3 network to the tenants. It exploits inexpensive commodity equipment to scale the network to several thousands of tenants and millions of virtual machines.

Through the encapsulation of a tenant's L2 (Ethernet) packets in its own IP packets, NetLord gains multiple advantages over prior solutions. NetLord's design can exploit commodity switches in a way that facilitates simple per-tenant traffic management in the network, greatly reducing FIB table size. NetLord's ARP protocol simplifies the design, by proactively pushing the location information of VMs to all servers. NetLord improves ease of operation by only requiring a static one-time configuration that is fully automated.

Evaluation shows that the NetLord architecture scales to several thousands of tenants, and hundreds of thousands of VMs. NetLord can achieve a 4x improvement in goodput over existing approaches, and imposes only modest overhead.
Discussion:

Stefan Saroiu (Microsoft Research): NetLord is optimized to work with COTS components and commodity switches, but how will it interact with other boxes such as intrusion detection systems (IDS), that may have issues because of IP encapsulation? Answer: Right, we would need an IP filter or decapsulator to sit in front of IDS-like components. This can be done efficiently, at low cost. Follow-up: Do you have an incremental deployment story? Answer: We've given it some thought. Incremental deployment seems readily feasible, although the FIB pressure may not come down drastically in incremental deployment, as more endpoints are hidden, so the benefits increase in direct proportion to deployment.

Dantong Yu (Brookhaven National Lab): How can you ensure that the interference between multiple tenants is reduced, as in the previous talk? Answer: Quantitative analysis of that aspect, getting an adequate amount of bandwidth, and separation of tenants, is not the focus of this work.

Kang Xi (NYU Poly): Can you provide a comparison between modifying the hypervisor or modifying the switches. For example, what about Cisco's FabricPath? Answer: The presenter gently expressed skepticism as to whether Cisco FabricPath achieves multi-tenancy, can be made to work with small FIB tables, or is cheap, i.e. commodity. Follow-up: How would virtualized address spaces talk to each other? Answer: This can be done with designated virtual interfaces that can reach other, which can be embedded in the hypervisor, and the details are in the paper.

Anja Feldmann (TU Berlin/T-Labs): Low-cost switches may not have layer-3 capability. What if your last-hop switch does not understand IP? Answer: To clarify, we are relying only on layer-3 forwarding, not routing. This basic data plane service is generally provided in commodity hardware; the expensive aspects of layer-3 lie on the control plane, such as OSPF implementation.

Srikanth Kandula (MSR): How does FIB size reduction increase throughput so significantly? Answer: FIB thrashing significantly reduces throughput. NetLord reduces FIB thrashing dramatically, and hence improves throughput. Follow-up: Shouldn't there be a significant fall-off then? Answer: Yes. We saw this (referred questioner to paper).

Session 3: Bulk Data Transfers

Inter-Datacenter Bulk Transfers with NetStitcher

Authors: Nikolaos Laoutaris, Michael Sirivianos, Xiaoyuan Yang, Pablo Rodriguez (Telefonica Research)
Presenter: Nikolaos Laoutaris
Notes by Jay Aikat
Summary:

NetStitcher is motivated by the need to move data between data centers, in order to replicate all data geographically. NetSticher exploits off-peak, otherwise-unused network bandwidth. It provides a "volume service" – a contract to move a certain amount of data over the pipe. The capacity of this pipe wanes and waxes over the course of a day, but NetStitcher guarantees delivery within a certain time (during the day). Since it is possible that not all of the links on the path between two data centers are available for use at the same time, NetStitcher uses "store and forward" via intermediate datacenters. (This is especially necessary when data is being moved across time zones). NetStitcher uses existing CDN servers as the intermediate storage servers. Their view is that bandwidth moves data across space (i.e., between data centers), while storage moves data across time (i.e., between opportunities to exploit spare bandwidth). They are trying to create a FedEx-like service for data transfer. Unlike mobile DTNs, this application has less churn, and relies on well-connected links for data transfer.
Discussion:

Rachit Agarwal (Urbana-Champaign): It seems that you need to be able to predict both the paths and their bandwidths – do you have any idea how to do this? Answer: At this time, we assume we are only using one AS, where we have full control and so we know the routing and the leftover capacity on each link. To extend this over multiple ASs, we would need to do prediction.

[Unknown] (Microsoft): Part of your goal is to save bandwidth costs, but building a petabyte storage node is not cheap, either. Answer: We are free-riding on the CDN infrastructure. But even if you did not have a CDN, the main idea is that you can exchange expensive bandwidth for much cheaper storage.

Anja Feldman (TU-Berlin/T-Labs): How do you recover if you have a network failure, and you need the bandwidth for your primary purposes – your predictions about free capacity are no longer correct? Answer: We re-execute the optimizations every 10 minutes, with updated time series. So if something goes wrong in the network, the system redirects the transfer along another path.

The Power of Prediction: Cloud Bandwidth and Cost Reduction

Authors: Eyal Zohar, Israel Cidon (Technion); Osnat Mokryn (Tel Aviv College)
Presenter: Eyal Zohar
Notes by Jay Aikat
Summary:

The authors present a Traffic Redundancy Elimination (TRE) system called PACK (Predictive ACK). They contend that existing solutions have significant limitations: high processing costs, poor scalability, and poor support for redundancy occurring over longer time scales (say, over days or weeks). The first main advantage of PACK is the use of client-side processing: chunks are monitored, and chains of redundant chunks are predicted by each client. Chunk prediction is facilitated by hashing: each chunk is associated with a SHA-1 signature, thus every every chunk has a universally unique ID. In this way, PACK minimizes processing costs at the server, which simply runs SHA-1 and is memoryless and stateless. In their evaluation, the authors demonstrated the advantages of PACK against two approaches: no-TRE (baseline), and sender-based TRE. Key observations include the fact that in a 24-hour study of ISP traffic, 30% e2e redundancy is observed. Longer-term redundancy (such as videos replayed on other days) occurs at a higher rate. There is also high email redundancy on longer time scales as people reread their emails and attachments. The efficient chunking solution and the use of client-side resources constitute the key advantages. The PACK software is freely available for download.

Discussion:

Dantong Yu (Brookhaven National Lab): If you have a very high-speed network, then how do you factor this speed into your prediction? Answer: This is a good question, and we handle it in the paper, where we show how to adjust the predictions to the speed of the network, and maybe even skip forward. Follow-up: Can you capture the context of a TCP flow from the client perspective? Answer: Yes, we handle that. The client needs to record the forward sequence to do these predictions.

Aditya Akella (Wisconsin): Remarked that the paper embodied a really cute idea, then wondered whether the scenario chosen to highlight the work is flawed, and asked whether there was an expectation that Google would run PACK in its datacenters (implying that this is misguided)? Answer: We are not implying that this is useful to everyone, but application developers would use this since it reduces cost for them. Follow-up: What is the client's incentive to use this service? Moving things away from the client into the cloud is a current trend, not vice versa. Answer: In a mobile environment, clients want to minimize bandwidth at all costs. Technologically complex solutions to derive small bandwidth savings in other ways exist, so this is the incentive.

Lihua Yuan (Microsoft): It seems that much redundancy may be due to improper caching. With a proper cache layer at the client, why do you need to recover in the first place? Answer: If the application is aware of TRE, then they can do what we did, but application guys are not redundancy-aware; this is not their core business.

Aditya Akella (Wisconsin): I think a compelling way to evaluate your approach would be to consider a hybrid approach, combining client-side and server-side features. Answer: We did it in the paper and evaluated it, and I agree that this makes sense.

Managing Data Transfers in Computer Clusters with Orchestra

Authors: Mosharaf Chowdhury, Matei Zaharia, Justin Ma, Michael I. Jordan, Ion Stoica (UC Berkeley)
Presenter: Mosharaf Chowdhury
Notes by Jay Aikat
Summary:

Data-intensive cluster applications spend a large fraction of their run time in data transfer. This can affect both performance and scalability. These applications exhibit several transfer patterns, such as "shuffle" (e.g., in Map-Reduce), one-to-many ("broadcast"), and many-to-one ("incast"). Each transfer acts as a barrier, so the most important metric is completion time for the entire transfer; this makes flow-level scheduling less effective at prioritization.

Core contributions: (1) Orchestra is a global management architecture that optimizes performance at the level of transfers rather than at the level of individual flows. To achieve this, Orchestra contains a transfer controller (TC) for each kind of pattern (shuffle, broadcast, incast), and thus, different applications can provide different TCs that are best suited to their needs. (2) Orchestra includes an inter-transfer controller (ITC) to coordinate among the TCs.

Their evaluations show large improvements in application completion times.

Discussion:

Hitesh Ballani (Microsoft): As you increase the number of flows in a system, the fair share of each flow will drop. Answer: We have an administrator-set upper limit on the number of flows that be can created for each node. The framework is meant for long transfers in data-intensive workloads, so TCP will have enough time to ramp up to get a fair share.

Vytautas Valancius (Georgia Tech): What assumptions do you make about the network topology? Answer: Our main use case is EC2, where we don't have much direct topology information.

Stefan Saroiu (Microsoft): In your evaluation, you use BitTornado – is it the case that once a peer downloads a file, it disconnects from the transfer? Answer: We had that implemention in the submission, but we updated to a version where the peers do not leave, based on the shepherd's suggestion.

