Advances in Distributed Storage

Andy Helland
Northwind Enterprises
12100 Center Avenue, San Martin CA 95046-9727
Phone:+1-408-316-6739   FAX: +1-408-683-2309
E-mail: andy@andyhelland.com

Presented at the THIC Meeting at the Sony Auditorium,
3300 Zanker Rd, San Jose CA 95134-1940

April 19-20,2005

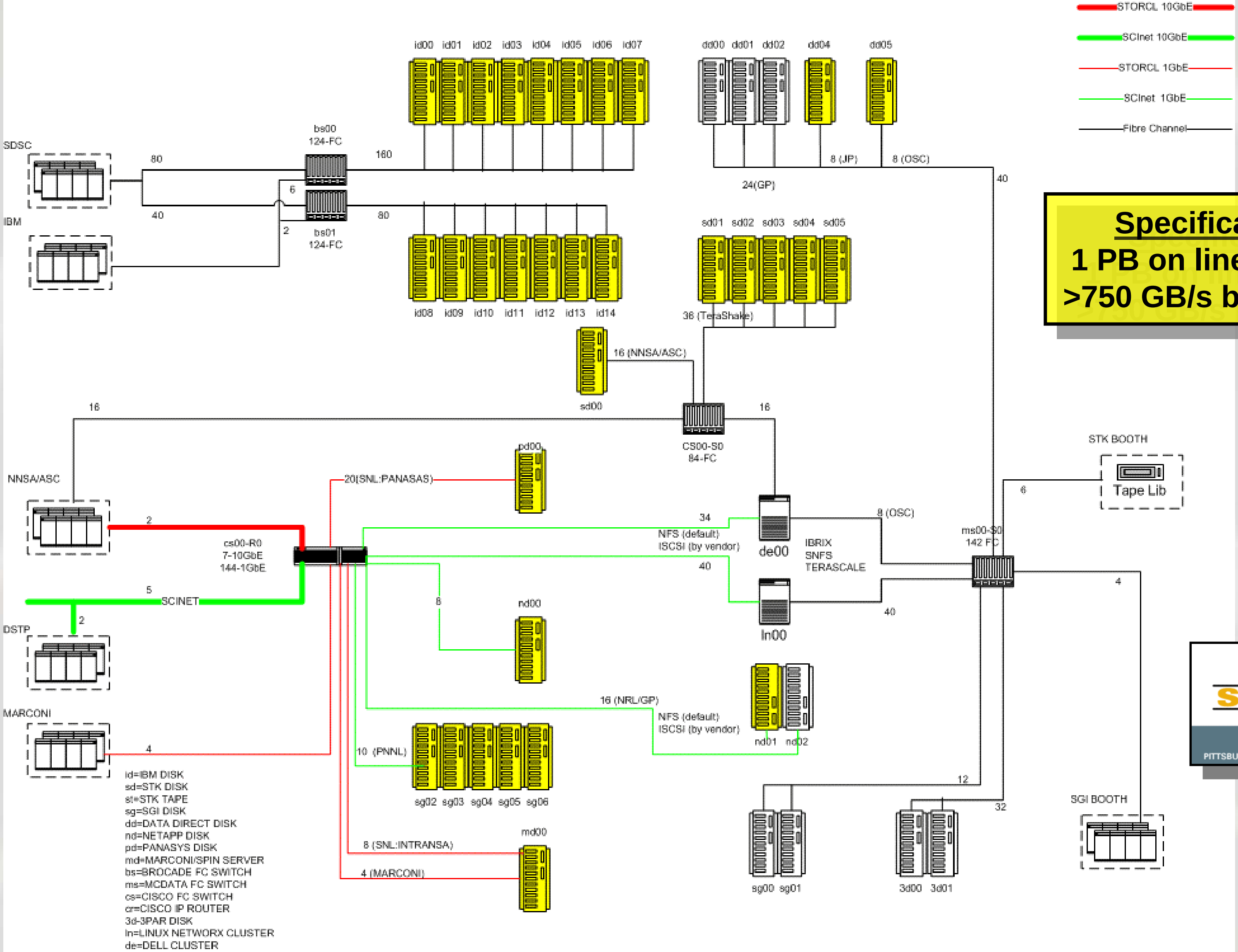
Overview

- The problems with distributed storage
  - Bandwidth
  - Latency
  - Protocols
- New technologies to enhance distributed storage networks
  - FC routing
  - SCSI fast write
  - Advanced TCP transport mechanisms
  - Distributed block caching

Bandwidth Inside the Data Center

- Bandwidth is “free”
- Multi-mode fiber is cheap
- Most Fibre Channel switches support trunking (higher bandwidth aggregate links)

SuperComputing 2004: StorCloud



id=IBM DISK
sd=STK DISK
st=STK TAPE
sg=SGI DISK
dd=DATA DIRECT DISK
nd=NETAPP DISK
pd=PANASYS DISK
md=MARCONI/SPIN SERVER
bs=BROCADE FC SWITCH
ms=MCDATA FC SWITCH
cs=CISCO FC SWITCH
cr=CISCO IP ROUTER
3d-3PAR DISK
ln=LINUX NETWORX CLUSTER
de=DELL CLUSTER

