Synchronous
Operations

Hosts

Record

1...Appl creates record

2...Write Record

Hosts

5...Release Host-OK

3...Mirror record

Record

Record

Record

4...Mirror OK

6...Disk Commit

Record

Primary (MCU)

Secondary (RCU)

I/O Consistent

Non-Synchronous & Synchronous DB I/O

Sequence Numbers and Time Stamps

Enterprise SAN Design
End-To-End SWAN Testing

Multiple DUAL Homed WIntel Servers Runing IOMeter Load Generator

I/O Flow

10/100Mbps (IP)

NT Servers

Dual-homed FC (2 Gbits/s)

XP Array Performance Monitor Station

Veritas Master/Media Server (UNIX)

Dual-homed FC (2 Gbits/s)

4x10/100Mbps (IP)

I/O Asynch. Acknowledgement Before Remote Write Complete

Brocade Performance Monitors Running On Each Blade

I/O Flow From WIntel Servers

Single-homed FC (1 Gbits/s)

IBM 3584 UltraScalable Tape Library

Brocade Silkworm 12000

Brocade ISL Testing

CNT SPOF I/O Stress Testing

Dual-homed FC (1 Gbits/s)

Continuous Available Data Replication I/O Flow

CNT UltraNet Edge Storage Router 1101

Dual-homed Gigabit Ethernet

IP IP

Cisco Catalyst 6509 + Sup I & MSM

Cisco Catalyst 6509 + Sup I & MSM

Dual-homed Gigabit Ethernet

Cisco Catalyst 6509 + Sup II & MSFC2

Dual-homed Gigabit Ethernet

Cisco ONS 15454

Continous Available Replication Port Fiber Channel 2 Gbs

HP SureStore XP 128

Continous Available Replication Port Fiber Channel 2 Gbs

8 Fiber Ports Configured For XP128 (2 Spares)

Replication load Balances Across Both Fiber Ports

8-Port Fiber Boards (6 Ports Active) Testing

CNT UltraNet Edge Storage Router 1101

IP IP

Cisco Catalyst 6509 + Sup II & MSFC2

Dual-homed Gigabit Ethernet

Cisco ONS 15454

AT&T Aura-Ring SONET OC-3

Asynch Write I/O Acknowledgement fromNYC XP Array

Single-homed FC (1 Gbits/s)

IBM 3584 UltraScalable Tape Library (3 Tape Units)

NT Servers

I/O Meter Generating Load I/Os

Dual-homed FC (2 Gbits/s)

IP

Veritas Master/Media Server (UNIX)

Brocade SilkWorm 3800 (16 port)

Brocade SilkWorm 3800 (16 port)

Enterprise IP Backbone Network

Enterprise IP Network

4x10/100Mbps (IP)

IP

UNIX Servers

HP SureStore XP 128

Multi-homed FC (2 Gbits/s)

Continous Available Replication Port Fiber Channel 2 Gbs

Replication load Balances Across Both Fiber Ports

8 Fiber Ports Configured For XP128 (2 Spares)

NOTES

1. SilkWorm 12000 comes with dual control units, multiple power-supplies and fans, and has two 16 port switches configured as separate fabrics.

2. CNT UltraNet Edge Storage Router 1101 are used exclusively for the communication between HP SureStore XP 128 devices.

3. Based on the design and the backbone bandwidth limitation, Asynchronous communication between the XP 128 devices was mandated by the original vendor.

Network Considerations For SAN's

Applications over the MAN/WAN

– Moving large amounts of data between sites

– File sharing – supercomputing applications

– Remote mirroring / disaster recovery

– Synchronous mirroring

– Asynchronous mirroring

❖ Tape backup

❖ Centralizing Storage Management

❖ Disk edge caching

❖ Virtualization

❖ High availability / clustering

❖ Video streaming

❖ SSP remote primary storage

Gateway   **MAN/WAN**   **Gateway**

Primary Storage

Remote Storage
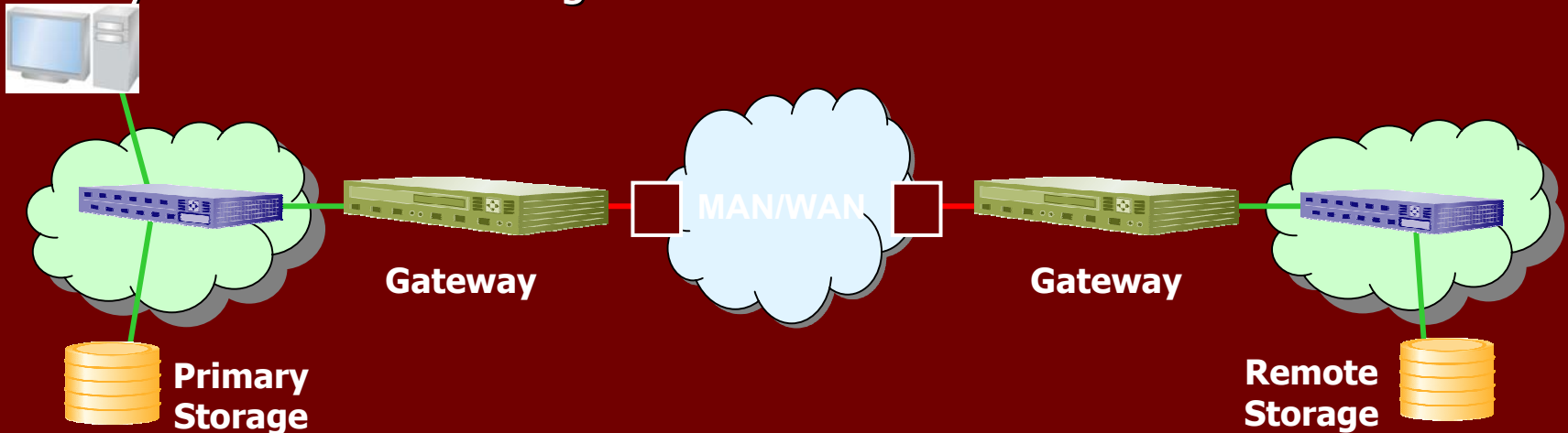
Moving Data Between Sites

- Fibre channel is the workhorse for moving data inside the data center
- Outside the data center there are two ways to move Terabytes between sites
  - Non real time – i.e. people power
  - Real time – direct connection
- Fundamentally, two things are necessary for a direct connection to efficiently move large amounts of data over distance
  - High Bandwidth interconnection
  - Flow Control - ability to fill and sustain bandwidth flow over time

Why Fibre Channel?

- Storage to storage transfers demands a high QoS
  - Needs "deterministic" communications
  - Generally very sensitive to latency
  - Requires in-order block delivery
- Fibre Channel is deterministic
  - Provides low latency and high bandwidth interconnect
- High performance SAN extension must maintain same QoS features over distance
  - High bandwidth
  - Low latency

How Much Bandwidth is Enough…?

- Moving 10 TB requires:
  - 2.25 hours using OC-192 (10Gb/s)
  - 9 Hours using OC-48 (2.5Gb/s)
  - 14 hours using "2G" FC (1600 Mb/s)
  - 28 hours using "1G" FC (800Mb/s)
  - 35.7 hours using OC-12 (622 Mb/s)
  - 6 days using OC-3 (155 Mb/s)
  - 20 days using T3 (45 Mb/s)
  - 1.6 years using T1 (1.5 Mb/s)

……**If** the pipe is fully utilized!

Data Movement Today – A Recent Case Study

Server

IP Network

Server

Los Alamos National Laboratory (LANL)

Sandia National Laboratory (SNL)

