High-Speed Data Transfer via HPSS using Striped Gigabit Ethernet Communications

_

Phil Andrews, Tom Sherwin, and Victor Hazlewood

San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
andrews@sdsc.edu, sherwint@sdsc.edu, victor@sdsc.edu

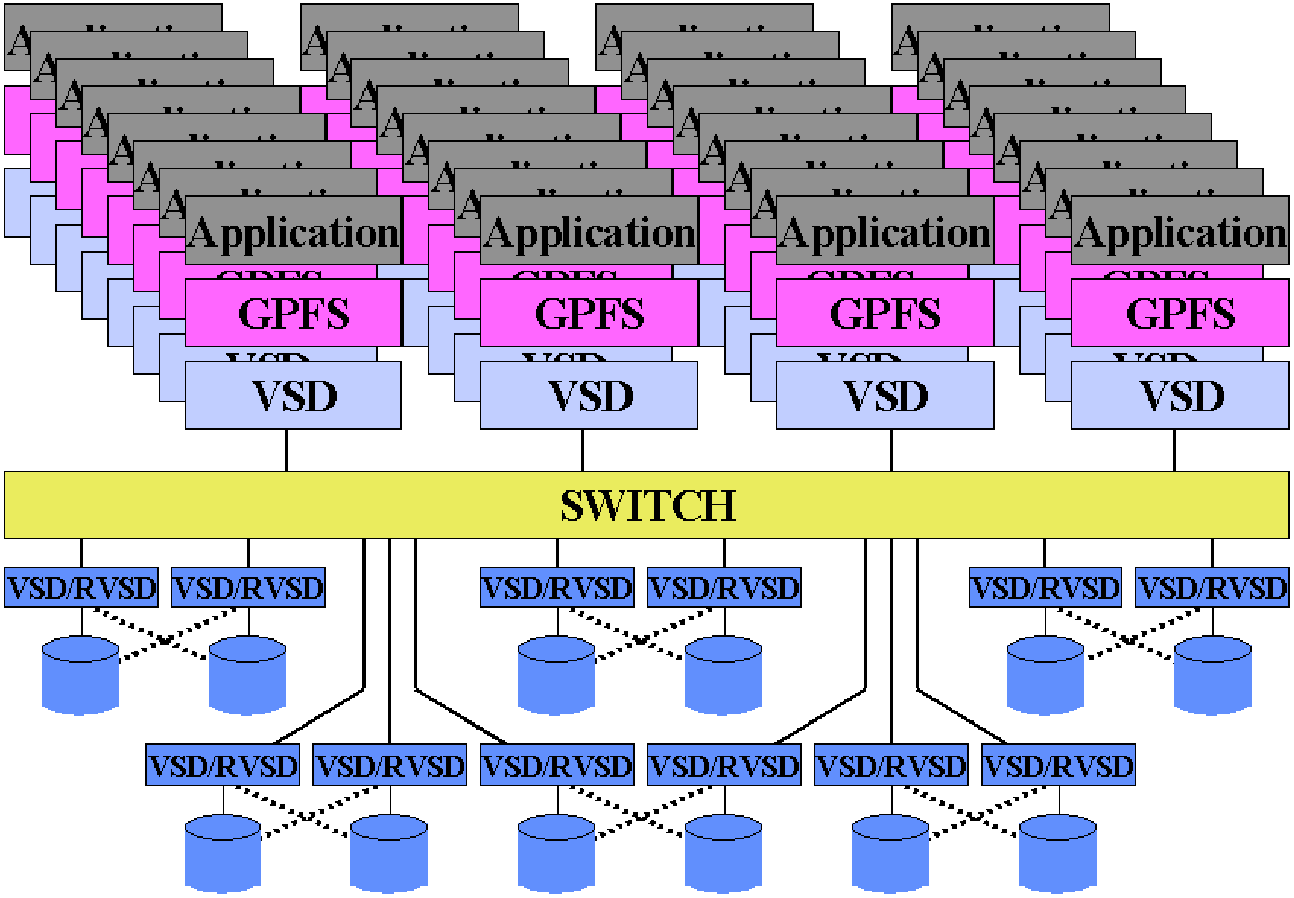
Why stripe?

- HPC no longer rules the roost: must use relatively cheap (good!), relatively slow(bad!) mass market components
- Remove badness via combining multiple slow components to make single fast one
- Disk stripes well accepted, RAID added failure mitigation, tape striping less common, RAIT and channel bonding are novel
- We used simple tape and network striping

What are we moving data between?