Specifications

1 PB on line storage

>750 GB/s bandwidth

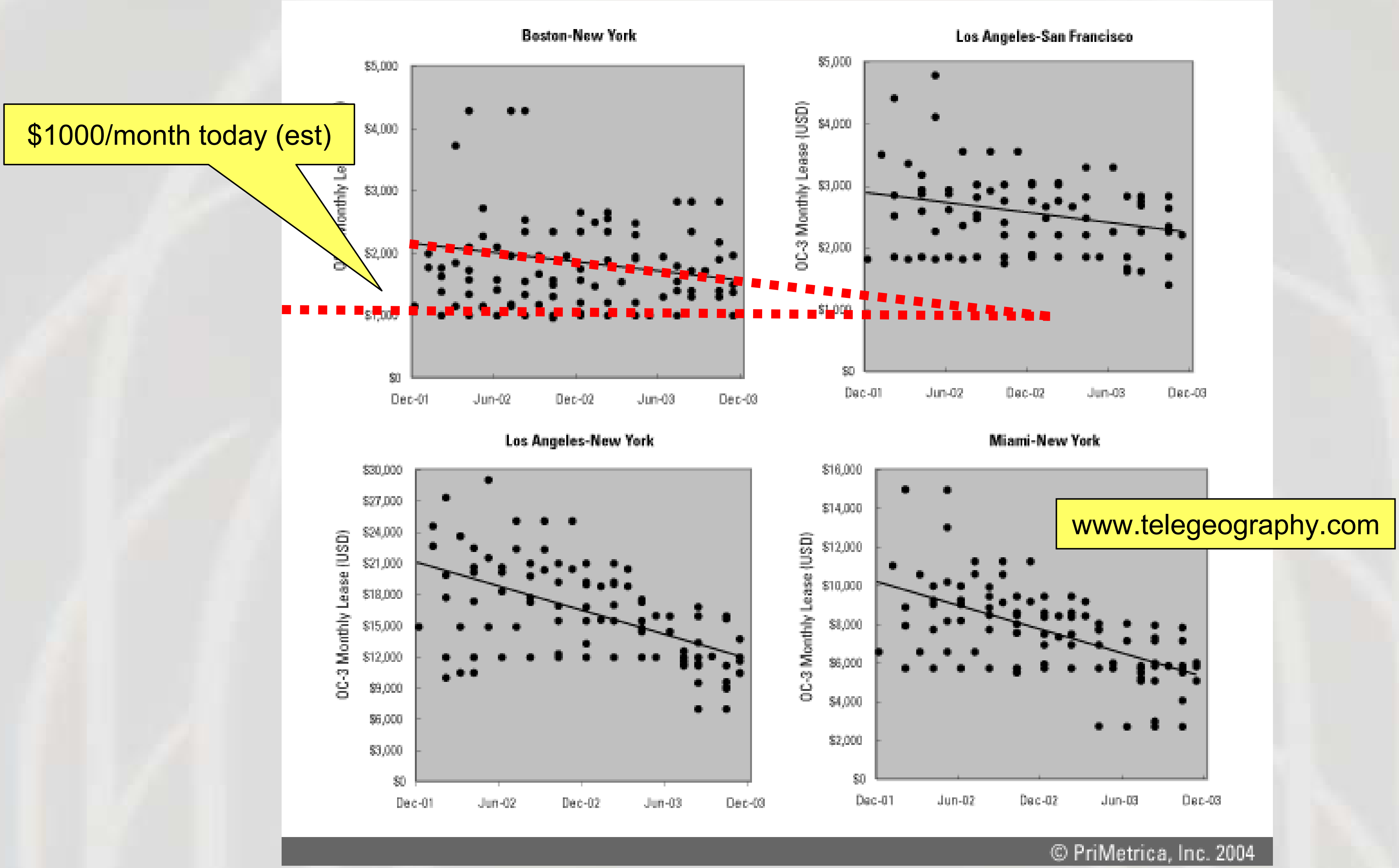
SC2004

PITTSBURGH PA NOVEMBER 6 - 12

Bandwidth Outside the Data Center

- Bandwidth is expensive
- Copper to the curb
  - T-1 (1.5 Mb/s)
    - $800 per month
  - T-3 (45 Mb/s)
- Fiber to the curb
  - SONET (OC-3, OC-12, OC-48)
  - Dedicated fiber
    - First fiber is expensive
    - Very cost effective (CWDM, DWDM)
- Metro Gigabit Ethernet offers very high value
  - $6000 per month
  - Metro only

POP to POP Pricing (SONET OC-3)



$1000/month today (est)

www.telegeography.com

© PriMetrica, Inc. 2004

Latency Inside the Data Center

- ❑ Fiber optic cable
  - ❖ ~5 ns per meter
  - ❖ Insignificant
  - ❖ Difficult to even measure!
- ❑ Fibre Channel switch
  - ❖ Typically 2 μs (cut through switching)
- ❑ Hard disk drive
  - ❖ 15,000 RPM FC/SCSI
  - ❖ 2 ms average latency

Latency Outside the Data Center

- Fibre optic cable
  - 5 µs/Km (add 30% - 50% to “crow fly” distance)
  - San Francisco - San Jose (1.2 ms, round trip)
  - San Francisco – Los Angeles (8.3 ms, round trip)
  - San Francisco – New York (62.3 ms, round trip)
- DWDM terminal equipment (10’s of ns)
  - ~4 meters of fiber
- SONET multiplexers (10’s of µs)
  - ~4 Kilometers of fiber
- Layer 3 (IP) devices
  - Layer 3 switches (10 µs or 4 Km of fiber)
  - Layer 3 routers (few ms or 400 Km of fiber)

Tannenbaum’s Famous Quote

> *“Never underestimate the bandwidth of a station wagon full of tapes hurling down the highway.”*

Bandwidth vs. Latency

Primary data center in San Jose, CA
Back-up site in Los Angeles, CA
Distance 400 miles

Infrastructure	**Bandwidth**	**Latency**
BROCADE CISCO SYSTEMS McDATA Fibre Channel	200 MB/s	3.2 ms (5µs/Km)
120 TB (!)	4.2 GB/s	8 Hours
Budget Moving Up in Truck Rental 1-800-Budget2 www.budgettruck.com 1.1 PB (!)	38 GB/s	8 Hours

Improving Distributed Storage Networks

Advances in transport technology (“the plumbing”)

Advances in Storage Connectivity

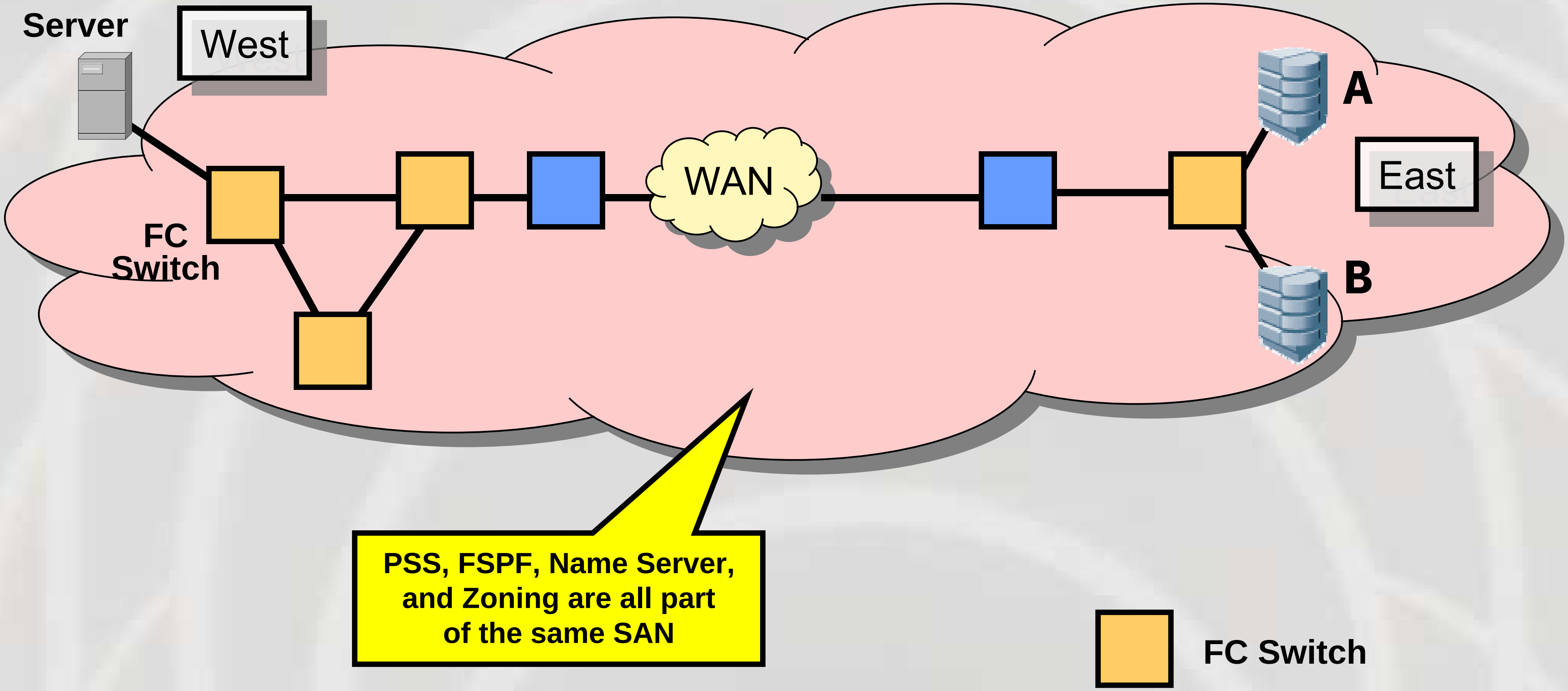
- FC routing (eliminate the “flat” fabric)
- Short cycle the SCSI write command
- Advanced transport technology
  - Fast TCP/IP
- Distributed block caching