[Didn't give name]: Is there a way to map customer-specified deadlines or SLAs onto the weights that you apply to individual transfers? Answer: Yes, it is possible, but we have not done that. Customers can give more info about their jobs. Right now, we try to optimize each transfer, but you could have some property on the overall job completion time, but we have not implemented that.

Srikanth Kandula (Microsoft): Your work is very focused on Map-Reduce-like workloads running in EC2-like virtualized settings. Can you generalize to a take-away for other patterns? Answer: We found four common patterns, the three we addressed plus random traffic matrices, which are hard to address and so we have not yet done so.

Session 4: Network Measurement I – Wide-Area Measurement

On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System

Authors: John S. Otto, Mario A. Sanchez (Northwestern University); David R. Choffnes (University of Washington); Fabian E. Bustamante (Northwestern University); Georgos Siganos (Telefonica Research)

Presenter: John Otto

Notes by Jay Aikat, Marshini Chetty, and Yan Shvartzshnaider

Summary:

In this paper, the authors present a view of BitTorrent usage sampled over a two year period, for 500,000 users, spanning 169 countries and 3150 networks. Using data collected from clients, including application-level data as well as active traceroute data, the authors investigated how is BitTorrent being used, where is this traffic flowing, and who is paying for it? The thesis is that BitTorrent traffic is generally more expensive than other traffic and the research goal was to determine what traffic characteristics result in high cost.

First, the authors made some basic observations about shifting traffic patterns, noting a 10% overall drop in users over the time period, but a 12% increase in overall system traffic, with Europe seeing a large drop but Asia and Africa both seeing significant increase. From an ISP standpoint, the authors find that most BitTorrent traffic is at the edge of the network, with the majority of traffic staying in Tier 2, and a small fraction going to Tier 1. To investigate the cost incurred by an ISP, the authors identified that for one (unnamed) ISP, their overall traffic tended to peak in the afternoon and evening, and peak BitTorrent use coincided with this peak usage – implying that the cost of BitTorrent to ISPs is high.

To conclude, the authors reiterated the value of their approach, whereby a broad view from the edge of the network is required to see a large distributed system's full spectrum of usage. This general approach can potentially be applied to understand other distributed systems like video streaming applications and peer to peer CDNs.

Discussion:

Sergey Gorinsky (IMDEA). What is the pricing function that you are using to map peak traffic to actual money? Answer: The cost matrix that we use is not absolute. We used relative cost considerations. For example, BitTorrent might contribute 30% of the traffic to the link but we estimate 60% of the cost, using relative metrics. In the paper we considered two different cost models: first, a volume-based model, and second, a common billing model, using 95th percentile exchanges between ISPs. We had two specific ISPs with known relationships.

Sergey Gorinsky (IMDEA): Did you consider the additive nature of pricing? The more you buy, the less the cost per Mbps? Answer: We're focusing only on the impact of the 95th percentile value, not other fixed costs.

Henning Schulzrinne (Columbia): Have you compared this data to the SamKnows data from a large number of ISPs? Answer: No we haven't made the comparison, but I'd be happy to look into that. Comment: They come up with slightly different conclusions.

Henning Schulzrinne (Columbia): When you talk about ISP X, for medium to large ISPs like Comcast, they don't pay for traffic generally. They get people to pay them for traffic. So ISP X must be someone other than a big provider or economically the data you are describing would not make sense. Answer: In what we are doing ISP X is paying its providers ISP A and B, and ISPs C through G are paying ISP X. Follow-up: Is your result based on your perception of the world or a real business model, since this differs from business models of North American ISPs? Answer: I should have made this more clear. Our model is based on ground truth information of specific ISPs.

Henning Schulzrinne (Columbia): Can you say anything on the nature of these ISPs? I have a hard time fitting it in into a classical type of ISP e.g., the tier 1 type, Level 3, Comcast, AT&T, etc. Answer: I can say that ISPs C to G are licensees for smaller access networks while ISP X, A and B are transit providers.

Predicting and Tracking Internet Path Changes

Authors: Italo Cunha (Technicolor and UPMC Sorbonne Universités); Renata Teixeira (CNRS and UPMC Sorbonne Universités); Darryl Veitch (University of Melbourne); Christophe Diot (Technicolor)

Presenter: Italo Cunha

Notes by Jay Aikat, Marshini Chetty, and Yan Shvartzshnaider

Summary:

In this paper, the authors focused on tracking internet path changes. Usually, to track large numbers of paths, `traceroute`-style measurements are used. However, these measurements are costly when conducted frequently, and are subject to network and system limitations. Other methods, like Tracetree and Doubletree, are more scalable, but have low accuracy. The focus of this paper is to improve the accuracy of measurements while on a measurement budget. The authors' contributions are NN4, which predicts path changes and distinguishes unstable and stable Internet paths; and DTrack, which separately tracks Internet path changes using a probing process. For predicting path changes, the prediction goals are to find the time until the next change, the number of changes within a time interval and to assess whether a path will change in a time interval. For feature selection, they compared NN4 against RuleFit, a machine learning technique to identify relative importance of features from traceroute measurements. Both identified similar features, but NN4 was considerably lighter-weight. Among the most important feature was past prevalence, or fraction of time a path was active in the past.

DTrack allocates probing rates per path using NN4's predictions and selectively targets probes with specific hop limits along each path in an attempt to reduce redundant probes at shared links. The presenter described extensive comparative evaluation between DTrack and a variety of alternative methods. The bottom line is that Dtrack detects more changes than current state of the art methods in a manner that is more up to date and more accurate. In future work, there are plans to extend Dtrack to reduce remapping cost by porting to a home gateway and coordinating probing across multiple monitors.

Discussion:

John Byers (BU): I have a question regarding your future work. Can you make inferences from path changes that you observed? For example, you might infer a large structural change from a small subset of changes. Answer: Yes, we look at the path changes. So we don't need to detect the same change on the other path; this is done purely with topology information. Follow-up: I was also wondering on periodicity in your dataset. Do you witness any periodicity? Answer: No we didn't observe any periodic data.

Sharon Goldberg (BU): Was the evaluation based on a trace or other specific dataset? Answer: The data we used was from using FastMap, we simulate our methods on top of this data with a lower probing budget.

Hyun-chul Kim (Seoul National University): [referring to feature selection slide] Which of these features was important using RuleFit? Answer: Path prevalence; beyond four most important, others did not help much. Follow-up [referring to future work

slide]: What extra info do you have from BGP updates? Answer: We could potentially use routing updates, but we haven't started working on this yet

Carey Williamson (U. Calgary): Can you comment on the applicability of two of your assumptions in practice: 1) the Poisson process of the probing, 2) independence of that process on each path. Answer: Independence probably does not hold, but we found that use of a Poisson process is probably better than uniform because path changes are bursty. We have not studied the accuracy of the Poisson process in practice.

Unknown (Hong Kong Polytechnic University): How do load balancers impact your probing methods, and what about manual path changes? Answers: We detect load balancers using Paris traceroute, and probe within the load balancers explicily. If a router configuration creates a path change, then yes, our system will detect these path changes.

Broadband Internet Performance: A View From the Gateway

Authors: Srikanth Sundaresan (Georgia Institute of Technology); Walter de Donato (University of Napoli Federico II); Nick Feamster (Georgia Institute of Technology); Renata Teixeira (CNRS and UPMC Sorbonne Universités); Sam Crawford (SamKnows); Antonio Pescapé (University of Napoli Federico II)
Presenter: Srikanth Sundaresan
Notes by Jay Aikat, Marshini Chetty, and Yan Shvartzshnaider
Summary:

The authors studied network-access link performance, based on measurements taken directly from home gateways (SamKnows/FCC data from 4,000 devices across multiple ISPs, plus BISMark data from a much smaller set). These measurements can be difficult; doing them at the gateway avoids the confounding effects of wireless networks, etc. They measured throughput for HTTP, TCP, and UDP.

They describe how cable providers use traffic shaping and PowerBoost, which provides short bursts of higher bandwidth. DSL, on the other hand, uses interleaving, which can increase latency but decreases loss, leading to higher throughput. They showed how cable and DSL customers can see different kinds of bandwidth and latency behaviors, and how excessive modem buffering ("bufferbloat") can create latencies of up to 10 seconds.
Discussion:

Henning Schulzrinne (Columbia/FCC): The FCC/SamKnows data is available online; anyone can reproduce the measurements or conduct their own analysis of the data.

Shyan Pooya (U. Alberta): The bufferbloat project says "there is no right buffer size" – can you comment? Answer: we believe large buffers are there to reduce loss, for packets going from the high-capacity home network into the lower-capacity access link. But these large buffers don't help very much, and they create other issues; these large buffers aren't justified.

Justin Ma (Berkeley): Are ISPs being honest about the services they are providing? Answer: We didn't have SLA info, but we compared average throughput to 95th percentile throughput. (Henning Schulzrinne adds: The SLA information is available in a separate report, and they found that it's between 80-90%, in terms of meeting advertised rates. In fact, FIOS recently used the FCC/SamKnows data to advertise their advantage over cable, saying that they deliver 110% of advertised speed.)

Kai Pui Mok (Hong Kong Polytechnic Univ.): You used ICMP to measure last-mile latency, but some routers may delay ICMPs; would this affect your measurments? Answer: we measured baseline latency when there was no cross-traffic, so we don't expect extra delays. It is possible that there were other delays, but we could not account for that.