Fibre Channel Storage Area Network

Fibre Channel Storage Area Network
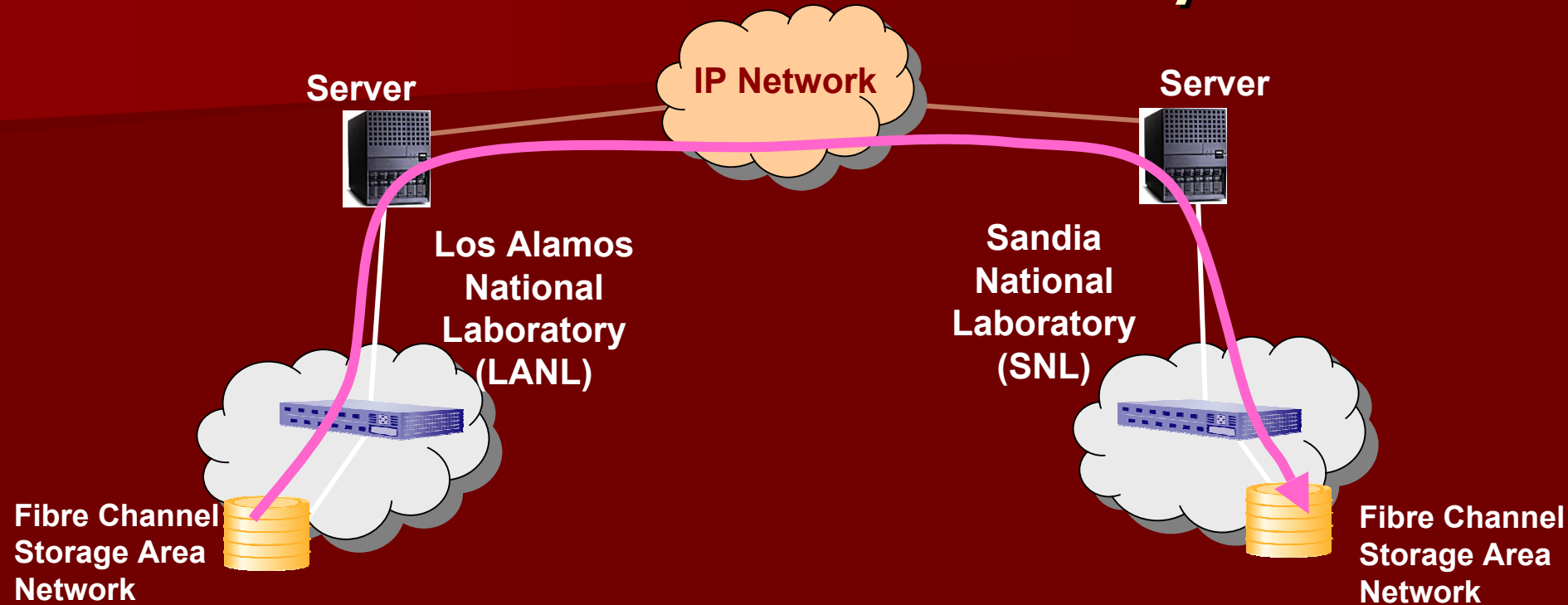
Scientists at LANL currently dump 100GB of supercomputing data to tape and FedEx it to SNL because it is faster than trying to use the existing 155Mb/s IP WAN connection

– Actual measured throughput of 16Mb/s! (10% bandwidth utilization)

The Better Way –
Directly Between Storage Systems

IP Network

Server

Server

Local Data Center

Remote Data Center

FC SAN

FC SAN

SAN Gateway

High QoS Link

SAN Gateway
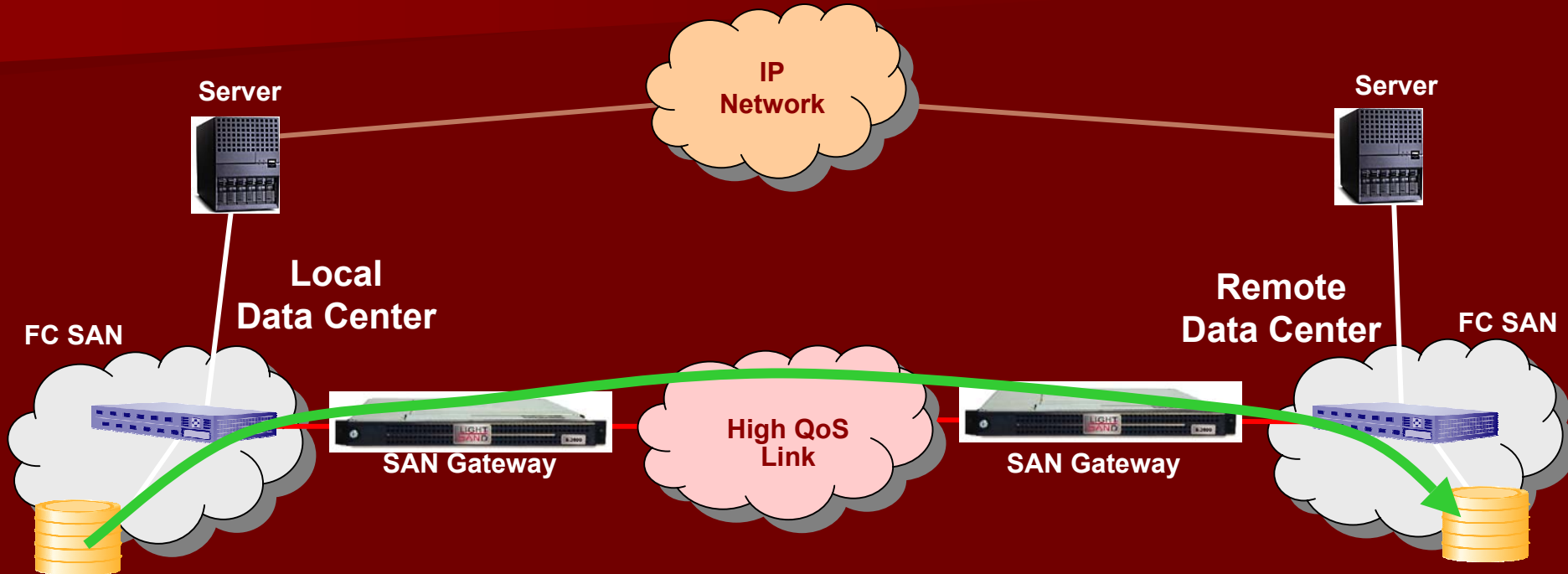
Utilizing the full bandwidth of this interconnection the data could have been moved in just over 80 minutes!

Transport Layer Extension

How do I move the signal?

- FC over Dark Fiber
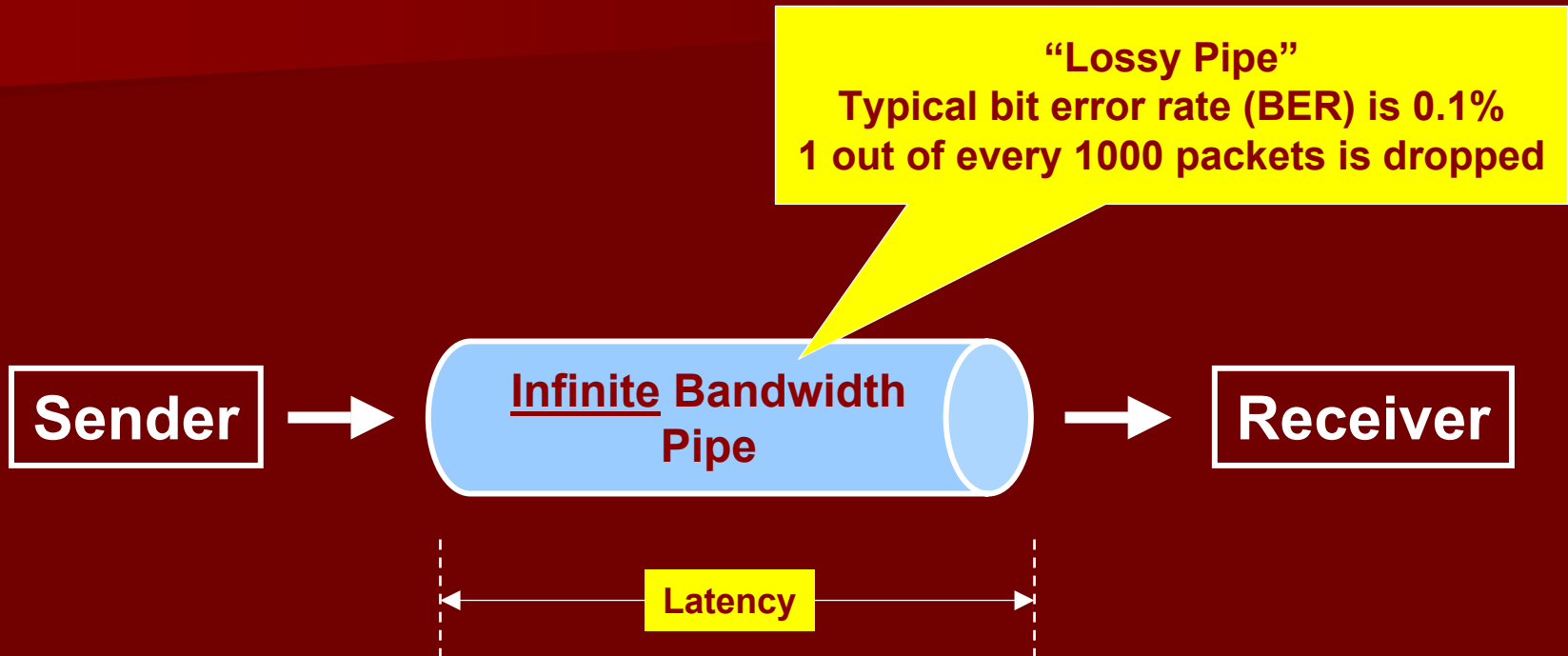- FC over DWDM
- FC over ATM
- FC over SONET
- FC over IP

