Elephants, Mice, and Lemmings! Oh My!

Making life better in data centers and high speed computing

Fred Baker
Fellow

3 July 2014

I’m going to touch on three things that are changing as we speak – and research needs to be in front of

- Incast and Map/Reduce
- Network Simplification
- **Routine Attacks, Backdoors, and Government Snooping**

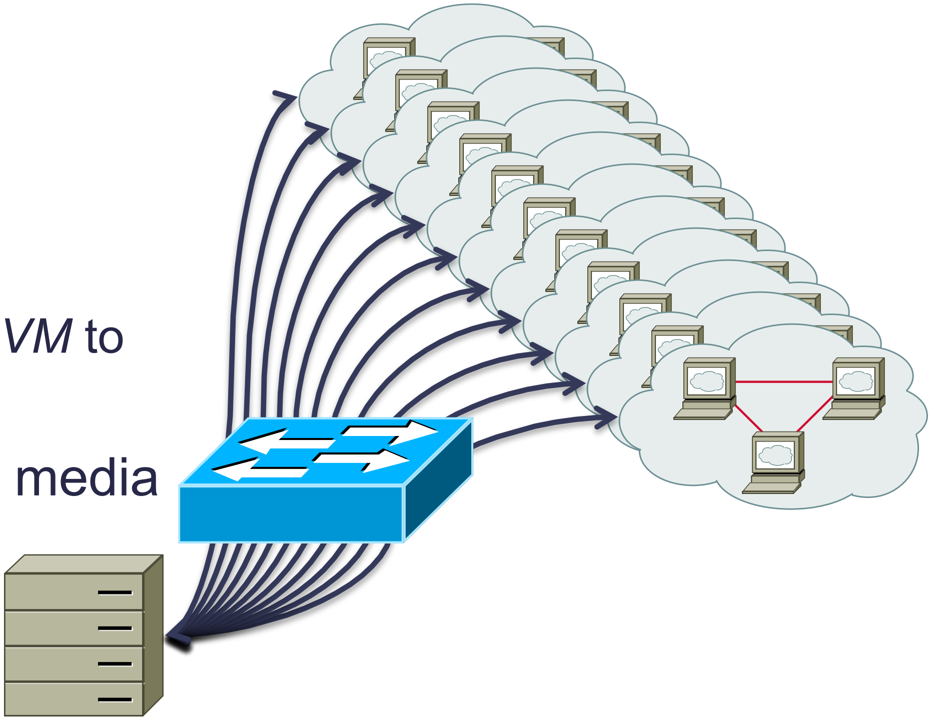
Data Center Latency Control

Persistent Deep Queues

- In access paths (Cable Modem, DSL, Mobile Internet)
  - Generally results from folks building a deep queue with permissive drop thresholds
  - One DSL Modem vendor provides ten seconds of queue depth
- In multi-layer networks (WiFi, Input-queued Switches)
  - ***Channel Acquisition Delay***
  - Systems not only wait for their own queue, but to access network
  - In WiFi, APs often try to accumulate traffic per neighbor to limit transition time
  - In Input-queued switches, multiple inputs feeding the same output appear as unpredictable delay sources to each other
  - In effect, **managing *delay* through queue**, not queue depth

Data Center Applications

- Names withheld for customer/vendor confidentiality reasons
- Common social networking applications might have
  - $O(10^3)$ racks in a data center
  - 42 1RU hosts per rack
  - A dozen Virtual Machines per host
  - $O(2^{19})$ virtual hosts per data center
  - $O(10^4)$ standing TCP connections *per VM* to other VMs in the data center
- When one opens a <pick your social media application> web page
  - Thread is created for the client
  - $O(10^4)$ requests go out for data
  - $O(10^4)$ 2-3 1460 byte responses come back
  - $O(45 \times 10^6)$ bytes in switch queues **instantaneously**
  - At 10 GBPS, **instant 36 ms queue depth**

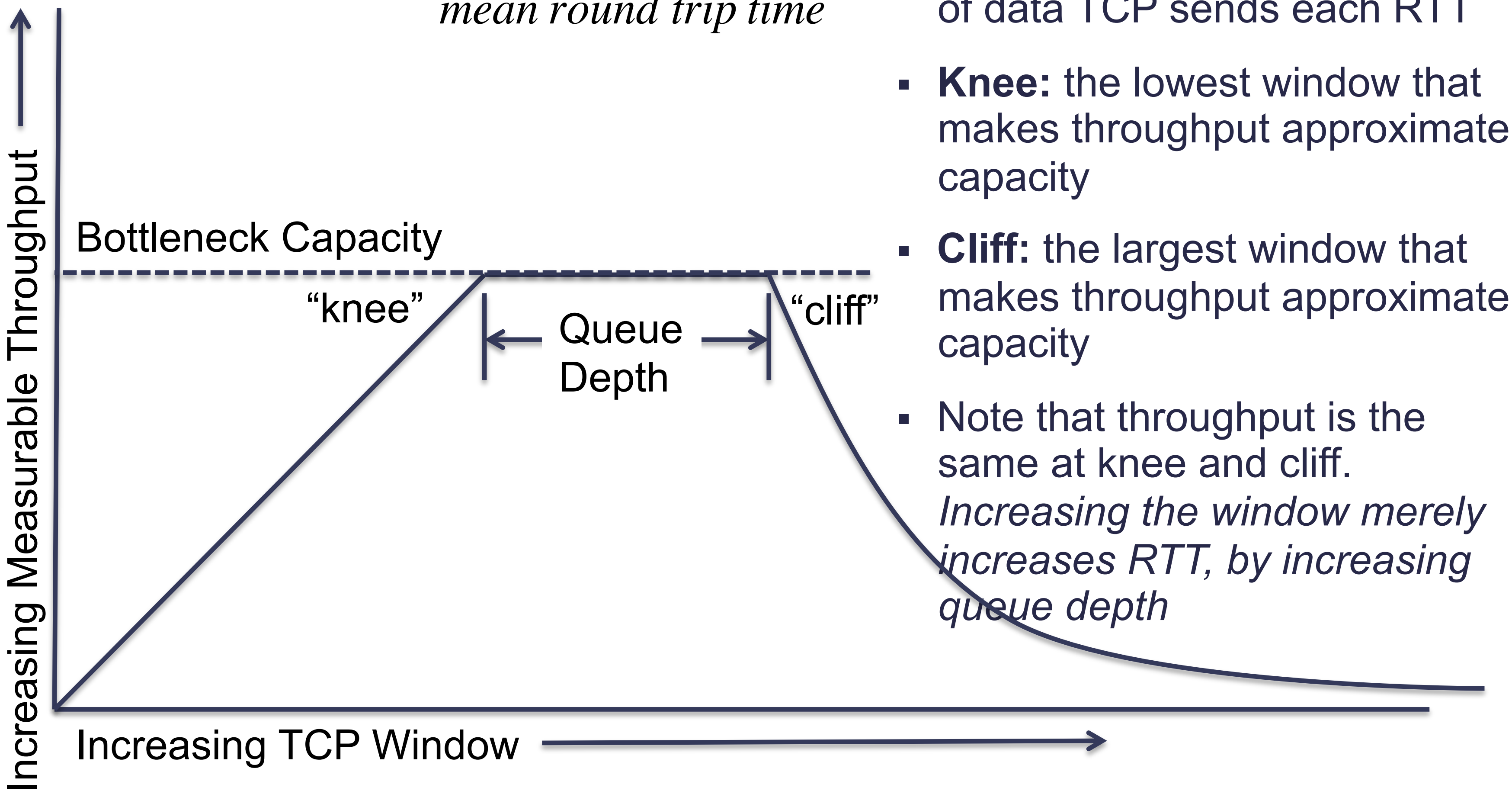


Taxonomy of data flows

- We are pretty comfortable with the concepts of mice and elephants
  - “mice”: small sessions, a few RTTs total
  - “elephants”: long sessions with many RTTs
- In Data Centers with Map/Reduce applications, we also have ***lemmings***
  - O(10^4) mice migrating together
- Solution premises
  - Mice: we don’t try to manage these
  - Elephants: if we can manage them, network works
  - Lemmings: Elephant-oriented congestion management results in HOL blocking

Simple model of TCP throughput dynamics

$$\textit{mean throughput} = \frac{\textit{effective window in bytes}}{\textit{mean round trip time}}$$

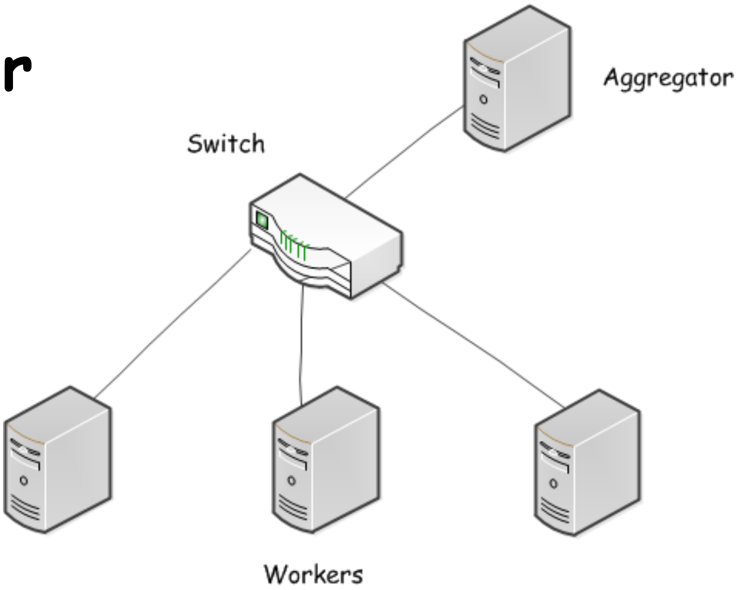


- **Effective Window:** the amount of data TCP sends each RTT
- **Knee:** the lowest window that makes throughput approximate capacity
- **Cliff:** the largest window that makes throughput approximate capacity
- Note that throughput is the same at knee and cliff. *Increasing the window merely increases RTT, by increasing queue depth*

Yes, there is a more complex equation that takes into account loss. It estimates throughput above the cliff.