Sharon Goldberg (BU): What kind of SLA structure would or should a home user have? Answer: Current SLAs just describe download and upload throughput, but these are not necessarily comprehensive; SLAs should also include latency and describe the effect that this would have on applications such as Web browsing. There needs to be a more comprehensive advertising model.

Anja Feldman (TU-Berlin/T-Labs): Why do you limit your measurements to the gateway instead of including the entire home network, and why don't you measure the pipelined download latency for real, multi-object web pages, instead of just throughput and ping latency? Answer: Those are tough questions; the point of this work was to first characterize the access link, then we could build on that to characterize the entire user experience. We can then move on to study home-network issues. There's also a lot of work left to do to translate these metrics into something the user can understand.

Session 5: Wireless

Random Access Heterogeneous MIMO Networks

Authors: Kate Ching-Ju Lin (Academia Sinica); Shyamnath Gollakota, Dina Katabi (MIT)
Presenter: Kate Ching-Ju Lin
Notes by Fadel Adib, Danai Chasaki, and Vijay Gabale
Summary:

The authors present 802.11n+, a protocol that allows MIMO nodes to fully participate in 802.11. The motivation behind the protocol is that, in current Wi-Fi networks, when an 802.11 single-antenna node is transmitting, all other nodes refrain from transmitting, even though, in practice, MIMO nodes with additional antennas have sufficient degrees of freedom to still be able to transmit additional streams.

Two fundamental challenges need to be addressed in such a scenario: (1) carrier sensing in the presence of ongoing transmissions, and (2) transmitting without interfering with ongoing transmissions. The authors address the first challenge by viewing transmissions as a mult-dimensional signal (one dimension for each antenna), and forcing all transmitters to contend for the medium in the null space orthogonal to all ongoing transmissions. The second goal is accomplished by using interference nulling and interference alignment. Interference nulling is exemplified by a 2x2 MIMO transmitter participating in parallel with a single-antenna stream. Here, the additional transmitter calculates its channel at the single-antenna receiver, and transmits on both its antennas such that its transmission is nulled at the single-antenna receiver. Interference alignment is exemplified by adding an additional 3-antenna transmitter to the mix; now more complexity is needed to retain sufficient degrees of freedom. The authors propose having this transmitter both null its signal at the single-antenna receiver, and align it with the existing transmission from the 1-antenna transmitter at the 2-antenna receiver. Since the latter alignment requires only one degree of freedom, the 3-antenna transmitter is still able to transmit one additional stream, using the remaining degree of freedom.

Experiments, for a simple scenario of one single-antenna node, one 2x2 MIMO node, and one 3x3 MIMO node, show that the throughput improvements of 802.11n+ over 802.11n can be as high as 2x on average.
Discussion:

Shivkumar Kalyanaraman (IBM Research, India): How do you estimate channels h_1 and h_2 without using centralized controllers? Also, how do you handle symbol level synchronization? Answer: For the first question, we can do it in a a fully distributed way

by annotating the packet header. For the second question, we can leverage previous work to overhear the channels and use the phase difference. Follow-up: How many concurrent interferers can you handle at the same time? Answer: We have experimented with up to 3 concurrent nodes decoding them online (as we were limited by hardware), and up to 4 concurrent nodes decoding them offline.

Chen Qian (UT Austin): Could you please explain the results figure? It seemed that 802.11n should have higher median throughput, i.e. 50Mbps rather than 25Mbps, for the status quo. Answer: USRP2 uses 10MHz bandwidth, which is half the bandwidth of Wi-Fi (20Mhz). Accordingly, one would expect half the Wi-Fi throughput; as you can see, our experiments for 802.11n achieve about 25Mbps median and thus a projected 50Mbps throughput for Wi-Fi and 100Mbps for our proposed 802.11n+. Follow-up: Why is the throughput improvement for 802.11n+ in your experiments double rather than triple whereas now you could use 3 streams? A: The reason is that, for example, in 802.11n it is equally probable for any of the nodes to win the medium after contending. Thus, for example, if initially the 2-antenna node won the medium, the 3-antenna AP will be able to transmit one extra stream in 802.11n+ as opposed to none in 802.11n; in this scenario, the improvement is 1.5x. In short, the 2x improvement is an average over many runs.

Abhishek Mishra (Lehigh U.): Interference alignment seems to require that the APs know how many users are using the channel. Is it a reasonable assumption to have only 3 users using the channel? Answer: If you mean how do we estimate the channel, in the paper we detail how to do this estimation.

Strider: Automatic Rate Adaptation and Collision Handling

Authors: Aditya Gudipati, Sachin Katti (Stanford University)
Presenter: Aditya Gudipati
Notes by Fadel Adib, Danai Chasaki, and Vijay Gabale
Summary:

Aditya presented Strider, an automatic rate adaption and collision handling protocol. The authors classified previous approaches into passive and active schemes, both of which have limitations. In the former, throughput-maximizing rate adaptation must explicitly infer channel characteristics like packet loss, but the wireless channel strength can vary rapidly with time, causing inaccuracy. In the latter, coordination mechanisms like exponential backoff or RTS/CTS mechanisms are used, which can incur substantial overhead. Thus, a key question asked was: Can we avoid the overheads? (Owing to time constraints, Aditya presented only the bit rate adaptation part of the work.) He presented a rateless encoding and decoding framework which achieves automatic rate adaptation. The collision-resilient encoding algorithm is based upon the use of constellation points of the signal where the minimum distance between the constellation points determines the error rate. The authors ask, can we adjust the minimum distance without explicitly estimating channel strength? The solution uses a fixed-channel code and a minimum distance transformer (MDT) which sits after modulation and processes coded symbols. To mitigate the worst-case exponential complexity of decoding, the authors decode one BPSK symbol at a time. For the decoding scheme to work optimally unequal power allocation is necessary, but the authors show they can do systematic power allocation on a per-transmission basis offline. The authors' conclusion is that Strider is capable of achieving throughput nearly as high as that of an omniscient collision-free scheduler with minimal feedback that still performs well in fast fading channels.

Discussion:

Aaron Schulman (University of Maryland): Is the choice of 1 modulation scheme, QPSK, based on the fact that it works well on most wireless networks, i.e. better than, say, 16QAM? And can you achieve rates as those of 16QAM and 64QAM? Answers: We use a fixed channel code and modulation and then combine the packets. Effectively, Strider's performance depends on the number of bits in the packet. Yes, we can achieve rates such as 4.4bits/transmission using the rate values of k and the channel code. Follow-up: Do you have channel feedback? Answer: Yes, 1 bit of feedback to acknowledge a whole chunk, but no feedback about channel state.

Jianping Pan (University of Victoria): Do you use fixed modulation, or what is the source of randomness? Answer: As far as phases are concerned, we can assume a pre-calculated, shared source of randomness between sender and receiver. Follow-up: Do you have to measure channel state for power allocation? Answer: No, we do it without knowing the state. Follow-up: But when do you stop? Answer: We stop whenever we receive the ACK for the entire batch.

Clearing the RF Smog: Making 802.11 Robust to Cross-Technology Interference

Authors: Shyamnath Gollakota, Fadel Adib, Dina Katabi (MIT); Srinivasan Seshan (Carnegie Mellon University)
Presenter: Shyamnath Gollakota
Notes by Fadel Adib, Danai Chasaki, and Vijay Gabale
Summary:

Shyam presented TIMO, which allows 802.11n to communicate in the presence of high-power interferers from different technologies. Recent studies show that most of the Wi-Fi problems in home networks are due to high-power interference from non-802.11 technologies using the ISM band: e.g., baby monitors, cordless phones, and microwave ovens.

Prior research either tries to address interference from a particular technology, or hops to other frequency bands. This paper shows how a 2x2 MIMO receiver may still decode an 802.11 stream in the presence of such interference. The fundamental challenge for this work is that the receiver cannot decode the contents of the interfering, non-802.11 packets. However, Shyam showed that it is enough to characterize the angle of an interferer in the antenna domain to be able to decode in its presence. The proposed algorithm follows an iterative approach in finding this angle and afterwards nulling the corresponding interferer's signal.

In the presence of cross-technology interference, TIMO was capable of achieving very high throughput in scenarios where ordinary 802.11 completely lost connectivity.
Discussion:

[unknown] (MSR India): Was the reason for relatively low performance at the first 2 scenarios that the interference was absolutely high or because the beta (channel ratio) estimate was wrong? Answer: It is mainly because the channel estimate cannot be accurate in the presence of high interference, thus resulting in high residual interference.

Aaron Schulman (U. Maryland): How do you decode if there is a high BER in the preamble? Answer: Even if we have bit errors in the preamble, we can decode and recover the data. Our implementation considers 2 cases: (1) if the interferer starts before the desired signal, the angle beta can be calculated easily, (2) if the interferer starts after the signal of interest, we estimate the channel by correlating with the preamble, then subtracting out the desired packet data so that we are left with the interferer alone, and accordingly are able to estimate its direction.

Jonathan Perry (MIT): In order to recover beta, you have to hypothesize a correct beta, then check if it is correct? or could you have used error correcting code? A: That's one option, but yes, we use an initial estimate, then converge.

Eric Anderson (CMU): What was your baseline setup: 1x1 or 2x2 MIMO? Answer: The client is transmitting a single stream, but the receiver is decoding using diversity.

JK Lee (HP labs): How do you detect the presence of 802.11 transmission in the presence of another transmission; do you use energy jumps? Answer: As Kate noted in the previous talk, if a transmission is already occurring, detecting the beginning of a packet using energy jumps is much more difficult than using the preamble. In fact, using a preamble to detect the presence of an 802.11 transmission is much more accurate than a change in the energy.