Transport Layer Extension through the MAN and WAN
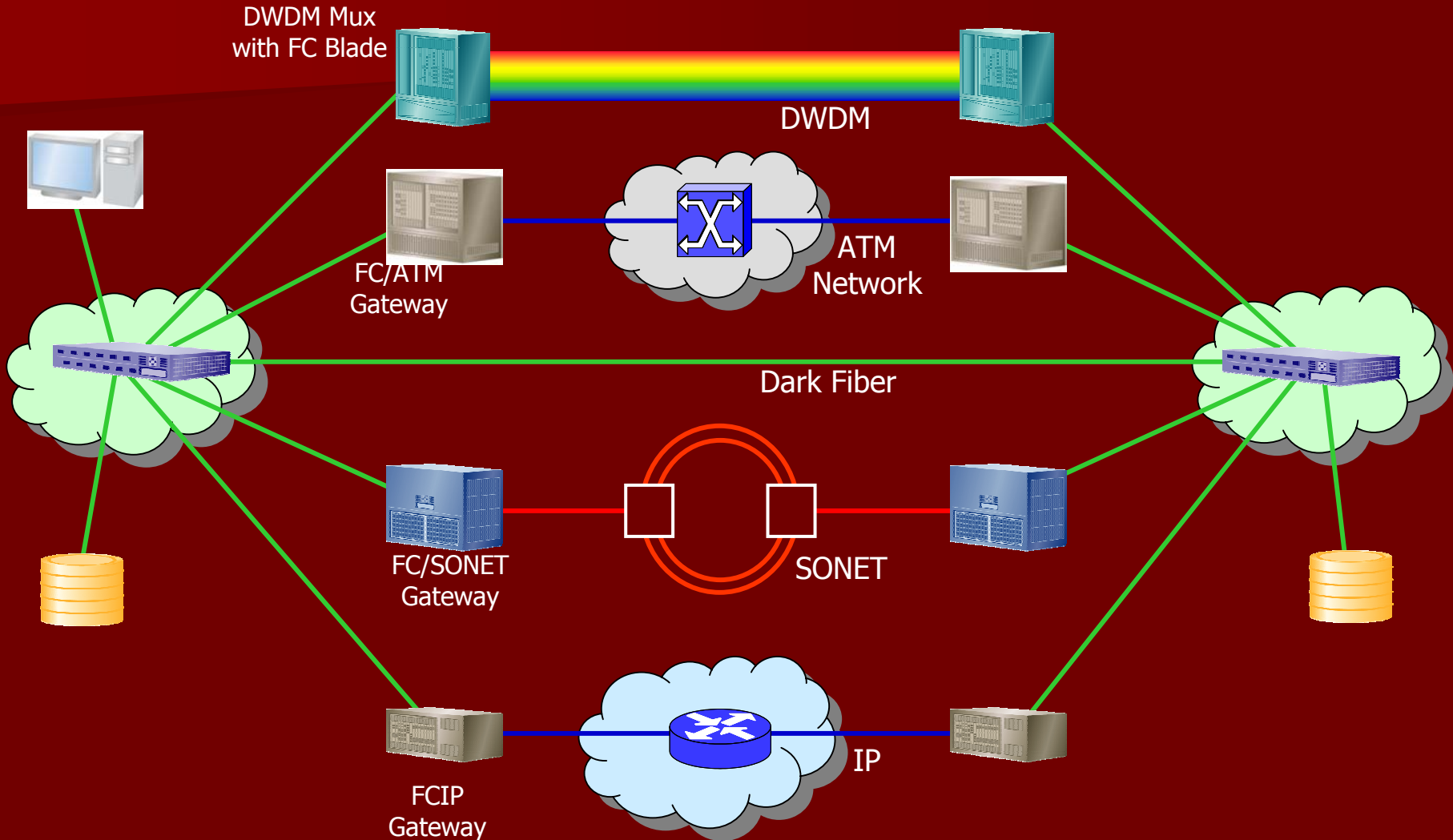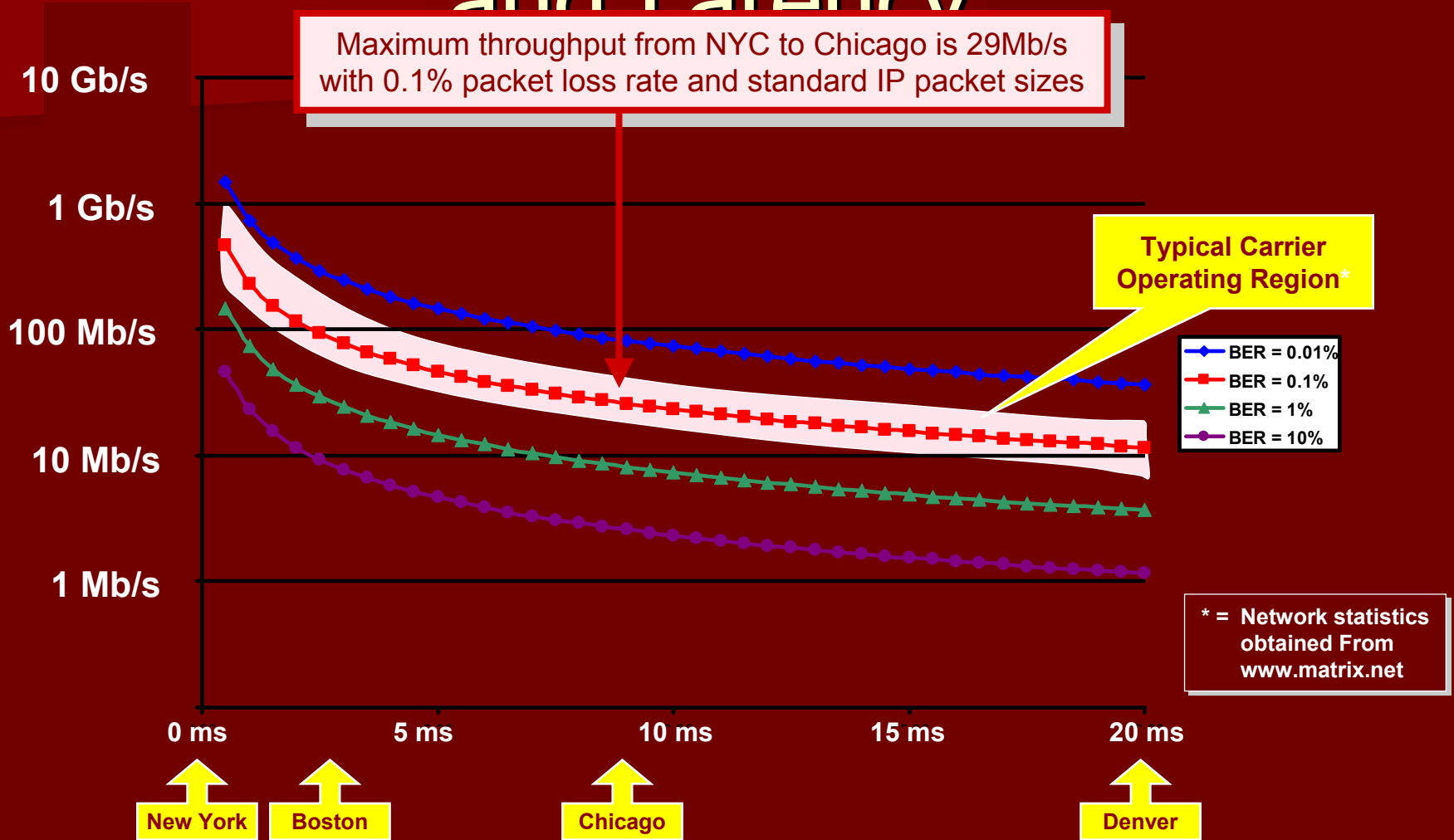
FC Data Encapsulation Efficiency

	Storage over IP	Storage over SONET	Storage over ATM
Mapping	FC FCIP TCP IP IPsec PPP SONET FCIP	FC SONET FC over SONET	FC AAL ATM SONET FC over ATM
Mapping Efficiency	**90%**	**98%**	**80%**

"Fat Pipes" Don't Guarantee High Throughput with TCP/IP

"Lossy Pipe"
Typical bit error rate (BER) is 0.1%
1 out of every 1000 packets is dropped

