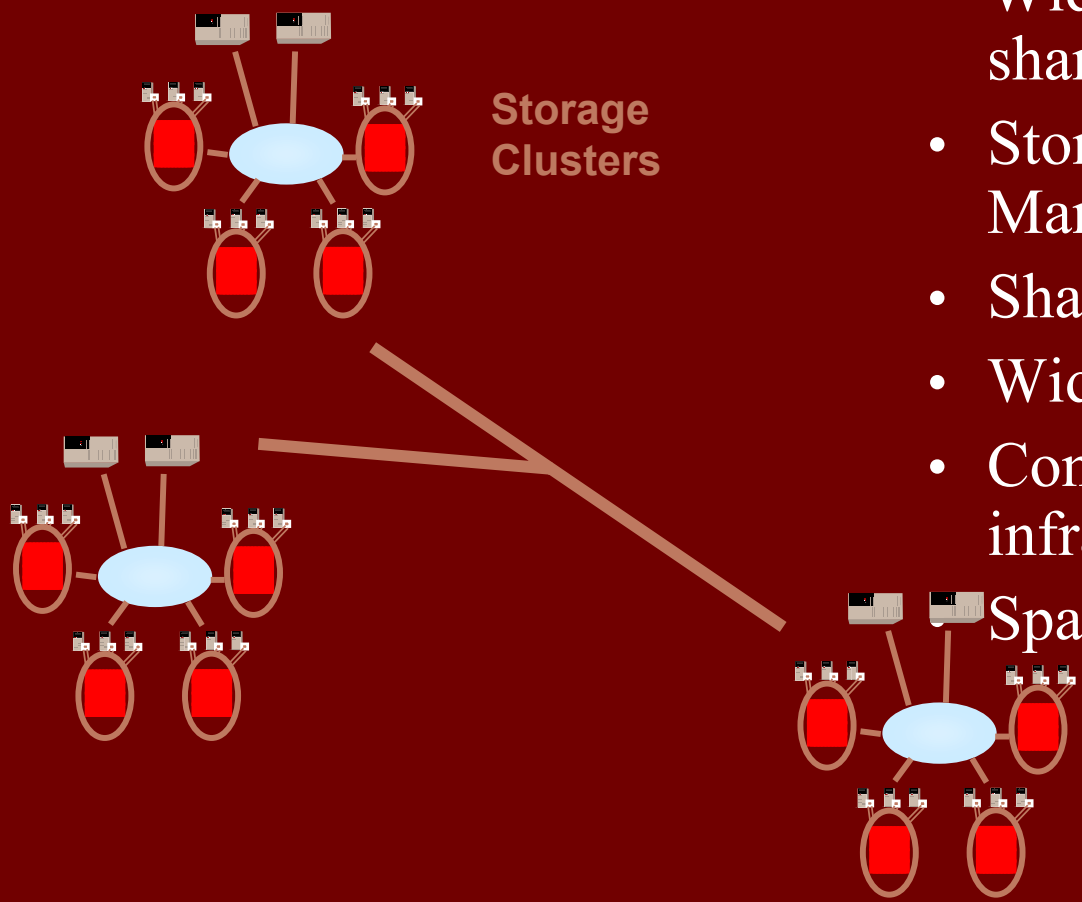
Storage Wide Area Network Design, Implementation and Performance Analysis

- Jim Gursha, President
- High Performance System Solutions, Inc.
  - 1735 York Avenue, 32H
  - New York, NY 10128

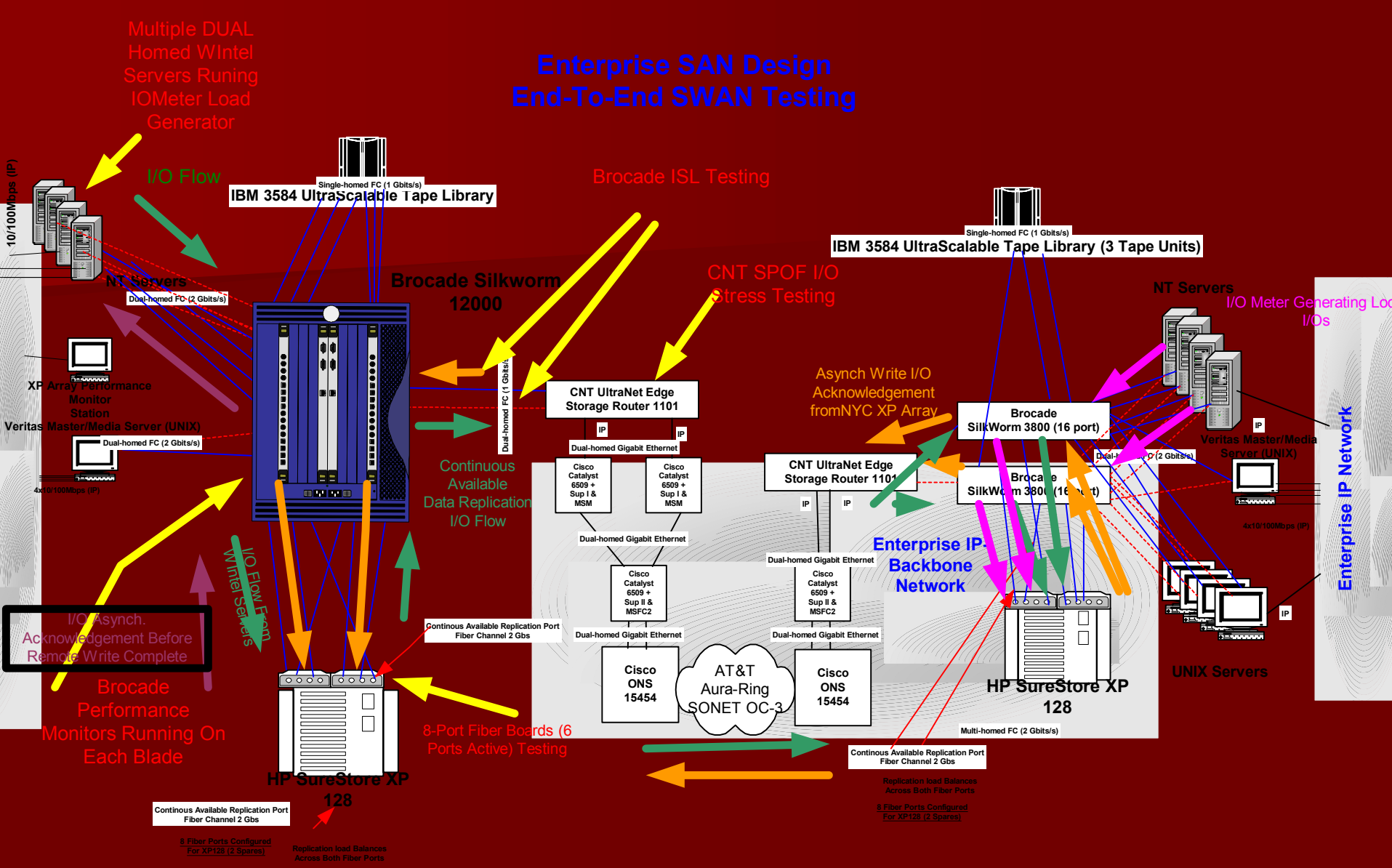
- HP Interex West Symposium
  Session # 3165 March 2004

HPSS
1735 York Avenue, 32H
New York, NY 10128
212-831-0291/917-359-2087
jimgursha@sandisks.com

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 standard infrastructure (Network)
  - Span: The World



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.

Host Bus Adapters

- Individual Component Utilization Is A Necessary Part SAN Architecture and Implementation.
- Selecting the Right Will Lower Overall SAN Costs.

Host Bus Adapters

- Connect the Server to the SAN.
- Alleviate the Server From Some I/O Processing.
- Typically, Assist in the Execution of Parts of Communications Protocol.
- Compatibility Across HBA's.