[unknown] (University of Arizona) Is beta constant? Answer: No, in the presentation, for simplification, I assumed beta is constant. However, in our experiments, we always track beta through averaging consecutive symbols in the iterative process.

Brad Karp (University College London): In your talk, you gave examples of unidirectional transmissions; how do you handle bidirectional links? Answer: One might use a classifier to identify the interferer and act accordingly.

Session 6: Network Modeling

Design Space Analysis for Modeling Incentives in Distributed Systems

Authors: Rameez Rahman, Tamas Vinko (Delft University of Technology); David Hales (The Open University, UK); Johan Pouwelse, Henk Sips (Delft University of Technology)
Presenter: Henk Sips
Notes by Marshini Chetty and László Gyarmati
Summary:

In this paper, the authors propose a different approach for conducting a large-scale design space analysis of BitTorrent. Rather than focus on standard game-theoretic models that cover at most a handful of point solutions, and conduct a competitive analysis of these, they strive to comprehensively consider a vast set of points in the design space by instantiating a large number of model parameters. Across the many protocols that result, they conduct a simulation-based analysis, tournament style, to assess the pairwise fitness of the protocols. Their model captures the repeated game flavor of BitTorrent, explicitly modeling heterogeneous bandwidth classes and optimistic unchokes. They characterize the protocols under study by three outcome metrics: performance, robustness, and aggressiveness. They argue that standard BitTorrent is not a Nash equilibrium, but that a more robust BitTorrent variant that comes out of their framework called 'Birds' (that explicitly sorts peers on the basis of proximity to its own upload speed) is a Nash. They acknowledge that different abstractions lead to different results and raise the following questions: if we include more details, would the Birds analysis still hold, and what would make such an analysis robust? Their results were validated with instrumented BitTorrent clients.
Discussion:

Ratul Mahajan (MSR): What class of distributed systems does this approach work for? Here you have competing protocols that have a common abstraction and you can consider certain variants. Could you do something with link-state vs. distance-vector, for example? Answer: You need to have common ground that protocols work on to compare between them, and for now we can only do this comparison within BitTorrent or another P2P application that works with a different protocol. Follow-up: What is this common base? Answer: The common base is the inherent workings of the protocols itself, such as peer selection, use of buffer space, etc.

Dah-Ming Chiu (Chinese University of Hong Kong): In traditional game theory, they do consider different strategies but you do not, why? Answer: It doesn't make it useless, but if you want to know how good your protocol is, you need some solid ground upon which to conduct simulations. Nash analysis helps with more information in the system to make other decisions.

Marco Canini (EPFL): Are you not shifting the work into being inventive with respect to the search space? Answer: Yes, to some extent this is true. The parameters that are brought in reflect earlier ideas, or factors that you think may be relevant. The difference is doing so in a systematic way, while in other approaches you try to prove it against one or two other approaches.

Carey Williamson (U. Calgary): Can you speculate on the sensitivity of your results with respect to the scale of the model? Answer: There are some scale effects in play. Two peers for example, is very different. Follow-up: I'm thinking about scaling up – how about scale effects in larger systems, thousands or millions? Answer: Right.

Randy Baden (U. Maryland): Your notion of robustness is pitting protocols against each other, but some protocols are very poorly performing, can you discard these and does that affect the performance of the other programs? Answer: Might it be wise to filter out protocols? Yes, it's true – much more detailed analysis could be done, and that may change the conclusions.

Brighten Godfrey (UIUC): Could you comment on doing a design space analysis over many protocols vs. game theoretic analysis of one protocol over a larger strategy space? Answer: Took question offline.

How Many Tiers? Pricing in the Internet Transit Market

Authors: Vytautas Valancius, Cristian Lumezanu, Nick Feamster (Georgia Institute of Technology); Ramesh Johari (Stanford University); Vijay V. Vazirani (Georgia Institute of Technology)
Presenter: Vytautas Valancius
Notes by Jay Aikat, Marshini Chetty, László Gyarmati, and Vijay Gabale
Summary:

The authors' goal is to analyze and evaluate various pricing practices in the wholesale transit market for ISPs, assess whether better tiered pricing strategies exist, and (as the title suggests), analyze how many tiers are present in an efficient market. This paper first makes technical contributions towards achieving this goal: (1) developing a novel way of mapping traffic and topology data to a demand and cost model; and (2) instantiating and fitting this model to three large real-world networks.

In the market modeled, the sellers are the large ISPs, and the buyers are smaller ISPs, content providers, etc. Connectivity today is sold at a bulk blended rate: a single price is set in $/mbps/month, and buyers are charged each month on aggregate throughput, even though some flows are costly to service, and others are cheaper. Undifferentiated pricing is Pareto inefficient, as clients lack incentives to avoid costly destinations, and ISPs lack incentives to invest. The alternative is tiered pricing, exemplified today by regional pricing, and paid peering.

The speaker asked, how efficient is tiered pricing, and do ISPs benefit from more tiers? Building a sophisticated model that addresses the complexity of this market, and driving it with real data, the authors estimate demand functions and servicing costs. With a data-driven analysis, the authors find that tiered pricing does in-

crease profit, but the returns diminish as the number of tiers increases. At the sweet spot of 3-4 tiers, 90-95% of all possible profit from tiering is captured. The analysis also shows that ISPs must judiciously choose how they divide traffic into pricing tiers, and being cognizant both of servicing costs (used today) as well as traffic demands (not used today) is effective.

Discussion:

Henning Schulzrinne (Columbia): First, competition wasn't mentioned – this sounds like monopoly pricing. Answer: [referring to backup slide] Yes, it's hard to model competition. This may requires a game-theoretic approach or more data. When you model demand, you may wish to assume inherent changes, or as a residual demand from competitive effects.

Henning: This problem seems substantially similar to physical delivery services such as USPS, FedEx, etc., where presumably crack economists have worked on tiered pricing. How is this result different? Answer: This is an important question. We started this work assuming the Internet would be similar to the USPS. Looking at the data, the very big difference is that 50% of total demand on the Internet is local – entering and leaving the same router. So both the cost structure and the demand structure is very different, and that drives our models.

Sergey Gorinsky (IMDEA): Geography is definitely relevant, but what about time of day or other variances in demand? Answer: Another very good question. We tried to look at time of day, but it was hard to come up with convincing cost models. Cost could be two times more if sent during day vs. night. But still an open research area.

Ramesh Sitaraman (U. Mass, Amherst): CDNs are the largest ISP buyers, getting tiered pricing from the ISPs. Have you thought about how such an overlay can change your underlay pricing models? Answer: Again a very good question. We talked to content providers, and they do solicit this type of tiered pricing behavior.

Nicolas Christin (CMU): Couldn't one or two providers moving to tiered pricing create feedback effects that would impact the model? Answer: Yes, and again this is difficult to model. But you could model demand as a residual demand, and then use our model to find the pricing strategies.

The Evolution of Layered Protocol Stacks Leads to an Hourglass-Shaped Architecture

Authors: Saamer Akhshabi, Constantine Dovrolis (Georgia Institute of Technology)

Presenter: Constantine Dovrolis

Notes by Marshini Chetty, Vijay Gabale, and László Gyarmati

Summary:

Why is the Internet Protocol stack an hourglass, why do the protocols at the waist of the hourglass appear to be difficult to replace, and how can a new protocol survive at the waist?

The authors explained that they wrote this paper as part of an ongoing clean-slate vs evolution debate and that they were inspired by biological models. They asked, what happens at waist of the stack compared to other parts? They find that most innovation is at the upper and lower layers of the stack, not in the middle, which tends to be ossified. Many waist-level protocols didn't survive.

They asked how we can make the Internet architectures more evolvable. They developed a model called EvoArch to abstract the evolution of protocol stacks, pointing out that this is not necessarily the correct or only way to approach this problem.

In the protocol stack, there are L layers. An item at layer X chooses an item at layer X-1 with some probability. (The nodes at layer X-1 are the "substrates" for layer X; the nodes in layer X are the "products" for layer X-1.) Protocols become less general towards the top. Each node has an evolutionary value, computed recursively based on the products of that node. TCP has high evolutionary value because it is used by many higher layer applications. With this value, the model also captures competition between two nodes at the same layer. For instance, HTTP competes with FTP, but TCP does not with UDP because they have fewer products in common, and so lie below a competition threshold.

Protocol births and deaths are modelled via various processes. EvoArch uses a discrete time model, where time advances in rounds; each round includes birth, competition, and mortality.

Why does EvoArch generate an hour-glass shape? At the waist the generality probability of the nodes is close to 0.5, and since variability in products is maximized, protocols compete intensely with each other, causing the death of their competitors.

The talk concluded with remarks about what this means for current and futures Internet architectures, and suggested that the model may influence how we teach this architecture. Other fields of science (biology, social structures, etc.) also see hourglass effects.

Discussion:

Walter Willinger (AT&T Labs Research): Where is the uncertainty in the model? Answer: The uncertainty in the model is in the selection of substrates, which is probabilistic. Follow-up: In reality this [randomness] is not the case; what can you say about this? Answer: The model is an abstraction of reality, but the model works probabilistically in practice. Follow-up: If you buy that line, you can claim that details don't matter, but your model has zero explanatory power. Since the answer to your question was known, did you start with assumptions and then the magic happened? Answer: We wanted a model to give structure to the known answers.

