Into the Exacloud

> *Why cloud computing requires a major expansion of wireless spectrum and investment*
> *An exaflood update: what Mobile, Video, Big Data, and Cloud mean for network traffic*
> *Plus, a new paradigm for online games, Web video, and cloud software*

BRET SWANSON > November 21, 2011

"Workers suffering from information overload, and companies drowning in the Internet-era exaflood of data? These are good problems indeed," writes *New York Times* technology reporter Steve Lohr, "if you are in the data storage business."

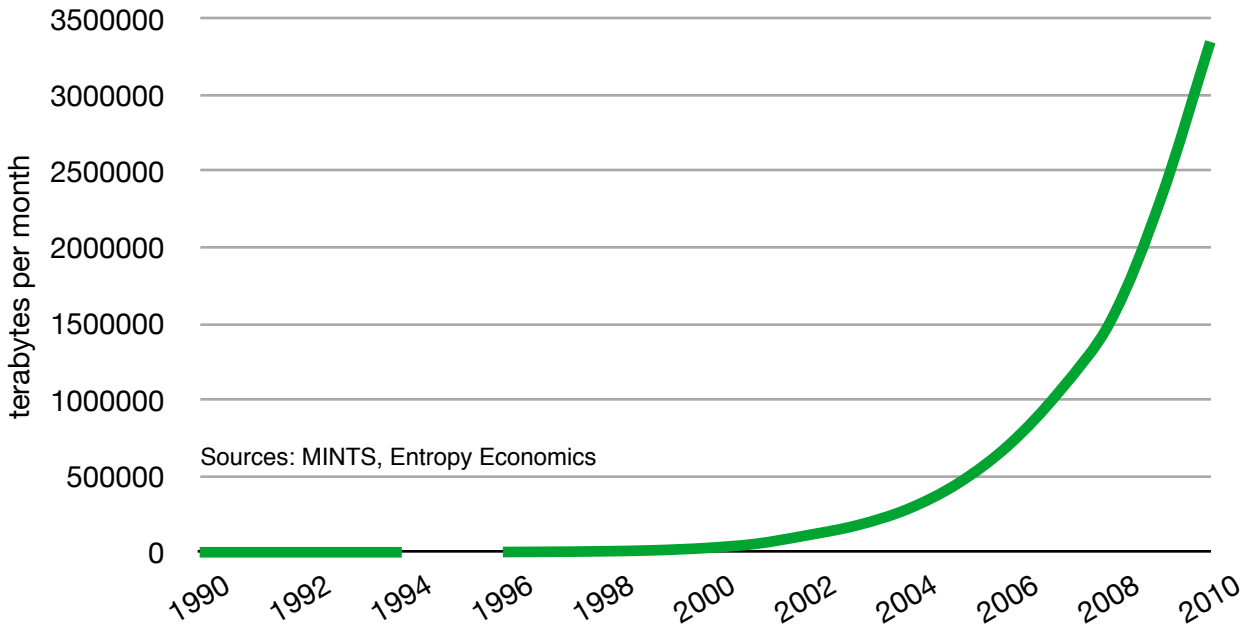
"In a shaky economy," Lohr continues, "companies are spending cautiously on most things, but computer storage in data centers is an exception. The most recent evidence came earlier this month, when IDC reported that sales of disk storage systems in the second quarter grew more than 10 percent, to $7.5 billion.

"The dollars understate the storage boom, since this is an industry working the way technology is supposed to — that is, you get more for less. Measured by the amount of data storage capacity, shipments jumped by 47 percent, compared with the year-earlier quarter."

Google reports that in 2010 its data centers, where many of these disk drives reside, consumed 2.26 terawatt-hours of power – that's two billion kW-hours. Thus the opening of its newest digital warehouse in chilly Hamina, Finland, a $273-million facility meant to take advantage of the cold air and seawater to cool its servers. Facebook is building a similar data center in Luleå, Sweden. Data center pioneer Equinix operates six million square feet across 98 facilities. Globally, Internet data centers now consume 1.5% of all electricity.

We have been chronicling the growth of the Internet for the last decade, and so these numbers do not surprise, though they still tend to amaze. The

Fig. 1 – One Estimate of U.S. Internet Traffic



ever-shifting nature of content, devices, network architectures and capabilities, and digital business models makes for a truly complex ecosystem. In recent years, studies measuring the growth of the digital universe have proliferated. Given these new data sources and analyses, we think it may be useful to update our previous reports.

Bottom Line

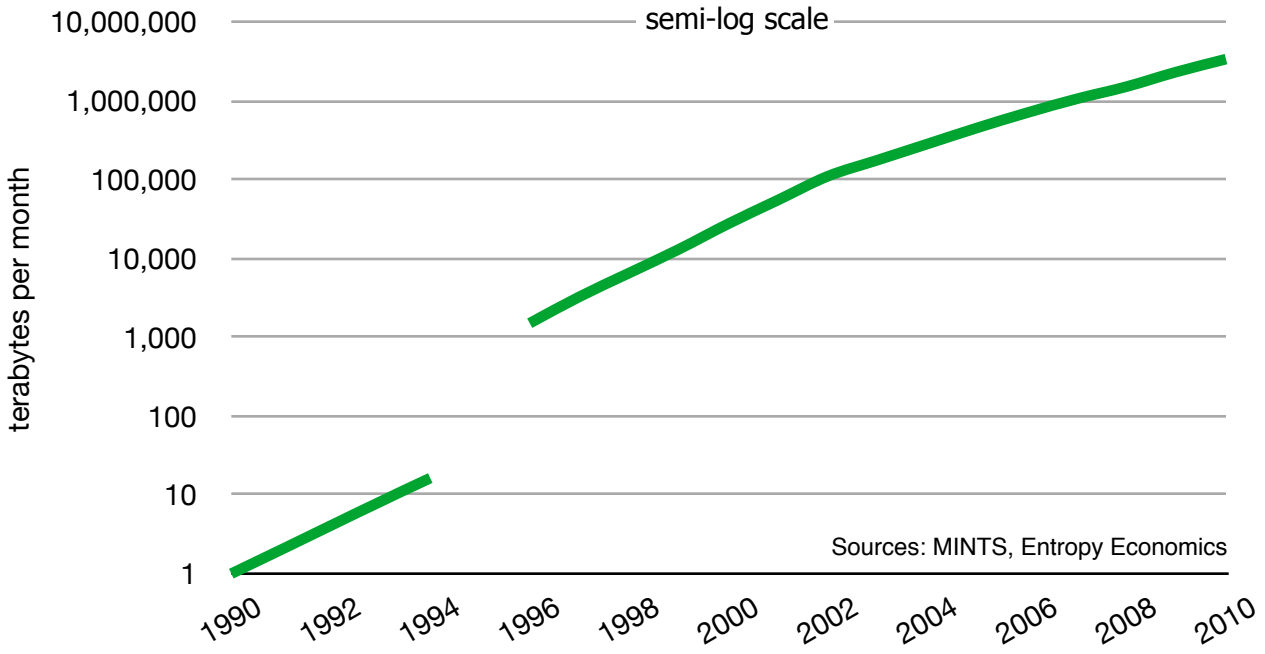
- Very large investments in info-tech infrastructure – including wireless – will need to continue for years to come.
- Wireless capacity, coverage, and flexibility is the chief bottleneck that must be addressed – and is today's chief public policy concern.
- Driven largely by Web video, network traffic continues to grow rapidly and may have accelerated in the last year or so.
- Networks are increasing in capacity, reach, and complexity, and content companies have become Internet infrastructure companies.
- Broadband connectivity enabled the rise of the cloud, and now the cloud requires ever more broadband – both wired and wireless.
- Enormous troves of data, both structured and unstructured, are piling up all over the world.
- The digital ecosystem, comprised of networks, devices, software, services, and the cloud is changing fast. Innovations are improving and disrupting most sectors of life and the economy, including entertainment, education, health, finance, retail, and government, not to mention our social fabric.
- The next generation of exacloud services will deliver unprecedented real-time content and software experiences and impose severe new demands on network capacity and speed.