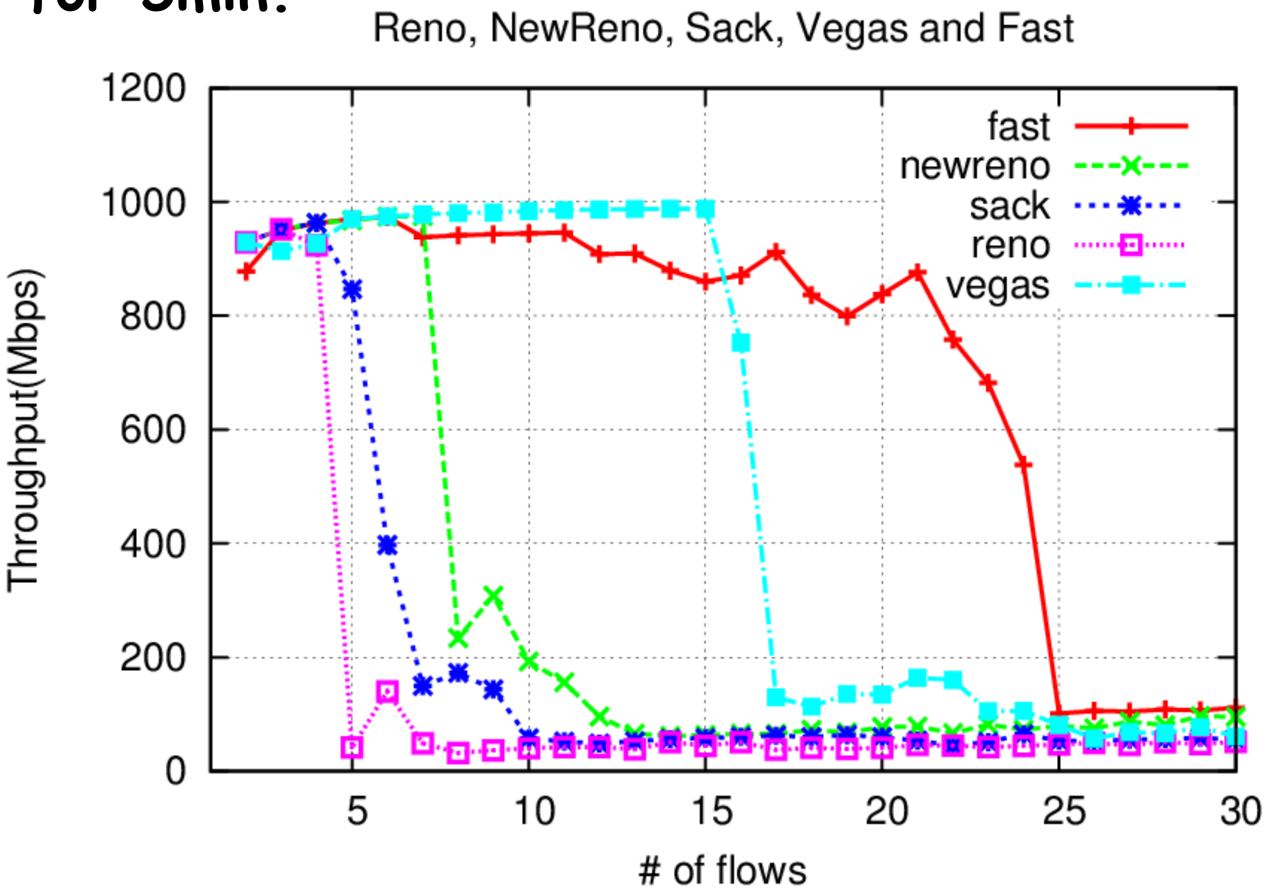
Technical Platform

- Machines
  - Hosts with 3.1GHz CPU, 2GB RAM and 1Gbps NIC (4)
  - NetFPGA
  - Freebsd 9.2-prerelease
- Multi-thread traffic generator
  - Each responses 64KB
  - Buffer: 128KB



TCP Performance on short RTT timeframes

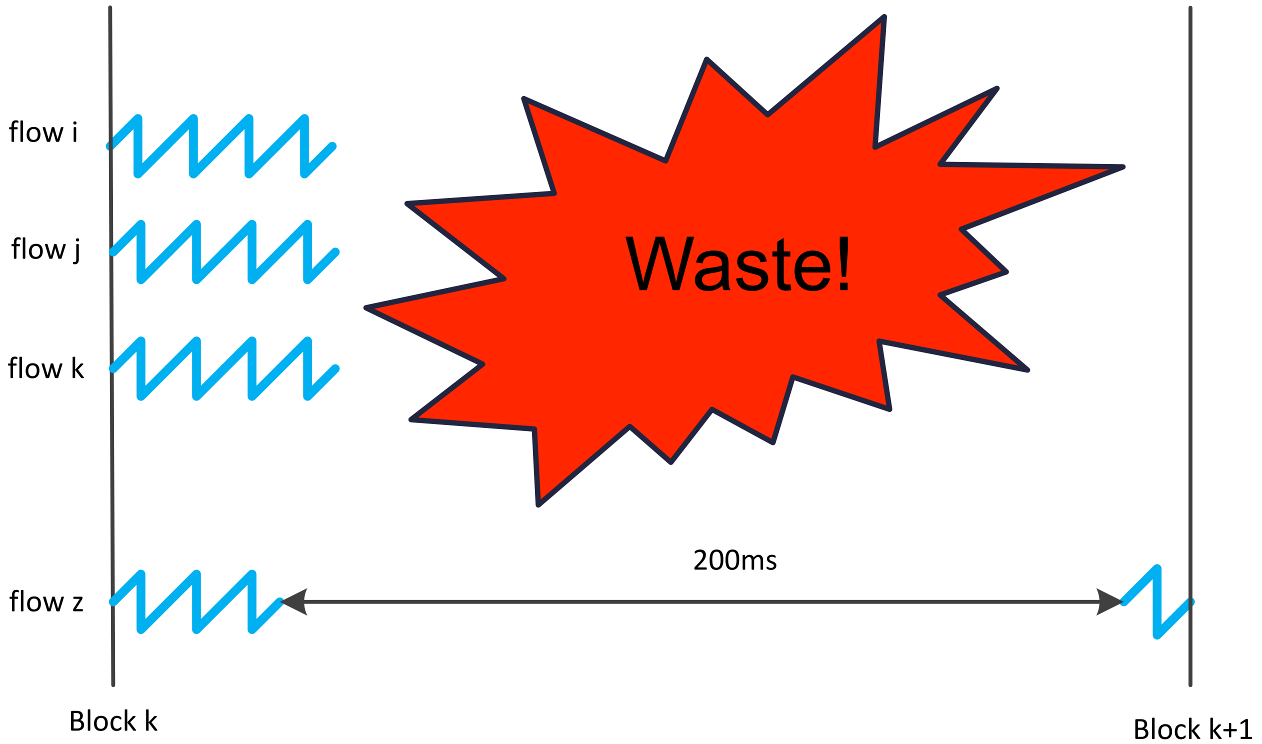
- Each flow responses 100KB data
  - Last for 5min.



Courtesy Tsinghua University Cisco/Tsinghua Joint Lab

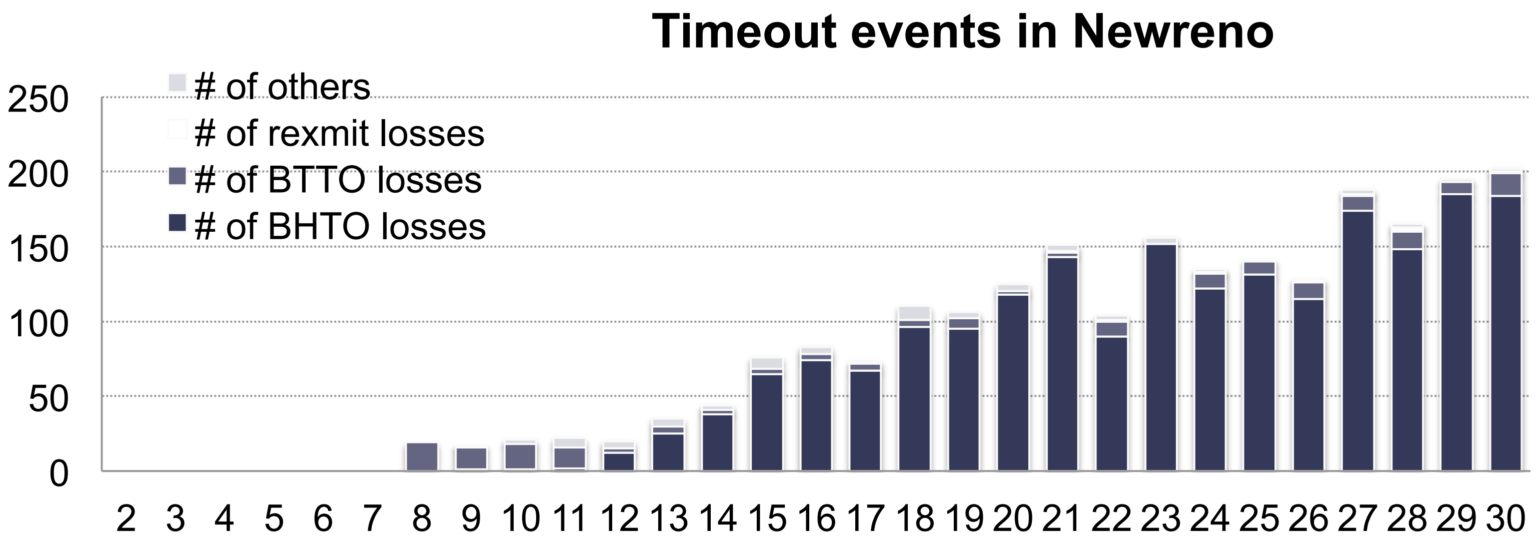
Effects of TCP Timeout

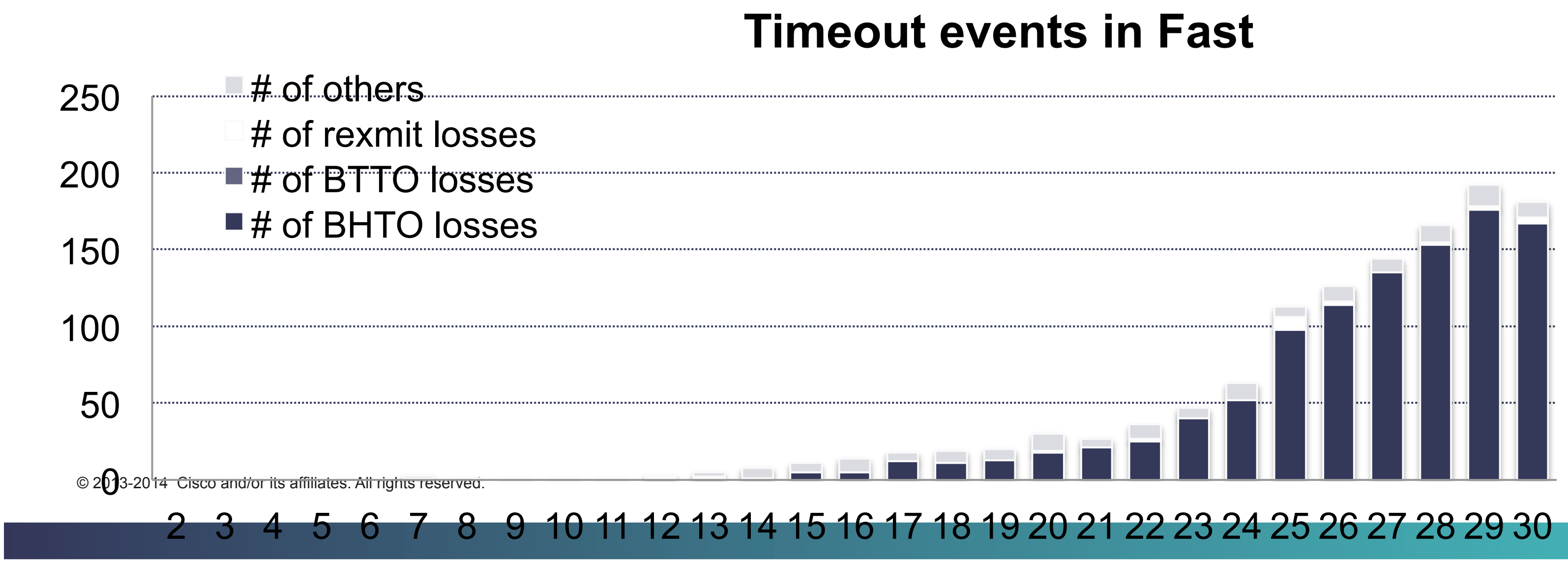
- The ultimate reason for throughput collapse in Incast is timeout.