Session 7: Neat Tricks

What's the Difference? Efficient Set Reconciliation without Prior Context

Authors: David Eppstein, Michael T. Goodrich (U.C. Irvine); Frank Uyeda (U.C. San Diego); George Varghese (U.C. San Diego and Yahoo! Research)

Presenter: Frank Uyeda

Notes by John Byers and Danai Chasaki

Summary: The authors consider settings in which distributed applications need to keep state information up to date and in sync, and need to do so efficiently, with minimal messaging overhead. When the state information is modeled as sets of items, then pairwise data synchronization becomes a set reconciliation problem: identifying and transmitting those items present in one set, but not the other. Existing schemes either require communication linear in n, the size of the union of the sets; only do approximate reconciliation; or have large computational overheads. The authors devise a new class of data structures called difference digests for set reconciliation built upon invertible Bloom filters and strata estimators that estimate the size of the differences. Difference digests use O(max(d, $\log n$)) communication, where d is the size of the set difference, and only linear computation. The authors demonstrate the effectiveness of their methods using large-scale simulations and also observe that a hybrid estimation scheme, coupling strata estimators with min-wise estimators, offers the best performance in practice.

Discussion:

John Byers (Boston University): Doesn't the connection to Tornado codes you describe indicate that non-uniform hashing might be reasonable, using a heavier-tailed degree distribution? Answer: Yes, we are looking at it further, and I agree that it seems to hold a lot of promise.

Jonathan Sobel (Cisco): You described a hybrid solution for strata. Do you also need a hybrid approach for differencing, in the unlikely event of inability to decode? Answer: Decoding may not succeed, but the probability is small. Depending on the size of differences you can boost the probability of decoding. For example, in the case of an error, you can re-run the algorithm, doubling the size of the filter.

Carey Williamson (U. Calgary): Do you require hash functions to hash to distinct cells? Answer: Yes. Right now, we enforce collision-freeness, but this may not be strictly required.

Ramesh Sitaraman (U. Mass, Amherst): Is there a natural generalization to a large number of sets? Answer: You can use pairwise comparisons, or a hierarchical comparison structure for synchronization. With just three of four sets, optimality is trickier. Follow-up: Is there a more elegant generalization; maybe for multisets? Answer: Multisets can put you in more complex situations, where there are more pitfalls. We are interested to see if there is a good way to do it, but we haven't investigated it much.

DOF: A Local Wireless Information Plane

Authors: Steven Hong, Sachin Katti (Stanford University)
Presenter: Steven Hong
Notes by Trang Cao Minh, Danai Chasaki, and Jeff Mogul
Summary:

This work looks at how a local wireless information plane, which provides detailed information about the RF environment, can enhance the design of future smart radios. Up to now, interference among devices using unlicensed spectrum has been managed "socially" – while current co-existence mechanisms (e.g., carrier sense, frequency hopping) prevent interference between devices of the same type, we have to use manual methods to avoid interference between different kinds of device types/protocols that share the unlicensed spectrum. (For example, WiFi is likely to disrupt nearby ZigBee communication.)

A smart protocol for coexistence would need to know what protocol types are operating nearby, how sensitive they are to interference, what bands they are operating on, and the spatial locations of various devices (to allow multi-antenna systems to null out interference). DOF (stands for *Degrees of Freedom*) builds a local wireless information plane, and provides these kinds of knowledge.

DOF is robust to signal SNR and to overlaps in time and frequency, and is computationally efficient. It uses raw time samples from the ADC, and does feature extraction to make its inferences. "Features" are repeating patterns, typical of any protocol, which are necessary for its operation and also are protocol-specific. DOF uses a spectral correlation function (SCF) to find these patterns efficiently, and support-vector machines for classification. This depends on a robust mechanism for inferring the number of interfering signals.

They ran experiments comparing DOF against several other methods, none of which can make all three inferences that DOF does. It outperforms each of these other methods, especially at lower SNRs, and delivers reasonably good accuracy.

Their smart-radio prototype uses DOF to make informed decisions at the MAC layer. It can implement policies such as "only use unoccupied spectrum," "also use spectrum occupied by a microwave oven," or "compete with other unlicensed protocols half the time." It gets higher throughput, with lower harm to other protocols, than a previous method (Jello).
Discussion:

[Didn't give name]: Have you tried to do this with wireless microphones? Answer: No. Follow-up: They only occupy 200KHz, so if you can detect them you can notch them out, but they move around a lot. Could you deal with this? Answer: Microphones tend not to use standardized protocols, so they are not a good match to DOF. Follow-up: when you have an adaptive frequency-hopping system, aren't you at cross-purposes? Answer: We do have to deal with this for Bluetooth, but we're using 1msec time windows, so we capture several hops, so we can identify these.

Biswaroop Mukherjee (RIM): Can you identify phase-modulated signals? Answer: Yes, the prior work suggests that you can. Follow-up: Not sure I agree; you rely on periodicity, but OFDM gets less cyclic when you get to large numbers of users. Answer: If you have more users, you can process smaller chunks of bandwidth.

Erran Li (Bell Labs): If you have multipath, do you have problems finding the angle? Answer: We found that you do need strong line-of-sight components to accurately determine the angle of arrival. You can still estimate the multipath components.

Anja Feldman (TU-Berlin/T-Labs): How accurate is DOF if you have thousands of devices in a small area, and they are moving around? Answer: We did our traces with devices moving at walking speeds. If they were moving at vehicle speeds, they would probably be out of range before you could adapt to them.

Session 8: Data Center Network Performance

Towards Predictable Datacenter Networks

Authors: Hitesh Ballani, Paolo Costa, Thomas Karagiannis, Ant Rowstron (Microsoft Research)
Presenter: Hitesh Ballani
Notes by Ming-Hung Chen, S. H. Shah Newaz, and Chun-Yu Yang
Summary:

In multi-tenant (cloud) datacenters, providers would like to offer each tenant the abstraction of a virtual network with predictable bandwidths and costs, while maintaining flexibility for the provider. The paper proposes two abstractions, Virtual Cluster (VC) and Virtual Oversubscribed Cluster (VOC). In a VC, there is no oversubscription of links in the tenant's virtual network; this is most suitable for data-intensive applications (e.g., MapReduce). In a VOC, a tenant's VMs are organized into groups, with full bandwidth between the VMs within a group, but with a maximum oversubscription ratio (O) on the links between the groups. The tenant can specify the group size and the ratio O. A VOC provides more flexibility for the provider, and hence lower costs for the tenant.

Oktopus is a system that implements these abstractions. It implements online allocation algorithms to place VMs so as to maintain the guarantees requested by the tenants. Rate limits are enforced in the hypervisors, based on shares set by a central controller.

The authors evaluated Oktopus on a 25-node two-tier testbed, as well as a larger-scale simulation, and showed that it improves job completion time and reduces the number of tenant requests that must be rejected, relative to a baseline design. They find that the abstractions can reduce tenant costs by up to 74%, while maintaining provider revenue neutrality.
Discussion:

Ramana Kompella (Purdue): Some people don't like the idea of bloating the hypervisor. Is it possible to implement this functionality on switches instead? Why are we afraid of putting functionality in the switches? Answer: Yes, but it would require future switches that can handle per-flow or per-tenant state. Existing hypervisors already support some of what we needed, and we think of end-hosts as integral parts of the network, with resources that we should use.

Robert Escriva (Cornell): This work assumes the providers are honest; what if they lie, and create more oversubscription than the tenants request? Answer: We don't think that will happen, because

if you can satisfy more tenants, you will get more revenue. We think verification of a provider's honesty is easier with Oktopus than with the status quo (of no guarantees).

Charlie Hu (Purdue): How did you decide that these two abstractions are necessary and sufficient? Answer: Our motivation was to strike a balance between tenant suitability and provider flexibility. We think our abstractions match expectations about existing network structures. However, one might need richer abstractions to map the needs of complex applications.

Ang Li (Duke): Have you considered other resources, such as CPU and disk I/O, that would affect job completion time? Answer: We have not given them explicit attention; we wanted to start with a building a predictable network layer. We think the virtual-network model might be a concise and elegant means for tenants to express their storage bandwidth needs.

Dantong Yu (Brookhaven): Can this handle the large and variable bandwidth requirements resulting from multi-stage MapReduce jobs? Answer: Not yet. Maybe it could be done by allocating first for one stage, and then again for migrating the data.

DevoFlow: Scaling Flow Management for High-Performance Networks

Authors: Andrew R. Curtis (University of Waterloo); Jeffrey C. Mogul, Jean Tourrilhes, Praveen Yalagandula, Puneet Sharma, Sujata Banerjee (HP Labs)

Presenter: Andrew Curtis

Notes by M.-H. Chen, C.-Y. Yang, and S. H. Shah Newaz

Summary:

The speaker first motivated the need for an improved version of OpenFlow. OpenFlow enables a programmable network, but its design imposes excessive overheads due to the presence of a centralized controller. Although many articles have tried to address this problem, those solutions are hard to scale to datacenter-sized networks. This paper characterizes the overheads of implementing OpenFlow in hardware, proposes DevoFlow to enable cost-effective, scalable flow management, then evaluates it on the problem of data center flow scheduling.