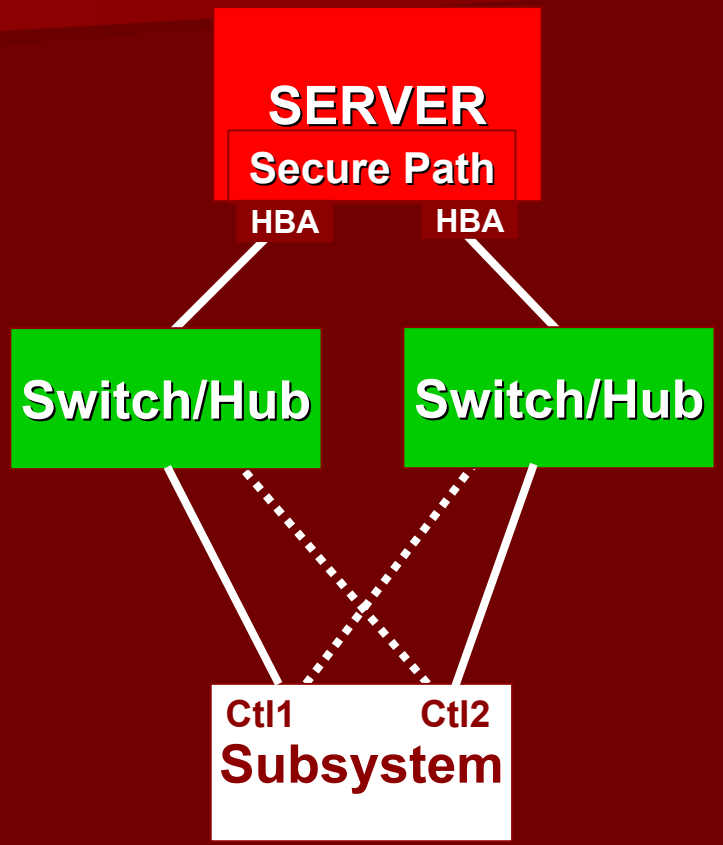
Emulex HBA's

- Dual Channel (LP9402DC)
  - High performance through the use of two Emulex Centaur ASICs,
  - Two 266MIPS onboard processors,high speed buffer memory.
  - Automatic speed negotiation capability which allows complete compatibility between 1GBS and 2 GBS.
  - switched fabric support,full-duplex data transfers.
  - high data integrity features,support for all Fibre Channel topologies
  - dual channel HBA. Channels deliver up to 800MB/s link bandwidth

EMULEX HBA's

- Single Channel (LP9802)
  - full duplex 2Gb/s Fibre Channel delivering up to 400MB/s
  - automatic speed negotiation
  - automatic topology detection
  - onboard hardware context cache for superior fabric performance
  - support for use of multiple concurrent protocols (SCSI and IP)
  - support for FC-Tape (FCP-2)devices
  - 66/100/133 MHz PCI-X 1.0a and PCI 2.2 compatibility
  - Buffered data architecture to support over 50km cabling at full 2Gb/s bandwidth
  - Windows 2000,Windows NT,HP-UX,Linux,NetWare,Solaris and AIX

HP *StorageWorks*™ Secure Path



Secure Path is Multi-path software

Benefits:

- Eliminates path as single point of failure
- Higher performance
- Static or dynamic I/O balancing
- Path failure detection

When Used:

- When highest availability needed
- When highest performance needed

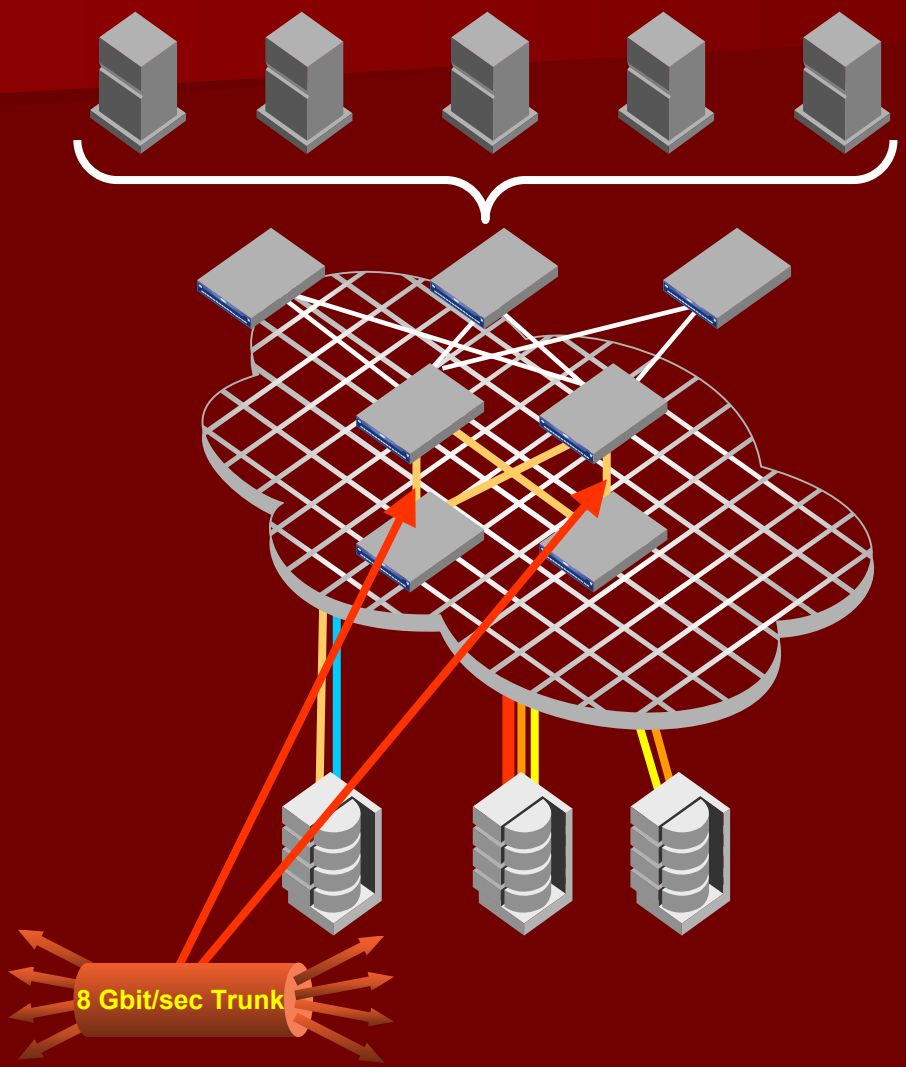
SilkWorm 12000 Core Fabric Switch

- Flexible, modular architecture
  - Scalable 64/128 port design
  - 2 Gbit/sec ports; Autosensing
  - 3rd generation Brocade ASIC
- 99.999% availability
  - Redundant, hot-swap elements
  - Non-disruptive software updates
  - Redundant 64 port switch config
- Intelligent fabric services
  - Interswitch link trunking
  - Frame filtering
  - Global performance analysis
- Multi-protocol architecture
  - 10 Gbit/sec fibre channel
  - IP storage interconnect
  - InfiniBand

Brocade Silkworm 12000

- Verify Hardware Components MTBF
  - Switch backplane 43,328,000 hours
  - Power backplane 10,722,000 hours
  - Control processor 177,000 hours
  - 16 port FC blade 153,000 hours
  - Power supply 500,000 hours
  - Blower FRU 473,000 hours
  - WWN card 2,153,000 hours

New Advanced Fabric Services: Inter-Switch Link Trunking

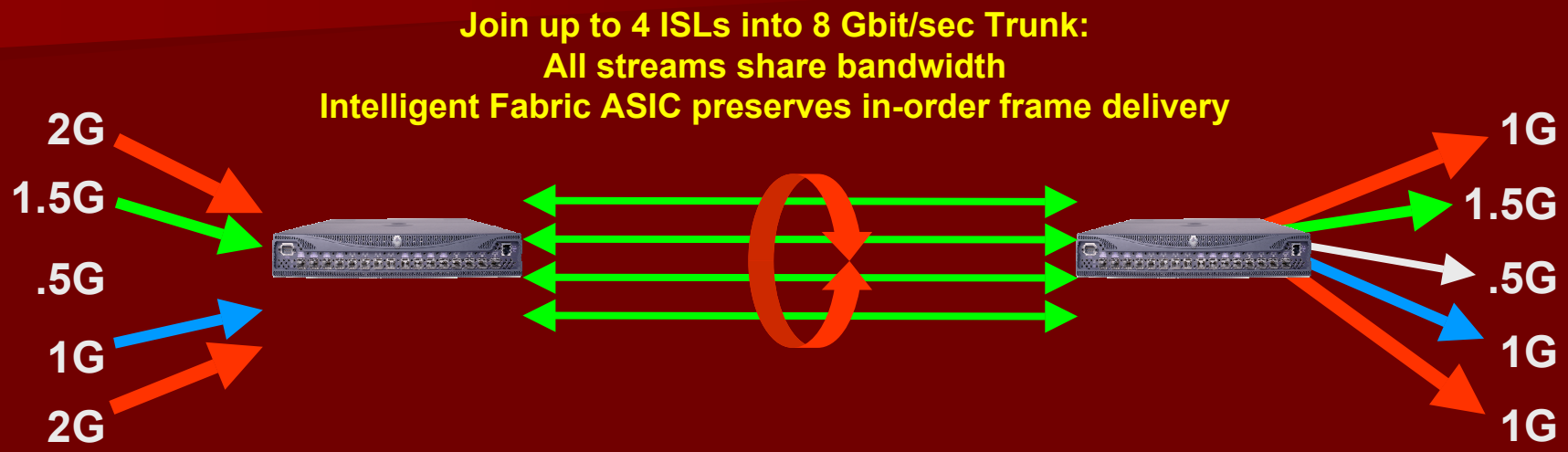


- ISL Trunking
  - 8 Gbit/sec logical links
  - Aggregate edge traffic
  - Zero management
- Simplify network design and management
- Simplifies ongoing administration (manage one link versus four links)
- Maximizes fabric performance
- Provides increased high availability in case of link failures

