COMPUTER & COMPUTATIONAL SCIENCES

Los Alamos National Laboratory
RADIANT
www.lanl.gov/radiant

Bridging the Disconnect Between the Network and Large-Scale Scientific Applications

Wu-chun (Wu) Feng
feng@lanl.gov

Research & Development in Advanced Network Technology (RADIANT)
Computer & Computational Sciences Division
Los Alamos National Laboratory

NNSA

Los Alamos Unlimited Release: LA-UR-03-6652

Los Alamos NATIONAL LABORATORY

Motivation

- Years of research on OS-bypass protocols (e.g., FM and ST) and RDMA engines (e.g., Elan3/Quadrics)
  - Success?
    - 7.2 Gb/s throughput and 4-μs end-to-end latency. MPI-to-MPI.
  - Problems:
    - Difficult for application scientists to use.
      - Some scientific applications, particularly TCP/IP-based ones, still felt "disconnected" from their computing and networking environment.
    - Reliable but sensitive to temperature changes.
    - Source-routed Elan3/Quadrics "*not compatible*" with IP-routed network.
    - More problematic issues in their computing environment.

Why Green Destiny, a supercomputing cluster running only Fast Ethernet?

"Green Destiny" Supercomputer

http://sss.lanl.gov

- A 240-Node Supercomputing Cluster in a "Telephone Booth"
- Each Node
  - 667-MHz Transmeta TM5600 CPU
    - Upgrade to 1-GHz Transmeta TM5800s. Top 500 Run?
  - 640-MB RAM
  - 20-GB hard disk
  - 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 160 Gflops peak (240 Gflops with upgrade)
  - 240 nodes
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)

"Developments to Watch: Innovations," *BusinessWeek*, 12/02/02.
"At Los Alamos, Two Visions of Supercomputing," *The New York Times*, 6/25/02.
"Two Directions for the Future of Supercomputing," *slashdot.org*, 6/25/02.

Virtues of Green Destiny

- Ease of Deployment, Management, and Use, i.e., Transparency
  - Though significantly easier, it still requires a "cluster wizard" to deploy and manage … but we're working on this.
- Reliability and Availability
- Computational Efficiency
  - No need for special infrastructure, e.g., machine room.
    - Performance/Power: Up to 10 times better.
    - Performance/Space: Up to 50 times better.
  - Environmental sustainability and friendliness.
- Good Performance

"Pseudo-Nirvana" for Large-Scale Scientific Applications

Better Performance Improved Transparency

Does The Network Have Similar Virtues?

- How many SIGCOMM'03 staff (i.e., light blue shirts) does it take to get a SIGCOMM attendee connected to the Internet?
  1. Read the instructions in your registration packet!
  2. Keep your "authorization web page" open after authorizing yourself to get free Internet access.
  3. You want to run IMAP? Go to the green-colored Ethernet cables for NAT-free access.
  4. Need to print out my slides. What IP address and netmask?
  5. Can't access anything. Oh, the DNS server is down?!
- What's my point?

  Do any of us *really* realize how much "network state" that we carry around with us. What's a poor "scientific applications" guy to do? What do your (grand)parents do?

Looking Over the Fence at Networks: A Neighbor's View of Networking …

- Research (David Clark, MIT and David Patterson, UC-Berkeley)
  - Too ad-hoc, too brittle, too little rigor, etc.
  - Is this still the case? Or never the case to begin with?
- The Internet
  - An amazing feat in scalability, but it is brittle (and in some cases, unusable) for the common user, e.g., ftp & passive.
  - How many of you have set-up DSL for your (grand)parents?
  - How many have had to answer any of the following questions?
    - Why do we need a "firewall"? Is something going to catch fire?
    - What are DHCP and IP?
    - Netmask? Are you disguising our network from viruses?
    - DNS? Is that short for "Do Not Start"?
    - Why is this taking so long? Aren't you a network expert?
    - Why can't I plug my cable into the wall to get onto the Internet?