The key to DevoFlow's operation is efficient devolved control and statistics collection, which avoids bottlenecks associated with centralized control, but preserves benefits such as near-optimal traffic management, and a simple switch mechanism. The authors propose two new mechanisms for devolving control to a switch. The first is "rule cloning," in which the switch locally clones a wildcard rule to create a new rule that instantiates the wildcarded fields by values matching a microflow, thereby avoiding most of the TCAM power cost. The second is "local actions," which relies on sampling and triggers to avoid invoking the central controller, for example by detecting elephant flows and handling them separately via thresholding.

The authors built a simulator and compare DevoFlow to ECMP and OpenFlow. Experimental results first confirm that DevoFlow performs as well as fine-grained flow management when load-balancing traffic in the data center. They then go on to show that DevoFlow improved throughput by 37-55% in two large test topologies, while reducing control traffic to 1% of that of OpenFlow, and achieving comparable reductions in flow table entries using multipath. The authors assert that DevoFlow can simplify the design of high-performance OpenFlow switches and envision enabling scalable management architectures to be built on OpenFlow for data center QoS, multicast, routing-as-a-service, network virtualization, and energy-aware routing.

Discussion:

Minlan Yu (Princeton): The more functions we add to switches, the better performance we get. Of the many design points between fully centralized and fully distributed, how can we pick the right one? Answer: This is definitely challenging. We focused on data centers and spoke to ASIC designers to understand what was possible there, but it is open to debate.

Simon Crosby (Bromium): If you do control plane separation from the forwarding plane, how quickly does the system degrade over time, making reasonable assumptions about how traffic changes. Answer: It depends on many things including workload, how quickly you collect flow statistics, etc. The takeaway is: as you set up one arbitrary deadline, that you would need two orders of magnitude more bandwidth (with prior methods).

Anja Feldmann (TU Berlin/T-Labs): In principle, there are many ways to define flows, but your definition seems to be over all 5-tuples? Answer: Yes, that is correct. Follow-up: Can you just do sampling using some of the inherent randomness in packet header fields, to create a reactive OpenFlow? Answer: It may be tempting to say that DevoFlow is just OpenFlow plus packet sampling, but we give more mechanisms for fine-grained control. Follow-up: Have you taken rather extreme cases to point out limitations in OpenFlow? Answer: That seems fair to say.

Amin Tootoonchian (U. Toronto): Based on quick calculations, your results seem to have changed since the HotNets version of your paper. Can you comment? Answer: Our numbers improved based on a better implementation. Let's look at details offline.

Improving Datacenter Performance and Robustness with Multipath TCP

Authors: Costin Raiciu (University College London & University Politehnica Bucharest); Sebastien Barre (Université catholique de Louvain); Christopher Pluntke, Adam Greenhalgh, Damon Wischik, Mark Handley (University College London)

Presenter: Costin Raiciu

Notes by Ming-Hung Chen, S. H. Shah Newaz, and Chun-Yu Yang

Summary:

Multipath TCP (MPTCP) aims to provide fair multipath transport in multipath topologies. The authors argue that the traditional approach, of mapping each flow to a single path, results in poor performance on a multipath topology. Experimental results show that single-path TCP flows that collide on a congested link reduce throughput, relative to what is possible, and it is impossible to be fair to all such flows.

MPTCP spreads the data from a socket-level connection onto a number of subflows, each of which follows a randomly-chosen path. Each subflow uses standard congestion-control mechanisms, so each path can be utilized as much as possible.

Experiments using several multipath topologies (EC2, Fat Tree, and Cisco) show that MPTCP can significantly improve aggregate network throughput, and that at most 8 subflows are needed. For the VL2 and BCube topologies, MPTCP also significantly improves fairness.

The authors also looked at which topologies would be best to use with MPTCP, to survive ToR switch failures and to reduce bottlenecks on the end-host links when the core is underloaded. They suggest modifying a Fat Tree by moving links from the core to the edge, and this Dual-Homed Fat Tree (together with MPTCP) can provide significant throughput improvement when the core is not overloaded.

Discussion:

Praveen Yalagandula (HP Labs): How do you decide how many sub-flows to use for each flow? Answer: This will be a constant

provided by the data center operator. Also, you should not enable multipath for short flows; you can use a timer to avoid multiple subflows until after (say) 100msec. Follow-up: is this constant proportional to the number of paths in the topology? Answer: yes, the worst case is Fat Tree because it has K^2 paths. For other topologies, you need a lot less.

Navendu Jain (MSR): On EC2, are you getting benefits due to multiple paths, or to spare capacity? That is, if you open multiple TCP flows on the same path would you get the same throughput? Answer: I don't know. My contacts in Amazon couldn't tell me. We haven't tried running multiple flows on the same path.

Jianping Pan (U. Victoria): How do you know whether Amazon is providing multiple paths on EC2? Answer: by running traceroute with different source and destination ports, which causes ECMP to choose different hops along the path.

Kang Xi (NYU-Poly): How does "one flow, many paths" affect security, since one-flow, one-path allows one middlebox to intercept the entire flow? Answer: You can't use middleboxes like this with MPTCP. In data centers, that might not be a big concern [Dave Oran interjects: some people would consider that a feature]. In general, the firewall would be close to one of the end-points; you could terminate the multiple flows at the firewall and run a single-flow connection to the destination. A firewall in the middle could disable MPTCP by removing those options in the SYN handshake.

Session 9: Network Management – Reasoning and Debugging

NetQuery: A Knowledge Plane for Reasoning about Network Properties

Authors: Alan Shieh, Emin Gün Sirer, Fred B. Schneider (Cornell University)

Presenter: Emin Gün Sirer

Notes by Katrina LaCurts and Trang Cao Minh

Summary:

In his talk, Emin Gün Sirer aimed to convince us that current protocols aren't expressive enough. Existing networks don't allow us to query properties of participants, and the commoditization of networks prohibits us from inferring how the data plane varies. NetQuery aims to solve these problems with a federated, distributed knowledge plane.

NetQuery's knowledge plane consists of a tuple-store that contains information about network elements, for instance $< type = host, OS = Linux >$. The tuple space is stored on disparate servers, and the information in these tuples can come from NetQuery itself, network administrators, or third parties. To solve the problem of trust, a TPM provides attestation chains, and applications can specify which principals they trust.

In addition to this distributed tuple space, NetQuery provides a logical framework, allowing users to reason about the tuples. NetQuery can provide an answer as well as a proof to questions such as "Do I have a good VoIP path?"

Gün and his students have built a prototype of NetQuery, and have shown its practicality for answering queries regarding topology quality, BGP configuration, and access control.

Discussion:

S. Keshav (Waterloo) inquired about the uniqueness of tuple IDs (which are of the form IP:port:ID), and Gün stated that they require unique IPs. It might make sense to change that if you were to redesign the system.

Stefan Sariou (MSR) brought up the point that many ISPs go to great lengths to lie about their services, and might not want to reveal information to NetQuery. Gün pointed out that there are many properties of a network that an ISP should be happy to reveal, to prove that they are better than other ISPs, and to support any PR claims. There might be other properties that ISPs would not want to reveal, of course. Moreover, we need a channel (for making queries about the network) to be present, before we have to worry about what would happen once we had the channel.

Stefan Sariou: But for ISPs to market themselves, do you really need all of this mechanism? Answer: It would be much better to have a technical basis for our decisions, rather than basing our decisions on PR. We've talked with ISPs who are proud of their network, and who have an incentive to compete based on facts, against other ISPs who compete based on PR.

Alex Gurney (U. Penn): You pointed out that there are certain kinds of cross-domain queries where the sanitization doesn't apply, and so to maintain confidentiality you could use secure multiparty computation. That seems very expensive. Answer: This is still ongoing work. Problems for our system occur when portions of the proof tree rely on different ASes, and parts of the proof tree have to be revealed. There might be special cases where this is OK; in the general case, we are down to SMC, which is expensive.

Karen Sollins (MIT): How would you make this work at global scale? Don't you need a global PKI, and a global ID space not tied to server IDs (since I might not have a specific physical location to send the query?) Answer: there is a distinction between where you get your tuples, and how you check them. They aren't secure because you got them from a particular address, they're secure because they've been certified via a PKI. [Session chair asked that the rest of the response be done offline.]

Debugging the Data Plane with Anteater

Authors: Haohui Mai, Ahmed Khurshid, Rachit Agarwal, Matthew Caesar, P. Brighten Godfrey, Samuel T. King (University of Illinois at Urbana-Champaign)

Presenter: Haohui Mai

Notes by Katrina LaCurts and Trang Cao Minh

Summary:

In this talk, Haohui Mai started off with some examples of why network debugging is hard. Even well-managed networks – such as the SIGCOMM conference network – can go down due to legacy devices, protocol inter-dependencies, security policies, etc., and tools like `ping` and `traceroute` only provide us with limited testing functionality. Previous work on configuration analysis is complicated, and misses implementation bugs.

Haohui's approach, Anteater, allows users to debug the data plane. This approach provides a unified analysis for multiple control planes, and can catch implementation bugs. At a high level, the network operator provides data plane state and invariants, which Anteater translates to SAT formulas, and then solves. The data plane is expressed as boolean functions $P(u, v)$, which specify the policy for packets traveling from u to v. Anteater can also express invariants as SAT formulas (e.g., reachability), as well as packet transformations.

On a real network at UIUC, Anteater was able to reveal 23 bugs within two hours, and the paper gives concrete examples of certain bugs (routing loops, for instance), which turned out to be relatively easy to fix. Overall, Anteater provides a network debugging approach that is practical for nightly tests on a network.