FC Routing

What’s Wrong with “Flat Fabrics”

- Fibre Channel (FC) only has 239 node addresses
- FC uses a flat routing protocol borrowed from IP routing
  - OSPF (Open Shortest Path First)
  - FSPF (Fibre Channel Shortest Path First)
- Link state protocols have three phases
  - Determination of link connectivity and “cost”
  - Flooding of all links and costs to all nodes (N^2 process)
  - Independent calculation of routing tables by each node
- N^2 processes do not scale well!
  - Link state protocols have trouble converging as N becomes large
  - Latency makes this problem worse

Simple SAN Extensions Create a Single Fabric



PSS, FSPF, Name Server, and Zoning are all part of the same SAN

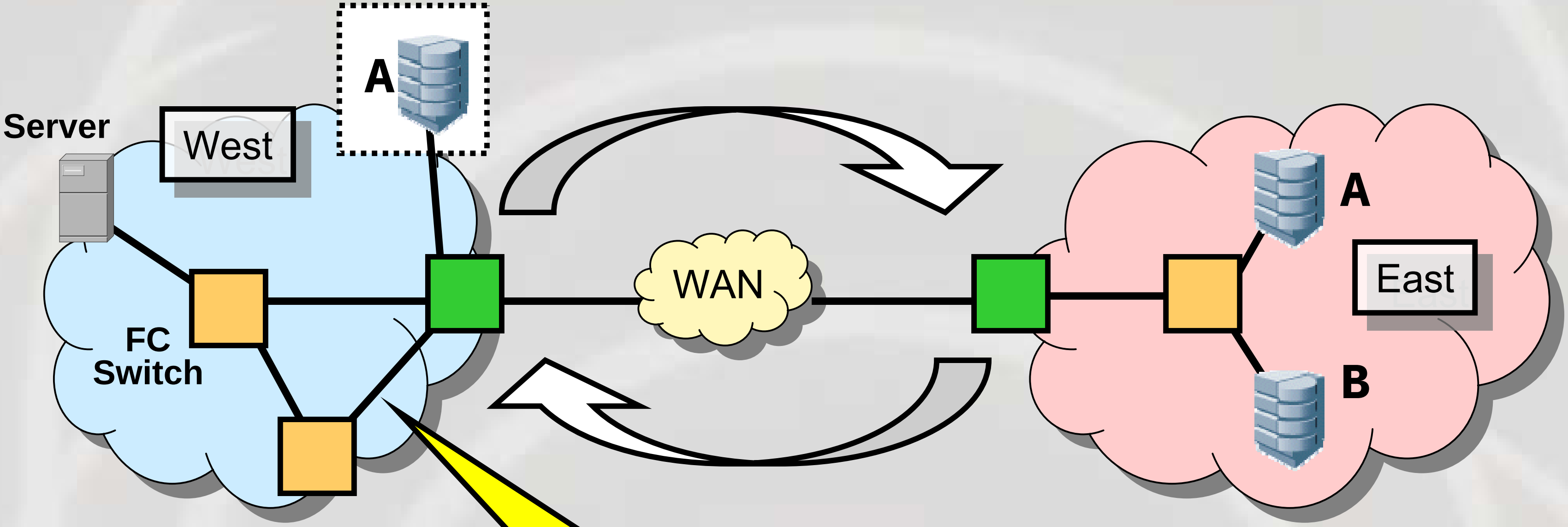
🟧 **FC Switch**

🟦 **Simple SAN gateway**

FC Routing Enables Scalable Fabrics

- ❑ FC routing does for FC what Border Gateway Protocol (BGP) did for IP) networks
- ❑ Local routing is not broadcast across the wide area
- ❑ Local SANs are connected to each other *hierarchically*
- ❑ Local disturbances to a SAN are *NOT* broadcast to other SANs

FC Routing Does NOT Merge Fabrics



PSS, FSPF, Name Server, and Zoning are all isolated in each SAN

🟧 FC Switch

🟩 Switching Gateway

Multiple Vendor Offerings for FC Routing

- LightSand first to market (SANcastle acquisition)
  - AR/DAT (Autonomous Region/Domain Address Translation)
  - Redundant address spaces and heterogeneous fabrics
- McDATA (Nishan acquisition)
  - iFCP protocol
  - Uses IP as the switching core
- Brocade
  - LSANs (Logical SANs) and FC Routing
  - Support coming for heterogeneous fabrics
- Cisco
  - VSANs (Virtual SANs) and IVR (Inter-VSAN Routing)
  - Can trunk multiple FCIP links together
  - Just announced support for multiple vendors

The problem with SCSI and latency

The Problem with Simply Extending Protocols

- ❑ SCSI was never designed for wide area operation
- ❑ Synchronous mirrors keep multiple disk arrays in lockstep with each other
- ❑ Local disk array performance is tied to performance of remote disk array
- ❑ As distance increases…
  - ❖ Security increases
  - ❖ Performance decreases
- ❑ At metro distance (or less), latency is not significant