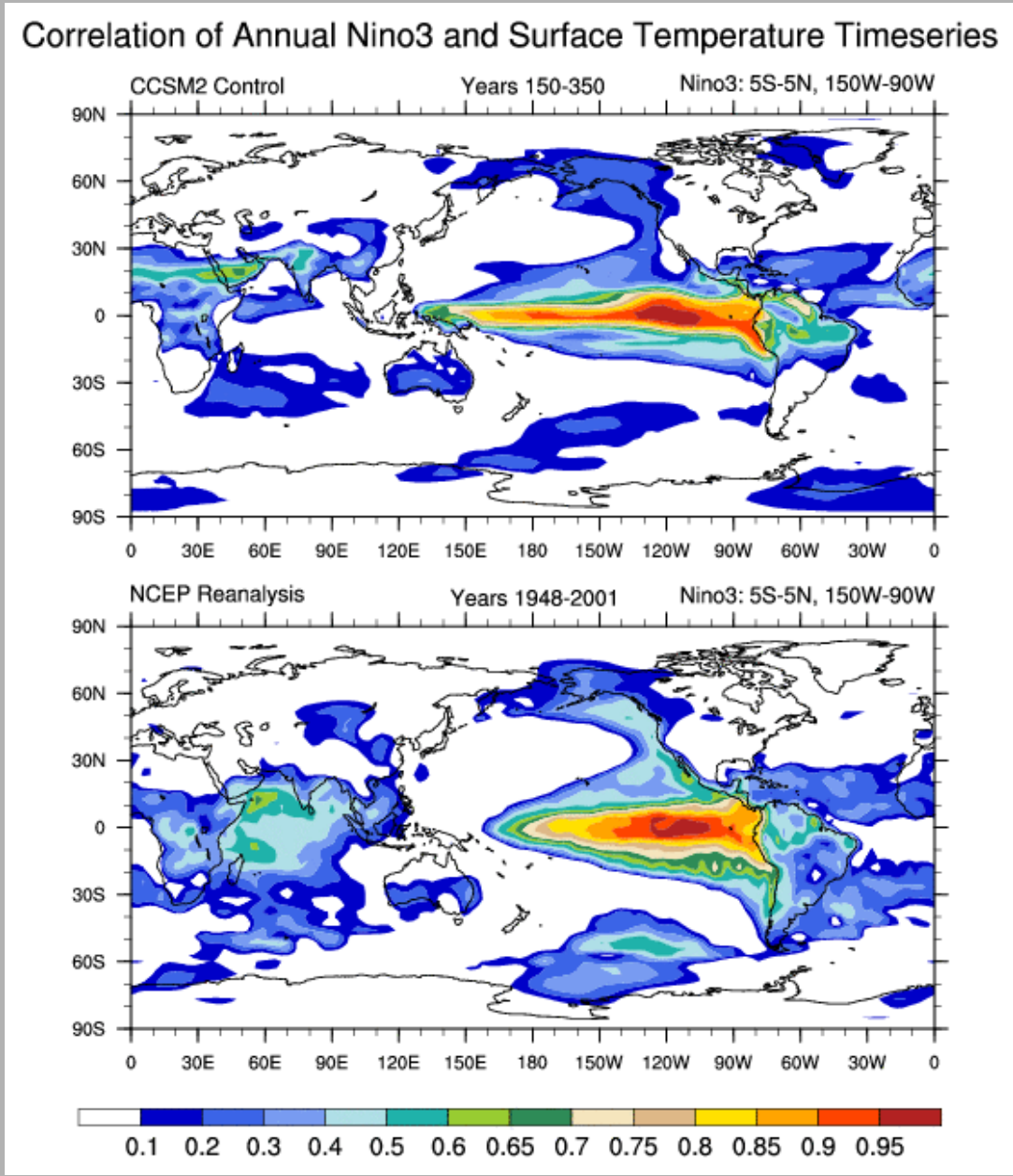
The Network: A Future Utility?

- Why can't I plug my cable into the wall to get onto the Internet?
- How many three and four-letter words (er, acronyms) do I really need to know?
- What is missing?
  - "Complete" transparency
  - Stable/fast performance
    - Smooth audio/video feed
    - Reliable transfer of data (from MP3 to high-energy physics data)
    - Always available.
  - Real security
    - See Cheriton's keynote talk.

Can the network be eventually viewed as an electrical utility?

What is the Disconnect?

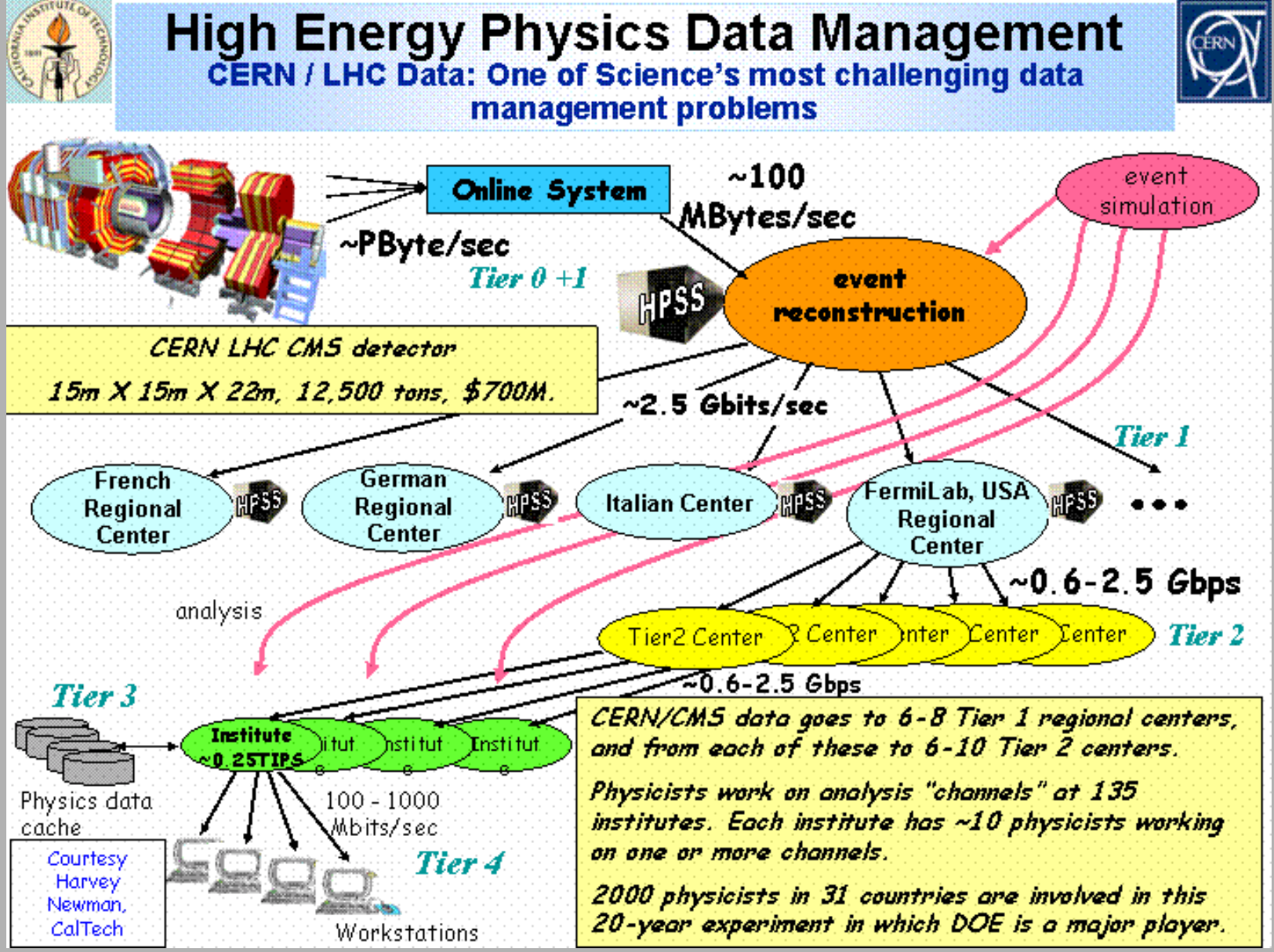
1. Large-scale scientific applications are *very* unhappy about their network and I/O performance.
   - Shouldn't "out-of-box" performance provide the best performance possible?
   - Why do we need "network wizards" to close the performance gap between what a novice can achieve and what an expert can achieve?
   - Why can't I transfer a petabyte of data reliably as a single FTP transfer?
2. The "network wizard" steps needed to improve performance (see http://www.psc.edu/networking/perf_tune.html) and network security should be *automatic* and *transparent* to the application.