- **Blue Horizon: IBM SP with 144 8-processor 375 MHz NightHawk2 nodes.**
- **Floating point rating = 144 X 8 X 375 MHz X 4 flops/cycle = 1.7 Tflops. One of the NSF Supercomputer resources.**
- **HPSS: High Performance Storage System, >260 TB in store. Combination of robotic tape storage and disk cache**
- **Data moved across Gigabit Ethernet (GbE) to either HPSS tape drives or HPSS disk cache**

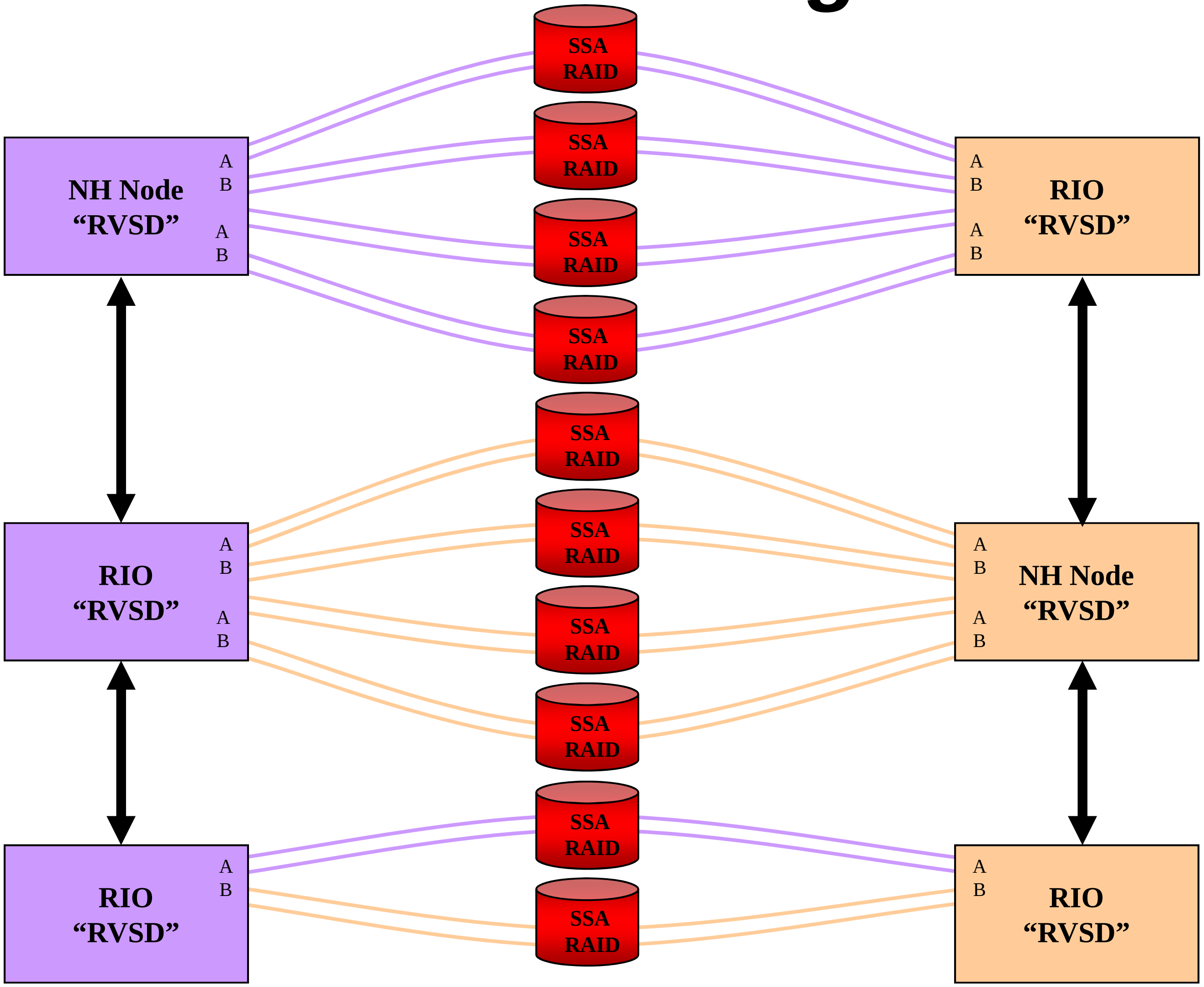
The GPFS file system layers for a 32-node application



