SAN Design Considerations for Optimizing Performance, Scalability, and Availability Over the Wide Area Network

Allan Bolding
Nishan Systems
3850 North First Street, San Jose, CA 95134
Phone: 408-519-3427 FAX 408-519-3705
E-mail: abolding@nishansystems.com

Presented at the THIC Meeting at the STK Bldg 8 Auditorium, 1 Storage Tek Dr, Louisville CO 80027-9451

July 22 - 23, 2003

Achieving High Performance

Achieving High Performance

- TCP Optimizations
  - Large window sizes
  - Traffic shaping or rate limiting
  - SACK
  - Jumbo Frames (Ethernet layer)
  - Sequence number wrapping
- Fibre Channel
  - FCP command spoofing
  - Block sizes, outstanding IOs

Achieving High Scalability and Availability

- SAN Routing
  - Fault containment
  - Broadcast Containment
  - Distributed Fibre Channel services
- Traffic Differentiation

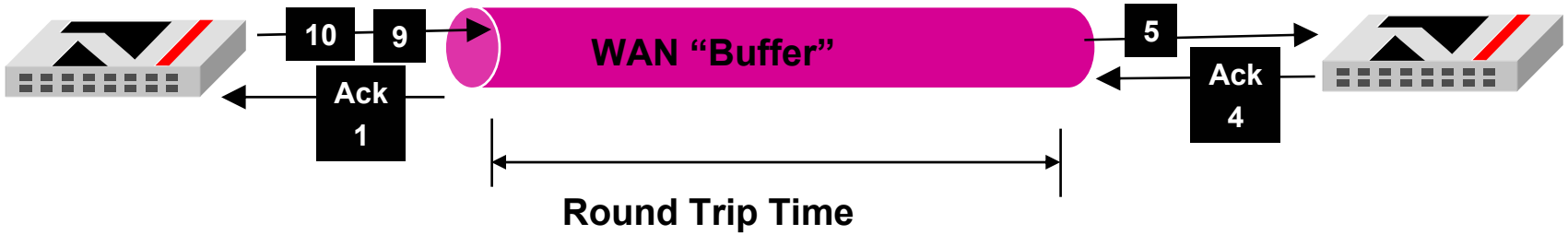
Achieving Higher Performance

- TCP Optimizations
- Fibre Channel

TCP Optimizations

Synchronous protocols need to keep the pipeline full.

- Lots of concurrent sessions
- Sending large data units (large TCP window size)



WAN Buffer Size = (RTT) * (bandwidth)

70ms * 1Gbps = 60 Mbps or 7.5 M byes

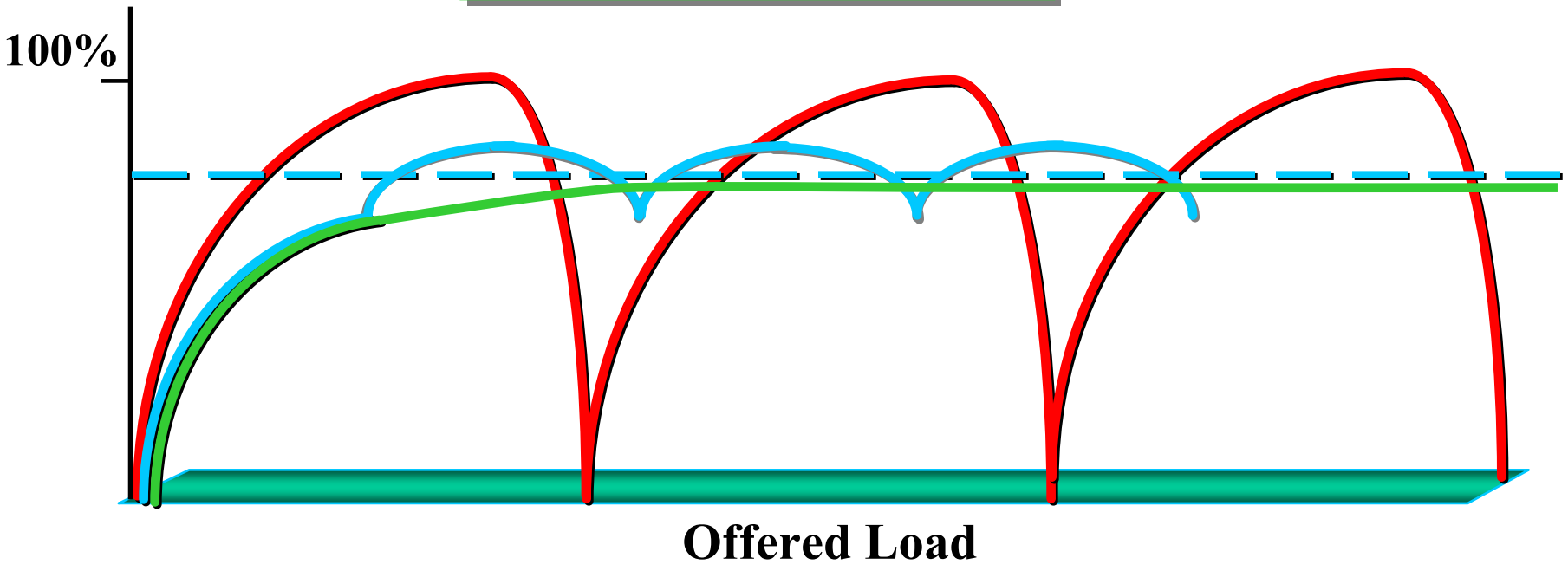
As the number of sessions increases the buffer is shifted from the WAN to the TCP sender's transmit buffers