Intelligent Bandwidth Utilization: Dynamic Load Sharing

- Load sharing across multiple ISL links
- Round robin assignment
- Can get “unlucky” with multiple high utilization traffic assigned to same link
- In our example, theoretical maximum is 8 Gbit/sec, but effective throughput is 5 Gbit/sec

Intelligent Bandwidth Utilization: Inter-switch Link Trunking



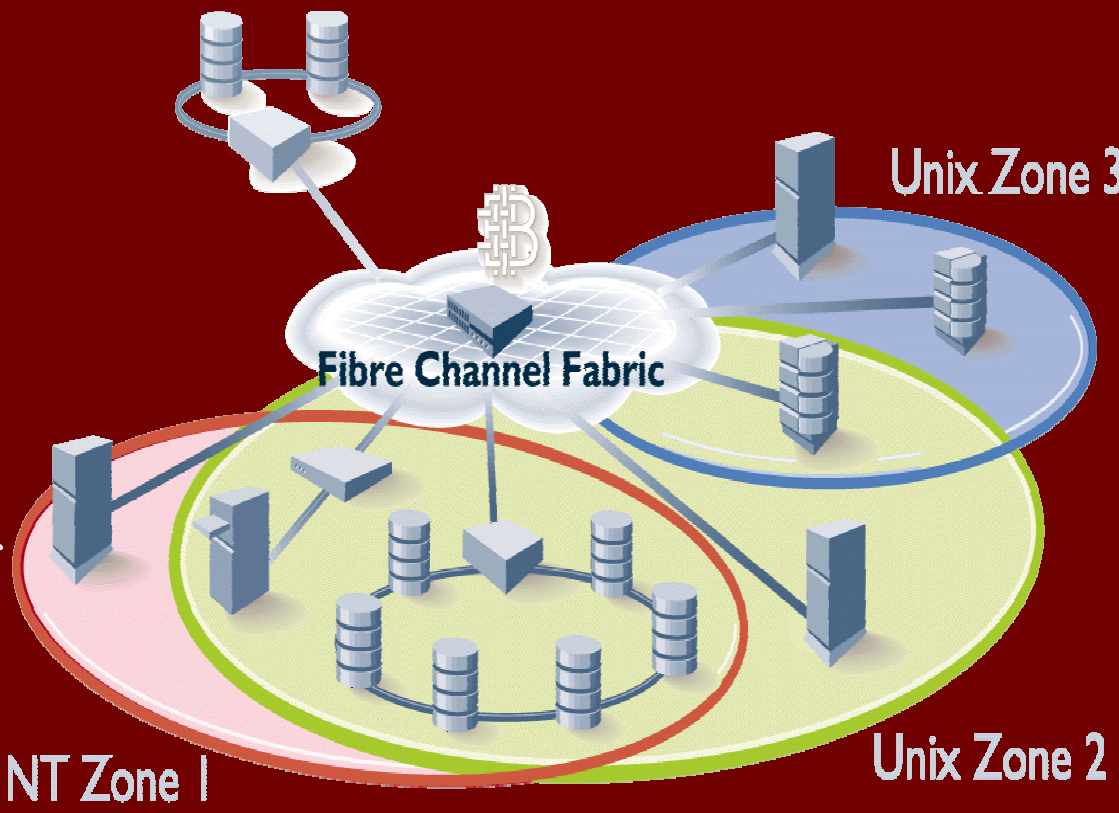
- Aggregate traffic onto fewer logical links
- Automatically created when switches are connected
- Managed as a single logical 8 Gbit/sec ISL
- Fault-tolerant – will withstand failure of individual links
- Supports redundant trunks between switches

Zoning Concepts

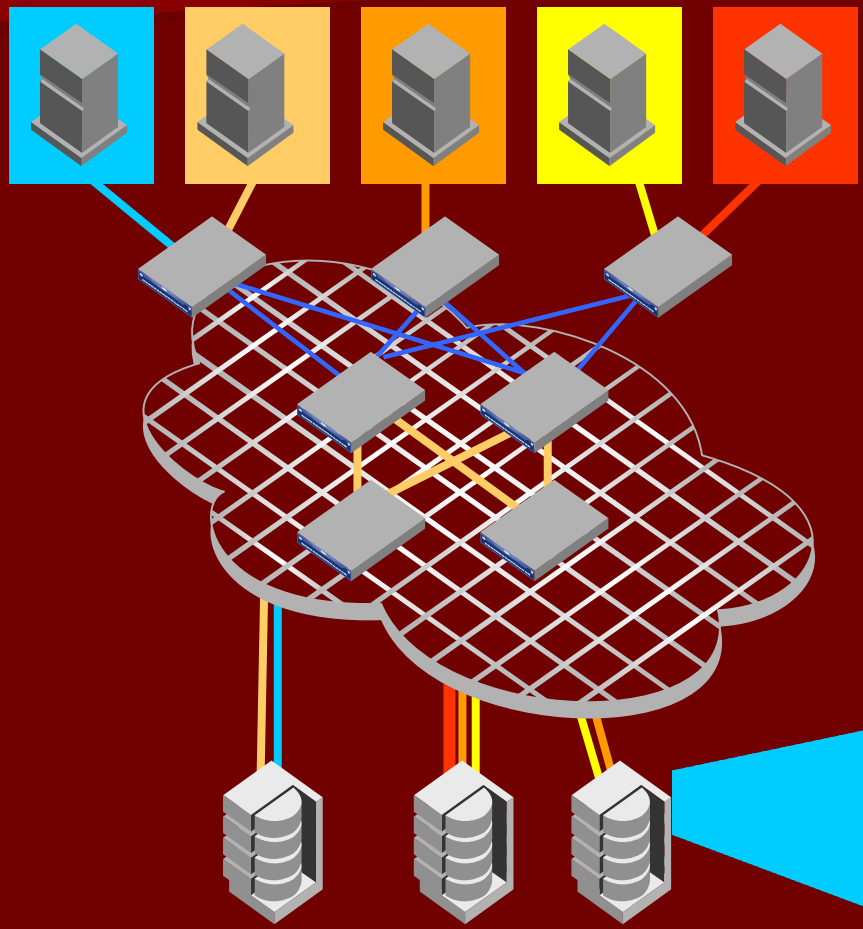
- Controls host access to logical devices connected to a fibre channel port via World Wide Name assignment.
- Ideal for multi-NT servers and SAN customers
  - Data protection in a multi-host environment
  - Prevent unauthorized access to LUN

Zoning

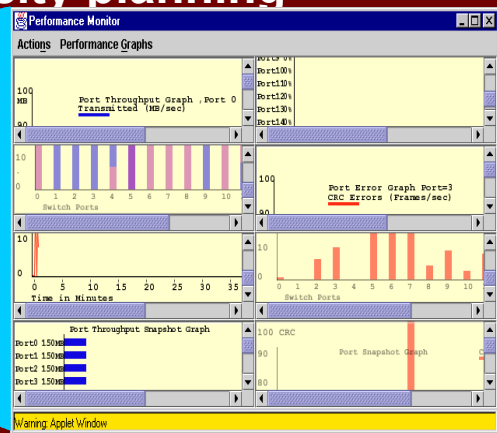
- Provides OS & storage isolation
- Store multiple zone configurations
- Zones based on port # or device WWN
- Updates distributed dynamically across the fabric
- Overlapping zones allowed
- No logical limit on the # of zones