Normal SCSI Write Operation

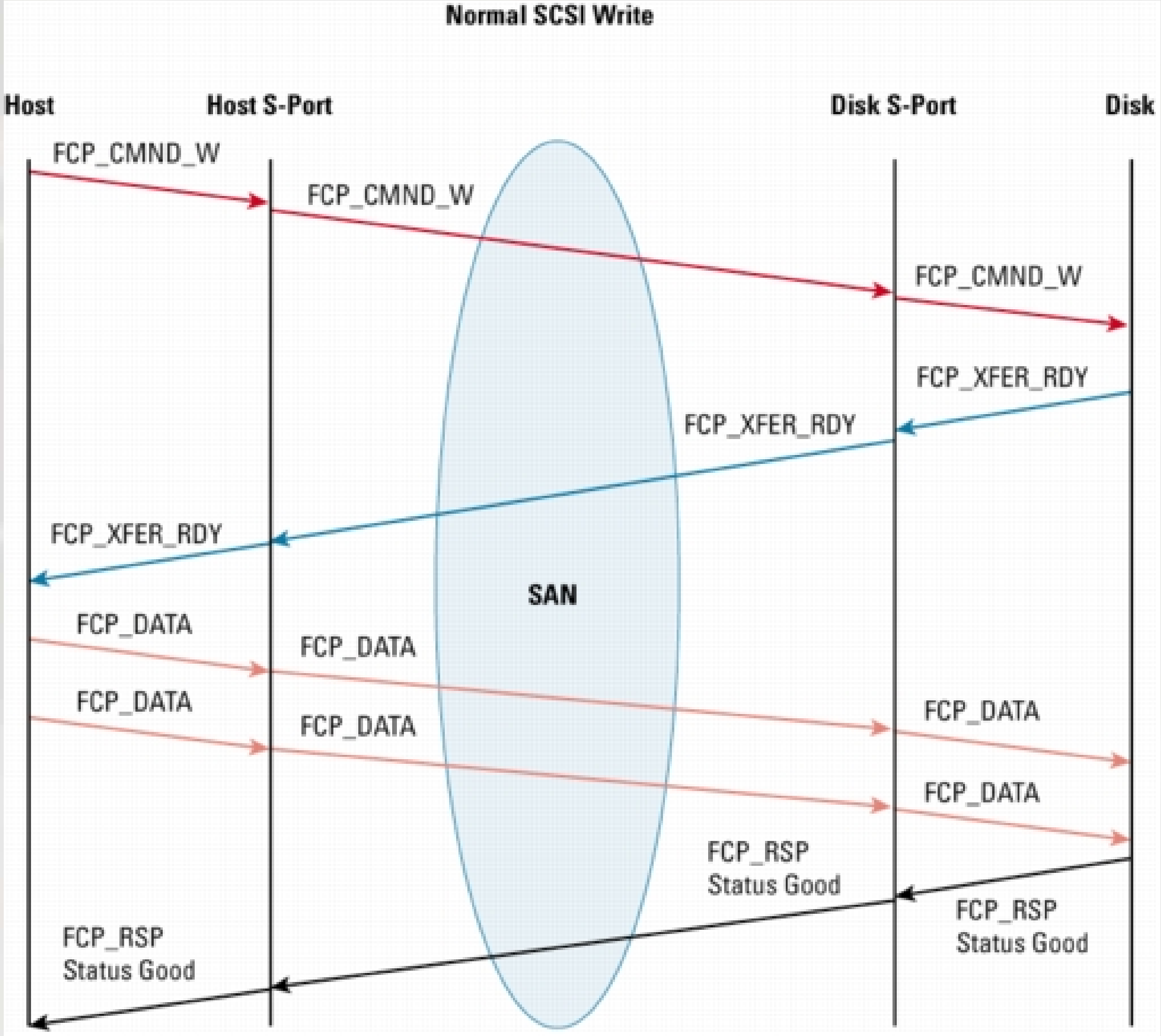


Diagram from Cisco

Short Cycling SCSI Reduces Impact from Latency

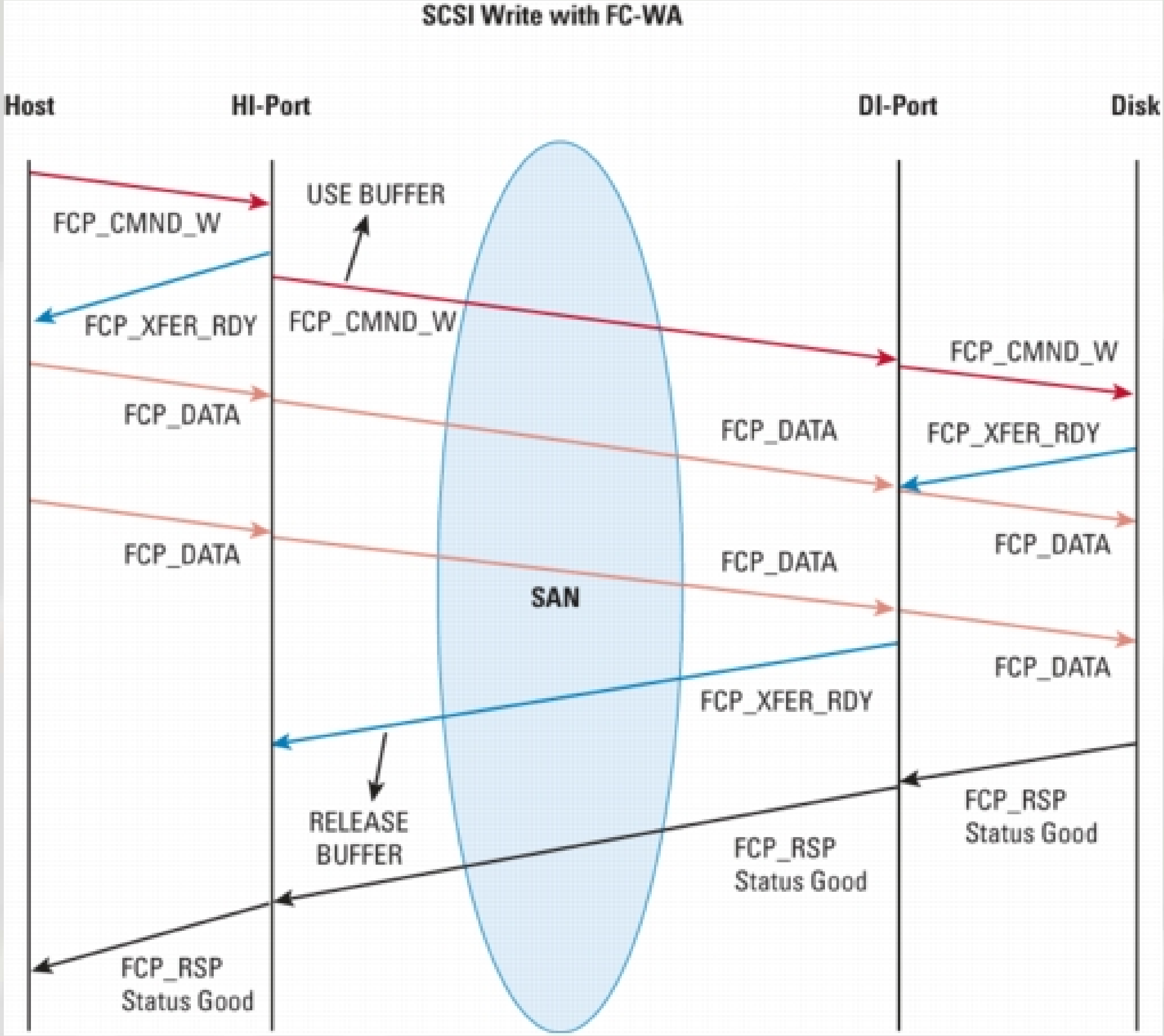


Diagram from Cisco (Write Acceleration).