Traffic Shaping or Rate Limiting Critical

Uncontrolled Congestion

Managed Congestion for Internet

Storage Traffic Management (shaped)



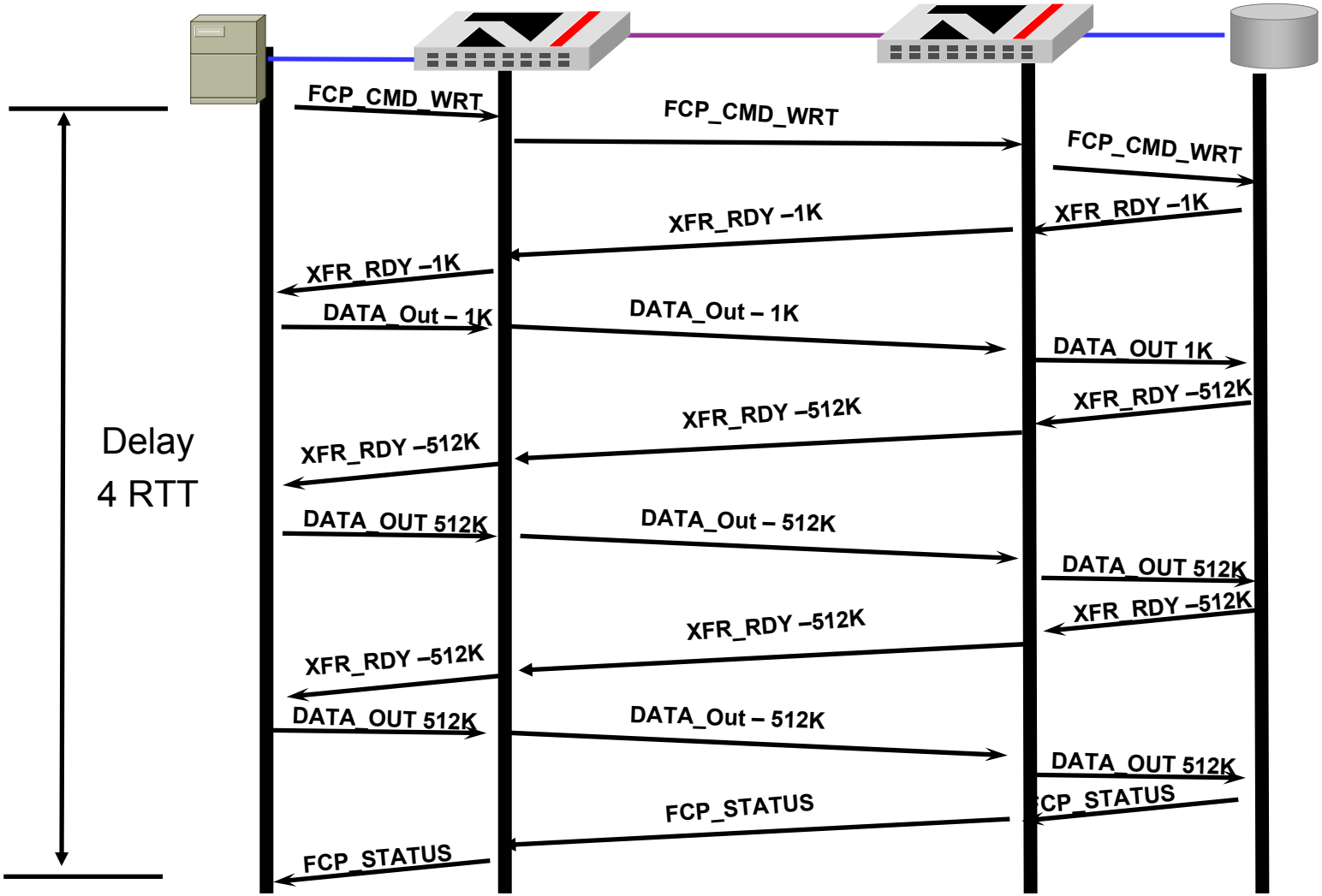
TCP Optimizations Cont'd

- **Traffic shaping or rate limiting:**
  - Critical to avoid packet drops
  - TCP retransmit mechanisms should not be relied upon for PERSISTENT packet drops
- **SACK: Selective Acknowledgements**
  - Useful when medium is dropping a percentage of the packets
- **TCP Window Scaling and Sequence wrapping (RFC 1323)**
  - 16 bits is not enough to specify an 8 MB window if each bit represents 1 byte.
  - 32 bit sequence field may not be large enough to ensure wrapping
- **Jumbo Frames**
  - Matches Fibre Channel frame size with Ethernet with TCP segment size