Courtesy Tsinghua University Cisco/Tsinghua Joint Lab

Prevalence of TCP Timeout

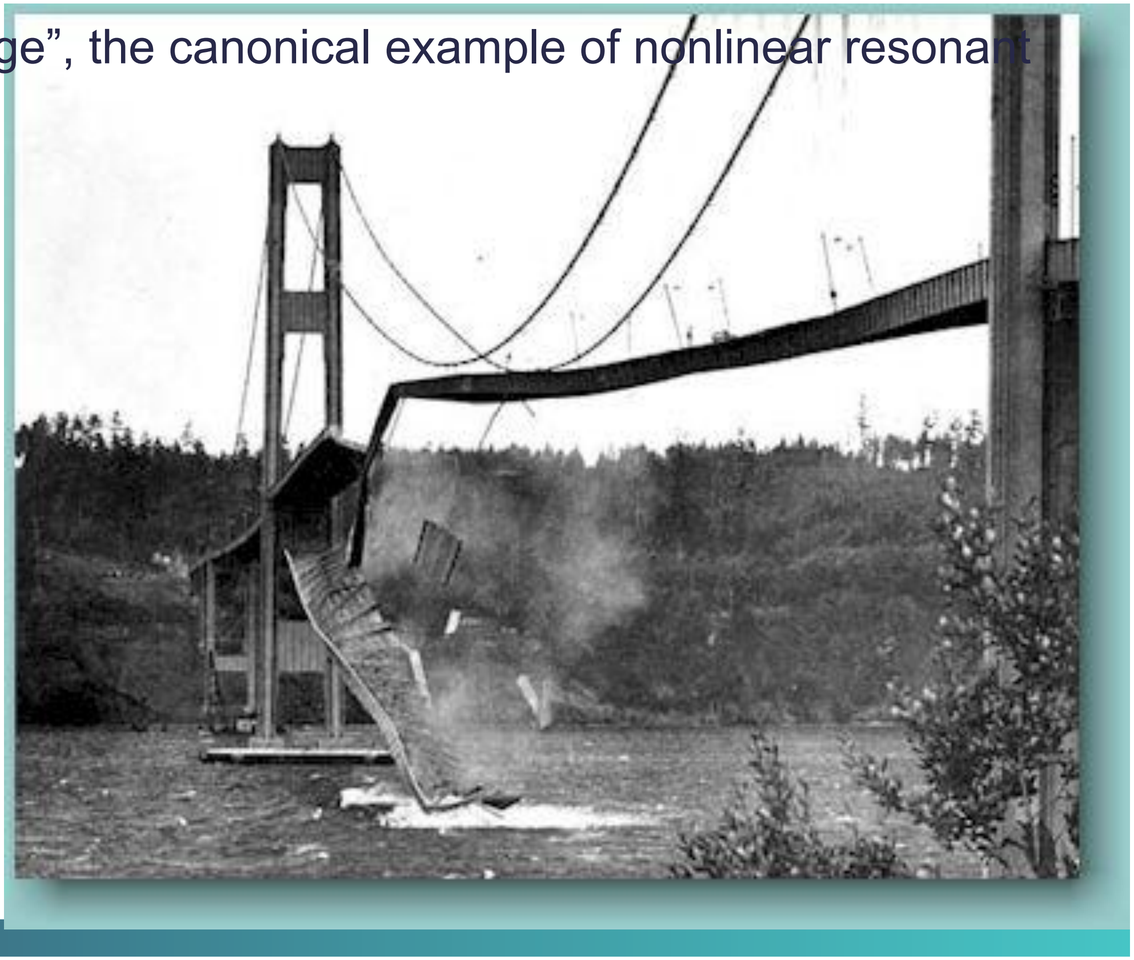




Courtesy Tsinghua University
Cisco/Tsinghua Joint Lab

What’s the other half of the incast problem?

- In two words, amplification and coupling.
- Amplification Principle
  - Non-linearities occur at large scale which do not occur at small to medium scale.
  - Think “Tocoma Narrows Bridge”, the canonical example of nonlinear resonant amplification in physics
  - RFC 3439



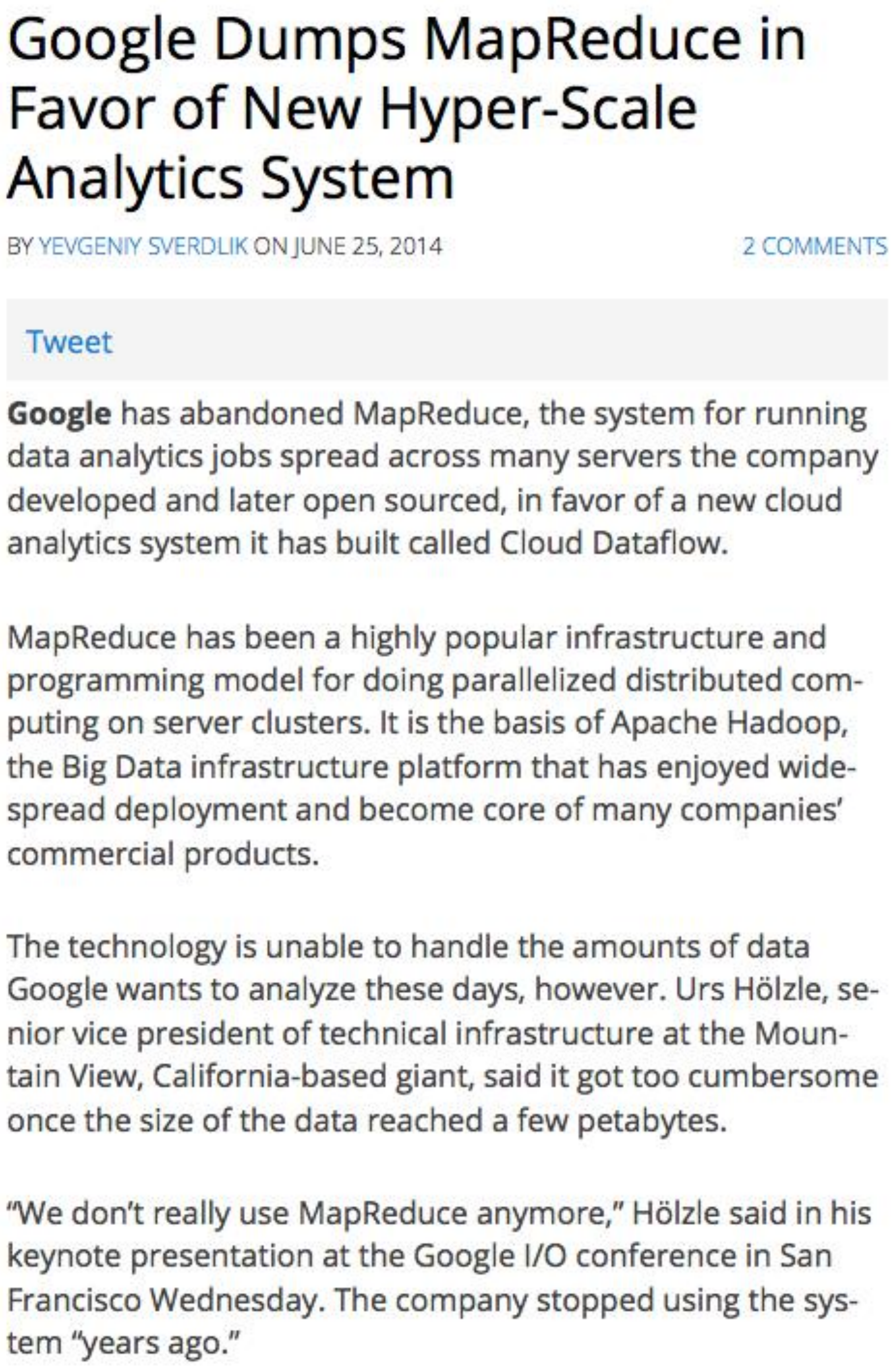
What’s the other half of the incast problem?

- Coupling Principle
  - As things get larger, they often exhibit increased interdependence between components.

- When a request is sent to O(10^4) other machines and they all respond
  - Bad things happen…

Large scale shared-nothing analytic engine

- *Time to start looking at next generation analytics*
- UCSD CNS – moving away from rotating storage to solid-state drives dramatically improves Tritonsort
- Facebook: uses Memcache as basic storage medium



Google Dumps MapReduce in Favor of New Hyper-Scale Analytics System

BY YEVGENIY SVERDLIK ON JUNE 25, 2014 2 COMMENTS

Tweet

Google has abandoned MapReduce, the system for running data analytics jobs spread across many servers the company developed and later open sourced, in favor of a new cloud analytics system it has built called Cloud Dataflow.

MapReduce has been a highly popular infrastructure and programming model for doing parallelized distributed computing on server clusters. It is the basis of Apache Hadoop, the Big Data infrastructure platform that has enjoyed widespread deployment and become core of many companies' commercial products.