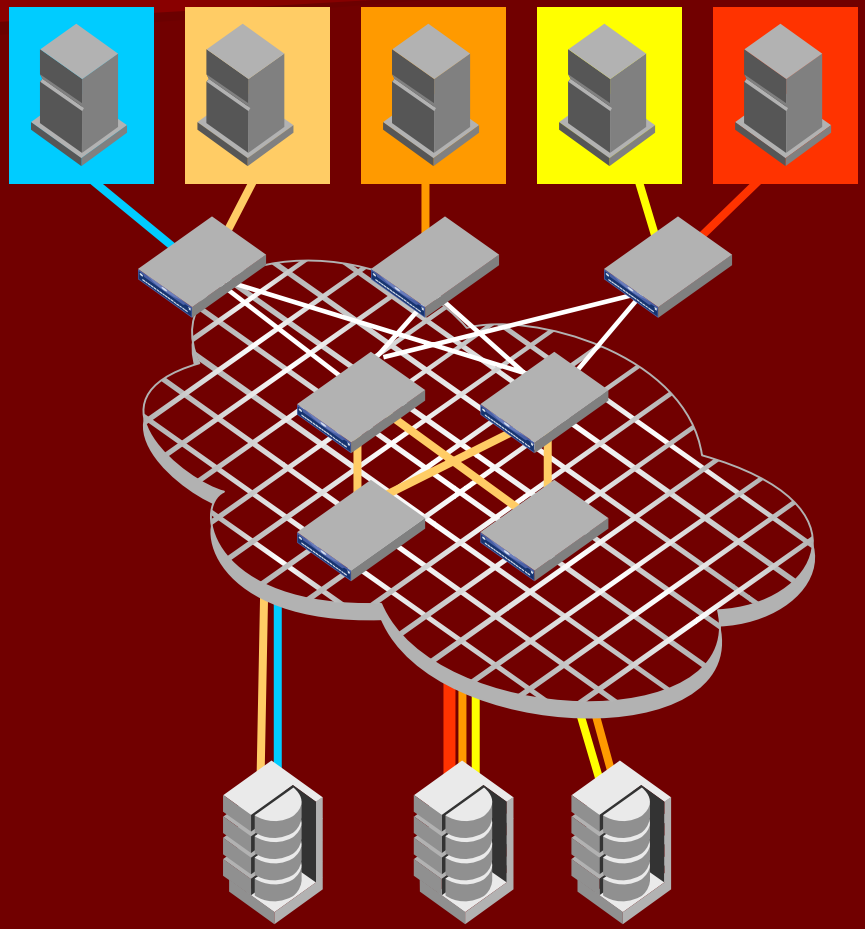
Advanced Fabric Services: New Global Performance Analysis



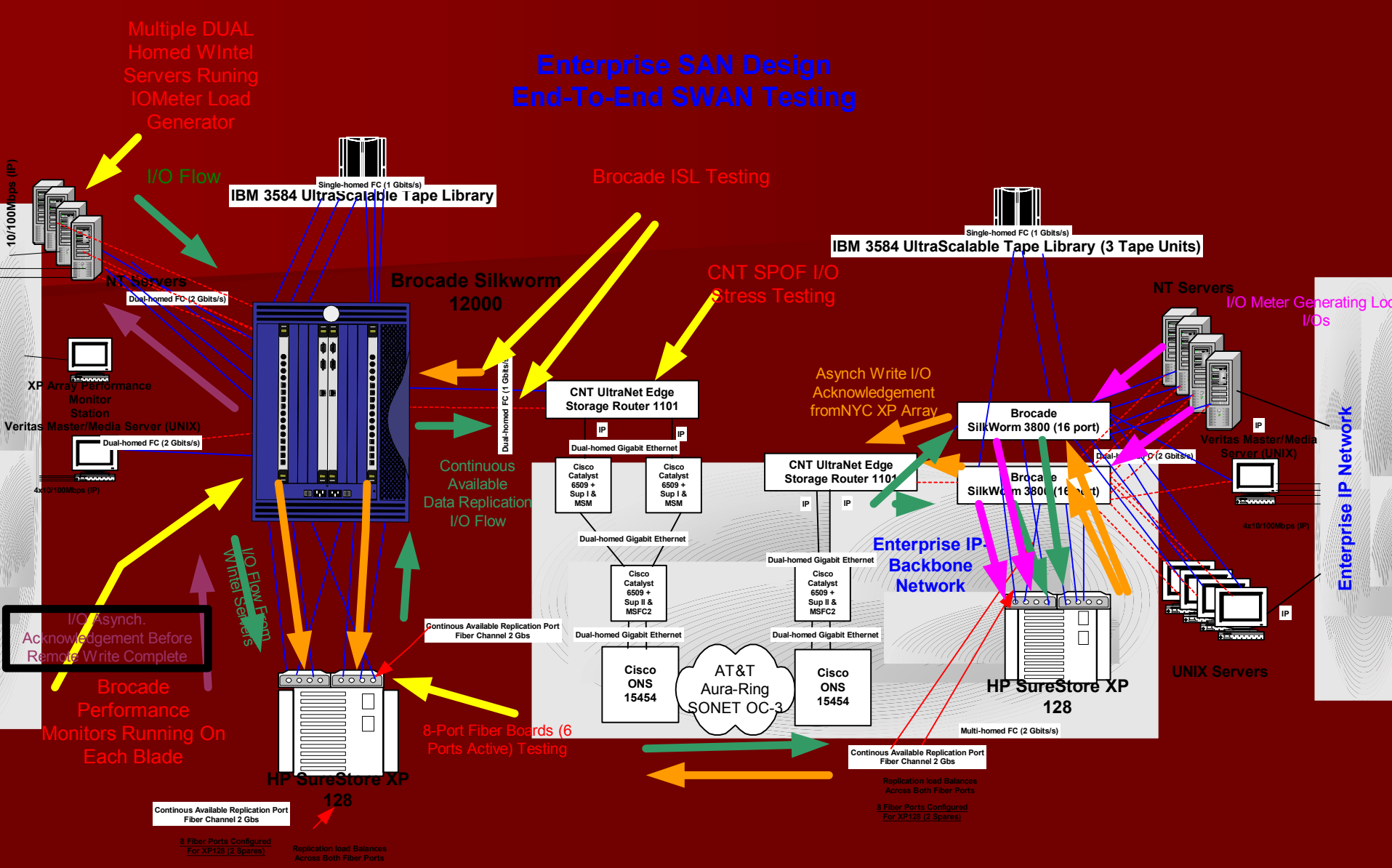
- Performance analysis
- Proactive SAN management
- Provides end-to-end performance measurement from the source to the destination target
- Optimize fabric resource allocation
- Maximizing performance tuning
- Reducing trouble-shooting time
- Improve capacity planning



New Advanced Fabric Services: Advanced Zoning



- WWN Zoning
  - Hardware enforced
  - Simple administration
  - Secure access control
- Provides a safer, more secure SAN network
- Hardware enforced access control
- Administrators can organize physical fabric into logical groups and prevent unauthorized access by devices outside of the zone

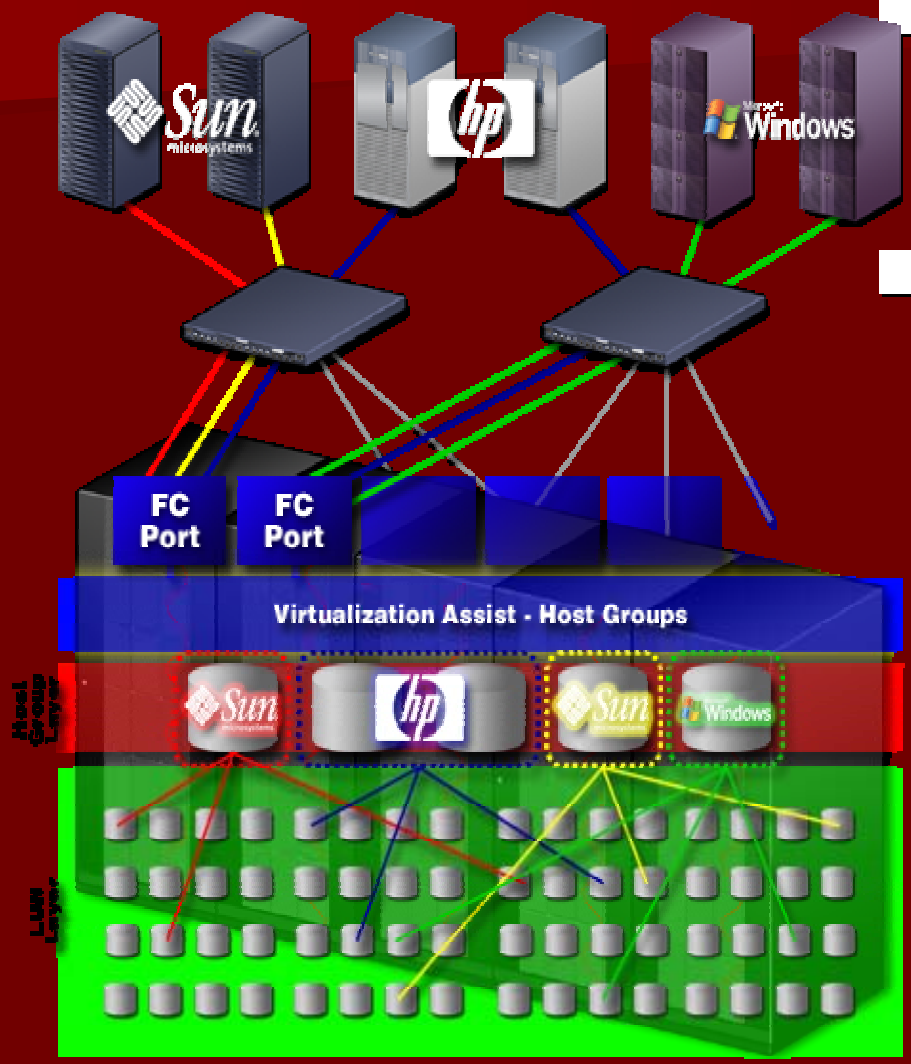


NOTES

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.