200 years of modeling El Niño events and surface temperatures on the Community Climate System Model (CCSM2) closely correlate with 50 years of actual climate data.

(Source: NERSC)

http://www.nersc.gov/news/thousandyear090402.html



A Hierarchical Data Grid as Envisioned for the Compact Muon Solenoid Collaboration.
The grid features generation, storage, computing, and network facilities, together with grid tools for scheduling, management, and security.

COMPUTER & COMPUTATIONAL SCIENCES

Application Requirements: Beat FedEx!

Application	Now / Near Term	5 Years Out	10 Years Out
Climate	Authenticated data streams thru firewalls. Ability to move 100-TB dataset over the WAN faster than FedEx.	Robust (reliable) access via multiple sites/paths. Ability to move 3-PB dataset over the WAN faster than FedEx.	Robust access w/ BW & latency for remote analysis & vis. QoS guarantees for distrib. sim. Ability to move 100-PB dataset over the WAN faster than FedEx.
SNS	(Facility comes on-line in 2006.)	80 Mb/s sustained. 320 Mb/s peak Ability to move 200-TB dataset over the WAN faster than FedEx.	1 Gb/s sustained.
MMC	100 Mb/s sustained. 200 Mb/s peak.	200 Mb/s sustained. 400 Mb/s peak.	2 Gb/s sustained. 4 Gb/s peak.
HEP	1 Gb/s & end-to-end QoS. Ability to move 300-TB dataset over the WAN faster than FedEx.	100 Gb/s over lambdas and real-time network monitoring. Ability to move 3-PB dataset over the WAN faster than FedEx.	1 Tb/s Ability to move 100-PB dataset over the WAN faster than FedEx.
FES	Authenticated data streams thru firewalls at 30 Mb/s sustained. Ability to move 20-TB dataset over the WAN faster than FedEx.	100 Mb/s sustained. 500 Mb/s peak (20 sec of 15 min, i.e., QoS). Ability to move 1-PB dataset over the WAN faster than FedEx.	QoS for network latency and reliability to support real-time remote experiments.
Chem Sci	Robust (reliable) access w/ security for *long* times. Ability to move 100-TB dataset over the WAN faster than FedEx.	10+ Gb/s sustained (collab. viz & data mining). Ability to move 2-PB dataset over the WAN faster than FedEx.	100+ Gb/s sustained (distributed simulations) Ability to move 40-PB dataset over the Wan faster than FedEx.
Bioinfo	Ability to reliably move 400-TB dataset in a “reasonable” amount of time.	Ability to reliably move 20-PB dataset in a “reasonable” amount of time.	?
Cosmology	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.

COMPUTER & COMPUTATIONAL SCIENCES

Application Requirements: Beat FedEx!

Application	Now / Near Term	5 Years Out	10 Years Out
Climate	Authenticated data streams thru firewalls. Ability to move 100-TB dataset over the WAN faster than FedEx.	Robust (reliable) access via multiple sites/paths. Ability to move 3-PB dataset over the WAN faster than FedEx.	Robust access w/ BW & latency for remote analysis & vis. QoS guarantees for distrib. sim. Ability to move 100-PB dataset over the WAN faster than FedEx.
SNS	(Facility comes on-line in 2006.)	80 Mb/s sustained. 320 Mb/s peak Ability to move 200-TB dataset over the WAN faster than FedEx.	1 Gb/s sustained.
MMC	100 Mb/s sustained. 200 Mb/s peak.	200 Mb/s sustained. 400 Mb/s peak.	2 Gb/s sustained. 4 Gb/s peak.
HEP	1 Gb/s & end-to-end QoS. Ability to move 300-TB dataset over the WAN faster than FedEx.	100 Gb/s over lambdas and real-time network monitoring. Ability to move 3-PB dataset over the WAN faster than FedEx.	1 Tb/s Ability to move 100-PB dataset over the WAN faster than FedEx.
FES	Authenticated data streams thru firewalls at 30 Mb/s sustained. Ability to move 20-TB dataset over the WAN faster than FedEx.	100 Mb/s sustained. 500 Mb/s peak (20 sec of 15 min, i.e., QoS). Ability to move 1-PB dataset over the WAN faster than FedEx.	QoS for network latency and reliability to support real-time remote experiments.
Chem Sci	Robust (reliable) access w/ security for *long* times. Ability to move 100-TB dataset over the WAN faster than FedEx.	10+ Gb/s sustained (collab. viz & data mining). Ability to move 2-PB dataset over the WAN faster than FedEx.	100+ Gb/s sustained (distributed simulations) Ability to move 40-PB dataset over the Wan faster than FedEx.
Bioinfo	Ability to reliably move 400-TB dataset in a “reasonable” amount of time.	Ability to reliably move 20-PB dataset in a “reasonable” amount of time.	?
Cosmology	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.	For every 1-MFLOP CPU, need 1-MB/s memory & 1-Mb/s network.