Sender → **Infinite Bandwidth Pipe** → **Receiver**

Latency

**With TCP/IP over the WAN,
BER and Latency affect throughput
more than the bandwidth of the pipes**

TCP/IP Performance vs. BER and Latency

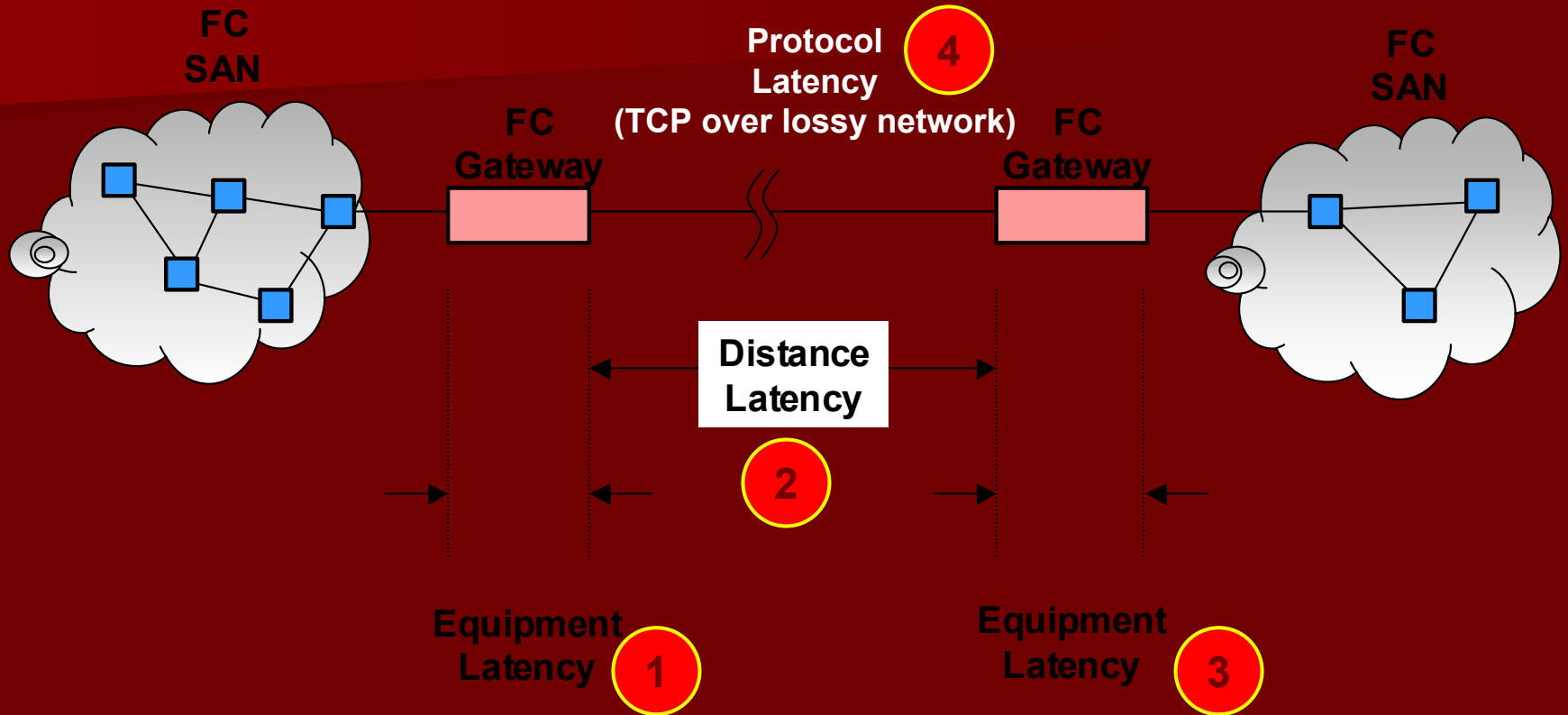**Maximum throughput from NYC to Chicago is 29Mb/s with 0.1% packet loss rate and standard IP packet sizes**

Typical Carrier Operating Region *

- BER = 0.01%
- BER = 0.1%
- BER = 1%
- BER = 10%

* = Network statistics obtained From www.matrix.net

- 10 Gb/s
- 1 Gb/s
- 100 Mb/s
- 10 Mb/s
- 1 Mb/s

0 ms 5 ms 10 ms 15 ms 20 ms

New York Boston Chicago Denver

What's Wrong with Routed IP in the WAN?

- Routed IP is designed for scalability and connectivity
- Routed IP is NOT designed for performance
- TCP algorithms are designed to make the end user back-off quickly at the first sign of congestion
- Throughput is controlled by latency (geographical distance) and packet loss rate
- IP Carriers will always have packet loss
  - Sell service by peak capacity
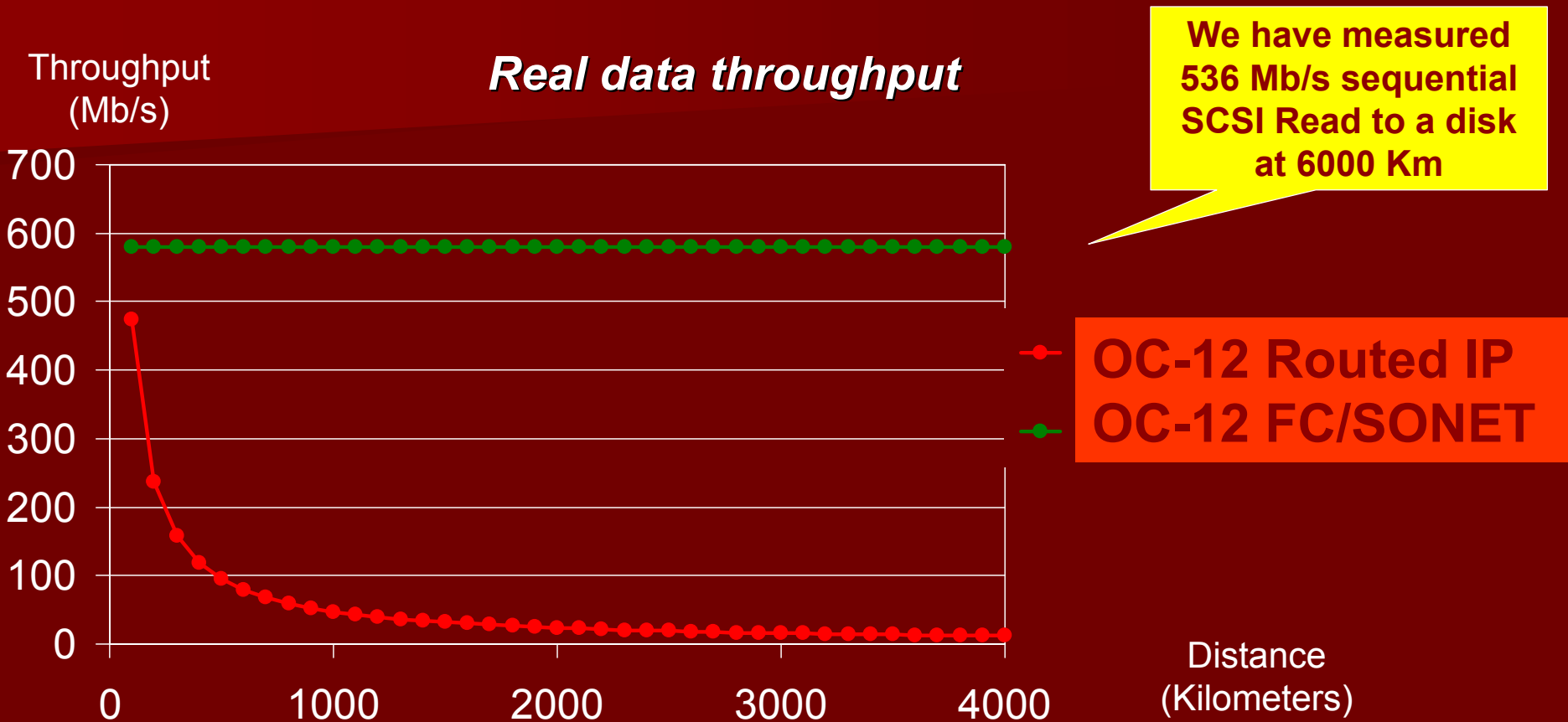 - Allocate equipment based on average capacity

Network Layer Flow Control via Credit Buffering

- Fibre Channel maintains throughput in the data center by using flow control via buffer to buffer credits
    - Nominally FC switches provide credit buffering up to 10km distance
- Any wide bandwidth, long distance movement of FC data must couple flow control over the WAN
    - System requires end-to-end credit buffering
- WAN gateway flow Control options
    - Transparent mode - relies on the FC switches to extend the credits (good up to 130-150km) This method is used in all DWDM gateways and some SONET, FCIP devices
    - Coupled Credit mode – The gateway handles the buffer to buffer credits over the WAN. This is extensible up to great distances (1000's of km). This method is used in most FC over SONET gateways
    - FCIP gateways must provide credit buffering in conjunction with PAUSE and TCP/IP