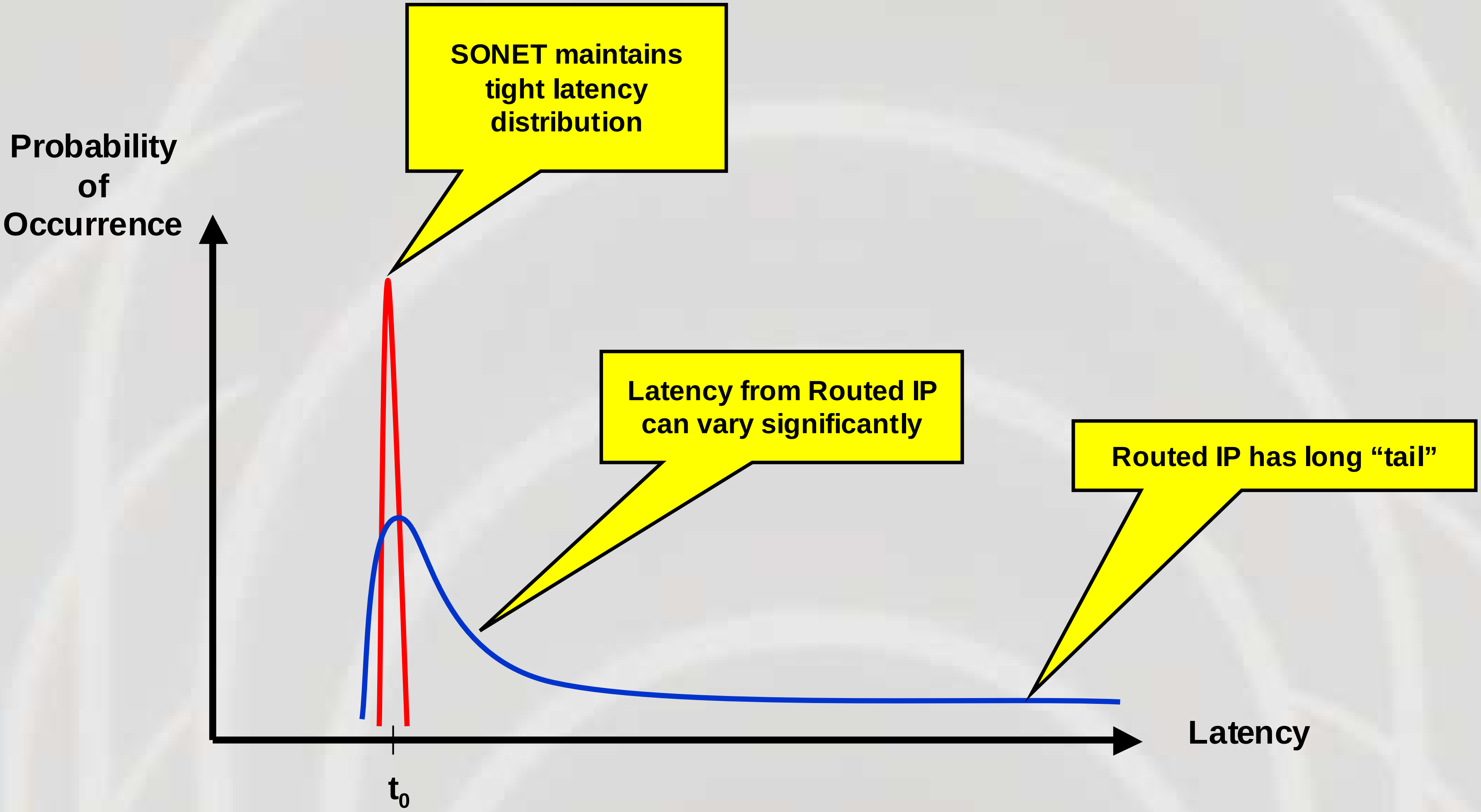
Similar technology available from McDATA (Fast Write)

Improvements in Transport Technology

Channels vs. Networks

- Channels are dedicated links
  - Highest performance
  - Fundamentally reliable
  - Do not scale well (dedicated links between all users)
- Networks (routed IP) are connectionless and unreliable
  - Great scalability
  - Fundamentally unreliable transport core
  - Problems with performance
    - Reliability must be added back
    - Classic TCP has problems even when the link is good (slow start)
    - TCP and other connection-oriented protocols create virtual channels

Latency Distribution for Channels and Networks

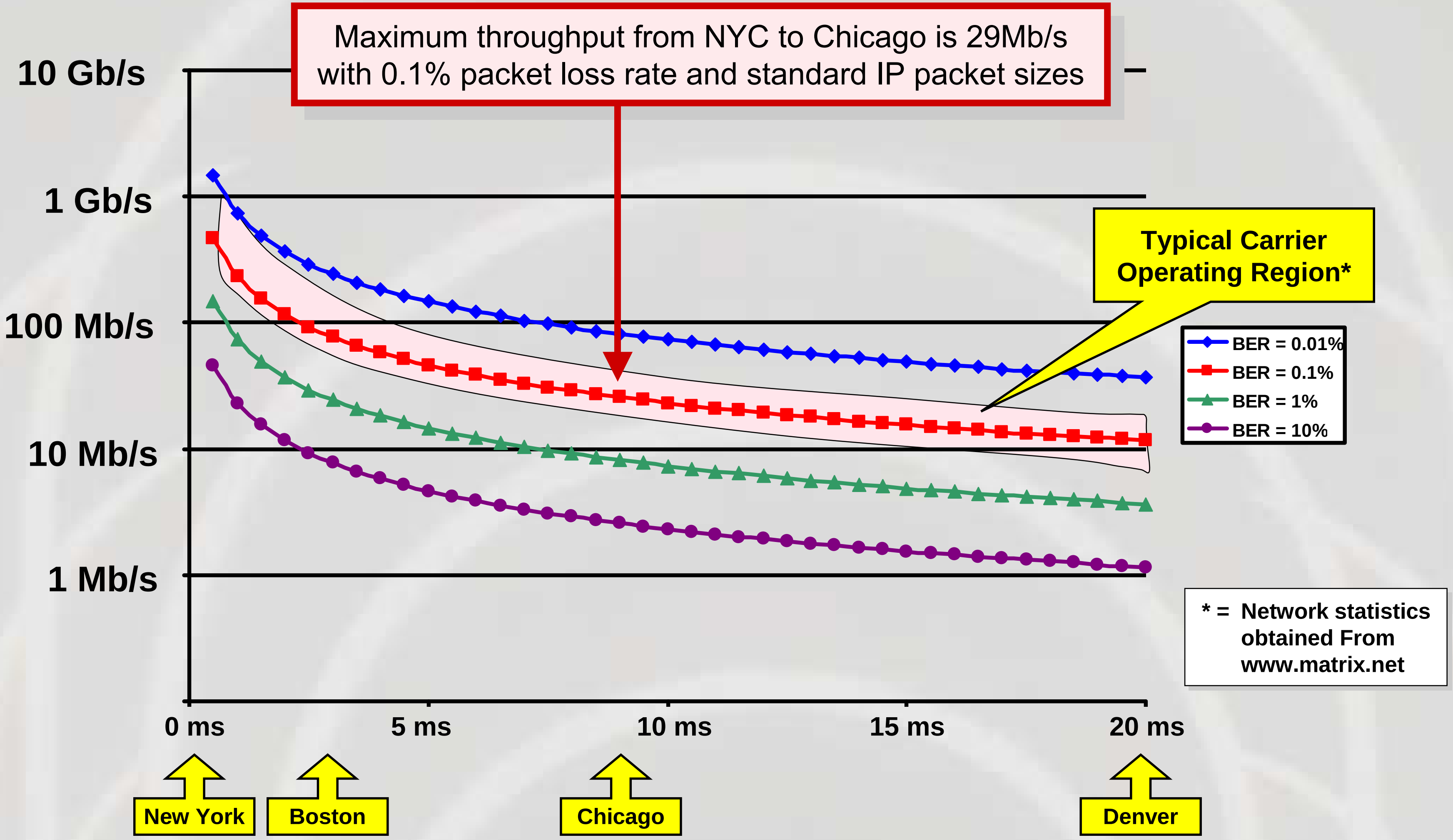


SONET maintains tight latency distribution

Latency from Routed IP can vary significantly

Routed IP has long “tail”