The "Wizard Gap"† Problem (Across All Network Environments)

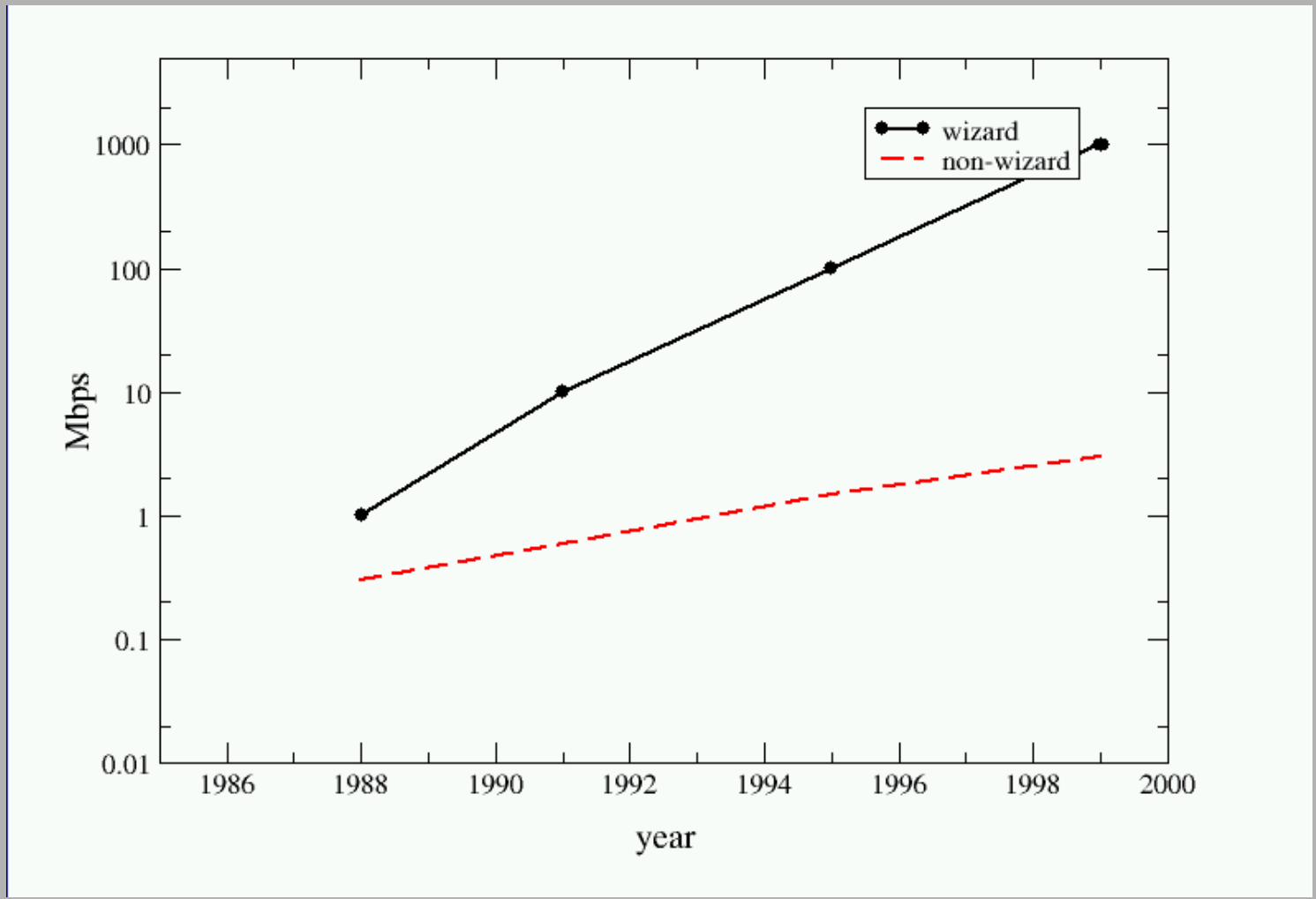
Performance Numbers from User Space to User Space

Environment	Typical	"State of the Art" w/ Network Wizards	Our Research
LAN with TCP/IP	300-400 Mb/s 100 μs	990 Mb/s → 2500 Mb/s 80 μs → 20 μs	4640 Mb/s → 7200 Mb/s 20 μs → 12 μs
SAN with OS-Bypass or RDMA		1920 Mb/s	2456 Mb/s 4.9 μs
			7200 Mb/s 4.0 μs
SAN with TCP/IP	1		7200 Mb/s 12 μs
	100 μs	32 μs	
WAN with TCP/IP (distance normalized)	0.007 Petabit-meters per second	0.270 Petabit-meters per second	23.888 Petabit-meters per second*

This beats FedEx in general, but how many "network wizards" did it require?