Multiple Sources for Latency
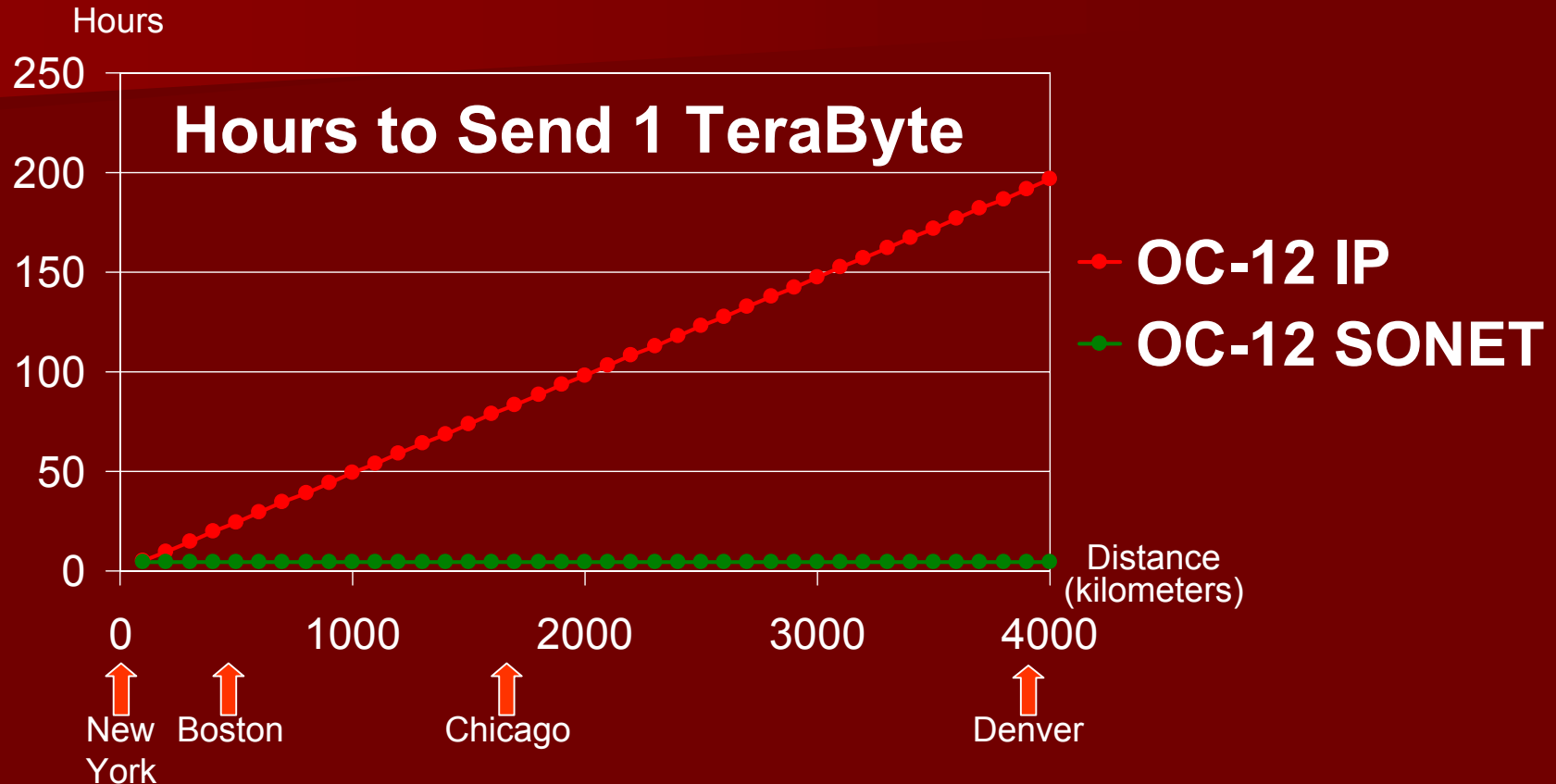
Fibre Channel Over SONET / WDM

Real data throughput

We have measured 536 Mb/s sequential SCSI Read to a disk at 6000 Km

Throughput (Mb/s)

700
600
500
400
300
200
100
0

Distance (Kilometers)

0   1000   2000   3000   4000

OC-12 Routed IP

OC-12 FC/SONET

Data re-transmission due to IP packet loss
limits actual IP throughput over

Fibre Channel over SONET
The High Efficiency, Long Distance Alternative

Hours

Hours to Send 1 TeraByte

250

200

150

100

50

0

- OC-12 IP
- OC-12 SONET

Distance
(kilometers)

0   1000   2000   3000   4000

New York   Boston   Chicago   Denver

Data re-transmission due to IP packet loss
limits actual IP throughput over

LIGHT SAND

Storage Over IP

Using CNT Solutions

Joint Solution Overview

- Continuous Access XP
  - UltraNet Edge, UltraNet Storage Director
  - HP XP Disk Arrays (XP48/XP256/XP512/XP1024)
  - HP Continuous Access software
  - Cluster Extension XP and Continental Cluster
  - All inter-networking options – IP, ATM, T3, Metro Fiber optics

- StorageWorks DRM
  - Enterprise Modular Array (Enterprise Virtual Array)
  - DRM (Data Replication Manager) Software
  - All inter-networking options – IP, ATM, T3, Metro Fiber optics

Fibre Channel SAN Issues - Extensibility

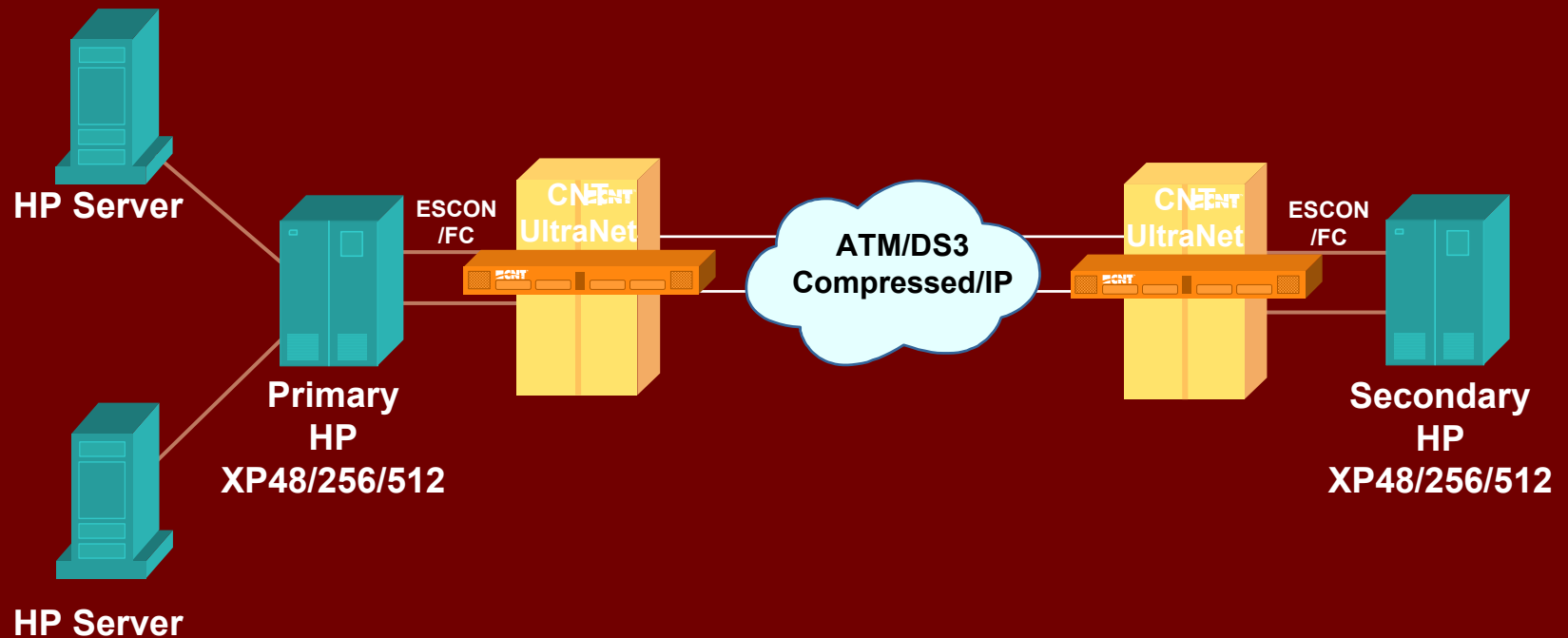Native FC <u>can not</u> be transported over most service providers infrastructure.
Wide area networks are based on IP or Ethernet and <u>not FC</u>, therefore, costly point-to-point Dark Fiber or DWDM is required to connect FC SAN.