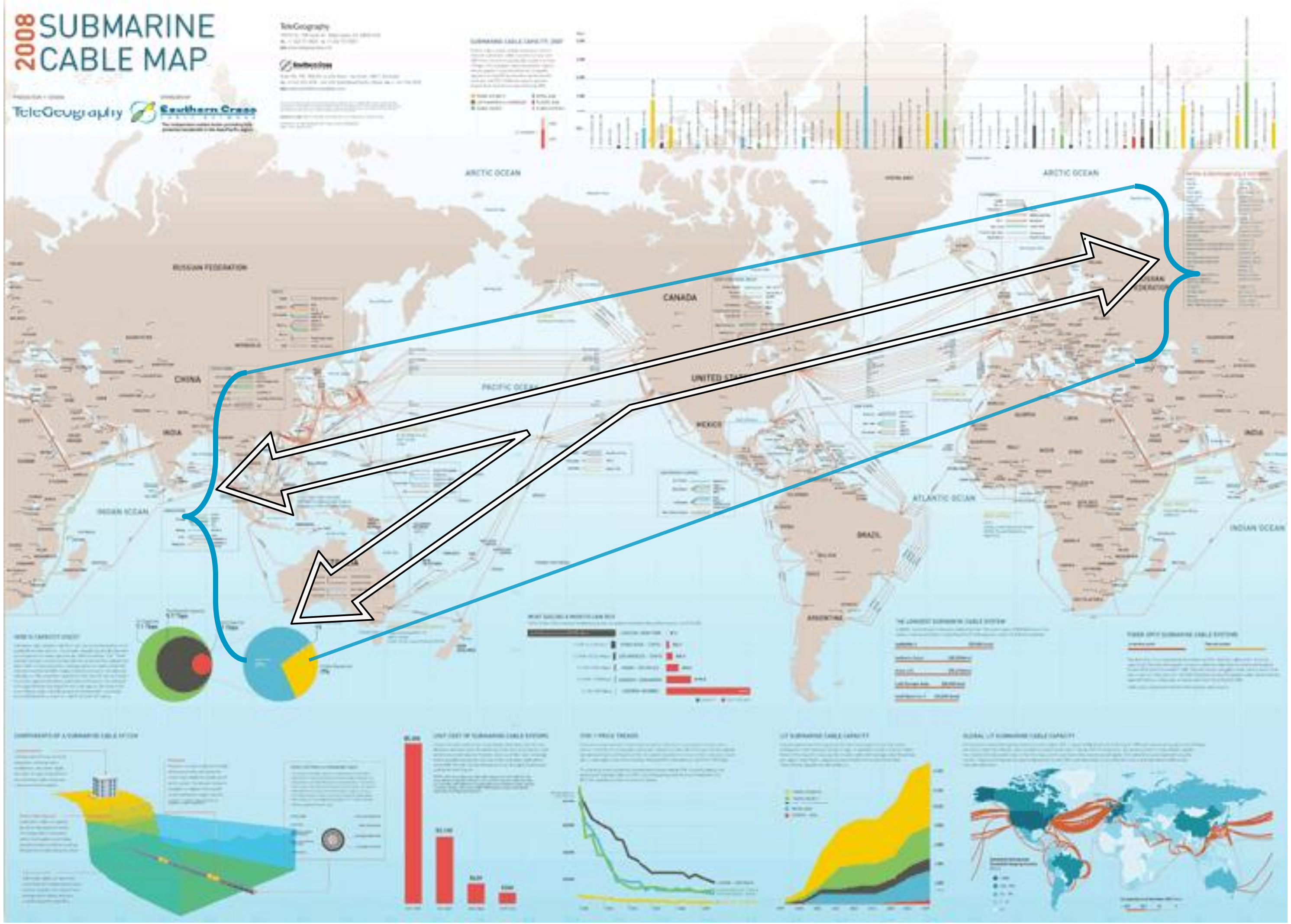
The technology is unable to handle the amounts of data Google wants to analyze these days, however. Urs Hölzle, senior vice president of technical infrastructure at the Mountain View, California-based giant, said it got too cumbersome once the size of the data reached a few petabytes.

"We don't really use MapReduce anymore," Hölzle said in his keynote presentation at the Google I/O conference in San Francisco Wednesday. The company stopped using the system "years ago."

Simplifying the Internet:
IPv4 and IPv6 in industry

Where Is the Broadband Internet Today?

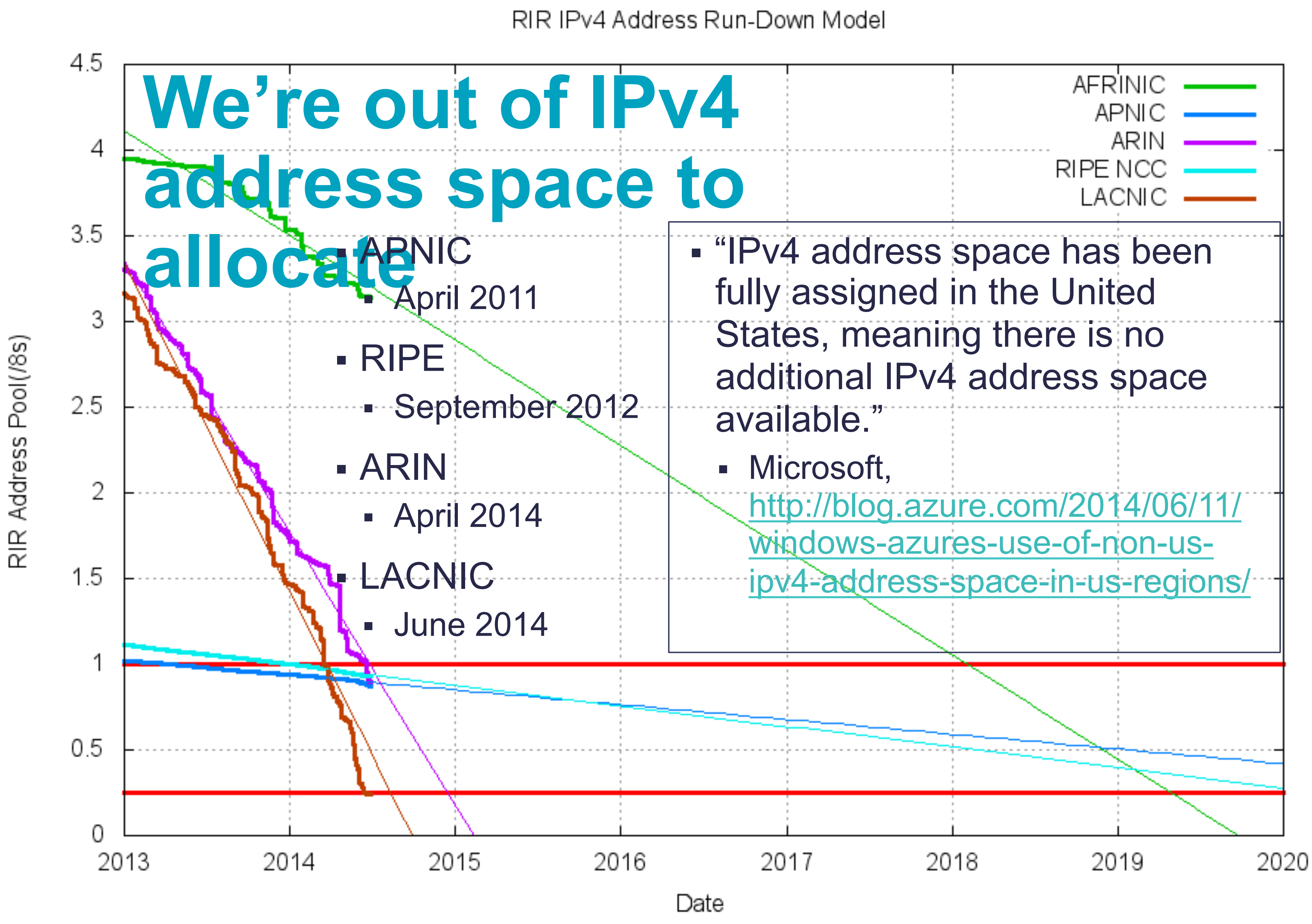
The Europe/America/East Asia/ANZ Fiber Corridor



Map copyright 2008 TeleGeography

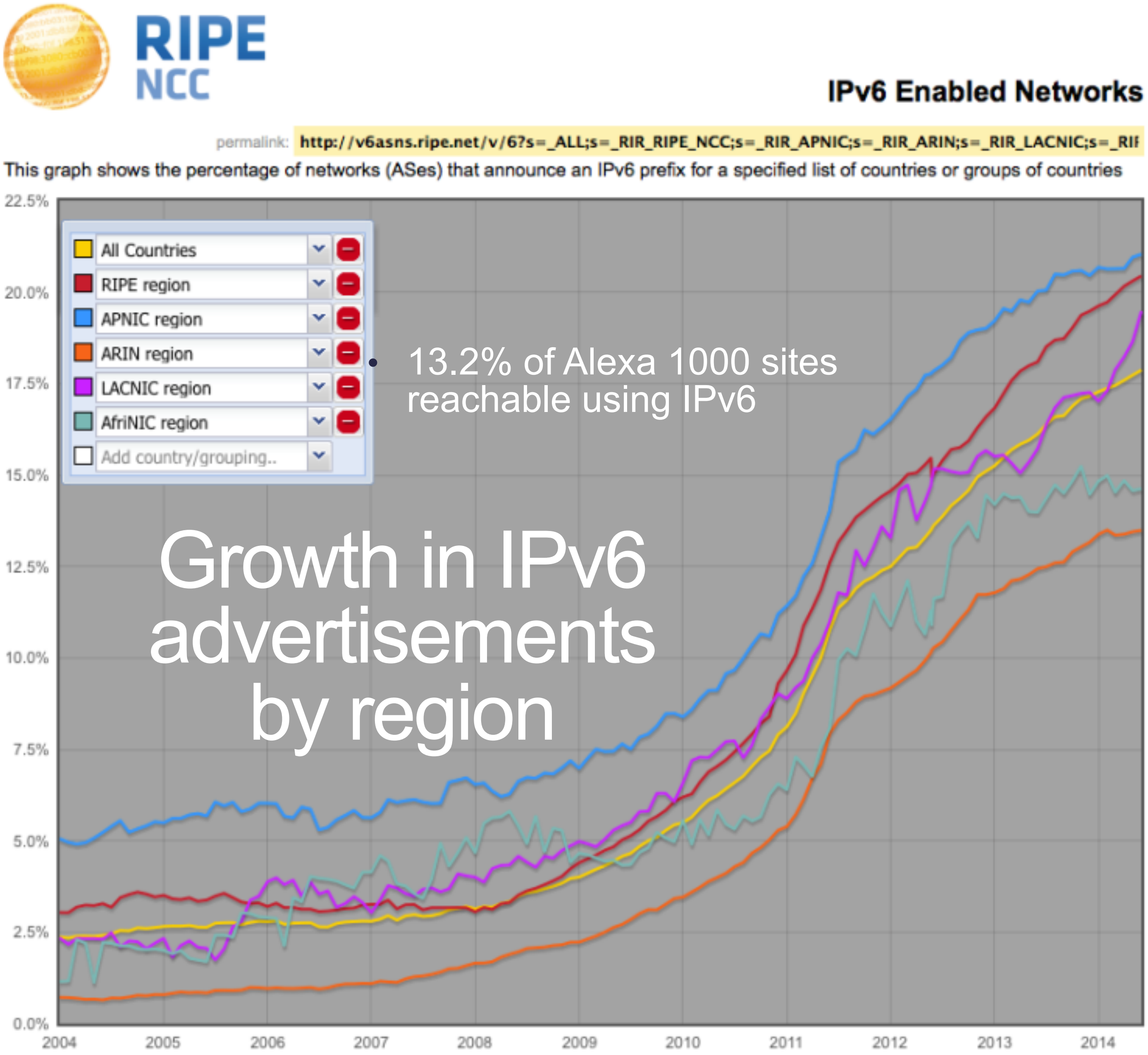
Today

We're out of IPv4 address space to allocate



- APNIC
  - April 2011
- RIPE
  - September 2012
- ARIN
  - April 2014
- LACNIC
  - June 2014

- "IPv4 address space has been fully assigned in the United States, meaning there is no additional IPv4 address space available."
  - Microsoft, http://blog.azure.com/2014/06/11/windows-azures-use-of-non-us-ipv4-address-space-in-us-regions/