Flood of New Traffic Research

At the time we published our initial articles and reports, few others were focused on Internet traffic research. University of Minnesota professor Andrew Odlyzko was the most prominent, and his MINTS group continues to collect and analyze traffic data from numerous sources around the world. Since that time, many academic and industry groups began measuring the digital universe:

- Cicso publishes semiannual Visual Networking Index reports, projecting traffic for the next four or five years.

Fig. 2 – One Estimate of U.S. Internet Traffic



- Akamai now publishes a quarterly State of the Internet report, which highlights security threats and traffic trends and ranks download speeds by region, nation, and city.

- In 2009, Craig Labovitz and U. Michigan/Atlas Observatory colleagues used a large, global, two-year sample of real Internet traffic to document (in a report and paper) the changing architecture of the Internet and its key sources and transmitters of traffic.

- UC-San Diego renewed the well-known "How Much Information?" study, previously conducted at Berkeley.

- EMC sponsors an annual series of IDC "Expanding Digital Universe" reports.

- The journal *Science* published a study of "The World's Technological Capacity to Store, Communicate, and Compute Information."

- McKinsey recently issued a "Big Data" study, linking the exaflood with specific beneficial economic impacts in health care, geolocation services, retail, and government.

- World Wide Web pioneer Tim Berners-Lee just received a million-dollar grant from Google to "index" the entire Web, an attempt to really measure how much content is connected to the Internet. The list goes on.

Retrospective

Looking back on the themes we thought would drive the Net may be a useful way to update and revise our quantitative and qualitative projections.

- In 2003, we said Web video, based on increased deployment and adoption of real broadband access networks (see Fig. 3), would take off like a rocket and result in a "new surge of Internet traffic." Today, YouTube alone receives 48 hours of video uploads each minute, or eight years of content uploaded every day. It streams three billion videos per day. Playbacks in 2010 reached 700 billion.

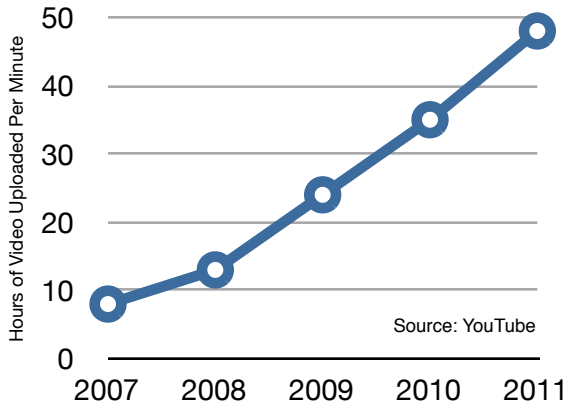
- In 2007, we said YouTube videos would increasingly be HD. Today, 10% of YouTube videos are HD.

- We expected the Mobile Revolution would significantly boost the time that people spend both creating and consuming content. YouTube reported that in 2010 its mobile video playbacks increased 200%, and most sources agreed that wireless data traffic overall grew more than 100%.

- We said that as part of the Mobile Revolution, the number and diversity of mobile device formfactors would grow. The rapid and widespread adoption of the iPad and other tablets is just one manifestation of this projection.

- In 2003, we wrote that inexpensive digital imaging chips (digital cameras) would increasingly be embedded in "every PC, laptop, Xbox, PlayStation, mobile phone, ATM, baby nursery, and auto bumper. Digital cameras will cover most angles of most amateur athletic, educational, theatrical, and family events." The ubiquity of cameras in mobile phones especially, we wrote, would result in a surge in wireless data traffic. Apple just reported that more photos uploaded to Flickr have been taken with the iPhone 4 than with any other camera or device. Facebook reports it receives 100 million new photos per day and now hosts around 100 billion photos. The photo-sharing app Instagram grew from 80 beta users a year ago to 10 million today.

YouTube Receives 48 Hours of Uploaded Video Each Minute



- In 2007, we said Netflix would move from a DVD-in-the-mail model to an Internet streaming model, and that these streams would account

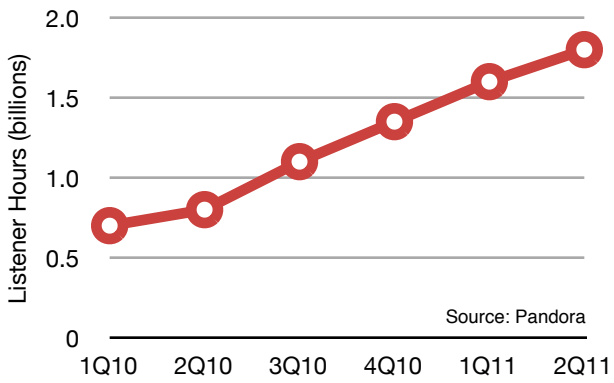
for a large increase in Internet traffic. Within just one month of introducing its streaming-only subscription plan in December 2010, Netflix streams jumped 38% to 200 million in January 2011. In May 2011, Sandvine reported that Netflix streams accounted for 29.7% of downstream traffic during peak evening hours in the U.S., increasing to 32.7% in October 2011.

- We projected that video calling and telepresence would, in the latter portion of the 2007-2015 period, yield massive traffic increases, along with the need for reduced latency and jitter. Although Skype, Apple's FaceTime, Citrix's GoToMeeting, and Cisco's Telepresence, among many other video chat applications, are gaining in usage, we have yet to witness mass adoption of these real-time conversational video tools, which we still anticipate.

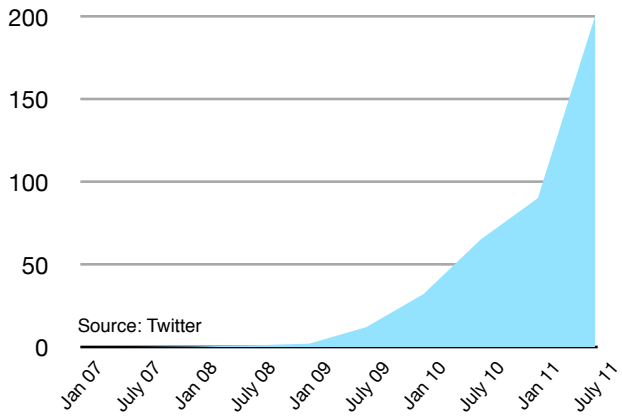
- We also said online gaming and virtual worlds would, toward the end of the period, boost traffic. These real-time rich visual applications do not yet account for a large portion of traffic but still appear poised for explosive growth.

- We said content delivery networks (CDNs), which cache content closer to end users, would grow dramatically in size and in their centrality to the architecture of the Net. Two CDNs, according to Atlas Observatory, are now the 7th and 8th largest "ISPs" on the planet. Google, moreover, which is in many ways a CDN, is the second largest "ISP."