Discussion:

Srinivasan Keshav (Waterloo) pointed out that the SIGCOMM network failed because a NAT table overflowed due to a large timeout, i.e., because of load and a misconfiguration; this problem does not seem like one that Anteater could solve. Haohui clarified that Anteater cannot solve problems that are not in the data plane, and

the session chair (Sachin Katti, Stanford) remarked that it would be nice to outline the types of problems that Anteater cannot solve.

[Unknown] recommended another IP reachability paper that uses a different solver and a FIB table analysis, but Haohui described the differences, stating that they handled packet transformations in general with a real implementation.

Aditya Akella (Wisconsin) asked why the emphasis on data plane analysis, given that static analysis has the benefit of testing before deployment? Haohui stated that the two methods look at different levels; here, it is easier to look at the data plane for routing loops, or other problems that are closer to the actual network behavior. Aditya pointed out that some of these things could be simulated beforehand if using static analysis.

Demystifying Configuration Challenges and Trade-Offs in Network-based ISP Services

Authors: Theophilus Benson, Aditya Akella (University of Wisconsin, Madison); Aman Shaikh (AT&T Labs - Research)
Presenter: Theophilus Benson
Notes by Katrina LaCurts and Trang Cao Minh
Summary:

Theo's goal in this talk was to explain the impediments that exist towards improving service migration and management for ISPs, given that ISPs have moved toward network-based services. By analyzing over 2.5 years of configuration files, he was able to answer questions about where the complexity lies and how it grows over time, using metrics from his 2009 NSDI paper.

Complexity starts at the provider's edge, due to the need to maintain consistency across devices. Over time, this complexity increases, mostly due to old devices becoming more complex. Configuration reuse does not help much, even at the customer's edge, since devices are becoming more specialized. In the core of the network, there is less complexity, particularly in the data plane. In order to mitigate complexity, Theo suggested that vendor selection should take this factor explicitly into account, not just more traditional metrics of functionality and cost. He noted that different languages impact complexity, and by finding customers implementing the same policies across providers, he was able to show that some vendors are consistently complex (across customers and policies).
Discussion:

Sachin Katti (Stanford) first commented that traffic has increased linearly over recent time, and then asked if complexity scales at the same rate as traffic. Answer: Traffic scales for two reasons: first, existing customers' usage increases over time, and second, ISPs gain new customers over time. He said that increasing complexity is largely due to the latter factor, and therefore, complexity may actually scale more slowly than traffic.

Nick Feamster (Georgia Tech) commented that this was an interesting and challenging problem, then asked how one could quantify that an implemented policy reduces complexity for customers. Answer: If the implementation only involves changing the high-level language, one could use the same or similar metrics as in this paper. However, a clean-slate redesign or a change in the high-level abstraction would probably need new techniques. Follow-up: While referential complexity is a useful metric, it seems like it is but a single metric, and identifying other metrics could be beneficial.

Srikanth Kandula (Microsoft) asked how these metrics related to semantic complexity (what operators have in their head) using an example of auto-configuration of bandwidth. Answer: In previous work (NSDI 2009), the authors showed that semantic complexity matches up with these metrics, using interviews with operators as one basis.

Session 10: ISPs and Wide-Area Networking

Seamless Network-Wide IGP Migrations

Authors: Laurent Vanbever (Université catholique de Louvain); Stefano Vissicchio (Roma Tre University); Cristel Pelsser (Internet Initiative Japan); Pierre Francois, Olivier Bonaventure (Université catholique de Louvain)
Presenter: Laurent Vanbever
Notes by Praveen Yalagandula
Summary:

Router reconfigurations are complex, because one needs to respect SLAs. Highly distributed changes need very coordinated modifications The paper addresses the specific problem of replacing an anomaly-free IGP configuration of a running network, router-by-router, without causing any routing anomalies.

A solution is possible if one follows a certain strict ordering, such that no forwarding loops happen during the reconfiguration. Finding and deciding if such an ordering exists is an NP-complete problem. They use a heuristic to find an ordering, which worked on most of the topologies that they considered, without packet loss. This heuristic also considers failures during reconfigurations. They have applied the method to networks with real routers.

Future work includes extending the work to handle reconfigurations of BGP, MPLS, etc.
Discussion:

Walter Willinger (AT&T Labs): Did you gain some insight into what types of problems are hard to solve this problem for, and which ones are easy? Answer: We describe these in the paper; the difficulty of finding solutions is related to the amount of path changes between the two configurations. Also, a good design (guidelines in the paper) should not lead to ordering issues.

[Unknown]: What about a really large topology? Is it hard to get the inputs? Answer: We have tested it on a really large Tier-1 network, with more than 1,000 routers. It's easy to get the inputs, you just need to compute shortest paths – not to simulate BGP decisions.

A Content Propagation Metric for Efficient Content Distribution

Authors: Ryan S. Peterson, Bernard Wong, Emin Gün Sirer (Cornell University)
Presenter: Ryan S. Peterson
Notes by Jeff Mogul
Summary:

The authors' goal is to build a high-performance P2P content-distribution service, which involves origin servers, end users organized into swarms, and cache servers. Most existing P2P protocols are not really optimized for including large servers into these swarms; their prior work on Antfarm was able to do optimize the use of origin-server resources. But Antfarm neglects the efficiency gain from using cache servers or overlapping swarms (which are somewhat common in BitTorrent).

The problem is to find an allocation of each host's bandwidth among swarms so as to obtain a global optimum. This is difficult because it is hard to define a metric for this global optimum. The paper is about a new metric, the Content Propagation Metric (CPM) that steers hosts towards a globally efficient allocation. Each host can make this measurement, and so can make local decisions that yield global efficiency – for each choice of a block to upload, it can choose the swarm for which the CPM is maximized.

CPM is based on historical data; it measures how fast ("block propagation bandwidth") the host's blocks propagate in each

swarm, as a rolling average for each swarm. Hosts proactively track blocks to deal with churn.

They built a system called V-Formation, based on Antfarm, that uses the CPM. It involves a Coordinator function, itself built as a distributed system, to track the propagation data. The evaluated V-Formation using a video-sharing service (`http://flixq.com`). They used 380 peers on PlanetLab, and a coordinator in EC2, with 20% of the peers in multiple swarms. With BitTorrent, they got about 3MB/sec of aggregate bandwidth; Antfarm converges to a higher bandwidth, but more slower. V-Formation converges almost as fast as BitTorrent, but to an even higher level (almost 6MB/sec). The logically-centralized coordinator scales linearly to thousands of peers.

Discussion:

Stefan Saroiu (MSR): How often do you have a case where there are two caches, and where some data is in one cache but not in another? Answer: Our focus was on how large service providers could use well-provisioned cache servers to help exploit bandwidth at the edge.

Walter Willinger (AT&T Labs): How would this apply to CDN brokers? Answer: Our efficient allocation might well span across multiple organizations; there is other work to do to figure what data to cache at each such server.

Ramesh Sitaraman (U. Mass. and Akamai): An industry perspective: with respect to video distribution, bandwidth is an important metric, but user experience is more important to the business side. That is why P2P has not been popular for enterprise streaming. Have you thought about metrics that capture user experience, reliability, re-buffering? Answer: We chose aggregate bandwidth because it has nice properties such as maximizing download completions, not specifically for streaming video.

John Otto (Northwestern): Did you notice any client-perceived performance improvement? Answer: Yes, especially in swarms that might otherwise be neglected.

Insomnia in the Access or How to Curb Access Network Related Energy Consumption

Authors: Eduard Goma (Telefonica Research); Marco Canini (EPFL); Alberto Lopez Toledo, Nikolaos Laoutaris (Telefonica Research); Dejan Kostić (EPFL); Pablo Rodriguez (Telefonica Research); Rade Stanojević (Institute IMDEA Networks); Pablo Yagüe Valentín (Telefonica Research)

Presenter: Marco Canini

Notes by Praveen Yalagandula

Summary:

The "access network" is the portion of the network from home gateways to the first hop of their ISP. Even though each gateway and DSLAM uses only a small amount of power, there are many of them, so their total power consumption adds up. Also, these devices have very high zero-utilization energy usage (base energy usage). Most of the traffic on these networks is very small, but non-zero; the Sleep-on-Idle technique does not solve fully the problem of reducing this energy consumption.

This paper introduces two techniques: (1) traffic aggregation at the user end; on average 5-6 home WiFi networks overlap, so one can put some gateways to sleep. They designed a distributed algorithm, broadband hitch-hiking (BH^2), to decide which gateways to turn off; and (2) line switching on the ISP end: they employ several 4-way switches to switch the subscriber lines to as few line cards as possible.

They also found that turning off some DSL modems improves the performance of others, by reducing cross-talk. Together, these techniques have the potential to save 66% to 80% of the access network energy.

Discussion:

Justin Ma (Berkeley): When you consider the manufacturing energy cost of those new switches that you put on the ISP side, does this really save overall energy? Answer: Good point, we have not considered that. But one thing is that ISPs already have many of those kind of boxes, that they use for management purposes, and might re-purpose. An ISP might also consider a static re-assignment of users to DSLAM ports, to save part of the energy.

Jukka Manner (Aalto Univ.): My intuition is that the reason my gateway cannot go to sleep is not my own packets, but the people doing port-scanning. Answer: We do not assume that downlink traffic would prevent the gateway from sleeping.