FC SAN

FC

FC

FC

FC

FC SAN
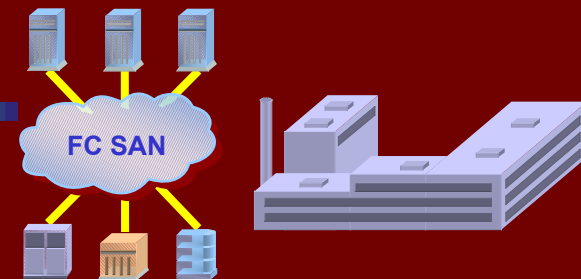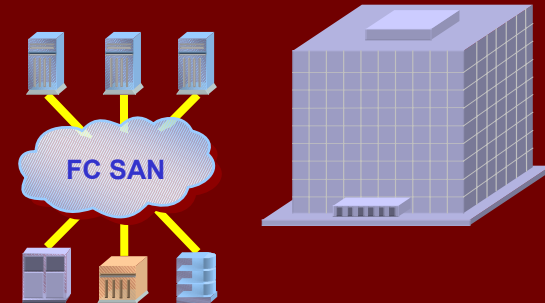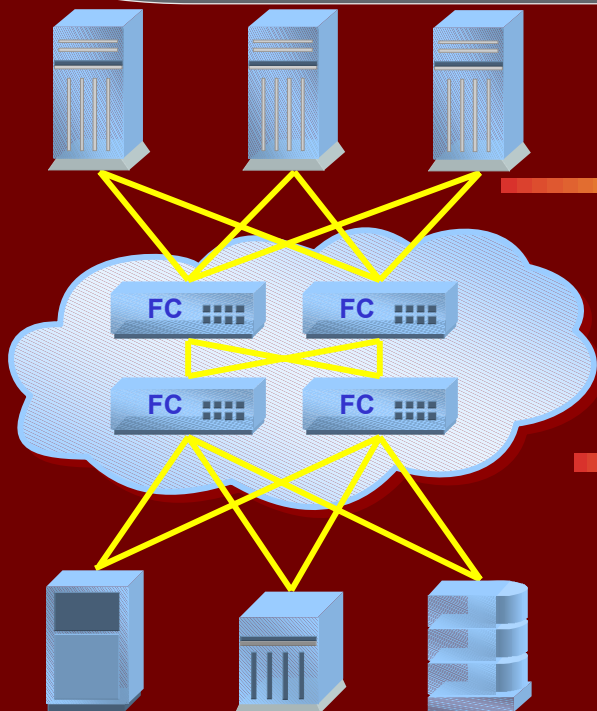
Fibre Channel SAN island

HP XP CA Remote Disk Mirroring Over WAN/IP

Storage Over IP

Using Nishan IP Storage Solutions

Asynchronous Mirroring – Going the

Asynchronous Mirroring

Production Site

IP Network

1000s of miles

Recovery Site
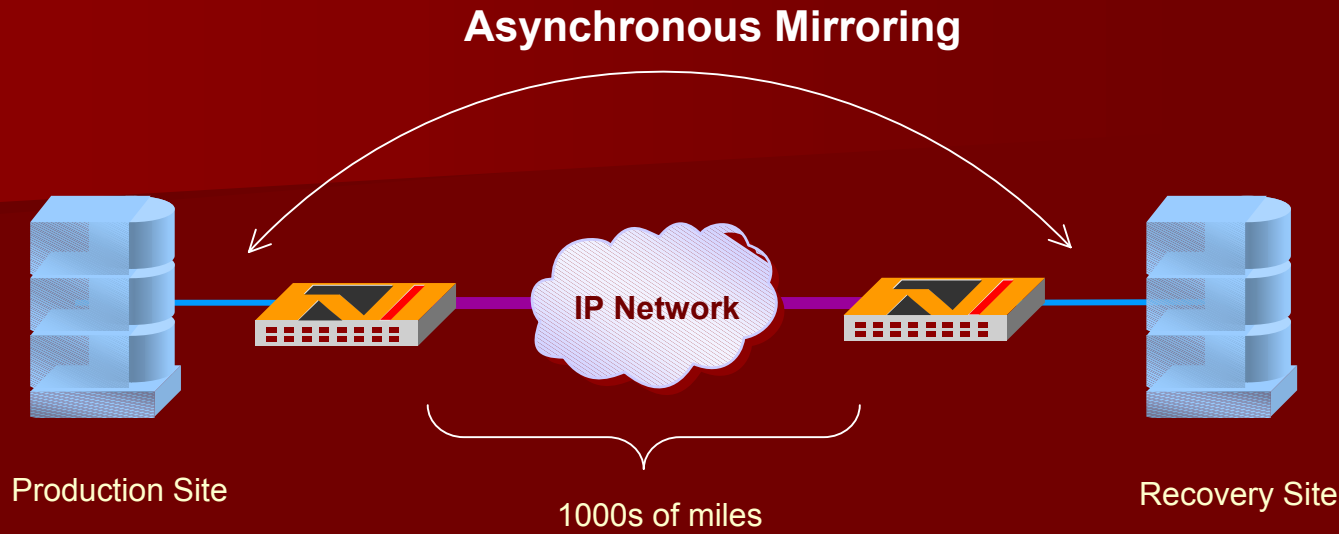
- **Customer required fast replication with 2 sites 2000 miles apart**

- **FASTWRITE technology excels at long-haul optimized performance**

Extending the Reach of Synchronous Replication

Vtesse Networks - Europe

Synchronous HDS TrueCopy
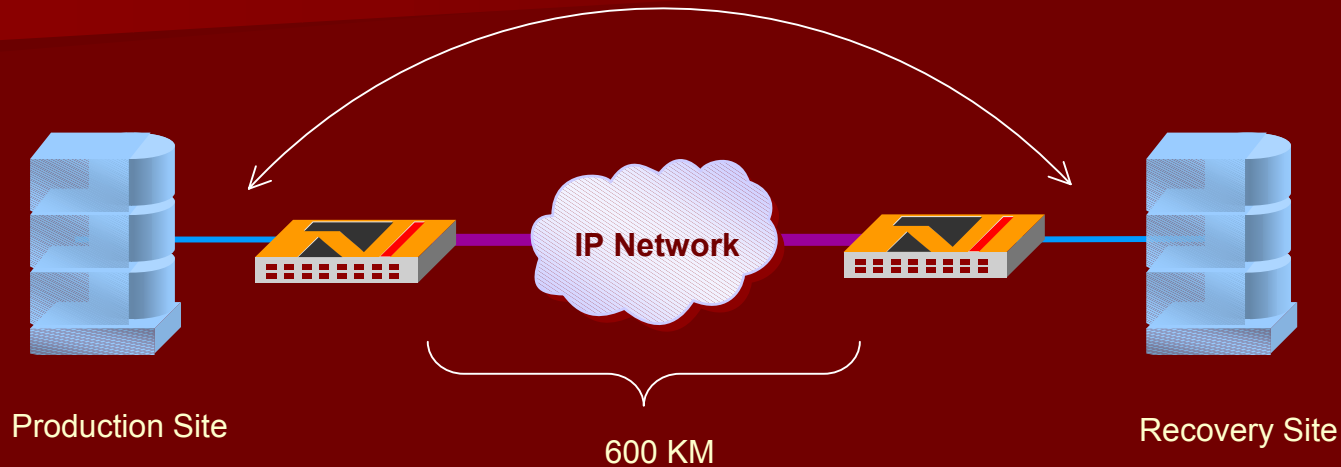
- **Nishan switches convert FC mirroring traffic to IP**
- **Synchronous data replication up to 600KM**
- **Expands the reach of Sync TrueCopy => more TC addressable market !!**