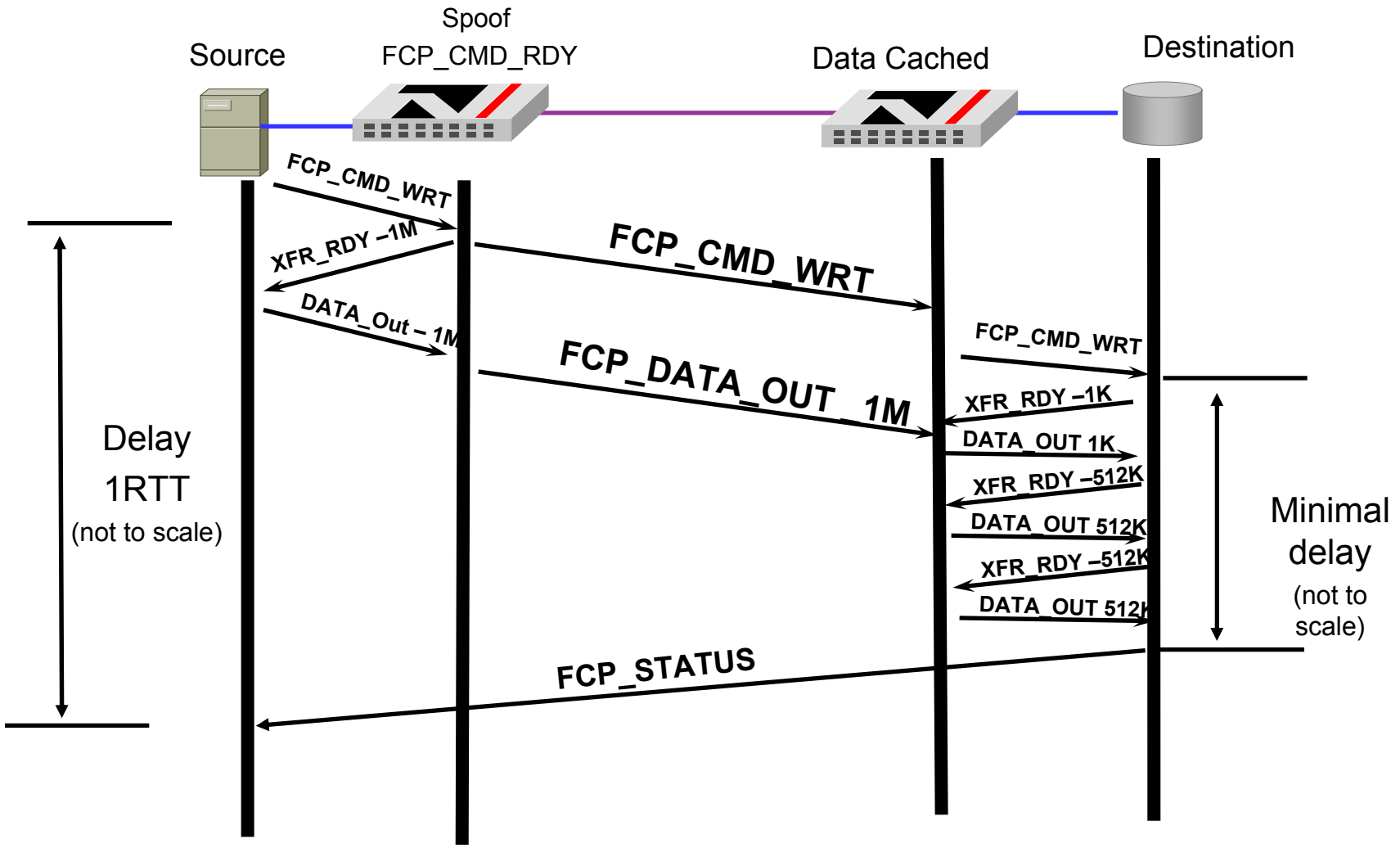
Fibre Channel Optimizations

- Keep the WAN link full
  - Large Fibre Channel Block sizes
  - Increase number of outstanding IOs
- Reduce the number of RTTs to complete a write transaction
  - Nishan Fast Write

Fibre Channel Optimizations (A 1MB write example)