† A term coined by Matt Mathis, Pittsburgh Supercomputing Center.
* Internet2 Land Speed Record. Achieved: 2/27/03. Certified: 3/27/03. Awarded: 4/11/03.

The Wizard Gap Graphically (for the Wide-Area Network)



Source: Matt Mathis, Pittsburgh Supercomputing Center

Doing Our Part?

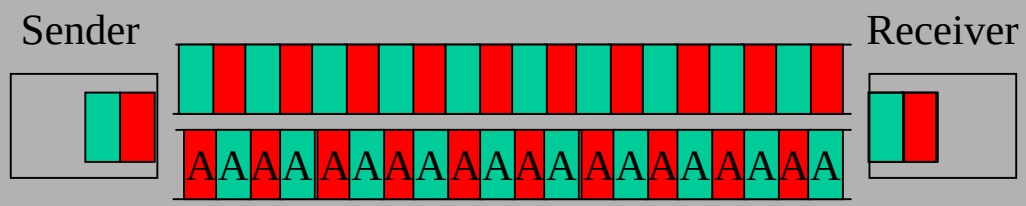
- Automate the "network wizard" steps needed to improve performance so that the "wizard gap" is eliminated.
  - LAN/SAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Networks of Workstations and Clusters. (Hurwitz, Feng: IEEE HotI'03).
    - "Evil TCP"
  - WAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Grids. (Feng et al.: ACM/IEEE SC 2003)
      - Fold experiences into existing software, e.g., operating system.
    - Automatic TCP Buffer Tuning (Mathis et al., SIGCOMM'98) → Dynamic Right-Sizing (Feng et al., 1999 tech report & J. of Grid Computing, 2003).

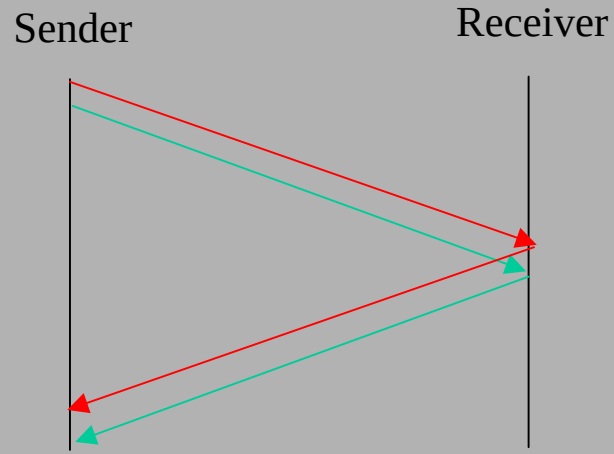
Doing Our Part?

- Automate the "network wizard" steps needed to improve performance so that the "wizard gap" is eliminated.
  - LAN/SAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Networks of Workstations and Clusters. (Hurwitz, Feng: IEEE HotI'03).
    - "Evil TCP"
  - WAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Grids. (Feng et al.: ACM/IEEE SC 2003)
      - Fold experiences into existing software, e.g., operating system.
    - Automatic TCP Buffer Tuning (Mathis et al., SIGCOMM'98) → Dynamic Right-Sizing (Feng et al., 1999 tech report & J. of Grid Computing, 2003).

The Need for *Transparent* Flow-Control Adaptation

- Without a "network wizard" to intervene and optimize network performance via flow control …
  - Wide-area transfer between SNL and LANL of a 150-GB dataset.
    - OC-3 (155 Mb/s): 8 Mb/s → 42 hours "Wizard Magic": 55 Mb/s
    - OC-12 (622 Mb/s): 8 Mb/s → 42 hours "Wizard Magic": 240 Mb/s
  - The bandwidth of a driving tapes of the data from SNL to LANL is a LOT better! 150 GB / 2 hours = 167 Mb/s.





Transparently provide end-to-end performance to the application, thus "eliminating" the need for network wizards.

Flow-Control Adaptation

- *Problems*
  - No adaptation currently being done in any "standard" TCP.
  - Default 32-KB is OK for LAN but not for WAN where BW*delay is *three orders of magnitude* larger.
- *Consequence:* As little as 3% of network pipe is filled.
- *Initial Solutions*
  - *Manual* tuning of buffers at send and receive end-hosts.
    - Too small → low bandwidth. Too large → waste memory (LAN).
  - *Automatic* tuning of buffers.
    - Auto-tuning: Sender-based flow control. [Mathis et al., SIGCOMM'98.]
    - Linux 2.4.x auto-tuning for web servers, *not* high-performance bulk data transfer.
  - *Network striping & pipelining* with default buffers.
    [UIC, 2000 & GridFTP @ ANL, 2001.]
- *Our Solution*
  - Dynamic right-sizing: Receiver-based flow control.

Dynamic Right-Sizing: Intelligent Flow-Control Adaptation

- Tricky Part
  - Modify TCP flow-control implementation *without* violating TCP protocol specification.
- Approach
  - Receiving host
    - "Measures" the rate at which the sender transmits.
    - Checks its available memory resources.
    - Advertises appropriate flow-control window (i.e, buffer size).
- Implementations
  - Kernel
    - Linux 2.4 patch to implement "dynamic right-sizing" of buffer sizes.
    - Typical speed-up in WAN: 15x – 30x.
  - User Space
    - drsFTP and DRS-enabled GridFTP prototypes
    - Typical speed-up in WAN: 6x – 8x.

Doing Our Part?