IPv6 Enabled Networks

permalink: http://v6asns.ripe.net/v/6?s=_ALL;s=_RIR_RIPE_NCC;s=_RIR_APNIC;s=_RIR_ARIN;s=_RIR_LACNIC;s=_RIf

This graph shows the percentage of networks (ASes) that announce an IPv6 prefix for a specified list of countries or groups of countries

- 13.2% of Alexa 1000 sites reachable using IPv6

Growth in IPv6 advertisements by region

APNIC

RIPE

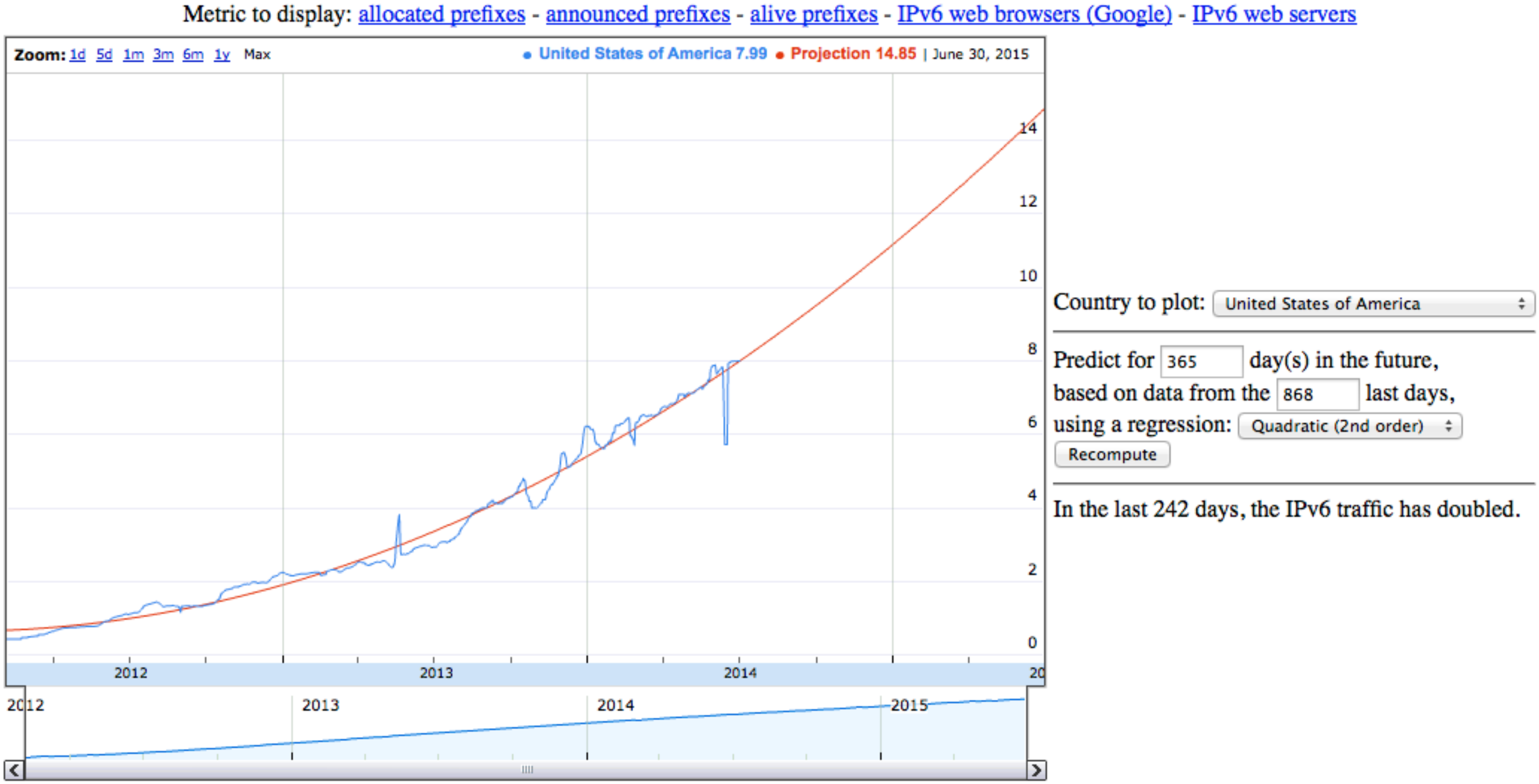
LACNIC

All Countries

AFRINIC

ARIN

Growth in US deployment:
Google percentage of IPv6 access to them



Metric to display: allocated prefixes - announced prefixes - alive prefixes - IPv6 web browsers (Google) - IPv6 web servers

Zoom: 1d 5d 1m 3m 6m 1y Max
● United States of America 7.99 ● Projection 14.85 | June 30, 2015
14
12
10
8
6
4
2
0
2012
2013
2014
20
2012
2013
2014
2015

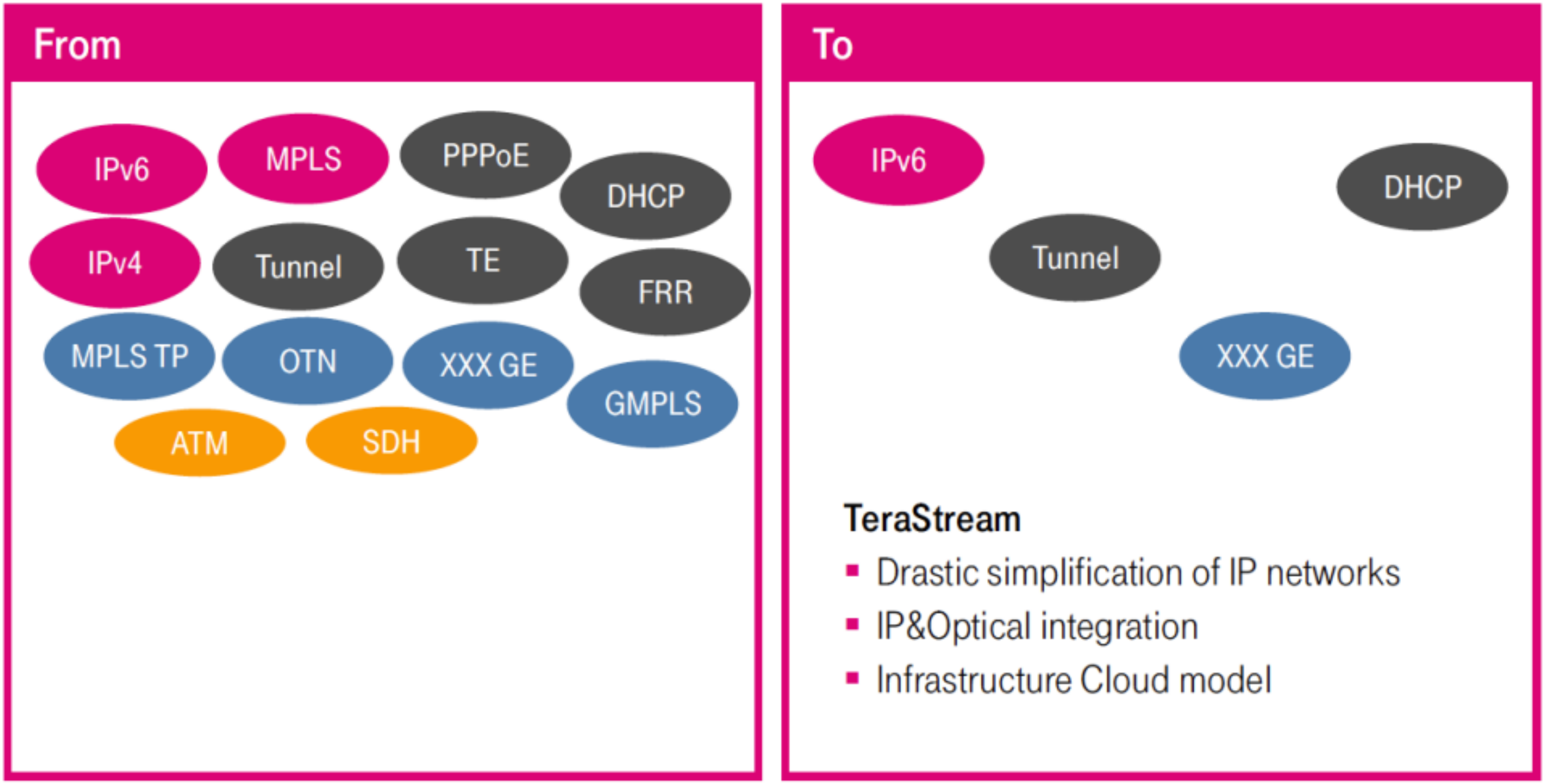
Country to plot: United States of America

Predict for 365 day(s) in the future, based on data from the 868 last days, using a regression: Quadratic (2nd order)

In the last 242 days, the IPv6 traffic has doubled.

https://www.vyncke.org/ipv6status/project.php?metric=p&timeforward=&timebackward=&country=us

KAIKAKU FOR IP NETWORKS
INDUSTRY LEADERSHIP

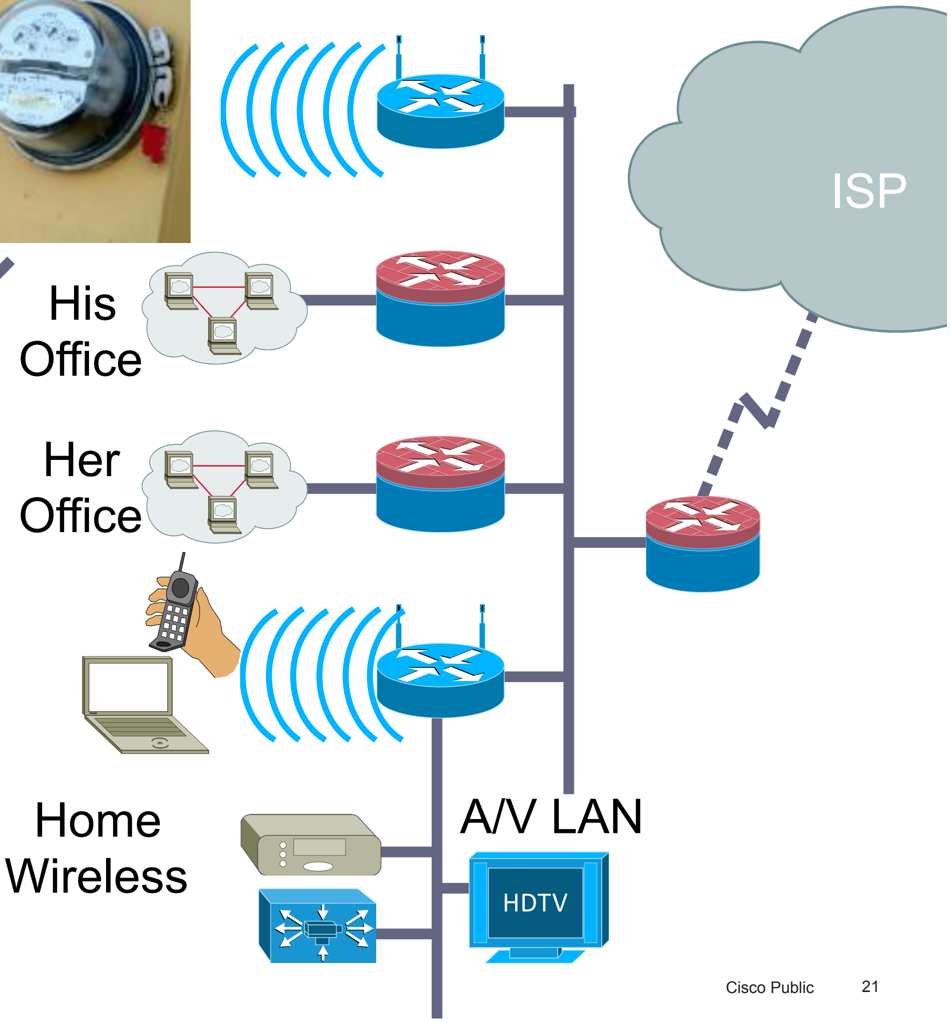


TeraStream

- Drastic simplification of IP networks
- IP&Optical integration
- Infrastructure Cloud model