Fibre Channel Optimizations With Fast Write

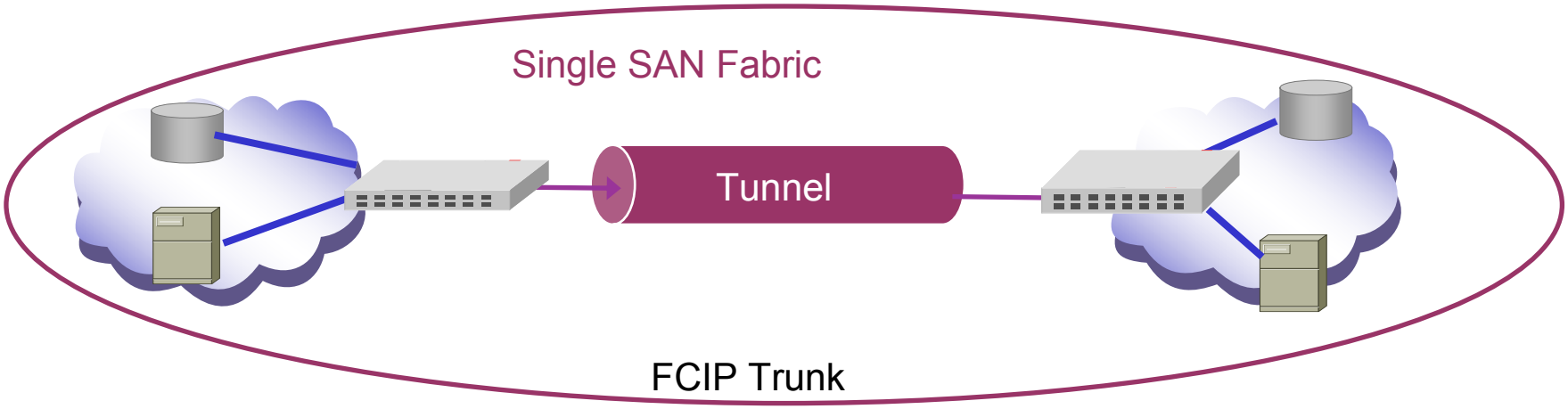


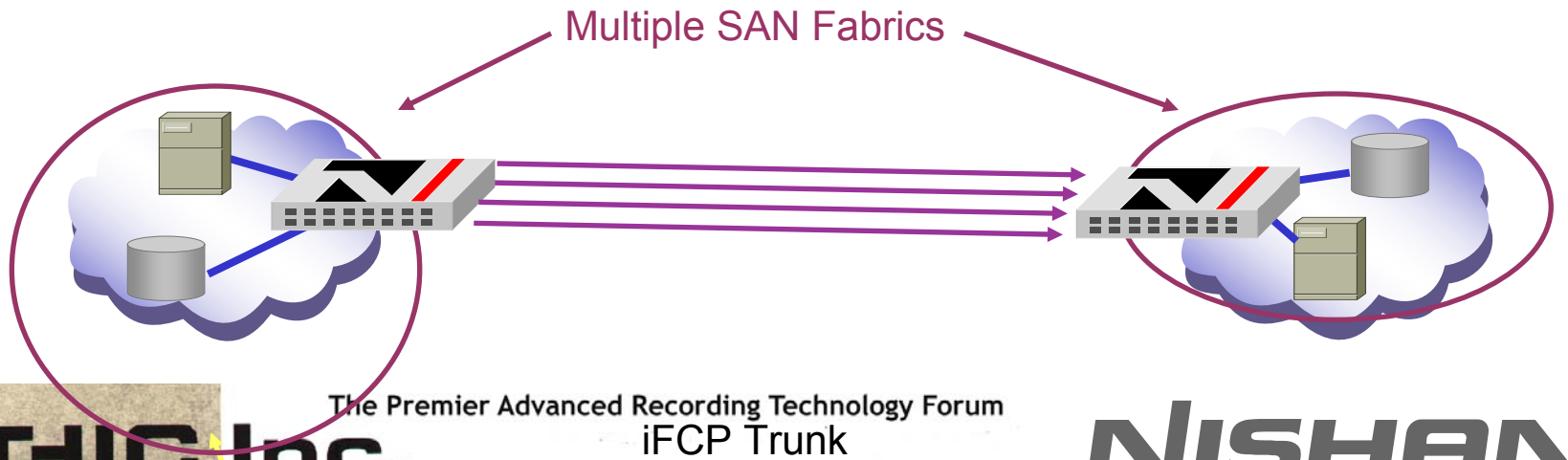
Achieving Higher Scalability

and

Availability

SAN Routing (iFCP) and SAN Bridging (FCIP)





Considerations: SAN Routing versus SAN Bridging

Single Fabric

- Topology changes advertised to all devices on F.C. fabric (SCN, RSCN Messages)
- Far-end and near-end must be part of the same F.C network. May require re-configuring HBAs
- SNS server shared for near-end and far-end

Multiple F.C. networks (iFCP, E_Port)

- **Broadcast Containment:**

  Near-end and far-end are isolated from topology change messages.