TCP Reno Behavior with Packet Loss and Latency

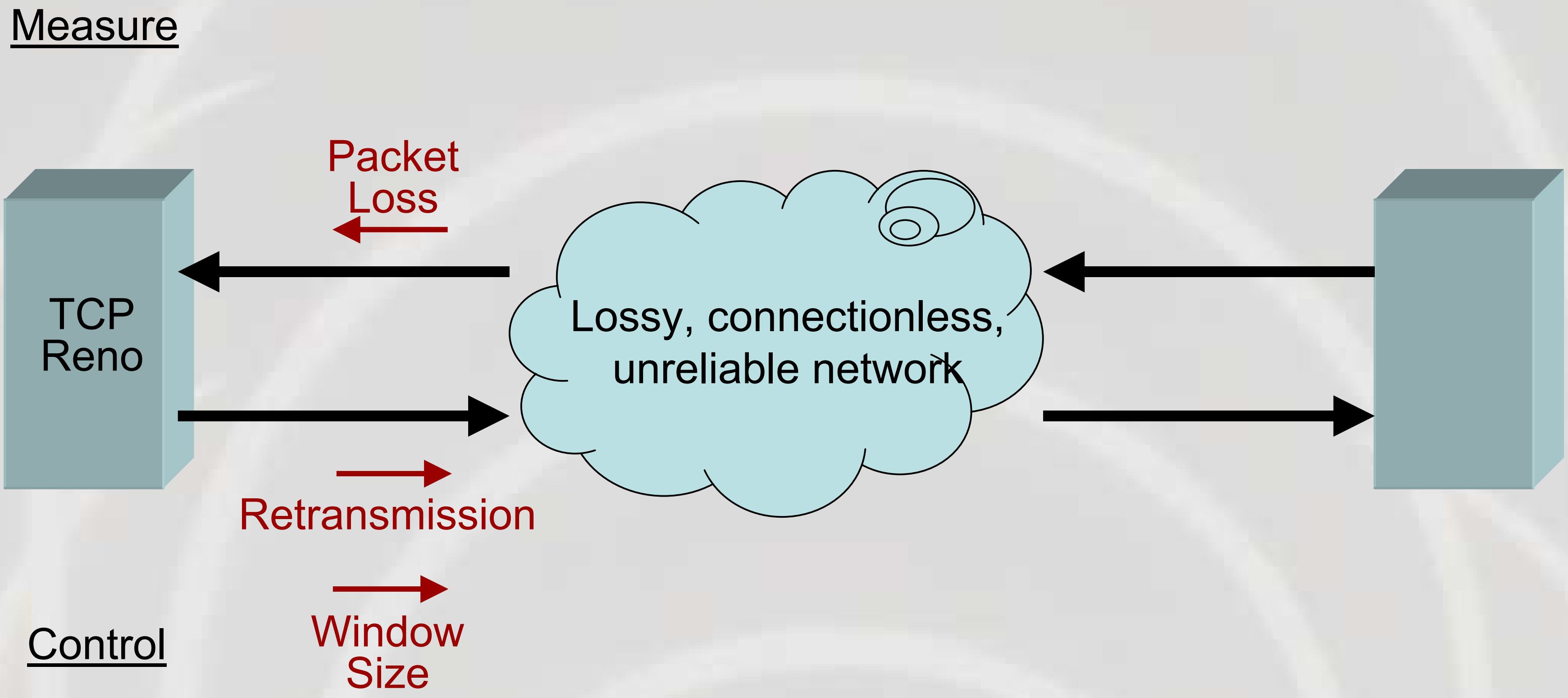


Maximum throughput from NYC to Chicago is 29Mb/s with 0.1% packet loss rate and standard IP packet sizes

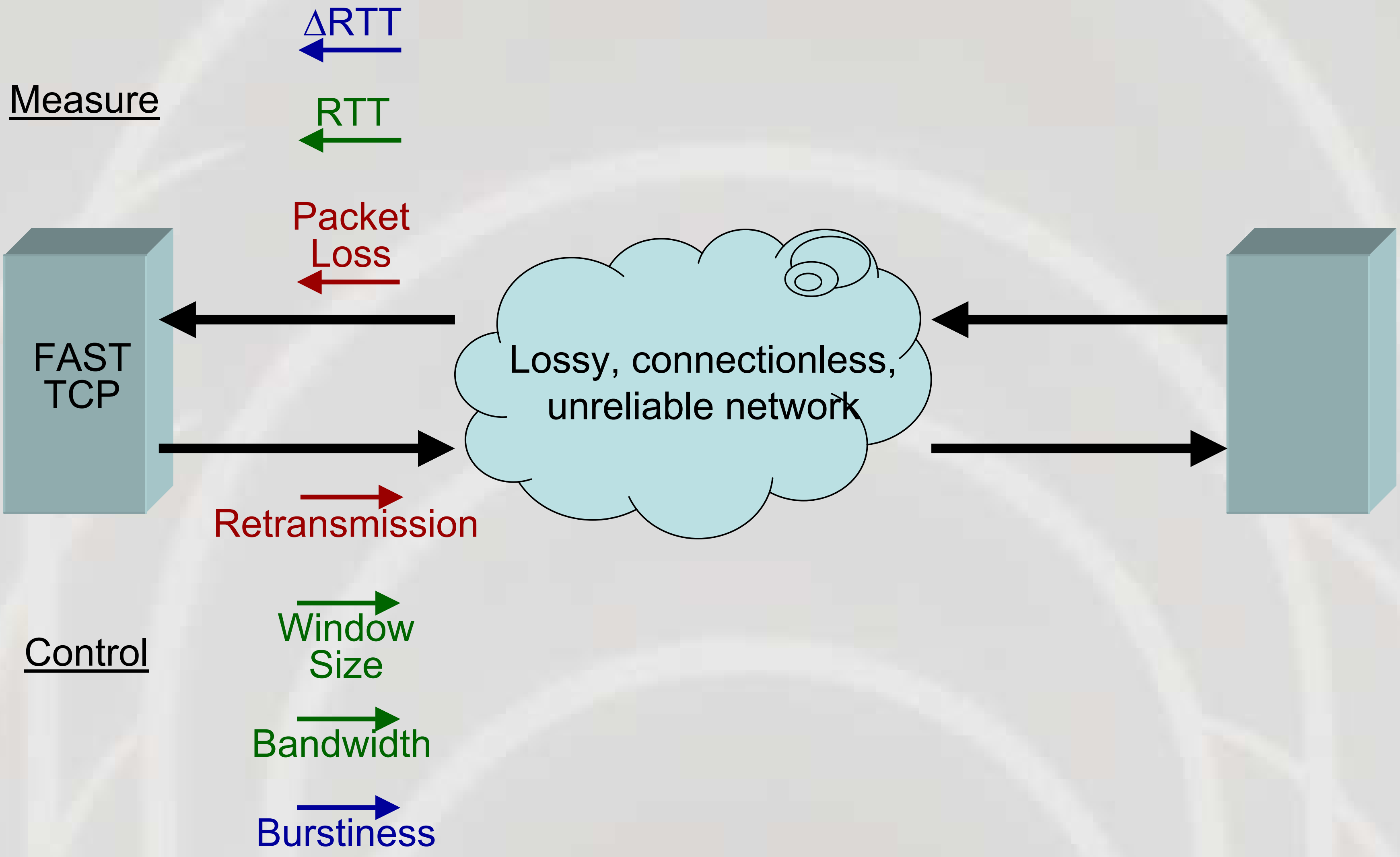
Typical Carrier Operating Region*

* = Network statistics obtained From www.matrix.net

TCP Reno (Classic TCP)