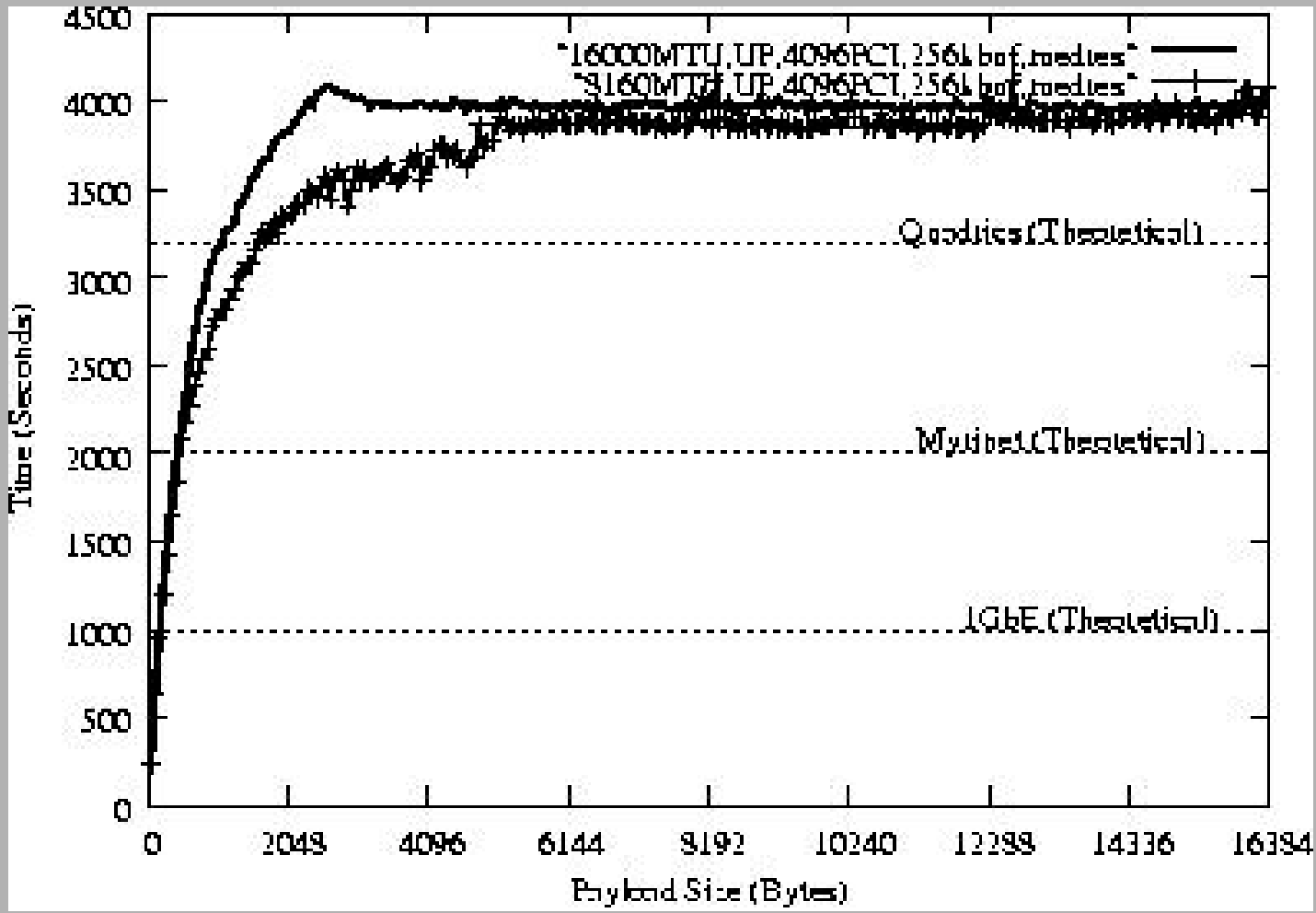
- Automate the "network wizard" steps needed to improve performance so that the "wizard gap" is eliminated.
  - LAN/SAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Networks of Workstations and Clusters. (Hurwitz, Feng: IEEE HotI'03).
    - "Evil TCP"
  - WAN
    - Optimizing Intel 10-Gigabit Ethernet Adapters for Grids. (Feng et al.: ACM/IEEE SC 2003)
      - Fold experiences into existing software, e.g., operating system.
    - Automatic TCP Buffer Tuning (Mathis et al., SIGCOMM'98) → Dynamic Right-Sizing (Feng et al., 1999 tech report & J. of Grid Computing, 2003).

10-Gigabit Ethernet for Clusters & Grids

- Manual Optimizations
  - Increase PCI-X burst size, uniprocessor kernel, MTU tuning.
- Expectation
  - Some of the above changes will be folded into 10GigE card to provide automatic "out-of-box" performance to application user.
- LAN & SAN End-to-End Performance
  - Back-to-back between a pair of dual 2.2-GHz Dell PE 2650s (with interrupt coalescing)
    - 8160-byte MTU: 4.11-Gb/s throughput, 21-µs end-to-end latency
    - 1500-byte MTU: 2.47-Gb/s throughput → CPU limited
  - Reverse multicast between GigE clients and Itanium-II server (without interrupt coalescing)
    - 8160-byte MTU: 7.2-Gb/s throughput, 12-µs end-to-end latency
- WAN End-to-End Performance (+ "hacked" DRS to get Vegas-like)
  - Internet2 Land Speed Record (2/27/03): 2.38-Gb/s single-stream TCP/IP between Sunnyvale, CA and Geneva, Switzerland. Terabyte in under an hour. (Bottleneck link: 2.45-Gb/s OC-48 trans-Atlantic link.)