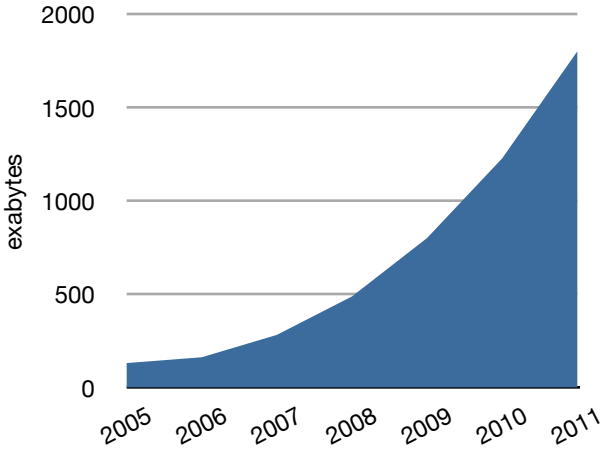
Pandora Users Listened to 1.8 Billion Hours of Streamed Music Last Quarter



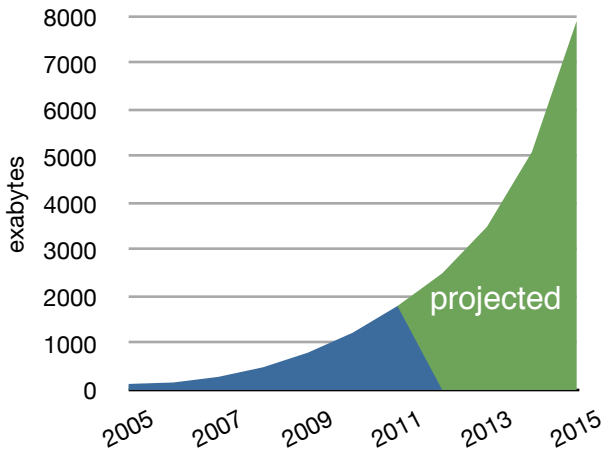
Millions of Tweets Per Day



Digital info created and replicated . . .



. . . could reach 8 zettabytes by 2015



Source: IDC - Digital Universe, 2011

- We thought remote back-up of files and photos would rapidly increase in popularity and that, far beyond simple remote back-up, our PCs and devices would become more intimately integrated with the cloud. Apple's iCloud service will lift this already fast-growing practice to a higher level of sophistication and market acceptance. Apple has built a new data center in North Carolina to support iCloud.

Video Is the Internet Star

Because of its data density, online video is *the* major driver of network traffic. It continues to set new records each month. In August 2011, according to comScore, 180 million unique U.S. viewers watched 6.9 billion sessions, for a monthly average of 1,080 minutes (18 hours) per viewer.

Google alone, mostly through its YouTube property, led the way in August with 162 million unique U.S. viewers, 3.5 billion viewing sessions, and 343.5 minutes per viewer. YouTube, according to one estimate by Sandvine, represents 11.04% of peak downstream U.S. traffic; Flash video is 4.88%; and Hulu is 1.09%.

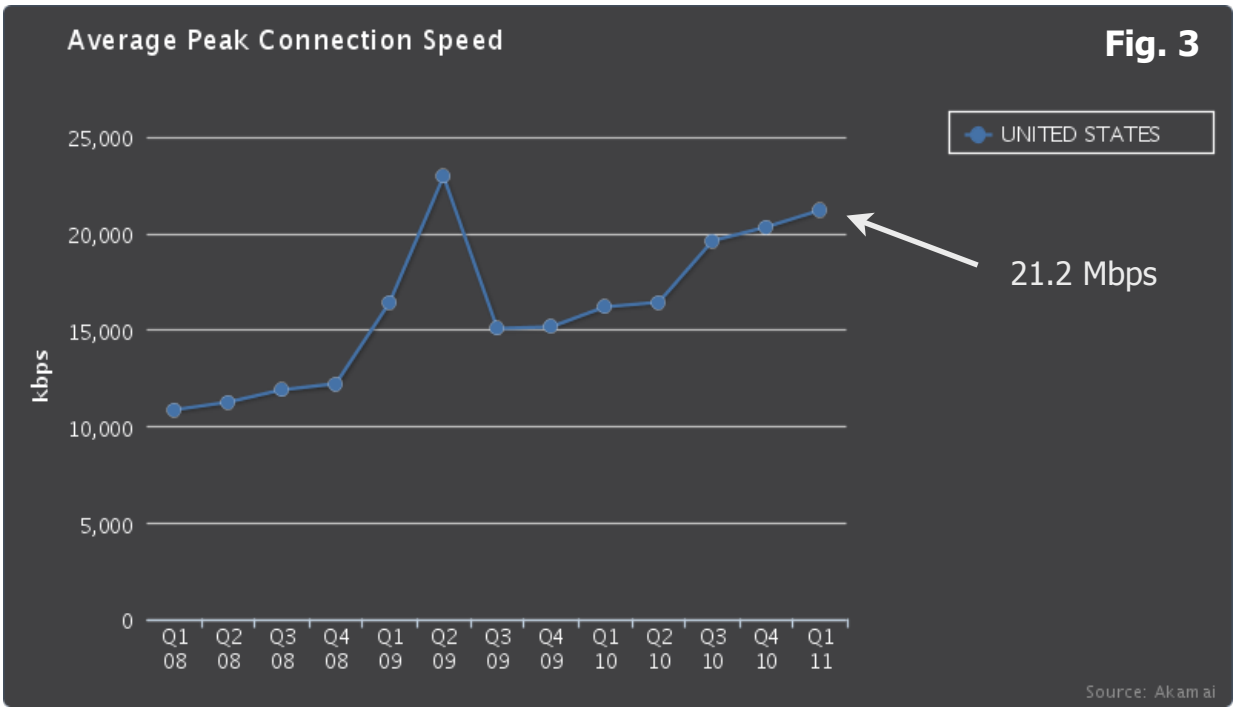
P2P traffic from the likes of BitTorrent remains a very large, if falling, portion of network traffic. Among other rationales, P2P is a technique to economize on scarce bandwidth. But as real broadband enabled real-time streaming, the relative need for P2P decreased.

Table 1 – U.S. Online Video – August 2011

Source: comScore	Total Unique Viewers (000)	Viewing Sessions (000)	Minutes per Viewer
Google sites	162,050	3,536,489	343.5
Vevo	62,285	519,702	60.9
Facebook	51,651	186,106	17.6
Viacom Digital	49,906	317,001	67.6
Microsoft sites	46,436	250,741	45.2
Yahoo! sites	45,475	237,973	46.3
AOL	40,671	260,666	54.7
Turner Digital	33,040	130,131	31.0
Hulu	26,413	166,500	192.4
NBC Univ.	24,994	71,491	14.6
Total U.S.	*180,379*	*6,908,009*	*1,080.0*

Skype, the voice-over-IP and video chat service, is now a significant portion of Internet traffic. Sandvine estimates that Skype is 1.29% of aggregate traffic and 3.81% of upstream traffic. (Because Skype is interactive and symmetrical, it creates proportionally more upstream traffic than other one-way video applications, which generate mostly downstream traffic.)

In recent days, Netflix signed a new content deal with Dreamworks; Amazon added Fox to its existing Prime Streaming lineup of CBS, NBC, Sony,



and Warner Bros. content; and YouTube is moving quickly to supplement its dominant position in free video by partnering to offer 100 dedicated channels of high-end professional content.

We also think Apple could enter the "TV" market in a much more substantial way. *BusinessWeek* and others report that Apple is moving beyond its existing tiny peripheral device, called AppleTV, and is readying an actual television display to be paired with a major upgrade of its Net-based video service. Apple has succeeded in the past with such integrated device-content offerings like the iPod and iTunes.

With its new video capabilities, HTML5 will bring much greater power and flexibility to the Web and to devices (like Apple's) that don't support video players like Flash. In addition, the upgrade from standard definition to High Definition (HD) video is a new source of traffic growth and is creating challenges for network operators.