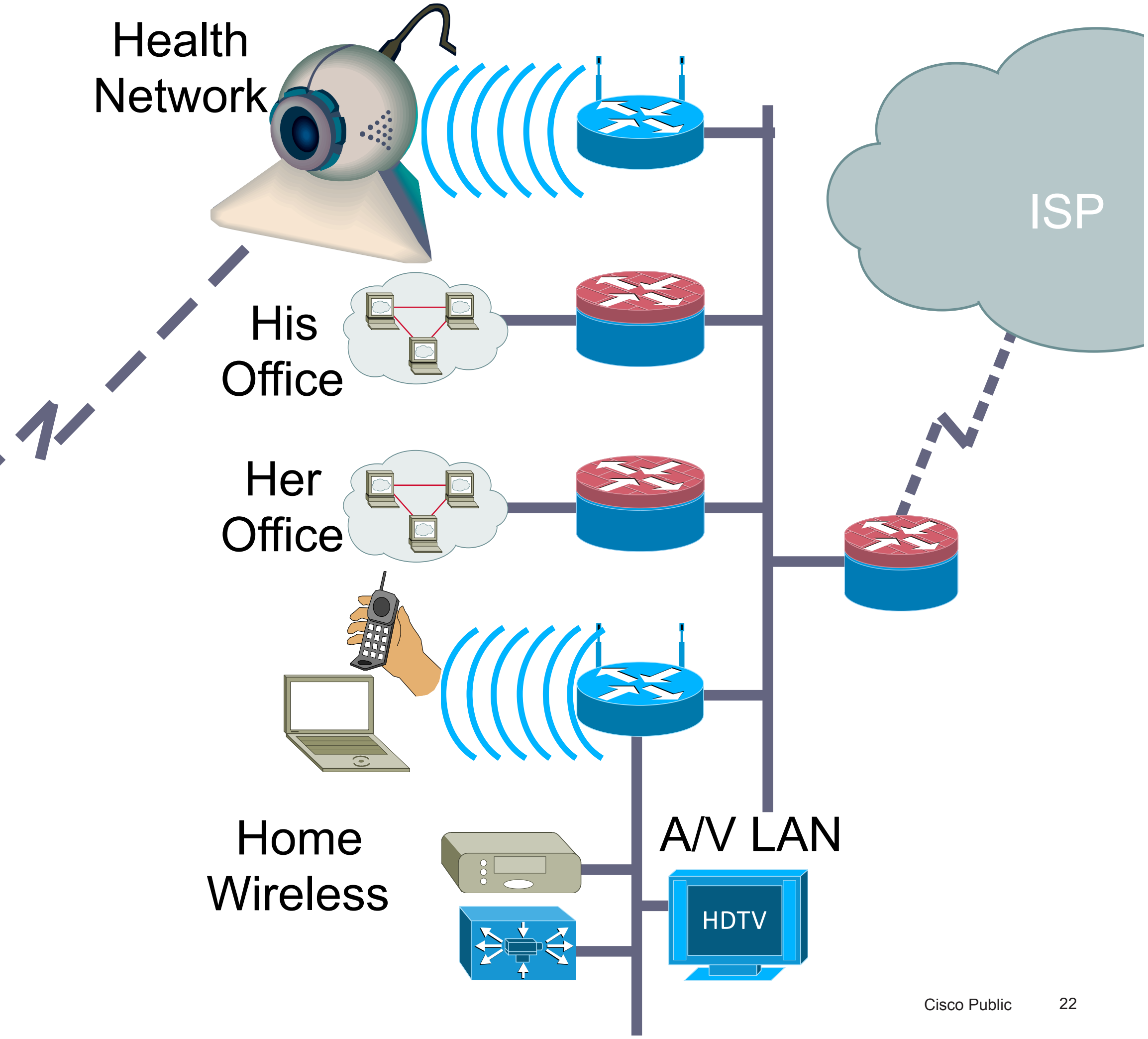
The changing ~~home~~ utility network

- Imagine a high end home network:
  - Audio/Video
  - Wireless
  - Telecommuting
  - Home Area Network
- What is the HAN?
  - Network connecting sensors in the home
  - Communications with utilities
  - Services to residents



Related to sensor networks for health…

- Infrared
- Motion sensors
- Heart Monitors
- Pedometers
- …

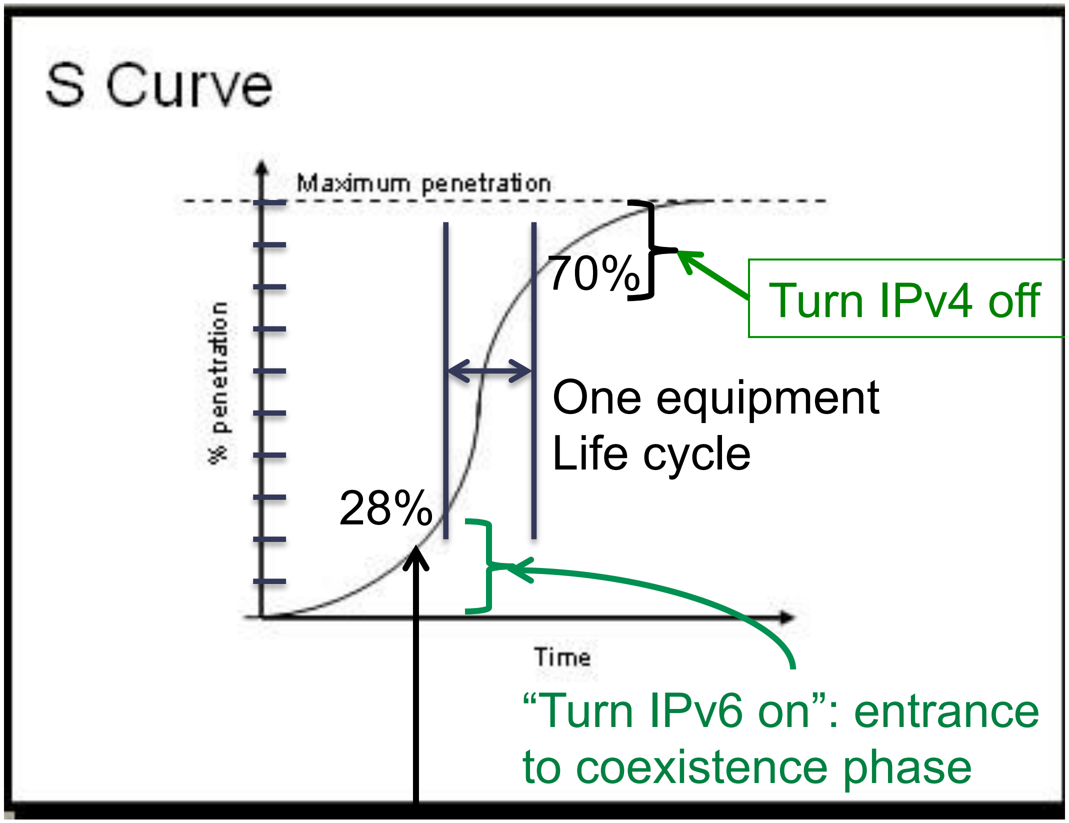


I help evaluate research proposals at http://www.cisco.com/research

- I see a lot of IPv4-only proposals.
- Excuse me, guys and gals. If your proposals don’t test IPv6, you’re missing a sea-change in progress.
- If you want to lead industry and do research that will affect the way the Internet is deployed, IPv6 needs to be part of it. *IPv4 doesn’t.*

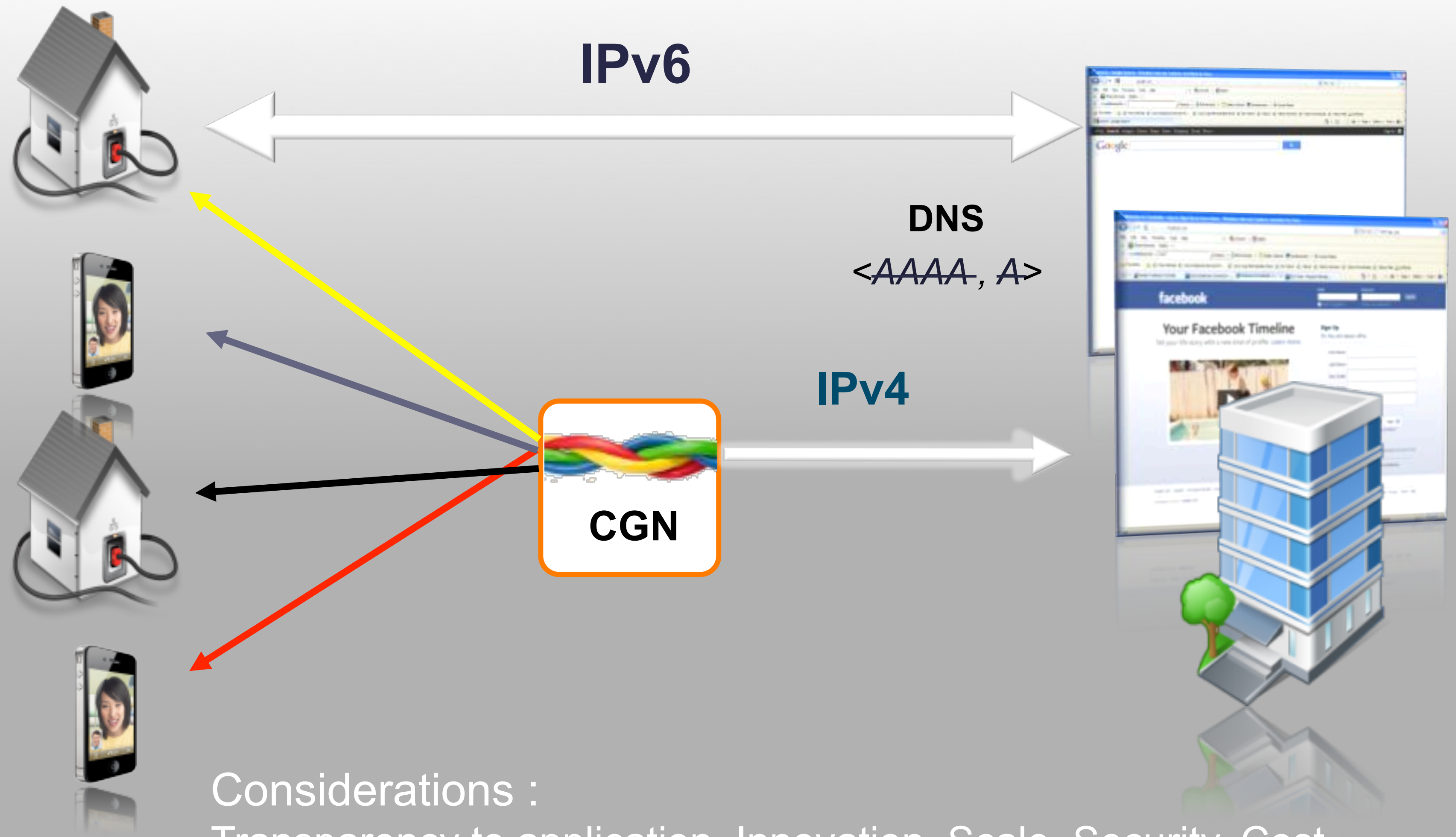
“When do you think IPv6 will be more important than IPv4?”