10-Gigabit Ethernet Bandwidth

Between a Pair of "Low-End" 2.2-GHz Dell PowerEdge 2650s



Success?

- Dynamic Right-Sizing (DRS)
  - Kernel
    - Implemented Linux 2.4.x patch. Ideal: Incorporated in Linux kernel.
    - Requested to implement DRS in FreeBSD 5.1.
  - User Space
    - Exported DRS technique from kernel space to user space: drsFTP.
    - Ported user-space DRS technique to GridFTP.

What's Next?

- RDMA over TCP/IP?
  - IETF RDDP effort over TCP/IP
- TCP Offload Engines (TOE)?
  - "A Bad Idea Whose Time Has Come?" Jeff Mogul, HP Labs.
- Adaptive RDMA or TOE?
  - Allow off-loading to be "load balanced" between host processor and NIC processor.
  - Highly application-dependent.
- Network-Aware Middleware
  - GridFTP: Middleware doing what the transport layer ought to be doing?

Middleware Perspective

- The World of Grid Computing
  - The large-scale scientific application is the "king" (or "queen") of high-performance computing & networking.
  - It's all about transparency, ease of use, and performance.
  - The grid community is creating middleware to address the "shortcomings" of network research
    - Better support for transparency, persistence, multiple streams, and plug-ability because network researchers "don't care."

GridFTP Feature Set

- GSI & Kerberos security
- Third-party transfers
- Parameter setting/negotiation
- Partial file access
- Reliability/restart
- Large file support
- Data channel reuse
- Defacto standard on the Grid

- Integrated instrumentation
- Logging/audit trail
- Parallel transfers
- Striping
- TCP buffer-size control (dynamic right-sizing)
- Policy-based access control
- Server-side computation
- Based on standards

Source: Bill Allcock, Ian Foster, Carl Kesselman, Miron Livny.

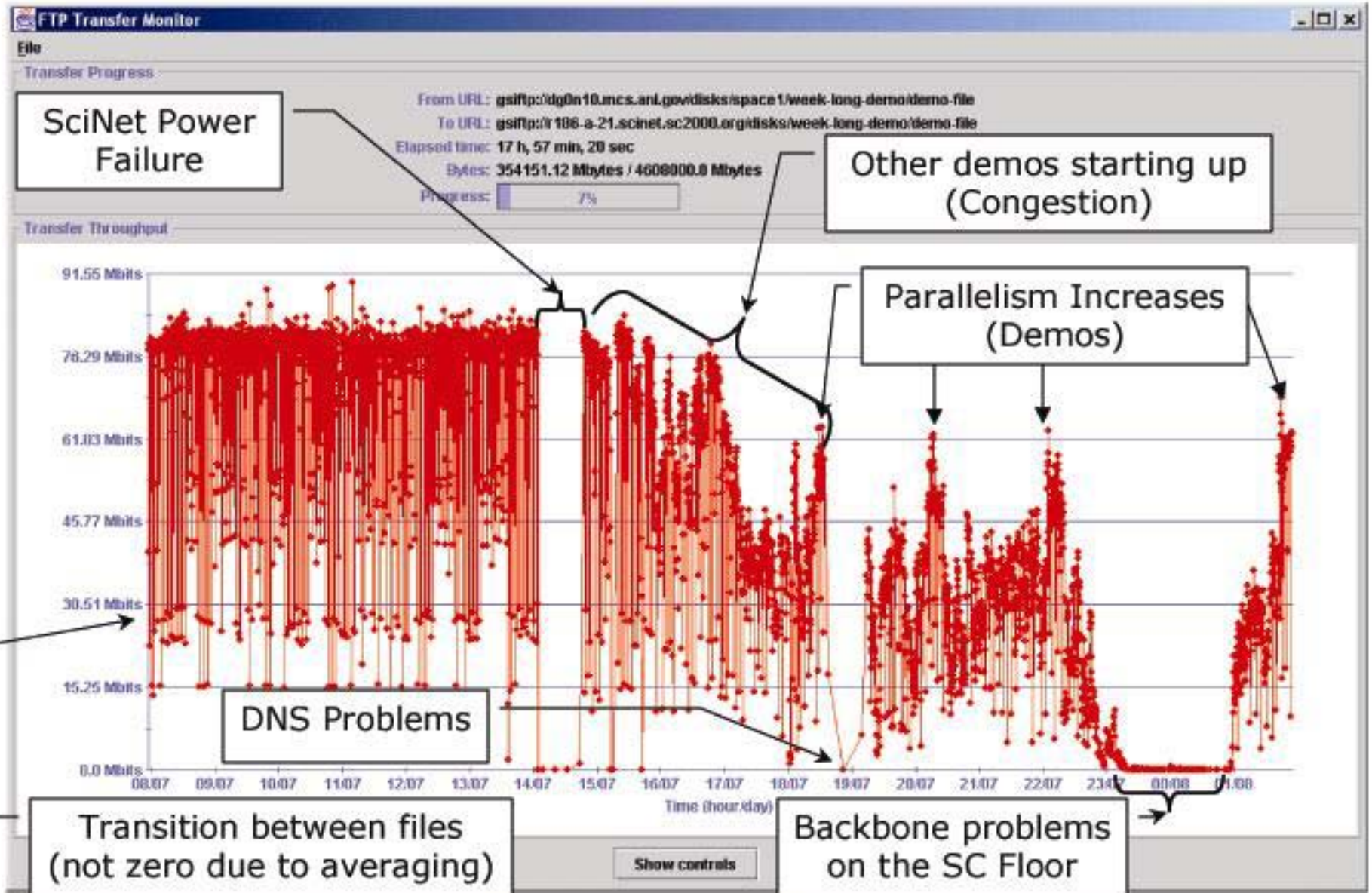
GridFTP: Robust and Reliable

- GridFTP "restart markers" allow recovery from failed transfers.
  - If any remote resource fails, the restart markers can be used to pick up the transfer, including "holey" transfers.
- Restart behavior is determined by a plug-in.
  - Default restart "plug-in" provided.
  - Can be modified for customized restart policy.
- Higher-level reliable file transfer service uses database for yet higher reliability.
  - e.g., 56-hour, 236-GB transfer survived multiple network failures and require no human intervention.