HP XP Series

"The Ultimate Consolidation Machine"

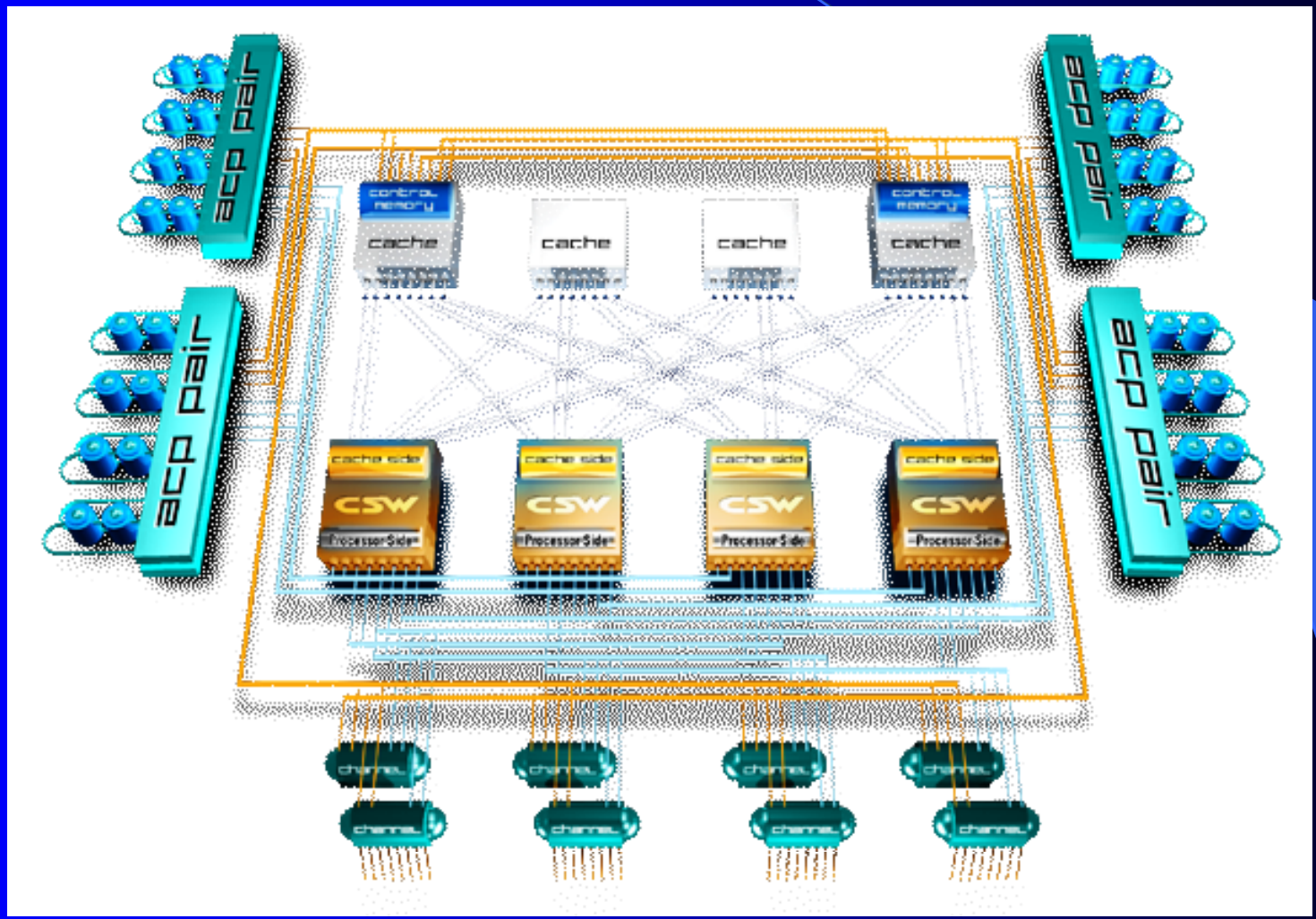


- Super consolidation has many tangible TCO benefits
- Super consolidation requires
  - Broad connectivity
  - Very high throughput
  - Large capacity
  - Capable management tools
    - Security, performance, allocation, availability

Lightning 9900 V Series Hardware Summary

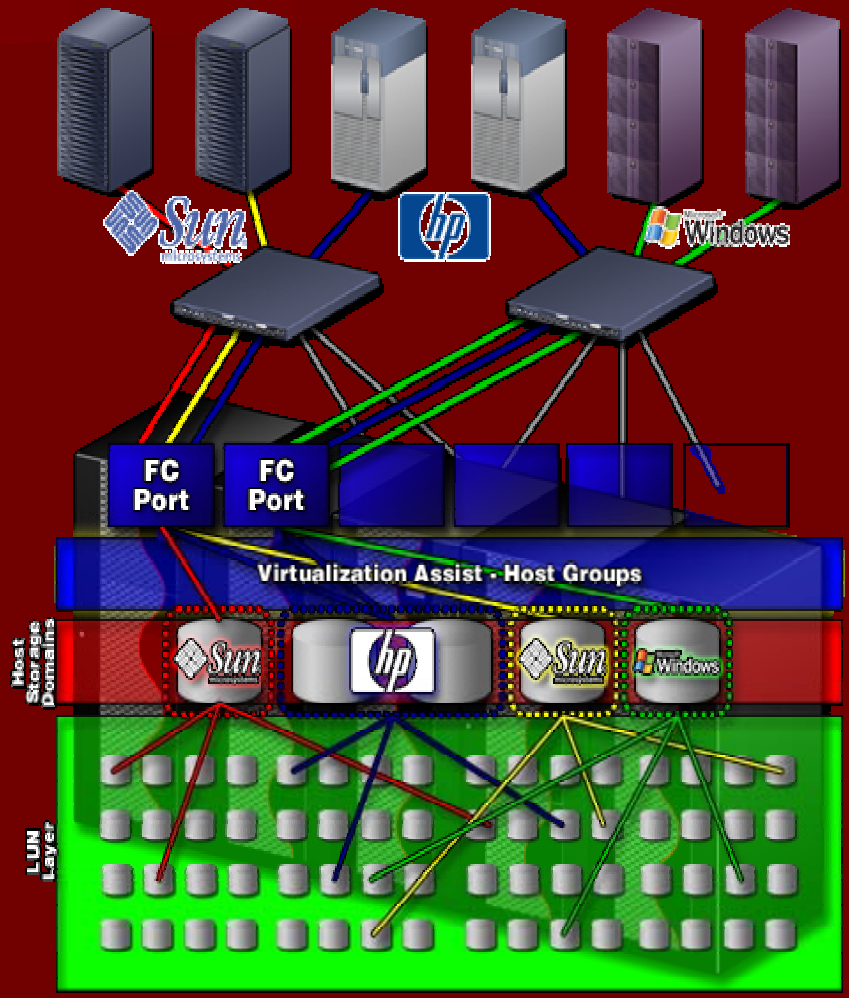
- **Lightning 9980V**
  - **32 ports**
    - **1 ACP std, 2nd, 3rd, 4th selectable**
  - **64GB cache**
  - **3GB control memory**
  - **8192 logical addresses**
  - **2+ throughput Lightning 9960**
  - **1024 HDD**
    - **36GB/73GB**
    - **75TB raw**
    - **Raid 5**
      - **3+1**
    - **Raid 1+**
      - **2+2, 4+4**
- **Lightning 9970V**
  - **24 ports**
    - **1 or 2 ACP selectable**
  - **32GB cache**
  - **3GB control memory**
  - **8192 logical addresses**
  - **Equal throughput to Lightning 9960**
  - **128 HDD**
    - **36GB/73GB**
    - **9TB raw**
    - **Raid 5**
      - **3+1**
    - **Raid 1+**
      - **2+2, 4+4**