Steinbach Credit Union – Wireless IP SAN

STEINBACH CREDIT UNION

Winnipeg, Manitoba

Steinbach, Manitoba

16 Miles

16 Miles

100 Meg
Full duplex

100 Meg
Full duplex

Proxim Wireless Bridge

Comm. Tower

Comm. Tower

Comm. Tower

Proxim Wireless Bridge

Cisco GE to 10/100 Switch

Cisco GE to 10/100 Switch

Nishan 4300

GE

Nishan 3300

GE

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

FC

Server

Server

Server

Server

Server

Server

Server

Server

Server

Server

Server

Servers

Servers
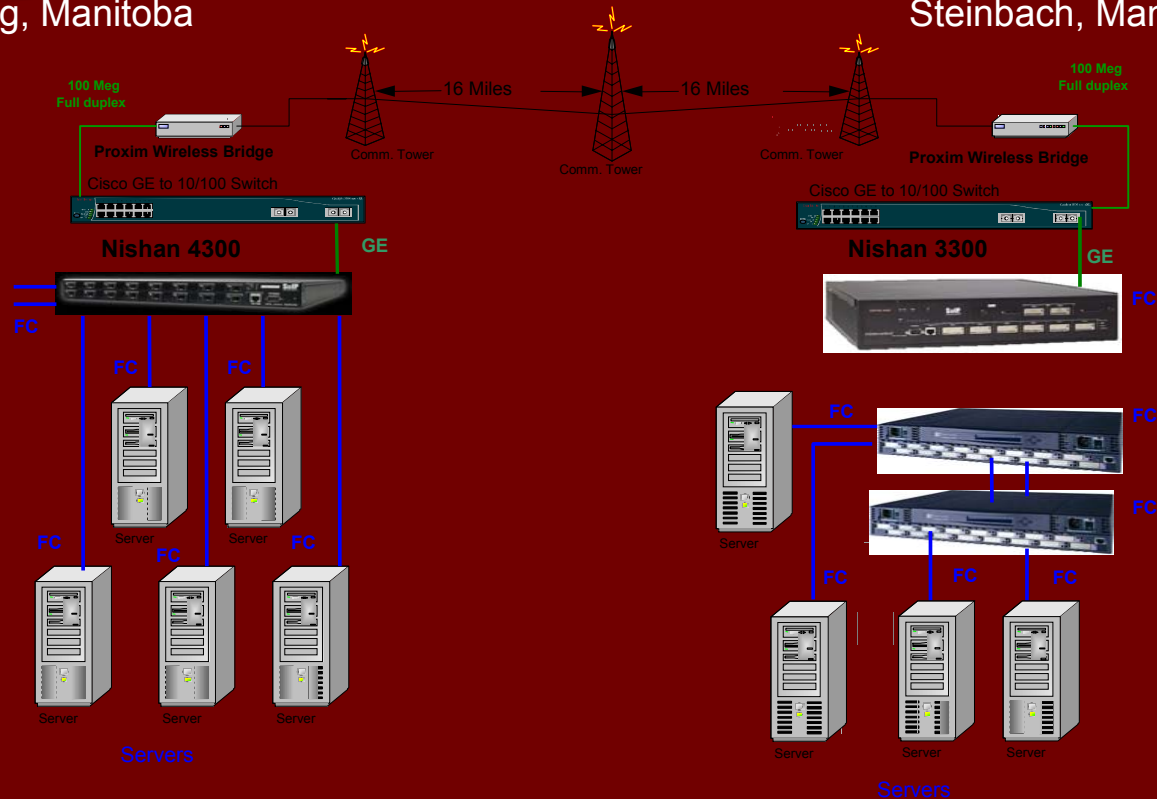
Enterprise-class Scalability via SAN Internetworking
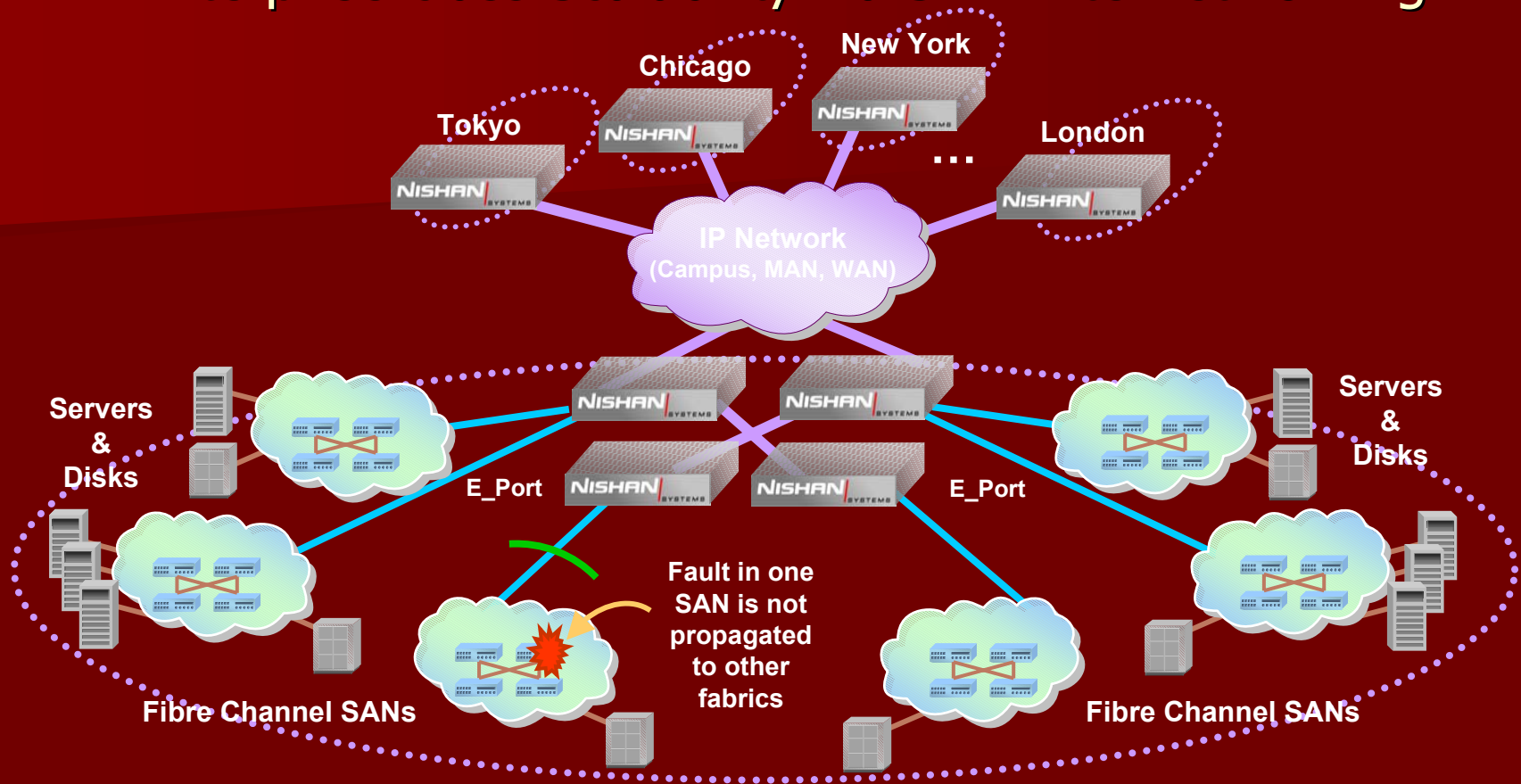
- **Fault isolation provides router-like scalability – Enterprise-class scalability**
- **Connectivity between Heterogeneous Fibre Channel switches**
- **Extensibility across any distance: campus, metro or WAN**

High Availability And Disaster Tolerant SAN Considerations

Storage WAN   (SWAN)

➢ Features

- Manage the WORLD as a single Entity

- Wide area data sharing/migration

- Storage Management/Reporting

- Shadow/Remote Backup

- Wide area DT

- Connectivity within <u>standard</u> infrastructure (Network)

- Span: The <u>World</u>

Fabric Extension

OC3 Link (155.52 mbs)*

Extended Buffer Credits

Extended Buffer Credits
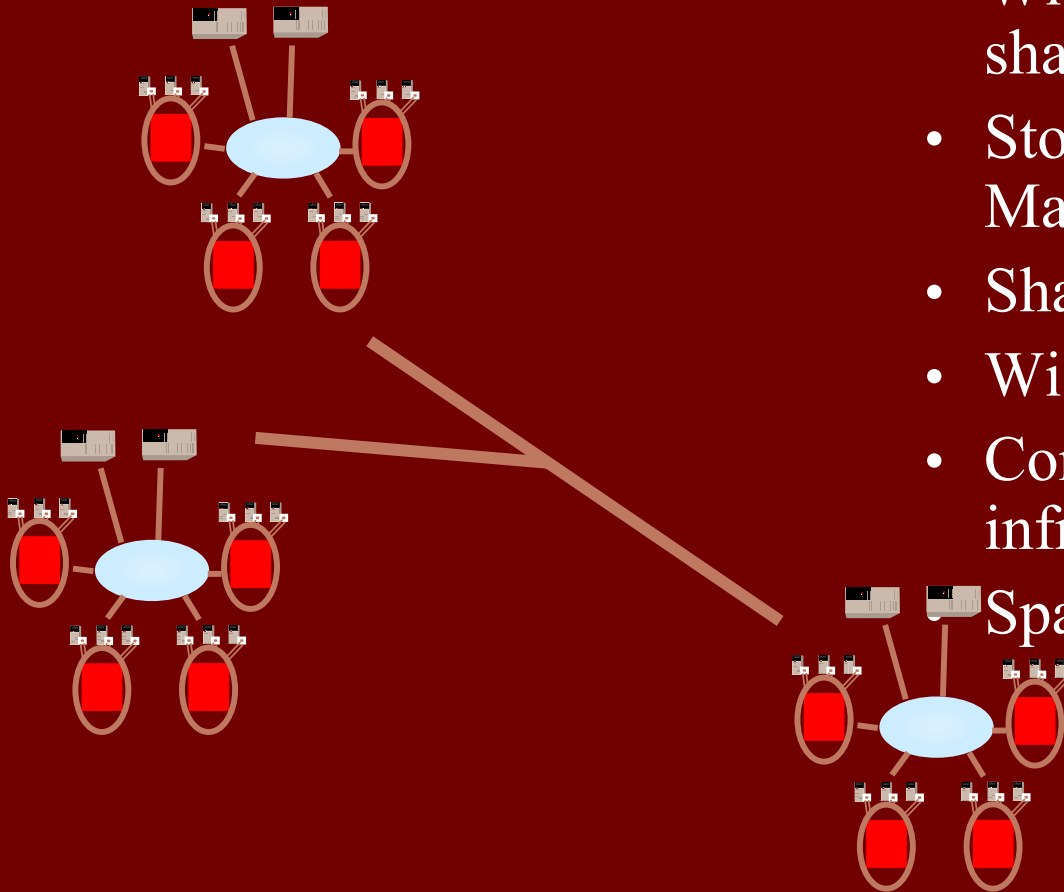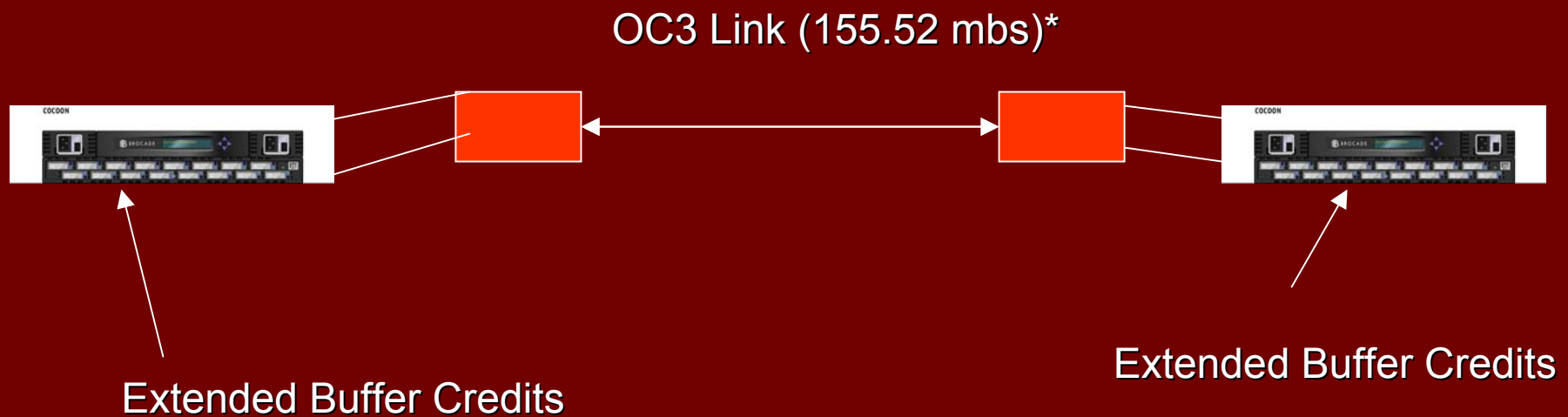
*Used to specify the speed of fiber optic networks. The base rate (OC-1) is 51.84 Mbps. OC-2 runs at twice the base rate, OC-3 at three times the base rate (155.52 Mbps), etc. Planned rates are: OC-1, OC-3, OC-12 (622.08 Mpbs), OC-24 (1.244 Gbps), and OC-48 (2.488 Gbps