Source: Bill Allcock, Ian Foster, Carl Kesselman, Miron Livny.

GridFTP at SC'2000: Long-Running Dallas-Chicago Transfer

Source: Bill Allcock



Quote from "The Art of Computer Systems Performance Analysis"

"If it is fast and ugly, they will use it and curse you; if it is slow, they will not use it."

- Who are "they"?
  - System administrators?
  - Network wizards?
  - Application users?
  - Scientific programmers?

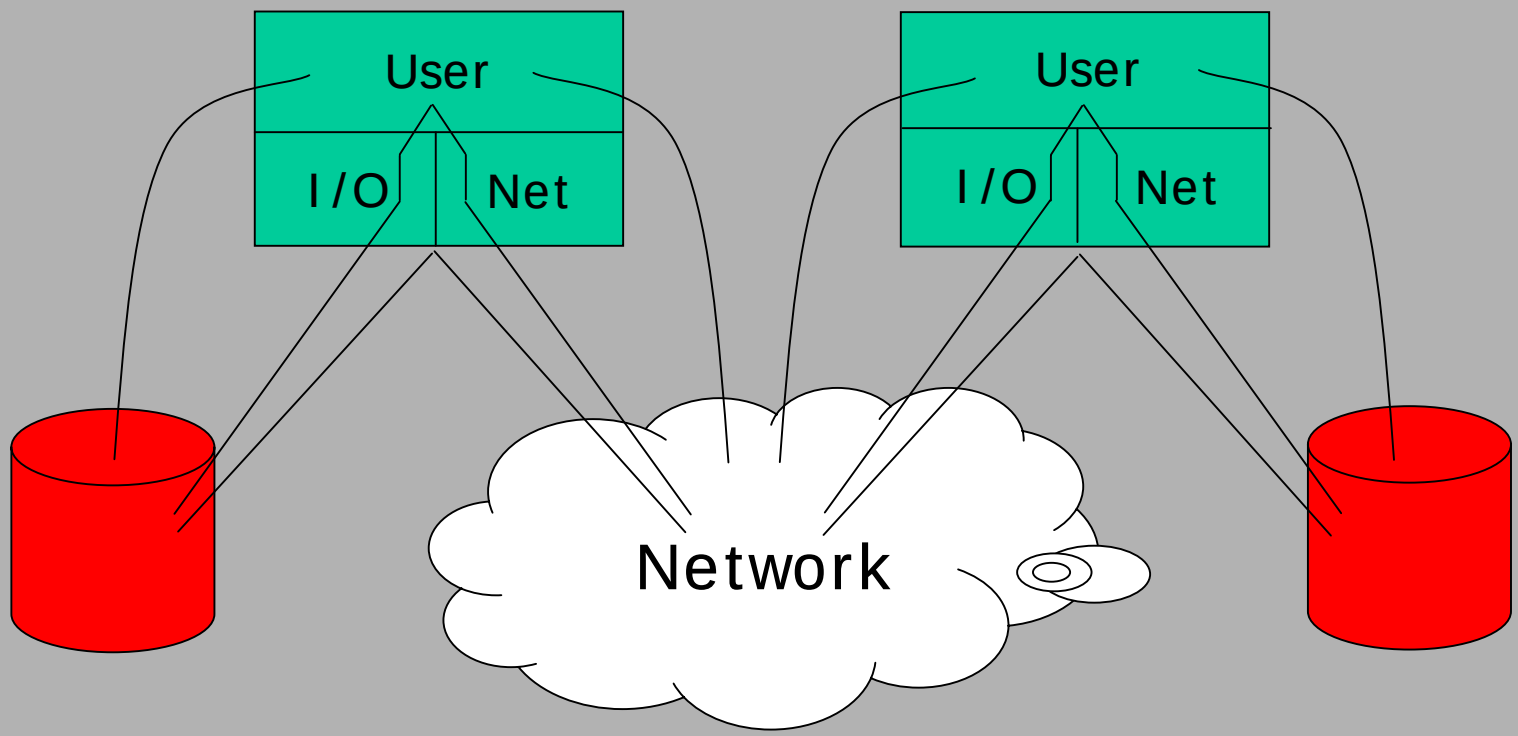
"they" = people with a thorough background in networking

Amended Quote:

"If it is fast and ugly, *networking folks* will use it and curse you, but *everyone else* will simply curse you. If it is slow but easy to use, networking folks will *not* use it, but everyone else will."

Next Step?

- A pathetic attempt to tie this talk into I/O :-)



Programming Model

- Assembly language is to sockets *as* high-level programming language is to ???
  - Quit manipulating flows at the socket level.
  - Handle flows as named objects that can be operated on, e.g., re-direction, flow merging, auto-downsampling of HTML files for mobile phones and PDAs.

"We engineered the Internet, and it works fine for e-mail and the web; but to do world-class scientific research, we need to develop a science of networking that delivers *usable* performance to DOE scientific applications."

- Allyn Romanow

Cisco Systems