Mobile Revolution

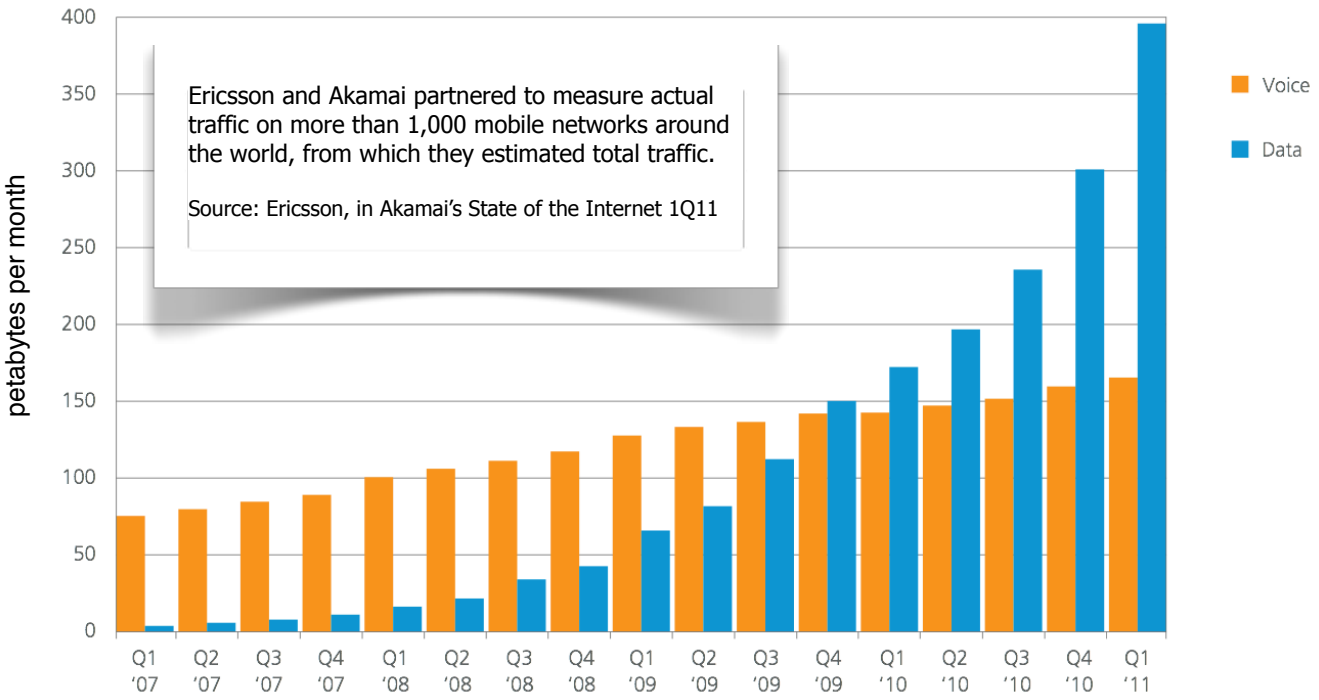
When we first started building 3G mobile networks in the mid-2000s, many thought it a silly and wasteful exercise. How would we ever use this capacity? Too much bandwidth at too much expense, not nearly enough applications and services. Mobile device screens were thought too small and too lifeless to watch video, surf the Web, or read, not to mention play games or video chat. There were no mobile "apps" as we know them today.

Just a few short years later, a 2011 Credit Suisse survey of U.S. wireless carriers found their networks running at 80% of capacity, meaning many network nodes are tapped out. The projected unusable abundance of 3G wireless capacity had, thanks to the iPhone and its smartphone cousins, turned into a severe shortage in many big cities.

As of October 2011, 500,000 distinct iOS apps had been downloaded 18 billion times on 250 million iOS devices. The competing Android OS marketplace of devices and apps is, by some measures, growing at an even faster rate and now powers some 43% of U.S. smartphones. Amazon announced in April 2011 that for every 100 paper books, it now sells 105 ebooks (delivered to mobile e-readers via wireless links).

The U.S. just surpassed the 100% penetration barrier – more wireless subscriptions (327.6 million) than people. Wireless Intelligence estimates nearly 1.5 billion 3G subscribers worldwide, and

Fig. 4 – Global Mobile Data Traffic Grew 130% Last Year



by 2015 3G subscribership will likely pass 3 billion. It estimates six billion mobile phone connections globally by the end of 2011, when Morgan Stanley estimates the worldwide total number of connected mobile devices will surpass 10 billion.

Ericsson and Akamai show that by the first quarter of 2011, wireless data transmitted over mobile phone networks approached 400 petabytes per month. This was a 130% increase from the first quarter of 2010 and was around 80 times more than monthly mobile data traffic in early 2007.

U.S. service providers invested $26 billion in wireless infrastructure in 2010. For the decade 2001-10, U.S. wireless investment was $232 billion. Investments in 4G networks are now in full swing. (For an overview of 4G and other mobile technologies, see this paper by Rysavy Research.)

I, Cloud

Hotmail, Yahoo! mail, and Gmail were early examples of mass-market applications hosted not on PCs or office servers but in the cloud. Consumer remote back-up providers like Mozy, Carbonite, and Dropbox gained widespread adoption in recent years.

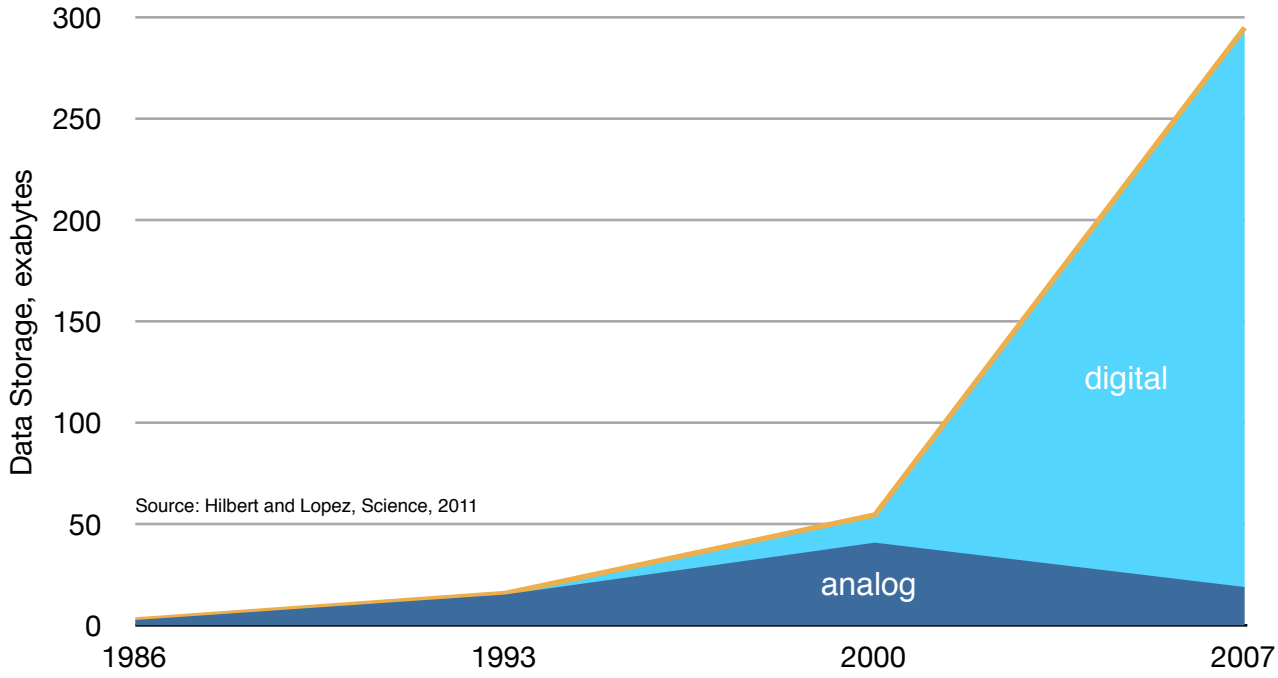
Salesforce.com revolutionized the customer relationship management (CRM) business with its cloud service. Moving a step beyond, Salesforce now serves as a sort of app store for the enterprise world.

The thousands of Web apps hosted in the cloud today are second nature. Cloud, like many big ideas, arrived with a bang but became a cliché rather quickly. Not for too much longer will we even think about "local" versus "cloud." Storage, bandwidth, and processing will increasingly be seamlessly integrated, making best use of the power of local devices and cloud resources.