XP Switch Architecture

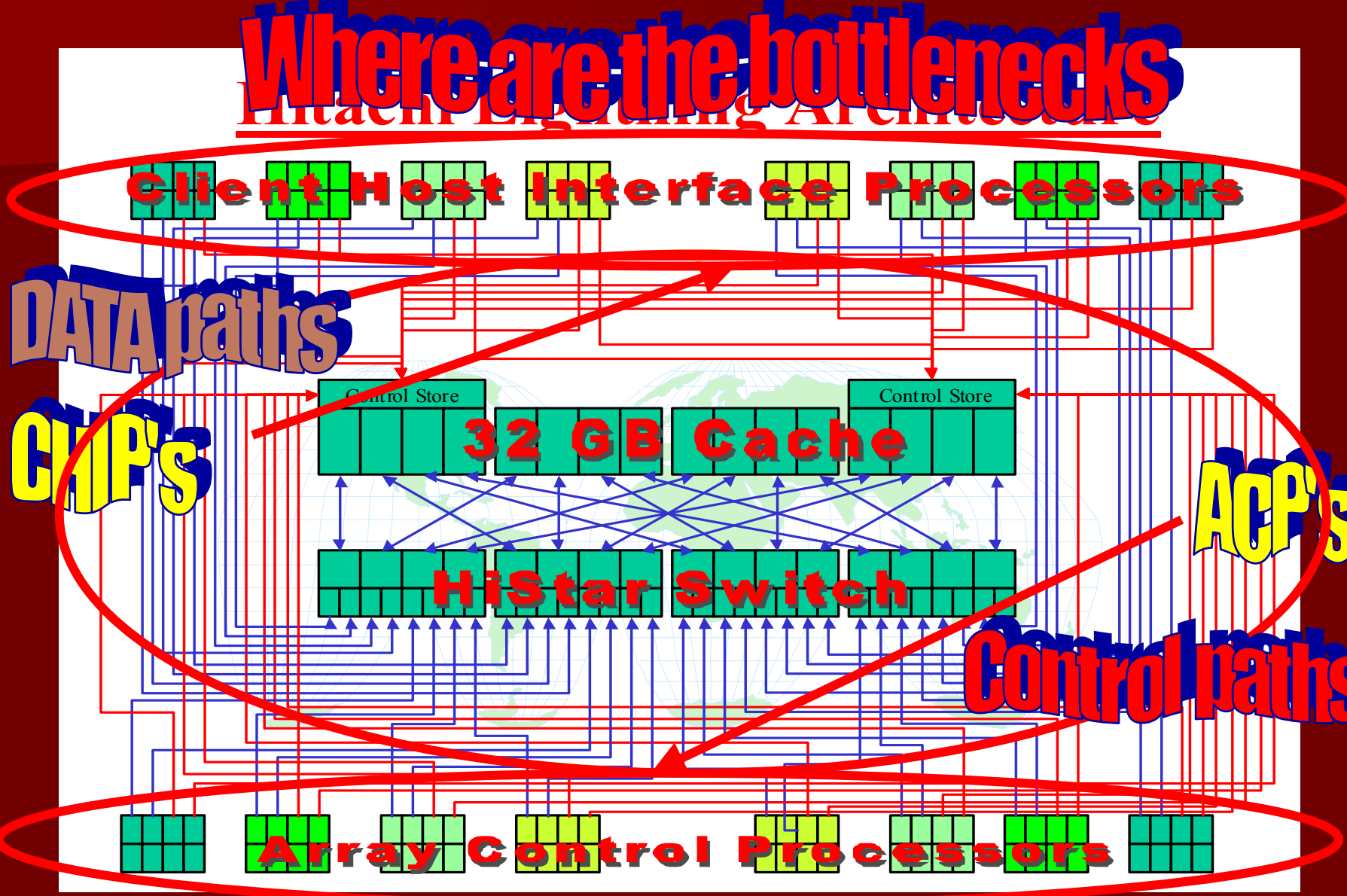


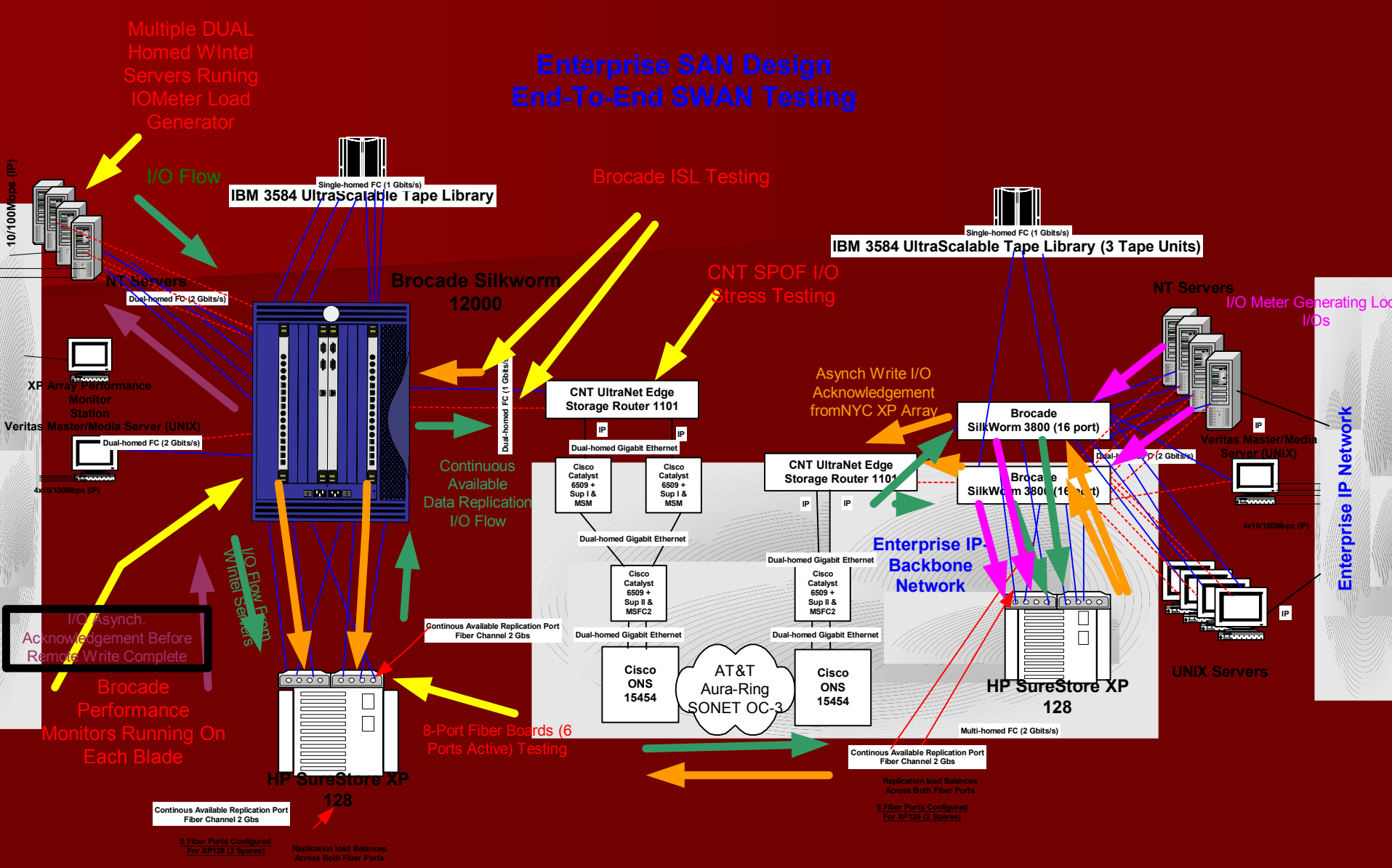
Virtualization Assist: Host Storage Domains

- Multiple Host Storage Domains can share same physical port.
- Each Host Storage Domain has its own logical FC port and its own independent set of LUNs.
  - Multiple LUN 0's
- Host connections routed to HSD based upon WWN.
- Fewer physical ports needed
  - Reduces complexity & cost
- More overall connections
- Enables consolidation



Overview

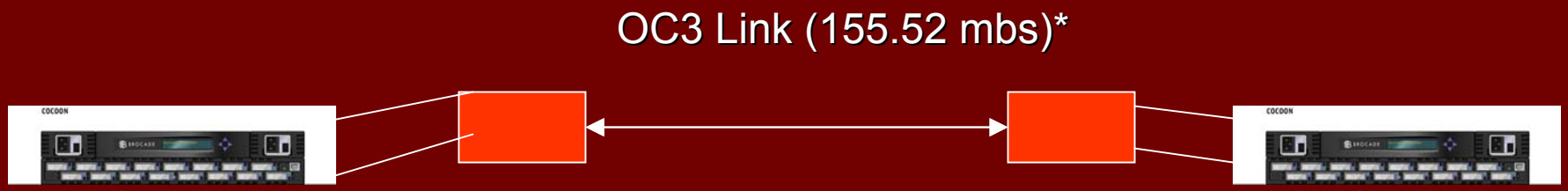

Where are the bottlenecks
Client Host Interface Processors
DATA paths
CHIP's
Control Store
Control Store
32 GB Cache
HiStar Switch
ACP's
Control paths
Array Control Processors



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.

Fabric Extension



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

Brocade Settings

- Fabric configuration need to have a separate and unique Domain ID. Example ID's in the Silkworm 12000 as 101 for the fabric switch named SAN1 and as 102 for the switch named SAN2

Brocade Settings (con't)

- License Key – each switch has the license key applied that was purchased (Web Tools, Zoning, Fabric, Fabric Watch, and Performance Monitor). In order to work with the CNT 1101 Edge Router, Remote Switch Licenses (RSL's) are required by CNT for each switch. (Two 16 port switches in the Silkworm 12000 and two 3800 switches for the remote site.)