- **Distributed Fibre Channel Services**

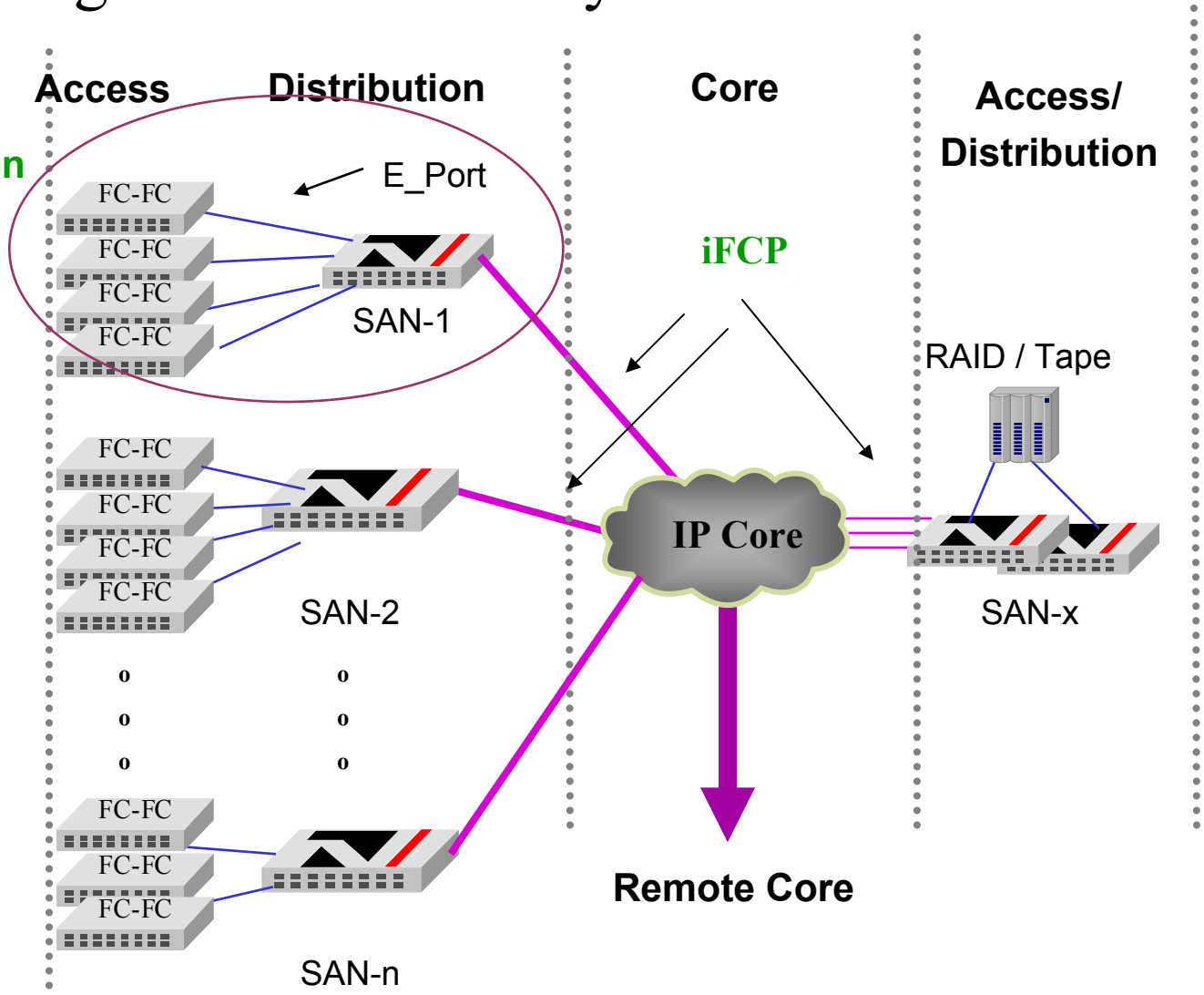
  Separate SNS servers, Separate Principle switches, No address reconfigurations (no re-configuring HBAs)