The cloud virtualizes everything: first it was servers and disks; now it is Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

For a (virtual) big box retailer, Amazon has been awfully innovative. Many tilted their heads when several years ago Amazon introduced its Web Services (AWS) and Elastic Compute Cloud (EC2), allowing Web companies and start-up developers to rent its mighty storage-compute-network infrastructure. What was Jeff Bezos doing with this supercomputing science project? Turns out, Amazon was amortizing its vast infrastructure that serves its traditional services over a much wider array of cloud offerings. "Each day," notes *BusinessWeek*, Amazon "adds enough computing muscle to power one whole Amazon.com circa

Fig. 5 – World's installed capacity to store information



2000, when it was a $2.8 billion business."

With the introduction of its Kindle Fire tablet, Amazon has even invented a new kind of browser, called Silk, that gets its power from Amazon's massive cloud assets. Silk offloads much of the processor-, bandwidth- and storewidth-intensive heavy lifting from the thin tablet itself and lets the AWS cloud do much of the work.

A typical webpage might consist of 80 objects (text, images, JavaScript, ads, etc.) that are often retrieved from around the Internet and then integrated and composed by your browser on your device. Silk lets the Amazon cloud collect the objects, assemble them, and then send a composed webpage to your Kindle Fire. Upon first look, some analysts even said Silk was more than a browser – maybe the first "cloud OS."

Facebook's new Open Graph paradigm will embed many rich media apps more deeply into the Facebook world. At the 2011 f8 conference, Netflix, News Corp., Spotify and others announced new deep integrations, further expanding Facebook as not just a social network but a cloud-based multimedia platform. Adobe, the maker of Photoshop, Flash, and other graphics tools, is moving most of its software to the Web via its new Creative Cloud. IDC thinks cloud services could reach 5 zettabytes by 2020.

Big Data

"Data is the new oil," says Andreas Weigend, former chief scientist at Amazon.com. "Oil needs to be refined before it can be useful. Big data start-ups are the new refineries." From tick-marks on stone thousands of years ago to hand-written ledger entries in centuries past, data has been around for a while. But the recent explosion in digital data – and our capacity to create, collect, store, transmit, massage, and analyze it – is something wholly new.

As recently as 2000, analog storage still trumped digital storage. But by 2007, a 2011 article published in the journal *Science* found, analog storage had actually declined in absolute terms and digital storage had grown 15 times larger than analog (see Fig. 5). IDC estimates the world will create or replicate 1,800 exabytes of data in 2011, up from 130 exabytes in 2005. It thinks we could approach almost 8,000 exabytes (8 zettabytes) by 2015.

Data has always driven financial markets. But new data sources will increasingly dive other industries. Examples: medical data, customer data, social network data, retail data, geolocation data, sports data, and sensor data from millions of cameras, machines, cars, planes, factories, weather stations, and network nodes. The PGA golf Tour records with lasers each and every shot hit and reports the results, down to the inch, in

Abundance, Not Apocalypse: The Exaflood and Its Discontents

When we began over a decade ago writing about the coming explosion of broadband connectivity and rich media delivered over the Internet, we viewed it as good thing. We still do.

We threw out a word – exaflood – to connote the enormous waves of data traffic that would flow over the world's networks. (The prefix exa means 10^{18}, or billion billion. One exabyte is a billion gigabytes.) We said broadband infrastructure and capacious data centers would drive new forms of traffic. We also said it would be a challenge accommodating new surges of data storage and transmission. In 2007 we wrote, "Today's networks are not remotely prepared to handle this exaflood." That was emphatically true.

Whether through misinterpretation or misrepresentation, some said this was a chicken-little prediction of Internet collapse. It was not. Nowhere had we said anything about crashing the Internet. We said that if we continued the process of building new fiber optic and wireless networks and new data center capacity, not to mention upgrading end-user devices at a Moore's law pace, we would both drive new traffic and manage this "flood." But it would be a process of never ending innovation, and sometimes there would be bottlenecks. Moreover, if we encouraged investment and flexibility in the digital arena, we would drive innovation at the fastest clip and also have the flexibility to adapt to digital world's unpredictability.

The iPhone phenomenon proved a good example. Apple designed a new device to exploit the newly capacious EDGE and 3G mobile networks. The iPhone had the first really good mobile Web browsing and video capabilities. The touch screen interactivity and software downloads from a newly conceived App Store drove traffic through the roof. In many urban areas, where iPhone and other smartphone penetration was high, traffic overwhelmed network capacity. There was no "crash," but clearly a flood of some sort.

In technology, abundance is a good thing. There is no end state. New abundances in one place create new bottlenecks elsewhere. Then further innovations come along and move the bottleneck yet again.

real-time. The sport with the richest history of collecting and analyzing information is baseball. The new movie "Moneyball," based on Michael Lewis's book, is a Big Data story.

Much of the analysis of Big Data is being performed using Hadoop, a software framework that leverages cluster computing to process large amounts of information. In July 2011, Facebook said it runs the largest Hadoop cluster in the world – some 30 petabytes. Email marketing companies secure access to the Twitter "firehose" – essentially a copy of all tweets across the globe – in order to spot trends and target consumers.

The McKinsey Global Institute looked at Big Data from an economic perspective. It estimates intensive collection, analysis, and implementation of fine-grained medical data boost annual economic value in the U.S. health care sector by $300 billion. McKinsey thinks personal geolocation services could expand annual consumer surplus by $600 billion globally.

What's Next

The rise of multimedia content delivered over the Web is a fundamental departure from the early days of email, data exchange, and simple websites. In our earlier reports, we outlined a new set of technologies that would take us well beyond existing notions of Web video and cloud computing.

Call it online gaming. Call it cloud streaming. We call it the "exacloud." It is cloud computing but of a scope and scale never seen before. Imagine a supercomputer built not of microprocessors (CPUs) but of thousands of graphics processors (GPUs). One of the world's most powerful supercomputer is IBM's one-petaflops Roadrunner at Los Alamos National Labs. But in 1% of the space and for 3% of the cost, we can build a graphics supercomputer that delivers three times Roadrunner's performance – three petaflops.

Connect this computer to the Internet, and you can stream any real-time interactive 3D video experience at any resolution to thousands of people using any browser on any device, from a home-theater to an iPhone. This "exacloud" will transform video games, movies, virtual worlds, business software, and most other media. Piracy goes away. So do DVDs, game boxes, and maybe even expensive personal computers. New content and software subscription models open up. Casual users gain access to services previously based on expensive, proprietary devices and platforms. Based in the cloud instead of on your device, interactivity thrives.

Firms like OnLive, Otoy, Gaikai, and others are now bringing this vision to life. OnLive has raised some $100 million in venture funding, is valued at $1.5 billion, and is now streaming hundreds of game titles from the major video game publishers. Los Angeles-based Otoy won an Academy Award for its work on "Avatar," "Benjamin Button," "Spi-

Table 2 – Top Global ISPs By Traffic – 5 New Between 2007 & 2010

Rank	2007		2009		2010	
	ISP Name	%	ISP Name	%	ISP Name	%
1	A	5.77	A	9.41	A	9.09
2	B	4.55	B	5.70	Google	7.00
3	C	3.35	Google	5.20	B	4.70
4	D	3.20	F	5.00	F	3.00
5	E	2.60	H	3.22	H	2.96
6	F	2.77	Comcast	3.12	K	2.89
7	G	2.24	D	3.08	L (CDN)	2.82
8	H	1.82	E	2.32	M (CDN)	2.60
9	I	1.35	C	2.05	E	2.30
10	J	1.35	G	1.89	Comcast	2.07
Top 10 % of Total Traffic		30%		41%		40%

Source: Craig Labovitz, et al. Atlas Observatory, 2011.