- Adoption follows an “Adoption curve”:
  - The more users adopt, the more users adopt, over a period of time
- Google, Akamai, Yahoo, Facebook report:
  - Google Fiber 76%
  - Verizon Wireless 50%
  - Free 37%
  - Comcast 25%
  - Deutsche Telekom 18%
  - AT&T 17%
  - T-Mobile USA 16%
  - Time Warner Cable 5%



We are here: 22% of AS’s Worldwide

IPv6 – “Full Spectrum” Internet

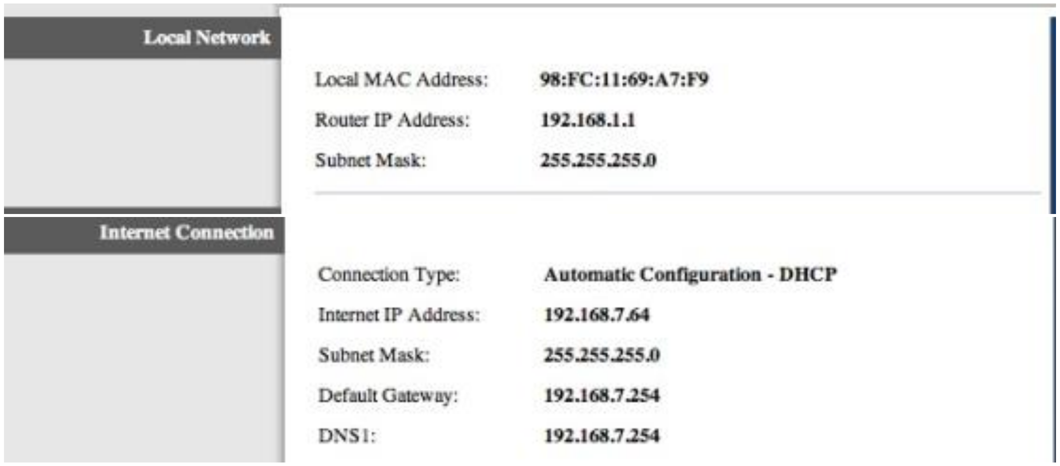


Considerations :

Transparency to application, Innovation, Scale, Security, Cost

What’s the problem with Carrier NAPT?

- My daughter’s house was broken into on 1 October 2012
  - A couple of days later, I bought a video surveillance system and her husband and I installed it.
  - Surveillance and recording – fine.
- The product includes a DDNS service:
  - Record a name and a NAPT translation in the home router
  - **iPhone, Android, Windows, and MacOSX Apps now advertised as being able to view the video record and manage the system**



Local Network

Local MAC Address:	98:FC:11:69:A7:F9
Router IP Address:	192.168.1.1
Subnet Mask:	255.255.255.0

Internet Connection

Connection Type:	Automatic Configuration - DHCP
Internet IP Address:	192.168.7.64
Subnet Mask:	255.255.255.0
Default Gateway:	192.168.7.254
DNS1:	192.168.7.254

- Oops: upstream NAPT.
  - Why do people mistake NAT for a security service?
  - No IPv6 service
- Implication
  - ***An advertised business service could not be delivered due to address multiplexing***

> " We do not think these customers will notice any difference at all in their broadband performance, but if any of these customers did have any resulting issues, we would be happy to restore their connection to an individual IP address."

However, it appears users are already noticing problems. "It's causing me a real headache, for a start none of my home servers are now accessible via the web, remote access to my PC is also blocked, and XBox Live requires NAT to be open to work correctly so has reduced multiplayer ability," said one user…

PC Pro, 3 May 2013

http://www.pcpro.co.uk/news/broadband/381646/customers-fume-as-bt-introduces-ip-sharing#ixzz2SFrOWK7d

“The next generation of IP address space is IPv6, which will enable far more addresses to be assigned than IPv4. Unfortunately, most servers and other Internet devices will not be speaking IPv6 for a while, so IPv4 will remain standard for some time to come.”

[In your IPv4 network access] “there are some applications such as online gaming, VPN access, FTP service, surveillance cameras, etc., that may not work when broadband service is provided via a CGN."

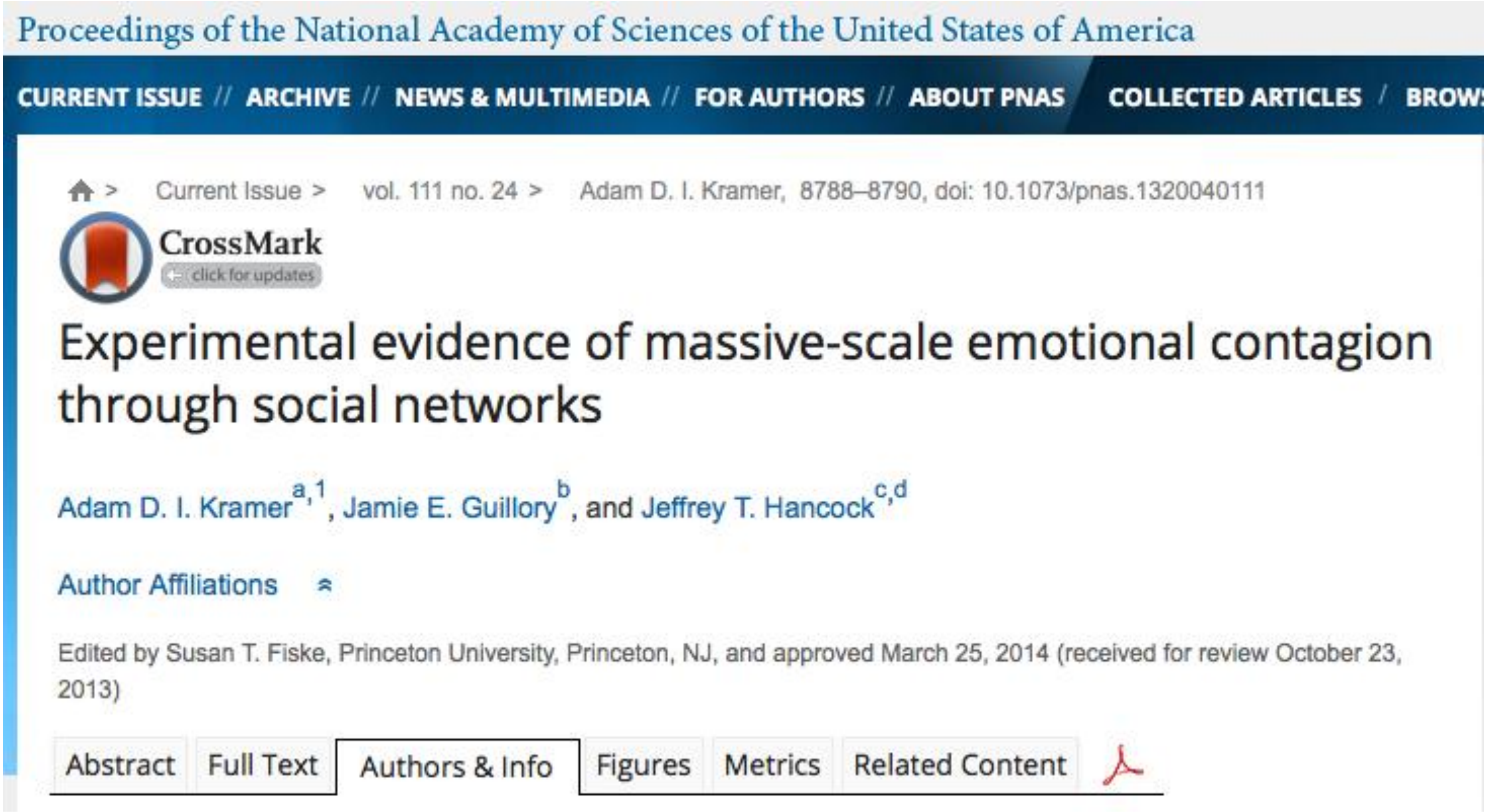
Verizon, April 2013

http://www.verizon.com/support/residential/internet/highspeedinternet/networking/troubleshooting/portforwarding/123897.htm

Routine (Routing?) Attacks, Backdoors, and Government Snooping

Routine (Routing?) Attacks, Backdoors, and ~~Government~~ Snooping

Facebook in the news


CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS
COLLECTED ARTICLES / BROW
CrossMark
click for updates
Abstract
Full Text
Authors & Info
Figures
Metrics
Related Content

Proceedings of the National Academy of Sciences of the United States of America

> Current Issue > vol. 111 no. 24 > Adam D. I. Kramer, 8788–8790, doi: 10.1073/pnas.1320040111

Experimental evidence of massive-scale emotional contagion through social networks

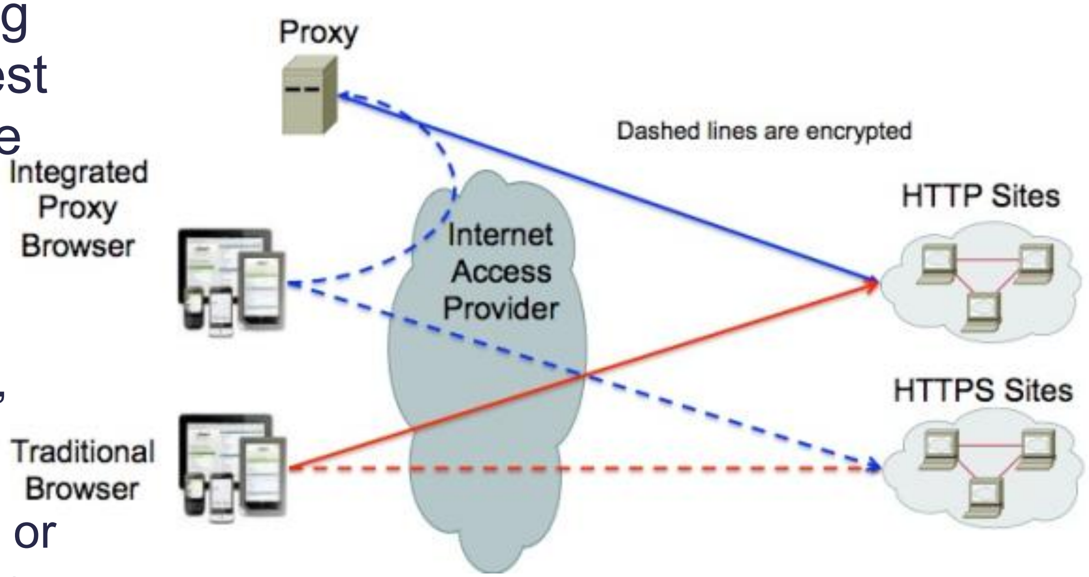
Adam D. I. Kramer[a,1], Jamie E. Guillory[b], and Jeffrey T. Hancock[c,d]