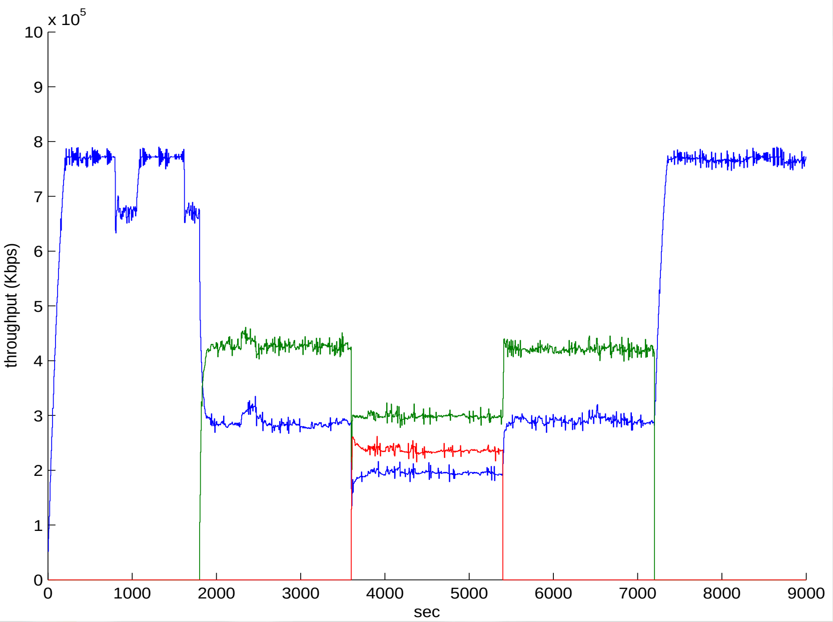


FAST TCP (Caltech et al)



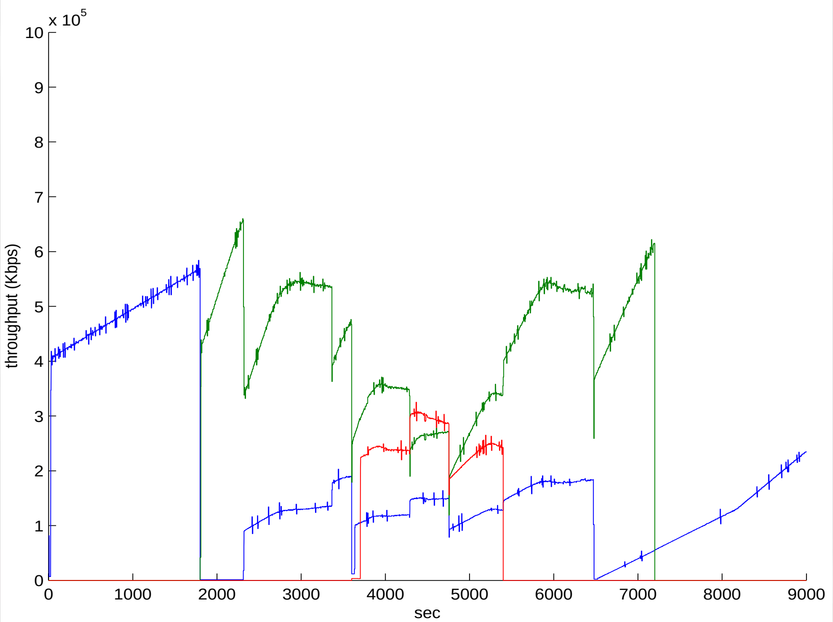
FAST TCP vs. TCP Reno (3 flows)

FAST TCP



1 flow | 2 flows | 3 flows | 2 flows | 1 flow

TCP Reno



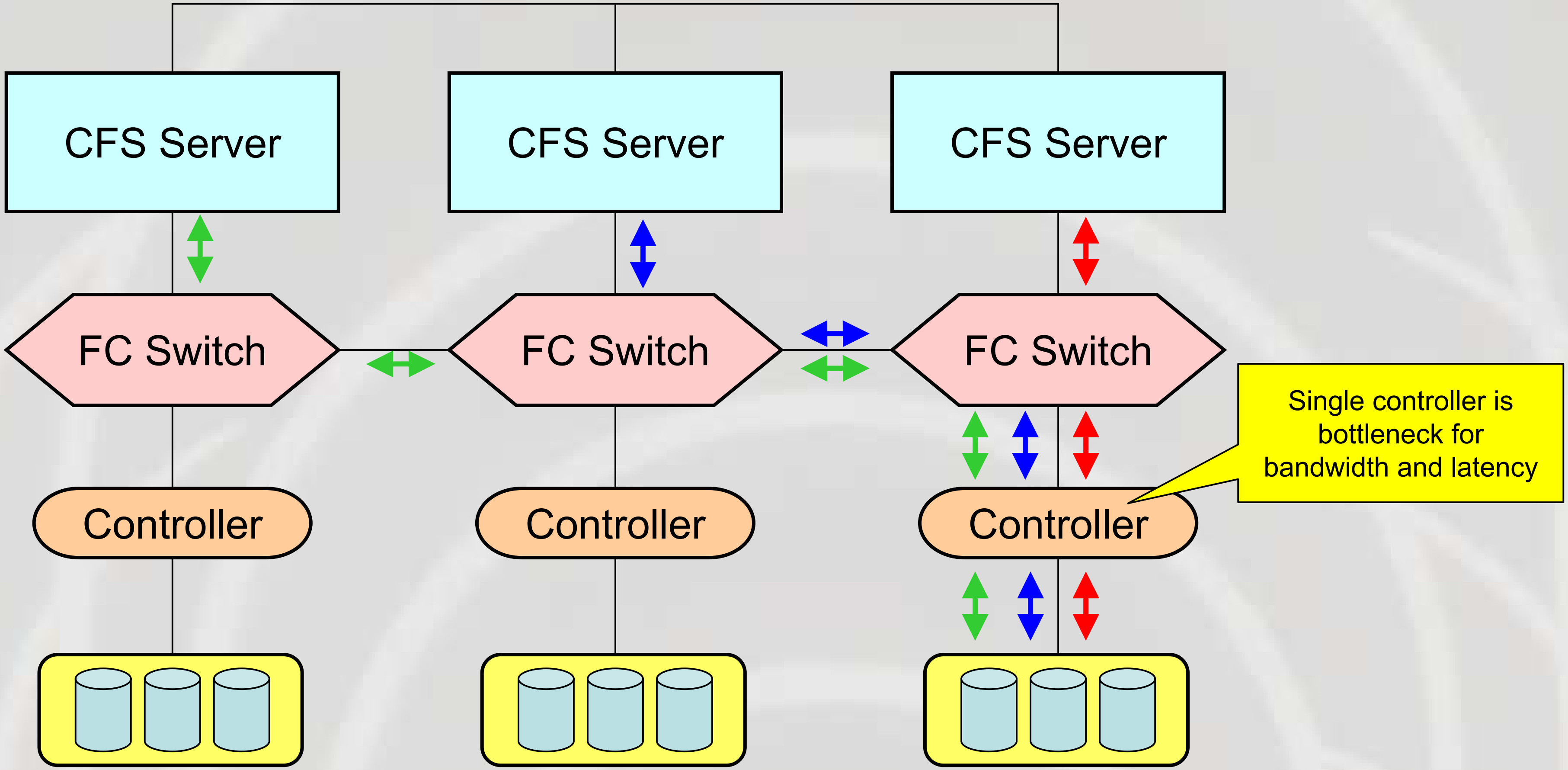
1 flow | 2 flows | 3 flows | 2 flows | 1 flow