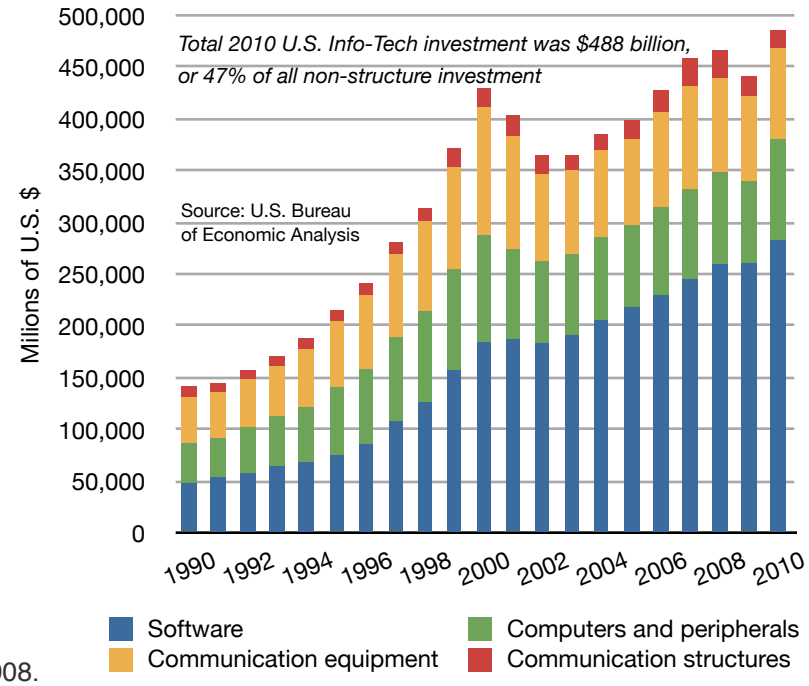
derman 2" and "3", and other feature films. Unlike OnLive, it does not offer consumer subscription gaming but instead provides its software and cloud infrastructure to third-party gaming, movie, and business software firms. Otoy is even creating a suite of software and services that empower individuals – from startup firms to hobbyist developers – with graphics tools that rival mighty studios like Industrial Light and Magic.

This new paradigm could generate enormous amounts of Internet traffic. High-definition video requires big bandwidth, and real-time applications tolerate very little delay. UC-San Diego estimates that 55% of total American information consumption, or 1,991 exabytes per year, is (brace yourself) video games. If just 10% of these games moved online, they would generate twice the worldwide Internet traffic of 2008. Video is not always the most important content on the Web, but it defines the architecture and capacity of (and often pays for) the networks, data centers, and software that make all the Web's wonders possible.

On one recent September evening, gaming, according to one traffic analysis source, accounted for around 2.69% of traffic. This was merely a snapshot – games are sometimes more, sometimes less – but it showed that gaming in its still-primitive state is already significant. As the "gaming" category moves toward streaming of rendered video, however, it could become the major source of new network traffic.

Beyond gaming, the exacloud will likely accommodate remote rendering of numerous apps and displayed content. Companies like SolidWorks and Autodesk make powerful software that runs on big-horsepower hardware to assist engineers with sophisticated 3D design and modeling. If high-powered apps, such as AutoCAD, can be hosted in the cloud, a tiny fraction of one supercomputer can replace hundreds of expensive workstations or an enterprise cluster. Although the first to make use of the exacloud's power will be games and engineering apps that require intensive graphics processing, cloud streaming will expand its scope across a wide range of applications and content.

Fig. 6 – U.S. Info-Tech Investment



The applications for real-time video are limitless. Video conferencing and chat will grow. Lots of other novel ideas will surprise us. A startup called Color has a new app that turns smartphones into real-time windows on the world – call it an "apperture." You can achieve much the same thing today via Apple's FaceTime, but Color promises deep integration with Facebook so that your friends can see where you are and, if they click on your stream, watch what you are seeing in real-time.

Because cloud-based applications are hosted remotely, they depend on ever more robust broadband and wireless links. Rich two-way multimedia and real-time apps require capacious, low-latency, nearly ubiquitous connectivity. For the cloud to work at the highest levels, it must perform as if the app is sitting on your desktop.

The Paralleladigm

"When the network becomes as fast as the processor," Eric Schmidt famously said, "the computer hollows out and spreads across the network."

That's an elegant prediction of what today we call broadband and cloud. But this epochal industrial transformation required a fundamental shift in

technology and information architecture. The old copper telephone lines, Von Neumann computer schematic, and client-server network model would not suffice for an era of real-time communication. Entirely new technologies both created the possibility of today's Internet and must advance at a furious pace to merely keep up with Web video, Big Data, and the exacloud.

The common denominator in this new technological paradigm is parallelism.

Over the past two decades, scientists noticed that the actual performance of microchips would not keep up with the addition of more silicon transistors and faster frequencies, growing at the pace of Moore's law. Slow access to memory meant that billions of transistors and clock cycles were left waiting, doing nothing much of the time. Chips running at ever higher frequencies, meanwhile, consumed way too much power and would melt without expensive cooling methods. This Von Neumann bottleneck meant that chips were getting larger and hotter but wouldn't deliver the bang for the buck promised by Moore's law.

Thus the rise of multi-core chips. The multi-core wave was previewed by the rise of graphics processors, or GPUs. Traditional microprocessors, or CPUs, couldn't deliver the parallel processing power needed for video games. Even the most powerful Intel CPU was not very good at accepting input from a teenager's joystick and then instantly rendering millions of pixels onto a video display, dozens of times per second. But the new GPUs, from Nvidia and ATI (now part of AMD), were massively parallel, containing dozens of individual specialized processors. Today GPUs are moving well beyond gaming into every digital field, from finance to oil exploration. Often now programmable, there is a new generation of general purpose graphic processors, or GPGPUs.

Neither could traditional microprocessors keep up with network traffic. For years the network equipment companies like Cisco would build their own highly specialized proprietary chips to power their routers and switches. Companies like AMCC and Motorola also built these network processors, or NPUs, but they were based on the conventional RISC computer architecture. An NPU is to a Cisco router what a CPU is to a Dell computer. Around the year 2000, however, a new company called EZchip led a new generation of network processors with a radically new architecture containing hundreds of parallel task optimized processors (TOPs). Only this architecture, or those similar, have been able to deliver fiber-speed routing and switching in a fiber-speed world of Internet multimedia. Like GPUs, the chief advantage of the new NPUs was achieving dramatically higher memory bandwidth – with much lower power consumption – thus diminishing much of the Von Neumann bottleneck.

Companies like Cavium were among the first to charge ahead with truly multi-core CPUs, now used in cloud data centers, where financial and transactional content, often encrypted, must be processed in real-time. Intel and AMD of course followed the lead of these parallel pioneers and now build mostly multi-core CPUs.

technology	parallel architecture
WDM (wave division multiplexing)	100s of wavelengths of light on single optical fiber
GPU (graphics processor)	512 stream processors on single chip
NPU (network processor)	100s of task-optimized processors on single chip
multicore CPU	2, 4, 8, 16 cores on single chip
cloud computing	massively parallel clusters with hundreds of thousands of servers
OFDM (4G LTE wireless)	100s of parallel sub-frequency bands
WiFi, femtocells	parallel wireless spectrum reuse
CDN (content delivery network)	parallel replication and delivery of static multimedia content
exacloud	a 1,000-GPU supercomputer delivering real-time dynamic content

The communications revolution would not have been possible, of course, without the singular contribution of parallel communications technologies like wavelength division multiplexing (WDM), which dramatically increased the capacity and flexibility of optical fiber by putting several, then dozens, now hundreds of separate communications streams onto a single thread of glass. Today, a single thread of optical fiber can transmit 69 terabits per second on 432 *parallel* wavelengths