Author Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

SPDY Proxies… and Mobile Networks

- Advance deployment of HTTP/2
  - Designed to encrypt and improve download efficiency.
- Multiple companies deploying SPDY proxies as a way to test the specification and help the Internet “go dark”
- Rising conflicts:
  - Mobile networks have been inserting ads and “optimizing” content for a while now.
  - That capability, whether used or not, appears to have moved to Google, Amazon, and others

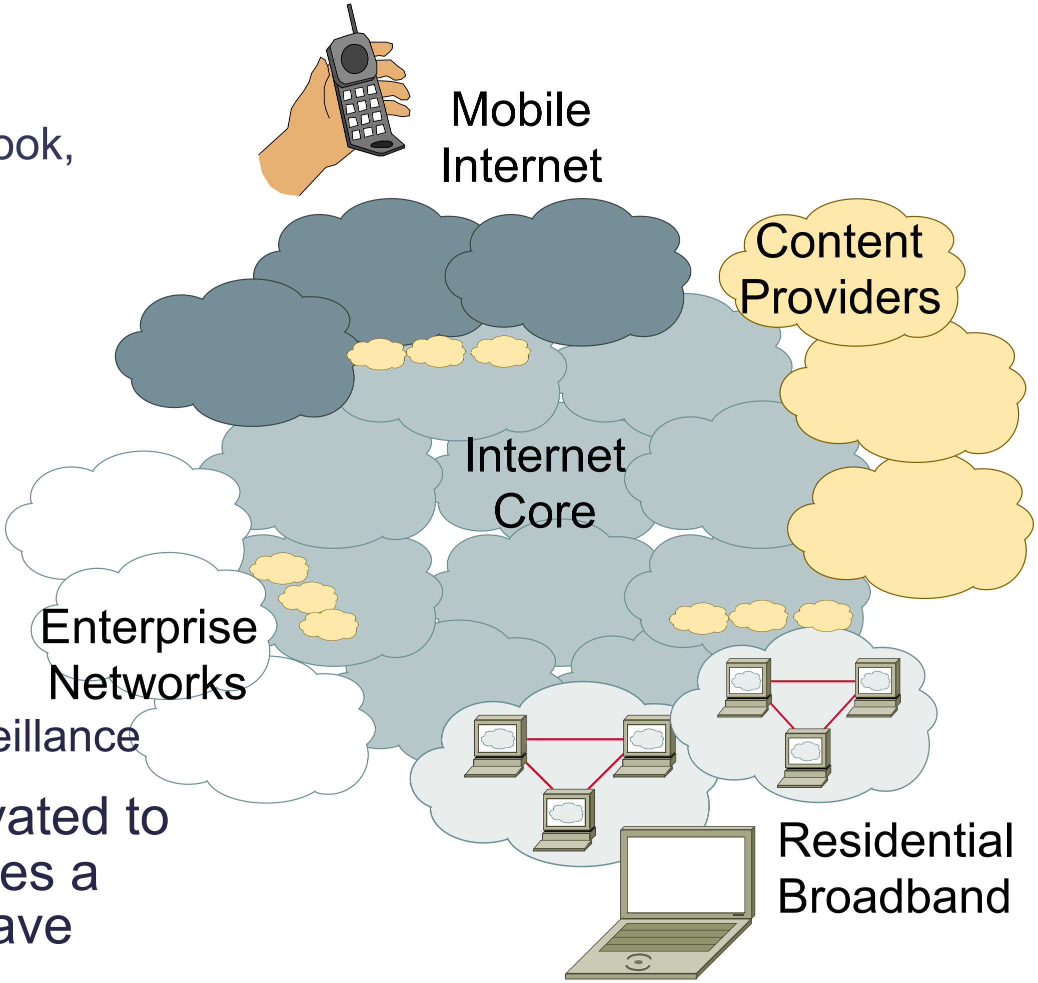


Are you receiving location-based or interest-based advertising?

It’s not just NSA

A truly simplistic model of the Internet from a business perspective

- Many components
  - Internet Core – transit providers
  - Content Providers – Google, Facebook, YouTube
  - Enterprise Networks
  - Residential Broadband
  - Mobile Internet/Telephone
- Different motivations
  - Source of revenue
  - Need for addresses
  - Location-based services
  - Simple access to entire Internet
  - Lawfully Authorized Electronic Surveillance
- People and companies are motivated to deploy a technology when it solves a problem that they believe they have



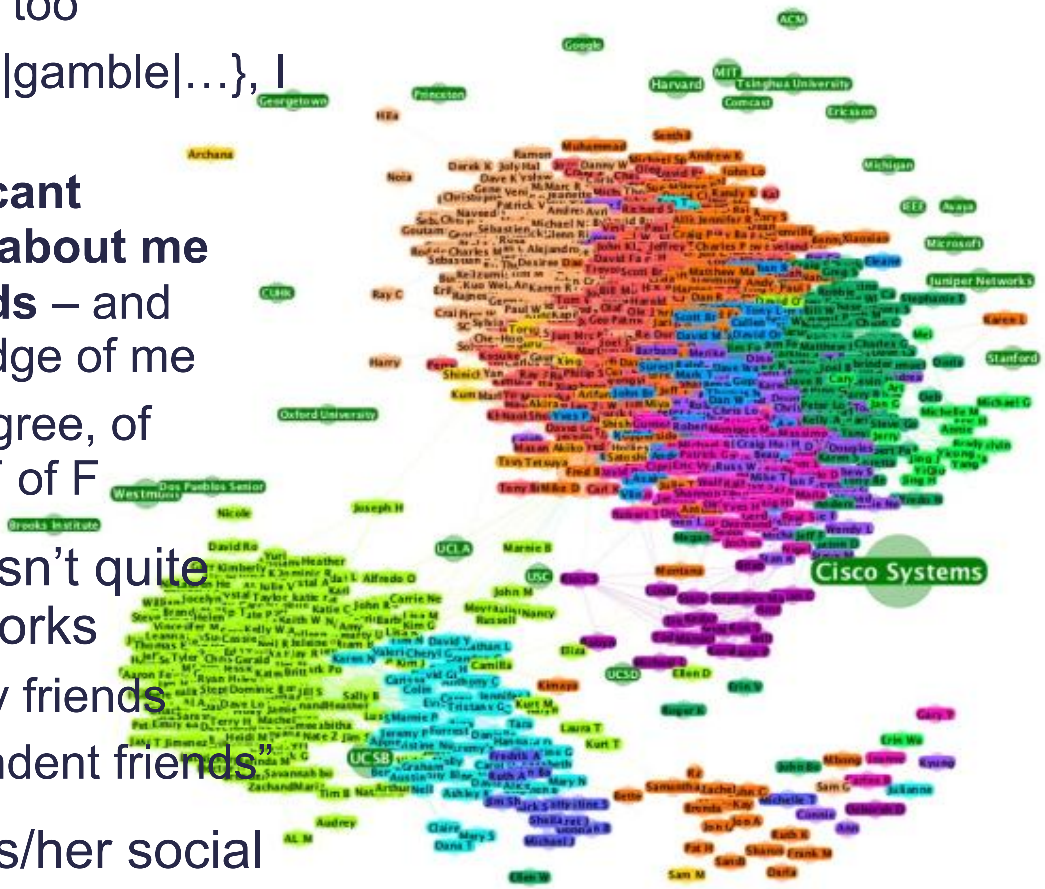
Content Providers

- The business of a content provider is…
  - To provide informative content to its customers
- Who is the customer?
  - If you're paying for the content, you're the customer
  - If it's free to you, you are integral to the product being sold

  - Eyeballs, **location** of eyeballs, marketing **statistics and insight**, **demographic** intelligence, search criteria, Digital Rights Management
- *Content providers need to know your geographic and topological **location** and **interests,** and associate that with **identity** and **relationship,** to deliver their product to their **customers***
  - ***IPv4 CGN obscures location, makes security diagnosis and service deployment harder***
  - IPv6 global addressing permits folks to determine topological location and by extension probable physical location

Identities? Relationships?

- Per sociological research
  - If all of my friends have a given opinion or medical situation, I probably do too
  - If all of my friends {smoke|drink|gamble|…}, I probably do too
  - **There are statistically significant inferences that can be made about me given knowledge of my friends** – and about my friends given knowledge of me
  - This is also true, to a lesser degree, of friends of friends, but not F of F of F
- Per UCSD Research: this doesn't quite follow in computer social networks
  - 10,000 FB friends are too many friends
  - Look at "photo friends", "respondent friends"
- Knowledge of a person and his/her social networks provides information useful for business purposes



http://connectedthebook.com/

Facebook TouchGraph

Mobile Internet (was: Mobile Telephone)

- Business models for communications
  - ❑ PSTN:
    - Location matters, distance matters, billable unit is the minute
  - ❑ Internet:
    - Location and distance irrelevant, billable unit is the month or the megabyte
  - ❑ Mobile Internet:
    - Billable unit is the month or the megabyte
    - Unless you're roaming (a different version of distance)
    - Location is important: cell location important to service, also a commodity to sell
    - So is the set of parties you call, or who call you.
- Value of IPv6 to Mobile Internet
  - ❑ Prior to Release 9, IPv4 and IPv6 require separate network attachments; pay accordingly
  - ❑ Simplify network – less internal NATs, simpler debugging

Some implications

In data centers and elsewhere, “this is the way we have always done it” isn’t good enough

- TCP Congestion Control doesn’t work well as latency control. We have better algorithms. We should consider using them.

- Map/Reduce is not the world’s best approach to computation. It has problems with amplification and correlation. With the improvement in mass storage, using silicon rather than rotating media, the company that invented it has moved on. Maybe research should be developing new paradigms.

IPv6 deployment is happening, and is needed

- We need it to continue Internet deployment worldwide. We’re out of IPv4 address space.
- We need it for new services, such as utility services and health care
- We need it simply to continue offering the services we have now in the way we have enjoyed them.
- In many ways, it can be thought of as IPv4 with larger addresses. There are other implications, though.

Go dark. Use PGP, TLS, https, http/2, and so on

But recognize that most invasion of privacy happens in places where the data is no longer encrypted. It depends on

- Cookies
- Authenticated relationships with the social media or other service
- Authenticated relationships among people
- Service logging
- Location tracking

And it's often done as a service, ostensibly to you.

Thank you.