Data Replication

> Replication can be done at many levels

> Replication can be done at many levels

Application	→	**Application**
Database	→	**Database**
Operating System	→	**Operating System**
Hardware	→	**Hardware**
Storage	→	**Storage**
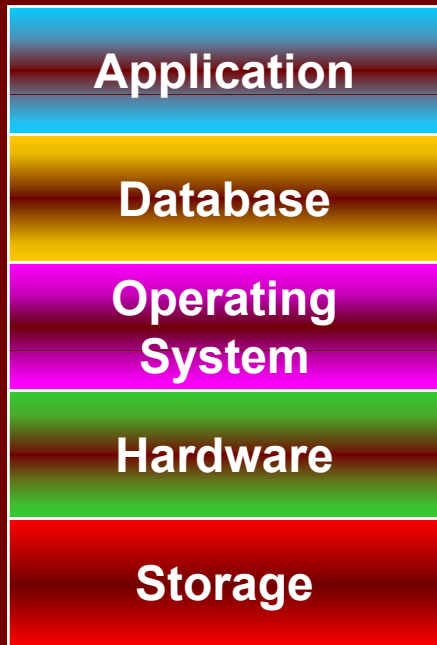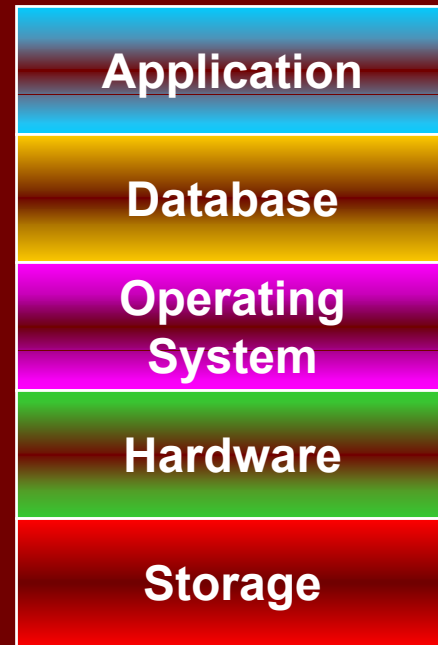
•Real Time COPY

- Provide Disaster Recovery
- **NOT to maintain two identical copies**
- Provide I/O consistent copy of data

Synchronous Or Asynchronous

- If Within ***Supported*** Distance Use Synchronous Because:
  - Data is more secure
  - Best overall performance
- Asynchronous is supported
  - Cases w/low I/O rates and some potentially lost data is acceptable if links are broken

Asynchronous
Operations

Hosts

Record

1...Appl creates record
2...Write record
3...Release Host-OK
4...SEND to network
5...XMIT over Network
6...Send to Disk Subsystem
7...Mirror OK
8...Disk Commit

Hosts

Record

Record

Record

Record

Record

Network

Network Directors

Primary 9900 (MCU)

Secondary 9900 (RCU)

Asynchronous
Update Sequence

Data (5) for C/TG_1
Data (4) for C/TG_0
Data (3) for C/TG_1
Data (2) for C/TG_0
Data (1) for C/TG_0

C/TG= Consistency group
SF= Side File

Attach
"Sequence Control Information"

TC Async

(2) (1) (3) (5) (4)

Asynchronous Transfer
with Out of Order

Sort

#3:Data (4)
#2:Data (2)
#1:Data (1)
[SF of C/TG_0]

#2:Data (5)
#1:Data (3)
[SF of C/TG_1]

Cache

#3:Data (4)
#2:Data (2)
#1:Data (1)
[SF of C/TG_0]

#2:Data (5)
#1:Data (3)
[SF of C/TG_1]

Cache

P-VOL 0 **P-VOL 1** **P-VOL 2** **C/T group_0** **S-VOL 0** **S-VOL 1** **S-VOL 2**

P-VOL 3 **P-VOL 4** **P-VOL 5** **C/T group_1** **S-VOL 3** **S-VOL 4** **S-VOL 5**

MCU (= Master subsystem)

RCU (= Remote subsystem)

Synchronous
Operations

Hosts
Hosts

Record

1...Appl creates record

2...Write Record

5...Release Host-OK

3...Mirror record

Record

Record

Record

Record

4...Mirror OK

6...Disk Commit

Primary (MCU)
Secondary (RCU)

**Enterprise SAN Design
End-To-End SWAN Testing**

Multiple DUAL Homed Wintel Servers Runing IOMeter Load Generator

I/O Flow

Brocade ISL Testing (Potential SPOF From Each Blade)

CNT SPOF I/O Stress Testing

10/100Mbps (IP)

NT Servers

Dual-homed FC (2 Gbits/s)

Brocade Silkworm 12000

XP Array Performance Monitor Station

Enterprise IP Network

Dual-homed FC (2 Gbits/s)

Dual-homed FC (1 Gbits/s)

CNT UltraNet Edge Storage Router 1101

IP   IP

Dual-homed Gigabit Ethernet

Cisco   Cisco

Dual-homed Gigabit Ethernet

Cisco

Dual-homed Gigabit Ethernet

Cisco

Continuous Available Data Replication I/O Flow

Asynch Write I/O Acknowledgement fromNYC XP Array

CNT UltraNet Edge Storage Router 1101

IP   IP

Dual-homed Gigabit Ethernet

Cisco

Dual-homed Gigabit Ethernet

Cisco

AT&T Aura-Ring SONET OC-3

NT Servers

I/O Meter Generating Local I/Os

IP

Brocade SilkWorm 3800 (16 port)

Brocade SilkWorm 3800 (16 port)

Enterprise IP Network

Dual-homed FC (2 Gbits/s)

Enterprise IP-Backbone Network

UNIX Servers

IP

I/O Asynch. Acknowledgement Before Remote Write Complete

I/O Flow From Wintel Servers

Brocade Performance Monitors Running On Each Blade

Continous Available Replication Port Fiber Channel 2 Gbs

8-Port Fiber Boards (6 Ports Active) Testing

HP XP Array

Replication load Balances Across Both Fiber Ports

Continous Available Replication Port Fiber Channel 2 Gbs

8 Fiber Ports Configured For XP128 (2 Spares)

Multi-homed FC (2 Gbits/s)

Continous Available Replication Port Fiber Channel 2 Gbs

Replication load Balances Across Both Fiber Ports

8 Fiber Ports Configured For XP128 (2 Spares)

HP SureStore XP 128

Continous Available Replication Port Fiber Channel 2 Gbs

8 Fiber Ports Configured For XP128 (2 Spares)

NOTES
1. SilkWorm 12000 comes with dual control units, multiple power-supplies and fans, and has two 16 port switches configured as separate fabrics.
2. CNT UltraNet Edge Storage Router 1101 are used exclusively for the communication between HP XP 128 SAN's.
3. Based on the design and the backbone bandwidth limitation, only Asynchronous communication between the XP 128 SAN's was recommended.
5. The design will work (in Async. mode) with existing Cisco's Enterprise IP-backbone without using QoS, Packet Prioritization, Traffic Shaping, or Layer 3 switching technology.

Questions??

Thank You