Brocade CNT Recommended Settings

- CNT recommended specific settings within each Brocade switch as follows:
  - BB Credit – the number of available buffer credits available on a per port basis are recommended to be set at 16.
  - R_A_TOV – the resource allocation time out value, which is set in milliseconds, is recommended to be 10000 (the default).
  - E_D_TOV – the error detect time out value, which is set in milliseconds, is recommended to be 5000 (the maximium).
  - Data Field Size – maximum size value in bytes recommended to be 2048 (64 bytes less than the default).
  -
  -

  [

Brocade Silkworm Settings

- In Order Packet Delivery (IOD) – when this parameter is set the frames will be delivered in order. However since the configuration only consists of two single blades each describing a fabric, the setting of this parameter, according to Brocade, will not have a major impact during a topology change.Thus this parameter for this configuration is recommended to be disabled to avoid a conflict with the Dynamic Load Sharing parameter

Brocade Silkworm Settings

- ➢ Dynamic Load Sharing (DLS) – this parameter is enabled and traffic to a remote switch will be shared among the available paths. However this parameter setting conflicts with the In Order Packet Delivery parameter setting. If both parameters are enabled, some frames maybe lost during a fabric recalculation of the routes in the fabric to guarantee that frames are not delivered out of order.

Brocade Silkworm

- Recommended that each port be individually tested and that the Brocade 12000 port statistics and performance tools, which are accessible directly through the web tool GUI interface, be used to monitor the results.
- The following example illustrates the results of a recent problem resolution. The error log for the switch had indicated that thresholds were being exceeded for a specific port within the fabric, and that the switch status had gone to marginal and back to healthy due to the error rates.

Brocade Silkworm

Error 23

0x301 (fabos): Oct 28 16:32:47
Switch: 0, Warning FW-STATUS_SWITCH, 3, Switch status changed from Marginal/Warn
ing to HEALTHY/OK

Error 20

0x301 (fabos): Oct 28 16:30:45
Switch: 0, Warning FW-STATUS_SWITCH, 3, Switch status changed from HEALTHY/OK to
Marginal/Warning

Error 17

0x301 (fabos): Oct 28 16:23:00
Switch: 0, Warning FW-BELOW, 3, fopportState014 (FOP Port State Changes 14) is b
elow low boundary. current value : 0 Change(s)/minute. (normal)

Error 16

0x301 (fabos): Oct 28 16:23:00
Switch: 0, Warning FW-BELOW, 3, fopportState000 (FOP Port State Changes 0) is be
low low boundary. current value : 0 Change(s)/minute. (normal)

Error 15

0x301 (fabos): Oct 28 16:23:00
Switch: 0, Warning FW-BELOW, 3, fopportLink014 (FOP Port Link Failures 14) is be
low low boundary. current value : 0 Error(s)/minute. (normal)

Error 14

0x301 (fabos): Oct 28 16:23:00
Switch: 0, Warning FW-BELOW, 3, fopportLink000 (FOP Port Link Failures 0) is bel
ow low boundary. current value : 0 Error(s)/minute. (normal)

Error 13

0x301 (fabos): Oct 28 16:22:11
Switch: 0, Warning FW-ABOVE, 3, fopportState014 (FOP Port State Changes 14) is a
bove high boundary. current value : 12 Change(s)/minute. (faulty)