RTT = 120 ms
BW_{max} = 800 Mbps

See www.netlab.caltech.edu

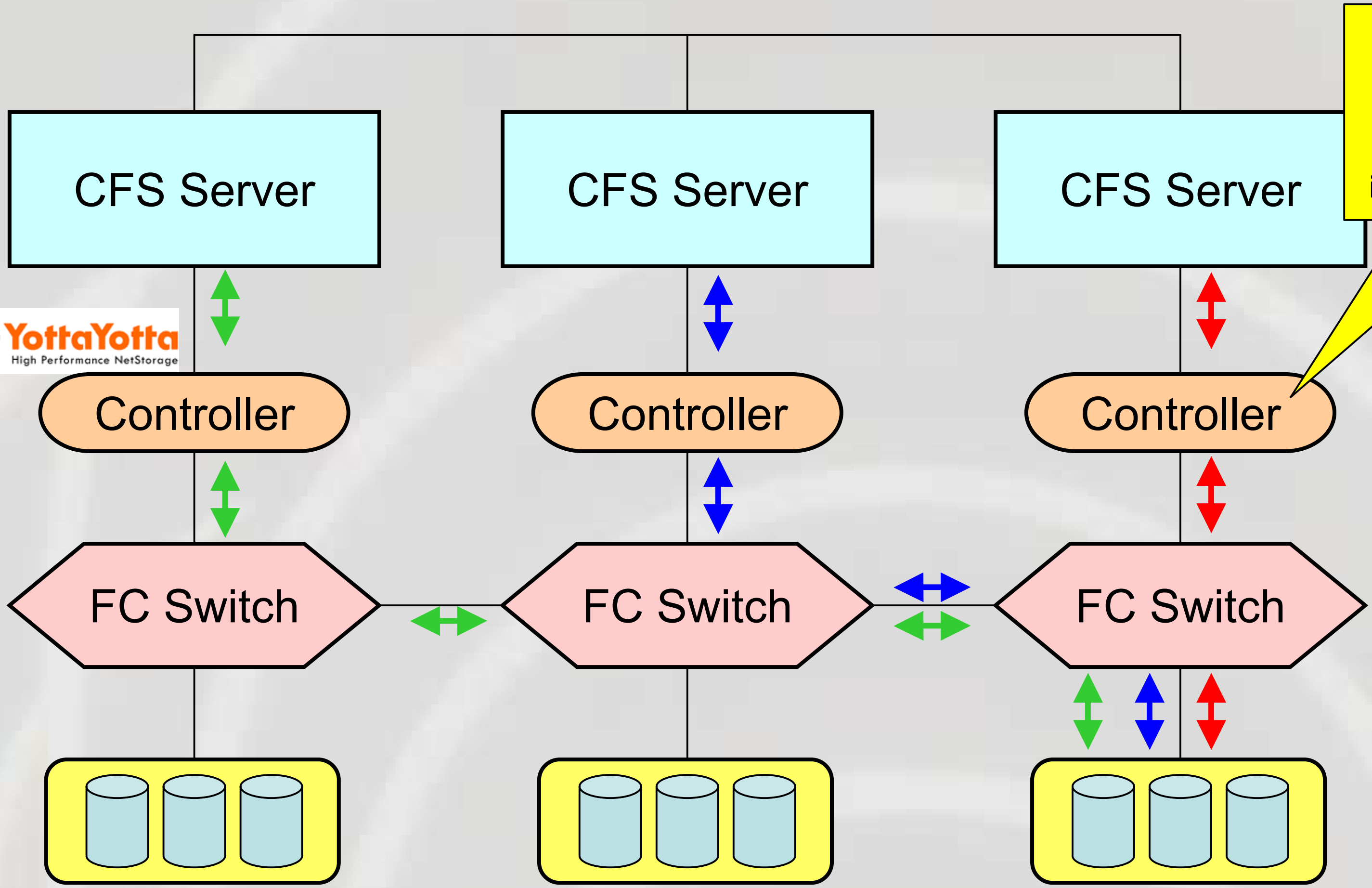
Managing Extreme Latency with Distributed Block Caching

Conventional Clustered File System (CFS)



Single controller is bottleneck for bandwidth and latency

CFS with Cache Coherent Block Controller



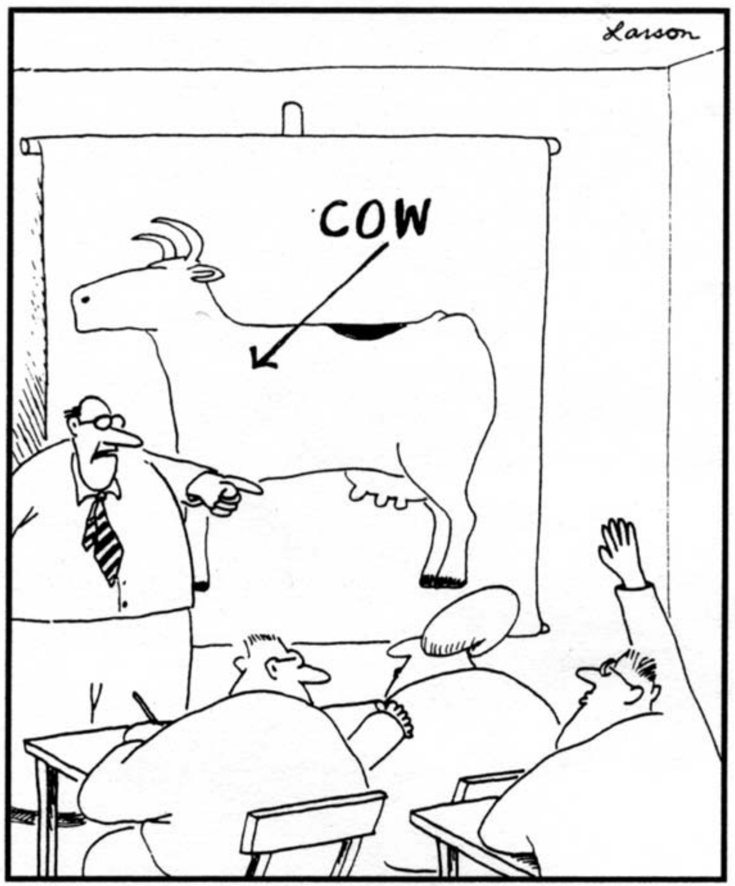
Cache coherent distributed controller dramatically reduces file lock latency and increases performance

See www.yottayotta.com

Summary

- We cannot repeal the laws of Physics
  - There will always be a price for distributed storage
  - $$$
  - Performance
- Many new technologies are being introduced to mitigate the impact of distance
  - FC routing
  - Fast TCP
  - Short cycling SCSI
  - Distributed block caching

Questions are Good!



"Yes … I believe there's a question there in the back."

especially

…~~even~~ if they seem obvious.