Fig. 7 – More Estimates of U.S. Internet Traffic

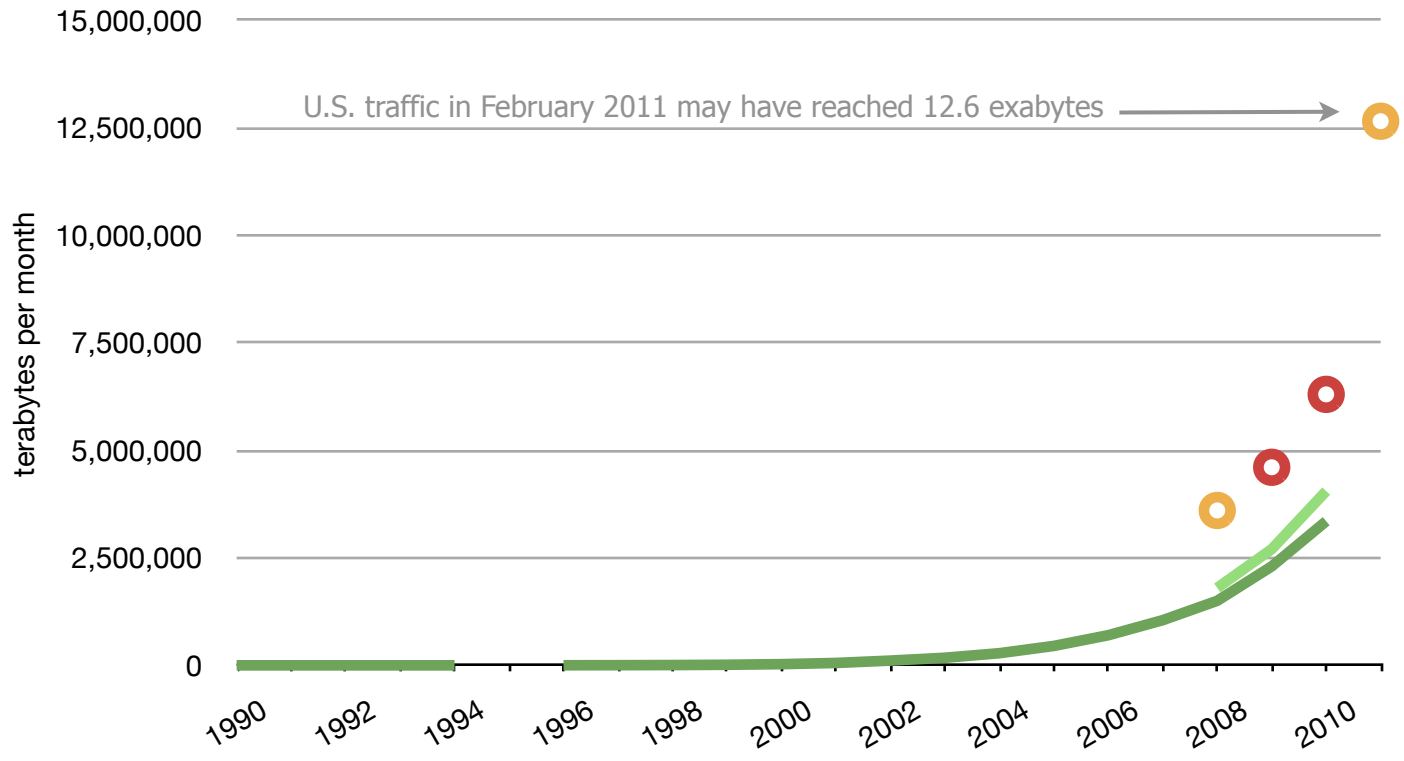
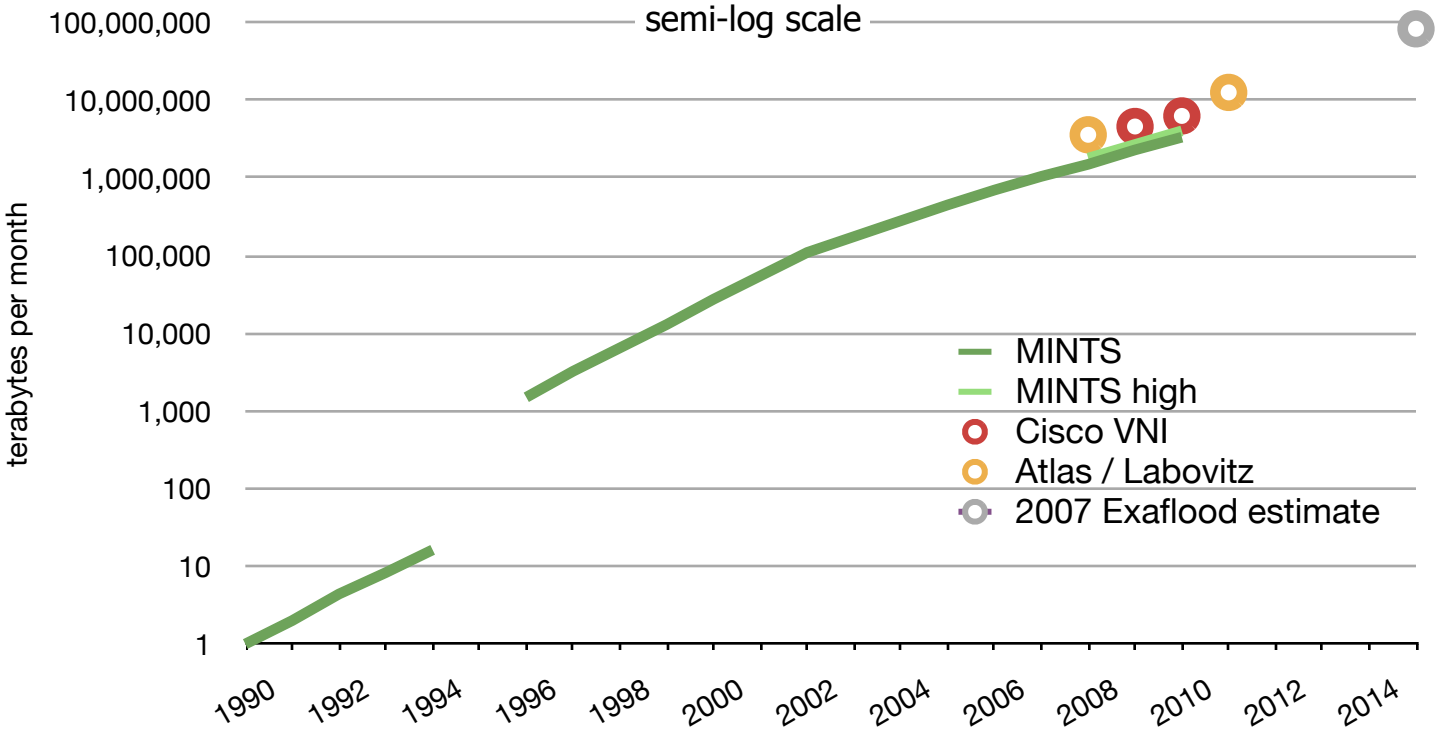


Fig. 8 – The traffic trend is consistent with our Exaflood estimate



("colors") of light over a distance of 240 kilometers.

Cloud computing itself is a massively parallel architecture that substitutes the virtualized resources of a vast pool of computing and storage power for the dedicated, localized power of your PC. Likewise, content delivery networks improve the performance of the multimedia Web by replicating, storing, and serving up content from thousands of parallel geographic locations around the world.

Like optical communications, the advance of wireless is chiefly a story of parallelism, where CDMA, OFDM, and femtocells rely on frequency and spatial parallelism to achieve their power.

The exacloud is the culmination of all these forces, where thousands of parallelized GPUs, themselves massively parallel in architecture, deliver thousands of simultaneous streams of rich content across optical and wireless networks, that rely on increasing parallelism to get these large amounts of data to end users, with as little latency as possible.

Today, academic and government supercomputer teams are building the fastest new machines using GPUs, thus imitating the commercial pioneers, OnLive and Otoy.