Brocade Silkworm

Researching the above errors in detail, the errors on the port were displayed by using the fabric command “porterrorshow” as follows:

```
                NY1SANX1LS1:admin> porterrshow
                frames  enc  crc  too  too  bad  enc disc link loss loss frjt fbsy
                   tx   rx   in  err shrt long  eof  out   c3 fail sync  sig
=============================================================================
                0:  574k 619k   0    0    0    0    0   39    0    4    0    1    0    0
                1:       0    0    0    0    0    0    0  1.0k   0    0    0    1    0    0
                2:       0    0    0    0    0    0    0  638    0    0    0    1    0    0
                3:       0    0    0    0    0    0    0  1.2k   0    0    0    1    0    0
                4:       0    0    0    0    0    0    0  3.8k   0    0    0    1    0    0
                5:       0    0    0    0    0    0    0  2.2k   0    0    0    1    0    0
                6:       0    0    0    0    0    0    0  825    0    0    0    1    0    0
                7:       0    0    0    0    0    0    0  6.0k   0    0    0    1    0    0
                8:       0    0    0    0    0    0    0  4.1k   0    0    0    1    0    0
                9:       0    0    0    0    0    0    0  2.3k   0    0    0    1    0    0
               10:       0    0    0    0    0    0    0  1.0k   0    0    0    1    0    0
               12:       0    0    0    0    0    0    0  423    0    0    0    1    0    0
               13:       0    0    0    0    0    0    0  3.8k   0    0    0    1    0    0
               14:615k 574k   0    0    0    0    0  1.1m   0    9   14    6    0    0
               15:       0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

Looking at port 14, under the “enc out” (encoding errors outside of frames) column, it is seen to have over one million errors. If the corresponding “crc err” (cyclic redundancy check) errors are low or zero, the problem may be a defective cable. Otherwise a corresponding moderate number of errors in the “crc err” column may indicate that the associated GBIC may be problematic.

Brocade Silkworm

A specific port show command after a fabric Telnet Login produced:

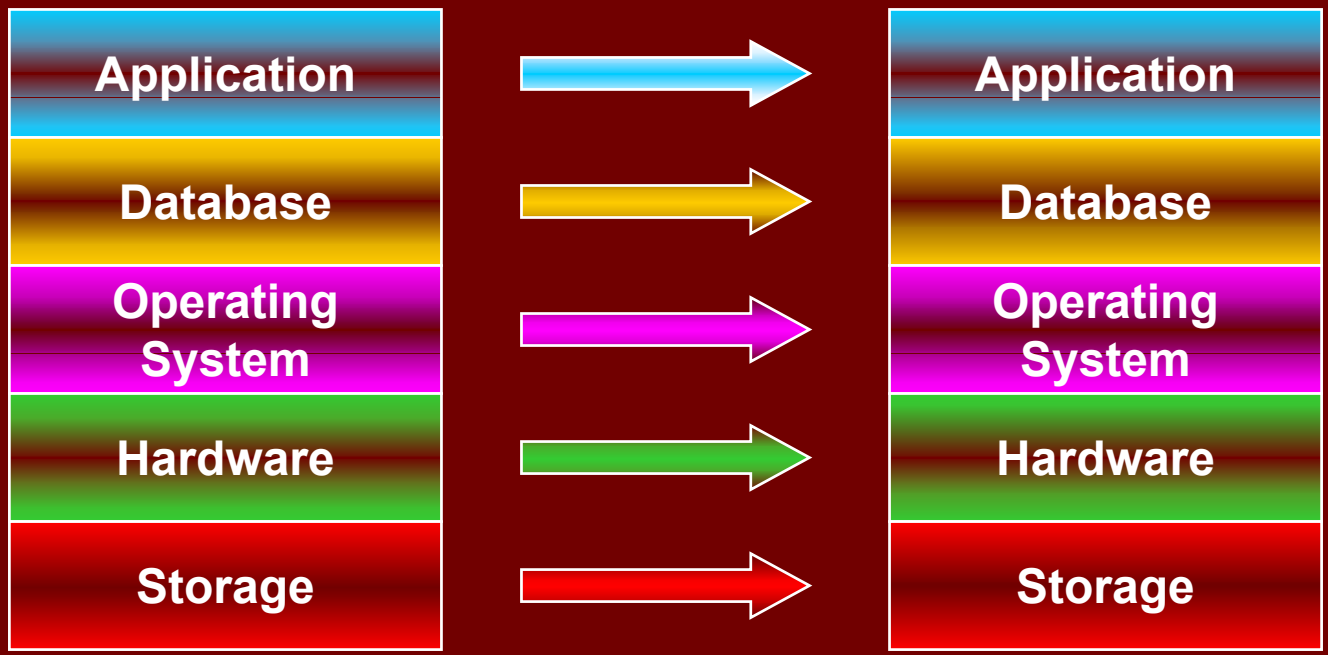
```
NY1SANX1LS1:admin> portshow 7/14
portCFlags: 0x1
portFlags: 0x23805b      PRESENT ACTIVE F_PORT G_PORT U_PORT LOGIN NOELP LED ACC
EPT
portType:  4.1
portState: 1    Online
portPhys:  6    In_Sync
portScn:   6    F_Port
portId:    660e00
portWwn:   20:0e:00:60:69:80:43:59
portWwn of device(s) connected:
10:00:00:00:c9:30:41:a2
Distance:  normal
portSpeed: N2Gbps
 Interrupts:        224       Link_failure: 9          Frjt:         0
Unknown:            27        Loss_of_sync: 14         Fbsy:         0
Lli:                77        Loss_of_sig:  6
Proc_rqrd:          127       Protocol_err: 0
Timed_out:          0         Invalid_word: 0
Rx_flushed:         0         Invalid_crc:  0
Tx_unavail:         0         Delim_err:    0
Free_buffer:        0         Address_err:  0
Overrun:            0         Lr_in:        8
Suspended:          0         Lr_out:       8
Parity_err:         0         Ols_in:       8
2_parity_err:       0         Ols_out:      0
CMI_bus_err:        0
```

Another possibility that may be causing excessive “Encoding Outside of Frame” is the HBA driver micro code version may not be compatible with the revision level that is supported by the vendor of the fabric switch and the disk array. For example HP only supports a driver version which is earlier than the latest released version for the Emulex HBA (9002L) that is currently being used. It is recommended that all microcode be the same revision level as the currently supported version to avoid potential problems.

Data Replication

- Replication can be done at many levels

- Replication can be done at many levels

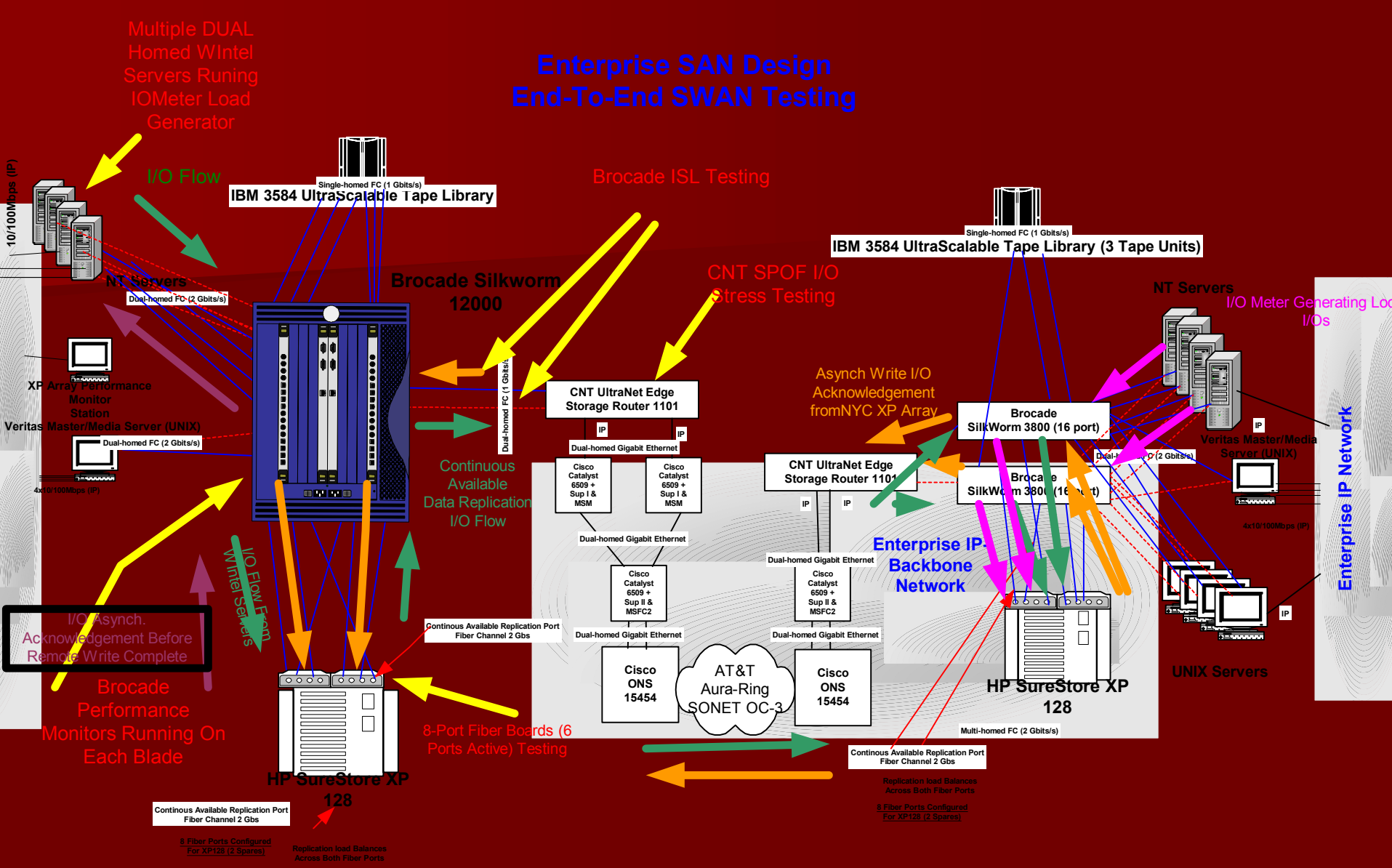


•Real Time COPY

- 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

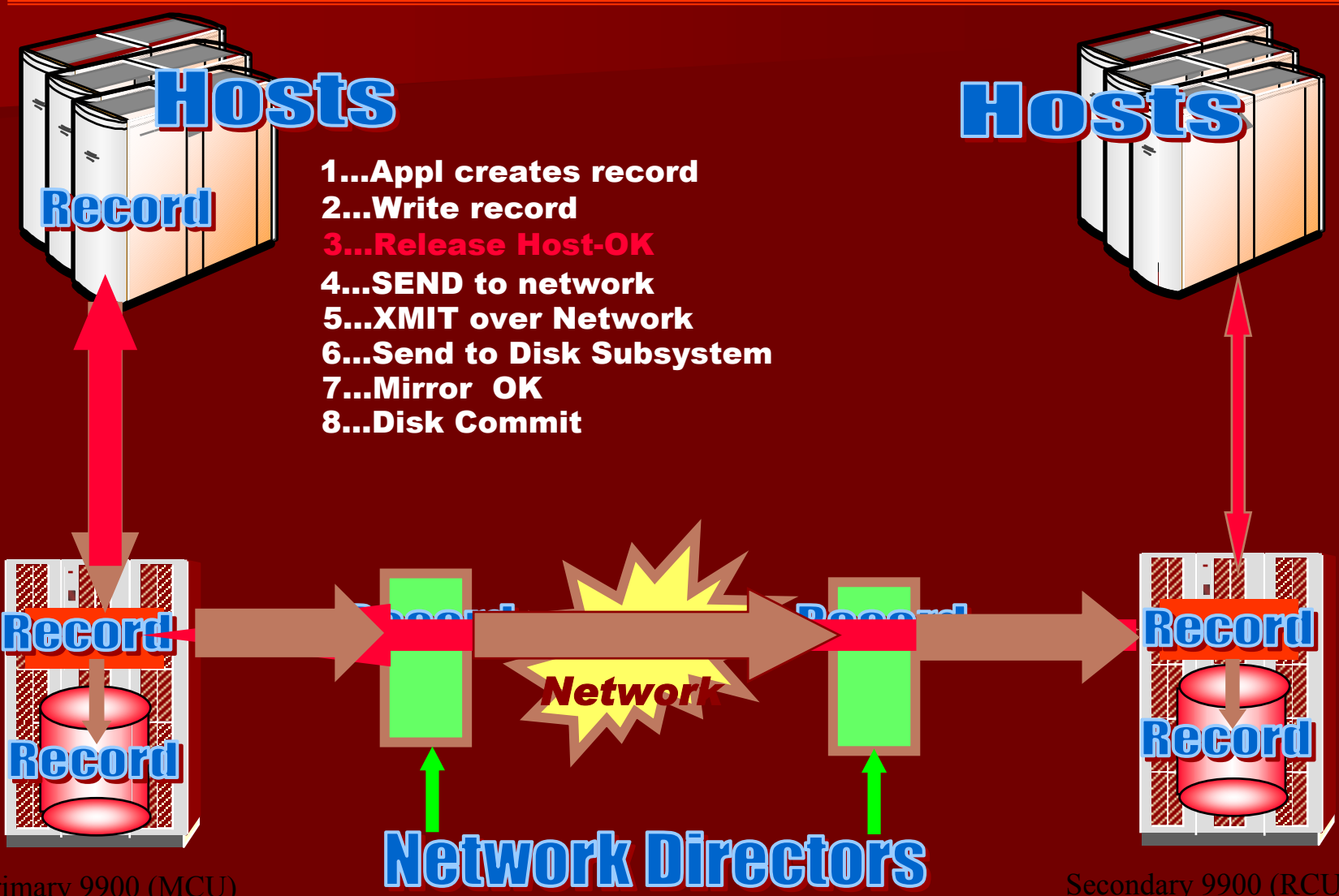


NOTES

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.

Asynchronous

Operations



Asynchronous

Update Sequence