- **Fault Containment:**

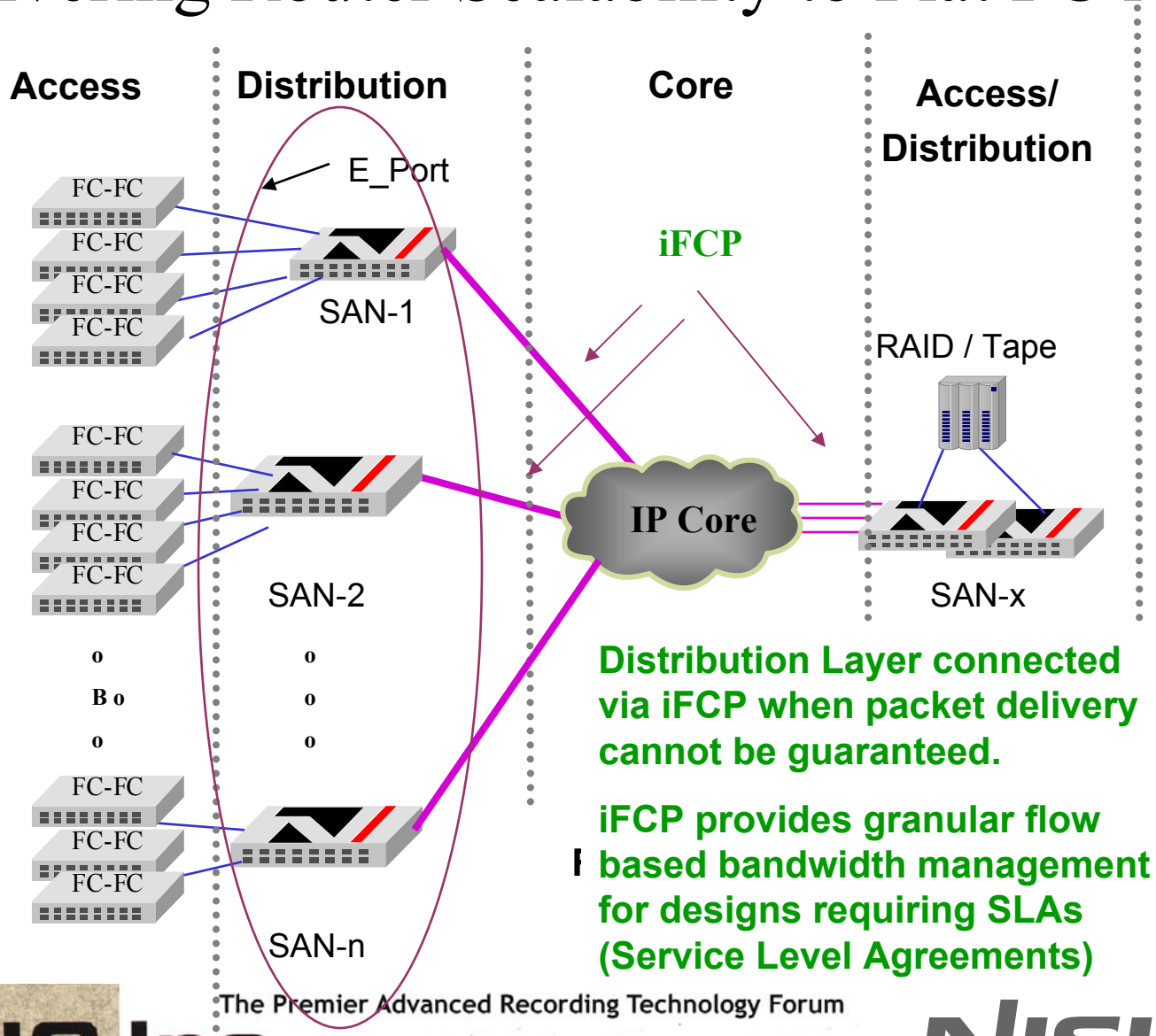
  Chattering HBA

Delivering Router Scalability to Flat FC Networks

Nishan E_Port provides Isolation between Fibre Channel Switches



Delivering Router Scalability to Flat FC Networks

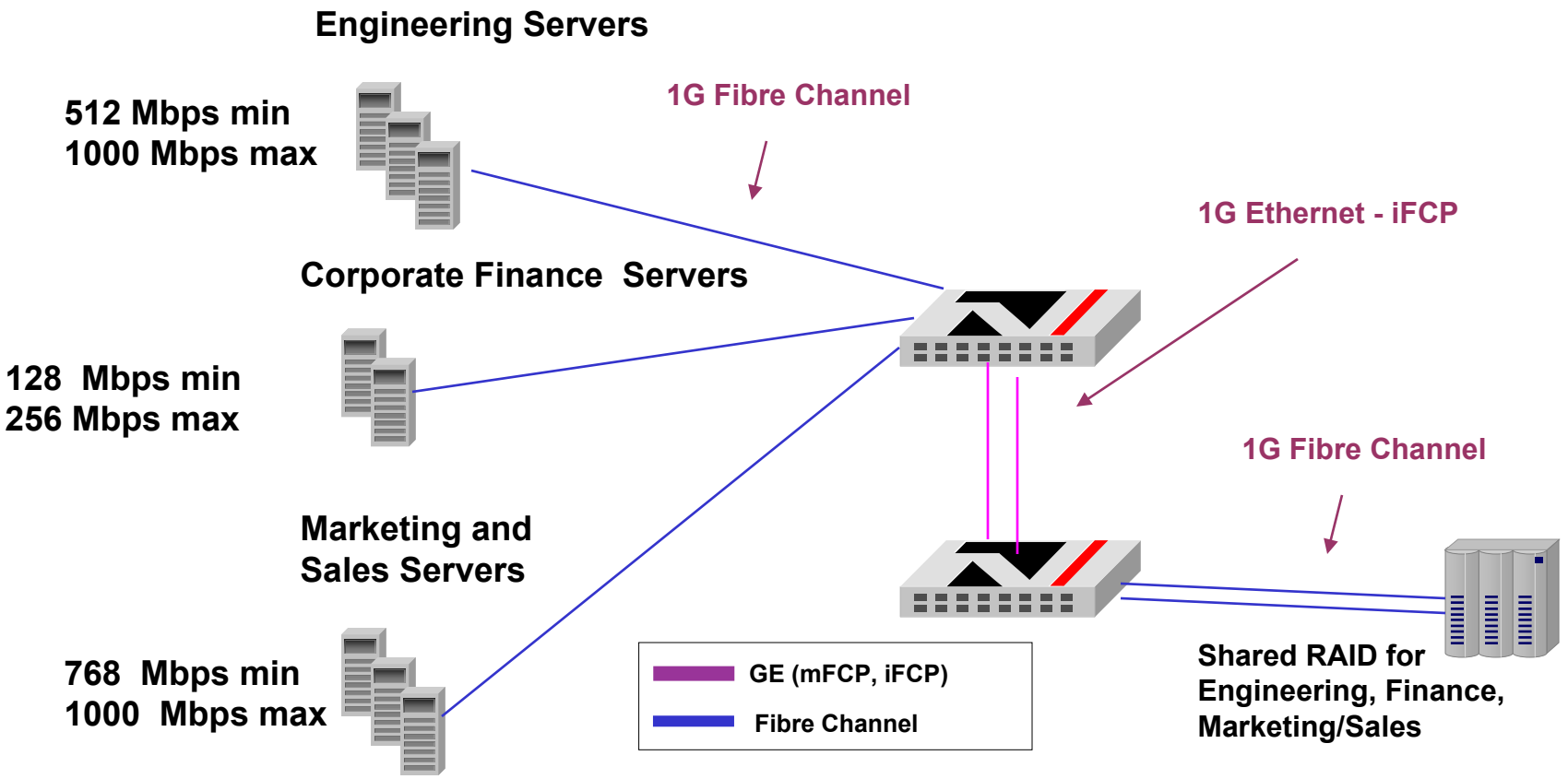


Distribution Layer connected via iFCP when packet delivery cannot be guaranteed.

iFCP provides granular flow based bandwidth management for designs requiring SLAs (Service Level Agreements)

Bandwidth Management - QoS

QoS to ensure deterministic behavior



Nishan Industry Firsts

- **First to provide SAN Routing**
  http://www.nwfusion.com/news/tech/2003/0217techupdate.html
  http://www.nwfusion.com/news/tech/2003/0714techupdate.html

- **August 2001: Transcontinental 215 mega bytes per second (Newark, NJ to Sunnyvale, CA) – Promontory project**
  http://www.nishansystems.com/techlib/techlib_papers.html
- **January 2002: First to validate wire-speed iSCSI – 219 mega bytes per second**
  http://www.nishansystems.com/iscsi/
- **November 2002 : San Diego Super Computer Center demonstrates 721 mega byte connectivity from La Jolla, CA to Baltimore**
  http://www.nishansystems.com/products/prod_downloads/SCDC_SupercomputingDemo.pdf
- **2002-2003: Carlson Companies deploys world first data center IP SAN**
  http://www.nishansystems.com/products/prod_downloads/CarlsonCaseStudy.pdf

Thank You