Traffic Analysis

Measuring one thing called "Internet traffic" is difficult, and becoming more so all the time. First, much of the traffic data is proprietary. Second, defining what is and isn't "the Internet" is tough. Much traffic is private (e.g., VPNs), and lots of networks like cable TV, IPTV, CDNs, and content companies that peer directly with broadband service providers may not interact with the traditional tier one backbone providers. For these and other reasons, Cisco instead estimates "IP traffic." Third, our networks are growing and changing so fast, collecting data and defining and comparing metrics over time is not easy.

Despite these challenges, a number of analysts have developed useful methods to estimate traffic levels and growth rates. Prof. Odlyzko at Minnesota Internet Traffic Studies (MINTS) collects data from a wide range of networks all over the world and derives an estimate.

Cisco projects future traffic by estimating consumer and business adoption patterns of broadband devices and services and the emergence of digital applications. Because its routers comprise a large portion of the world's network infrastructure, Cisco also has deep, empirical insight into real-time traffic loads and patterns.

Craig Labovitz, formerly of Arbor Networks and now with startup Deepfield Networks, conducted a major study over the last several years using a very large sample of real network traffic. Between 2007 and 2009, Labovitz and his Atlas team collected and analyzed 264 exabytes of traffic. They estimated traffic levels that were similar to Cisco's and slightly higher than MINTS'. As important, they found the Internet's architecture and its main players changing in profound ways.

Atlas demonstrated the growing centrality of Web video and CDNs. They also documented the rise of the "hyper giants" – content companies like Google and Facebook that, through their own data centers and direct peering with broadband service providers, had become network companies themselves. (As far back as 2003, we said Google was becoming an Internet infrastructure company.)

In April 2011, Labovitz updated his 2009 study, using data from February 2011. He estimated average peak Internet traffic of 90-110 terabits per second. In addition, the data showed the U.S. proportion of traffic actually increasing toward 50% of world traffic. This finding surprised because of the rapid growth of data centers and broadband usage in the rest of the world, where lesser developed nations are catching up from a lower base.

Analyzing Labovitz's data using typical diurnal traffic patterns, we arrive at a rough estimate for February 2011 U.S. traffic of 12.6 exabytes (see Fig. 7).

There is some reason to believe this estimate could be high. Some expert observers thought the February 2011 Atlas estimate could be 10% too much. The Atlas analysis also found the U.S. share of total traffic to be higher than other estimates. On the other hand, these measurements and estimates do not include private traffic like VPNs and pseudowire links, etc. It's not unreasonable to think we might be missing 10% of traffic through these and other sources.

Some think traffic growth has accelerated a touch in the last year or two. AT&T researcher Alexandre Gerber notices an uptick in traffic growth among AT&T DSL customers. Perhaps the pronounced shift toward Netflix's streaming service produced the bump? Or maybe it was just the overall increase in Web video viewing.

Gerber's overall estimate of the compound growth rate over the last decade is lower than the MINTS, Cisco, or Atlas estimates. But the measurement was of slower DSL networks, not across the broader Internet. As newer VDSL and fiber-to-the-x networks replace older DSL links, traffic on these networks may catch up with overall growth.

It's also important to remember that networks are build to accommodate *peak* traffic levels. Our estimate of total traffic, therefore, may offer insight to aggregate activity across the Net but isn't the most important metric for network architecture.

In 2007, we looked ahead and estimated how new rich-media and cloud applications might affect network traffic. Adding up the video, gaming, cloud, and mobile applications we saw emerging, we arrived at a very rough estimate of what the U.S. network would look like in 2015. We said U.S. IP traffic could hit 1,000 exabytes, or one zettabyte, for the year (see Fig. 8). Using a 2007 Cisco estimate as our baseline, that translated into a 56% rate of compound growth over the period – higher than some estimates, like Cisco's own, but not as high as others, such as Nemertes'.

MINTS continues to believe U.S. traffic is growing between 40% and 50% annually. Atlas agrees. The newest Atlas/Labovitz data show that traffic continues to grow briskly – something like compound annual growth of 52% between its 2008 and 2011 estimates. Our original 2007 "exaflood estimate" for 2015 traffic is thus not inconsistent with the current trend.

Video Drives Traffic

Most households *today* receive a continuous gigabit-per-second stream of video in the form of cable or satellite TV channels. These shared network architectures deliver broadcast content very efficiently. They send you everything they've got, and you tune into the channel you want.

These networks, however, lack the flexibility and interactivity of switched IP networks. We use local appliances, such as DVRs, to increase the flexibility of the broadcast networks, time-shifting content so we can watch it on demand. DVRs, however, can't match the Internet in flexibility, and notwithstanding 500 channels, video programmers still cannot match the Internet's size, diversity, choice, and interactivity.

If we measure broadcast content as it enters each home and office, it dwarfs current Internet traffic. If we measure broadcast content at the source, Internet traffic wins by a large margin.

Broadcast economizes on switching. Narrowcast economizes on bandwidth. There will be a mix of broadcast, narrowcast, multicast, and symmetrical interactive networks and services to match consumer preferences and the era's technological and economic constraints.

The pace at which individualized and interactive streams of rich video grow will in large measure determine the growth of overall network IP traffic.

Public Policy

Among McKinsey's six key takeaways for policymakers in its "Big Data" report was this: "Ensure investments in underlying communications and information technology infrastructure."

For half a century, the U.S. has led the world in digital computer and communications technology. Scientists and entrepreneurs have built our digital economy through experimentation and rapid innovation, spurred by venture capital and enabled by very large digital infrastructure projects. The entrepreneurship and investment that has sustained such fast growth for so long is due, in substantial part, to light touch government policies (at least compared to other industries). There have been mistakes, but for the most part scientists, entrepreneurs, and big investors have been allowed to build new things, try new products, challenge the status quo, cooperate and compete. They have also been allowed to fail.

The FCC's recent Net Neutrality order is a potential break from this basic hands-off approach, but it is now being challenged in the courts, and there is reason to believe a heavy-handed, hard-edged Neutrality regime will be avoided. Many other policy questions, from privacy and behavioral advertising to cybersecurity, are important. *For the ca-*

pacity topics covered in this report, the chief policy concern is in wireless.

The digital ecosystem is ever-evolving. We build new software, hardware, and network components to provide new services and to relieve bottlenecks created by increased usage, made possible through previous abundance. Broadband enabled the rise of cloud computing, for example, and now the cloud demands ever faster and widespread broadband. It's a never ending process.

In 2010, investment in U.S. fixed info-tech infrastructure totaled $488 billion (see Fig. 6). That was almost 47% of all U.S. non-structure fixed investment. The broadband and mobile service providers alone invested around $65 billion.

But to run real-time apps from the cloud and to accommodate high-definition interactive video, we will need another decade's worth of broadband and wireless innovation and investment. Cloud and video, essentially, require a new network – ever more robust fiber optic links connecting data centers with homes, businesses, and a much wider array of wireless access points.

Mobile devices will increasingly rely on the cloud for content, computing, and storage. Video chat will be mostly mobile. But these services require much faster, more robust, and more ubiquitous connectivity than exists today.

Today's crucial scarcity is thus wireless capacity. Part of this scarcity can be relieved through investment in new 4G networks and femtocells. A substantial portion of the scarcity, however, is due to a lack of available clean radio spectrum – the type of spectrum that can support 4G networks and the volumes and diversity of future traffic. Indeed, the macro-, micro-, pico-, and femtocells that will make up the HetNets (heterogeneous networks) of the future are vastly more powerful and flexible when using wider spectrum bands.

The Federal government, however, owns 61% of the best airwaves between 174 MHz and 4 GHz, while private mobile broadband providers control just 10%. Much of the remaining capacity in private hands is the old broadcast TV spectrum, which is trapped in a technology time capsule and is severely underutilized. Unleashing this spectrum through auctions and allowing greater flexibility to use, buy, and sell existing private spectrum is a paramount concern – if we want to survive and thrive in the exaflood era. EE

Fig. 9 – Cisco: mobile will grow as portion of total traffic