Jitu Padhye (MSR): You force users to associate with an AP that is further away. Does this affect other metrics for users? For example, will it affect low-latency applications? Or are you moving their connections? Answer: Note that the user will use a neighbor's gateway only when there is low traffic.

Jitu Padhye: What about low-bandwidth applications, such as VOIP, that are very sensitive to latency and packet drops, which would increase if the AP is further away? Answer: In our experiments, we did not observe any problems with video applications and ssh connections, but maybe certain applications need to be white-listed.

Srinivas Narayana (Princeton): Since users will be using other people's gateways, does this require changes to the ISP's charging infrastructure? Answer: We suggest a technique that requires modifying the user's wireless driver, which might be done as part of the home gateway.

S. Keshav (Waterloo): An observation: the absolute power saving is a very small fraction of overall power usage of these users. Answer: Our focus is on the overprovisioning of the access network, and making this more energy-aware and power-proportional.

Keshav: You're assuming that the gateway power is a given; perhaps it would be better just to use a more efficient chip set? Answer: Potentially yes, but at what cost to replace these gateways?

[Unknown]: Have you considered DVFS techniques to reduce energy use and increase proportionality? Answer: Yes, but in CMOS today, the leakage problem is making DVFS less effective.

Marwan Fayed (U. of Stirling): There are many devices in a home which are communicating with each other, such as audio components. Will they end up connecting through different gateways? Answer: Again, this suggests the use of a white-list for applications that should not use other APs.

Session 11: Network Measurement II – What's Going On?

Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications

Authors: Phillipa Gill (University of Toronto); Navendu Jain, Nachi Nagappan (Microsoft Research)

Presenter: Phillipa Gill

Notes by Marc Mendonca

Summary:

The presenter started from the premise that data center outages, when they do occur, have very high impact and are costly (estimates run into thousands of dollars per minute). We can improve reliability by understanding failures: characterizing them, quantifying their impact, and measuring the effectiveness of redundancy.

The authors' study was based on one year of network event logs (2009-2010) – a combination of syslogs, SNMP traps, trou-

bleshooting tickets, and resolution events. The failures of interest were device failures, where devices stop forwarding traffic; and link failures, where connections between two interfaces went down. One of the technical contributions was to reliably extract impactful failures from a large dataset that included many unimportant events as well as spurious alerts. In characterizing failures, the authors observed that load balancers experienced a large number of failures, but relatively low overall downtime. Top of rack devices had lower failure rate (5% probability of failure per annum), but very high downtime. Hardware and software faults equally contributed to overall failure rate, but hardware failures had much higher downtime. To quantify the effectiveness of redundancy, the authors used heuristics to compare traffic before and after failures in the presence of redundancy to estimate a normalized traffic ratio. Using this measure, they found redundancy to be very effective in masking link failures in the core, and least effective at the aggregation switch, due to errors like configuration errors that redundancy cannot remedy. Overall, the authors estimated a traffic increase of 40% due to redundancy.

Discussion:

Henning Schulzrinne (Columbia): Are your statistics sufficient to differentiate different types of devices? Answer: I can't name specific vendors, but we do have the capability to isolate measurements across vendors. For example, we observed a newer software version of a load balancer leading to a reduction in failures.

Karen Sollins (MIT): Is there a way to measure increased fragility in the system as given elements providing redundancy fail? Answer: This is a very good point, and we tried to look at it, but we didn't have a large enough set of redundant devices to gather significant data.

Jeff Mogul (HP Labs): Did you have enough syslog information to detect failures that occurred shortly after reconfiguration, which might indicate certain causes of failures. Answer: No, we didn't have access to scheduled reconfigurations, just triggered reconfigurations.

Understanding the Impact of Video Quality on User Engagement

Authors: Florin Dobrian, Asad Awan (Conviva); Ion Stoica (U.C. Berkeley/Conviva); Vyas Sekar (Intel Berkeley); Aditya Ganjam, Dilip Joseph, Jibin Zhan (Conviva); Hui Zhang (CMU/Conviva)

Presenter: Hui Zhang

Notes by Vijay Gabale and Marc Mendonca

Summary:

The speaker started by stressing that we are at the beginning of the Internet video era and motivated the benefits of understanding which technological factors (beyond the intrinsic quality of the content) most significantly impact user engagement.

The authors conducted a measurement study, using one week of data from multiple premium video sites. The authors studied three genres: live, live VoD, and stored VoD, using five quality metrics: buffering ratio, rate of buffering, join time, rendering quality, and average bit rate. Video quality was measured at the video player, as close to the end user as possible. The study found that buffering metrics were the most critical: when buffering events (such as playout interruptions or jitter) occurred at a rate of 0.2 events per minute, user engagement fell from 45 minutes to 10 minutes, on average. Live video users were the most sensitive to buffering issues: a 1% increase in buffering ratio reduced their play time by 3 minutes on average. Across all video genres, average bit rate and buffering rate also significantly impacted engagement. Based on this large-scale study, the authors were able to quantify which metrics most impacted engagement, and how these impacts manifested themselves for different genres of video. The presenter closed with the statement that in the near future, viewers will have zero tolerance for poor quality.

Discussion:

Henning Schulzrinne (Columbia): Did you see evidence that individual users experienced significant variabilities in quality, and should perhaps upgrade their network bandwidth? Question: This study was mostly about individual users, but we do have all the data to potentially isolate and localize where the issues were, for example a city-wide issue.

John Otto (Northwestern): Did you consider user reactions to a bad event, such as a user giving up after an outage? Answer: We have the user-level data, so we can study user behavior in isolation, Our follow-up work will look into issues like what triggers an event like a user leaving.

Dah-Ming Chiu (Chinese University of Hong Kong): Traditionally, people do these studies through subjective tests, but how do you calibrate your more quantitative metrics? Answer: Excellent question. Opinion scores are hard to do at large-scale. One common factor for us is that we instrument and measure the views of the same piece of content.

David Oran (Cisco): Did you instrument all types of video players? And were there any differences? Answer: Yes. First, there are different platforms (Flash, HTML5), and within those platforms, there are many players, which exhibit slight variations but the overall trend was the same.

Marco Marchisio (Telecom Italia): Did you perform tests with mobile users? Answer: Most of our data was from 2010, but in 2011, we do see many more mobile users. We see similar results, but this is very preliminary.

Bernard Wong (Univ. of Waterloo): Many metrics seem to be highly correlated. What would be the ideal or most revealing metric to user engagement? Answer: Our study focused on network-centric, packet-level metrics. There are different coding techniques, and other video-specific issues that would be interesting to study. In the paper we do talk about information gain from each metric, but it is all correlation-based, and we cannot infer causality.

An Untold Story of Middleboxes in Cellular Networks

Authors: Zhaoguang Wang, Zhiyun Qian, Qiang Xu, Zhuoqing Morley Mao (University of Michigan); Ming Zhang (Microsoft Research

Presenter: Zhaoguang Wang

Notes by Marc Mendonca

Summary:

The goal of this work was to develop a client tool, NetPiculet (in this case, for Android) that accurately infers NAT and firewall policies in cellular networks, and to understand the impact and implications of these policies. The policies that they looked at include IP spoofing, TCP timeouts, out-of-order packet buffering, NAT mapping type, endpoint filtering, TCP state tracking, filtering response, and packet mangling.

They studied 393 users from 107 carriers over a period of two weeks. They found that 4 out of 60 carriers allow IP spoofing (a potential security vulnerability). Some cellular firewalls buffer out-of-order packets, and therefore disallow fast retransmit; this can lead to performance degradation.

The tool can be downloaded from `http://mobiperf.com`.

Discussion:

Tatsuya Mori (NTT): Were you able to eliminate clients connected to WiFi networks? Answer: Yes, we can see which network the client is using.

Dina Papagiannaki (Telefonica Research): As you are using a single server at Michigan, which may be far away from the clients, you might be crossing multiple middleboxes. How do you know that you are actually studying the cellular middlebox, rather than a random one along the path? Answer: We have no middlebox on our university's upstream path, though there might be other middleboxes on the path. But we see the same results from users at many locations, which gives us some confidence in the results.

Dirk Kutscher (NEC Labs): There are other kinds of middleboxes besides firewalls and NAT. Some of your results (especially TCP reordering) may have been caused by TCP proxies, used by operators to optimize performance.

[Unknown]: Do providers give out public IP addresses to mobile clients, and do you see direct mobile-to-mobile communications? Answer: Yes, some carriers do assign public IPs to phones; in the US, only Sprint does this. There can be a direct phone-to-phone IP path, those packets are routed via a gateway. Some networks block these paths, though.

Jukka Manner (Aalto Univ.): Related to your power consumption figures, what were the T1 and T2 timers in the UMTS setup? The figure is totally dependent on the network and the device. Answer: We generalized by using the power model of ATT's network. The first timer is ca. 11 seconds and the second timer is ca. 6 seconds. We looked at published papers and used that.

Lars Eggert (Nokia): In your spoofing test, your mobile sends spoofed packets to your server, but you don't test the other direction. You might have missed providers who filter in the direction that would matter for a battery-exhaustion attack. Answer: we expect the anti-spoofing rules would be symmetric (yes, we should actually measure that.)

Lars Eggert: How can you tell if middlebox buffered out-of-order TCP packets or simply dropped them? You might not be able to tell. Answer: We know because when lost packet was retransmitted, the lost packets were released.

Lars Eggert: a lot of what you are seeing is explained by the requirements for "legal intercept."