RVSD Server Configuration

- **12 RVSD servers in the system in 6 redundant pairs (2-processor, 222 MHz NightHawk1 nodes)**
- **Each server drives 2 16-disk drawers of disk as primary with 2 additional during failover**
- **Each drawer configured as three 4+p RAID 5 arrays with a hot spare disk**
- **Total of 24 drawers of 18GB disks (384 disks, 288 spindles contribute) 5.2 TB net**

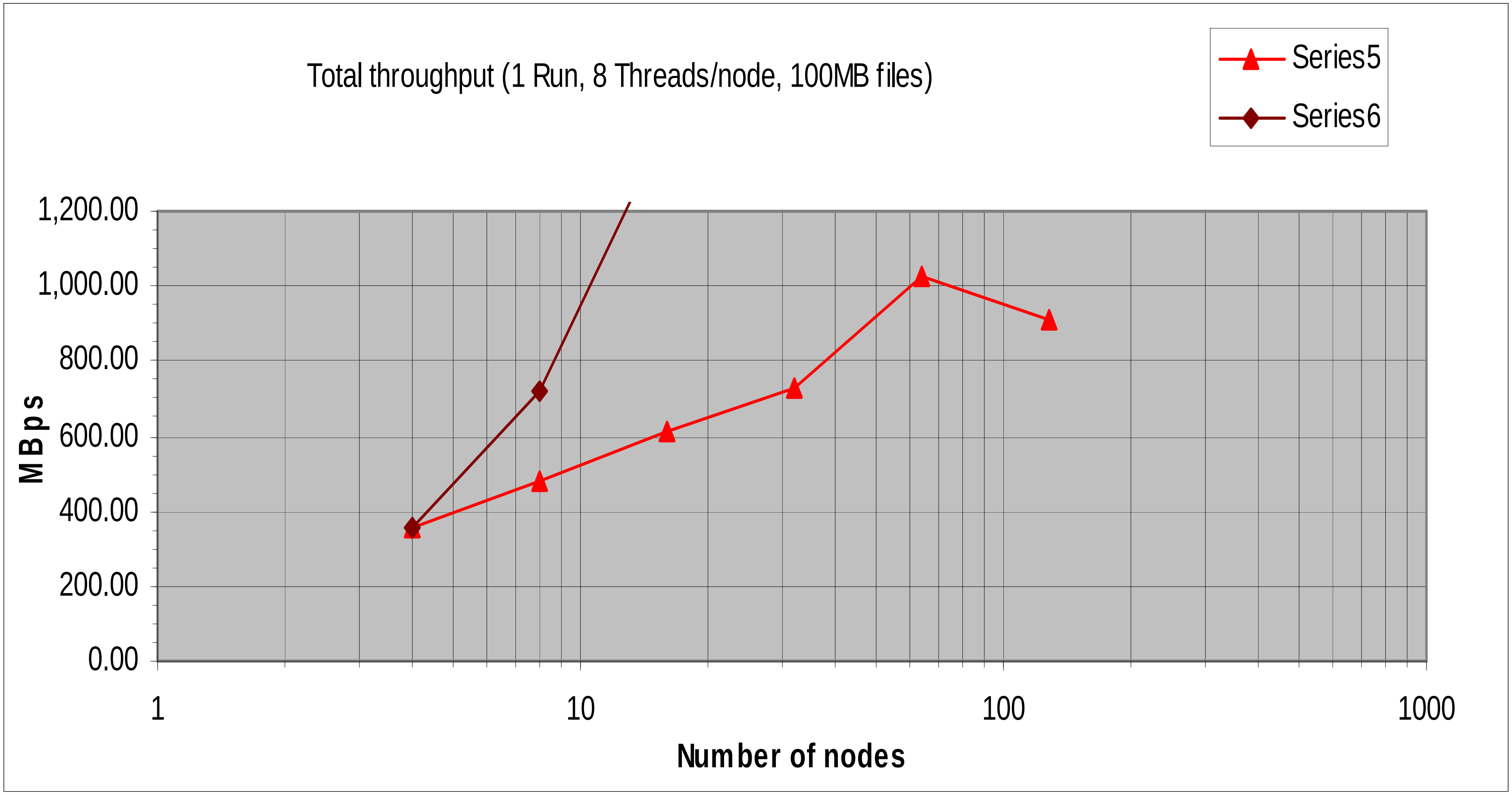
RVSD server configuration



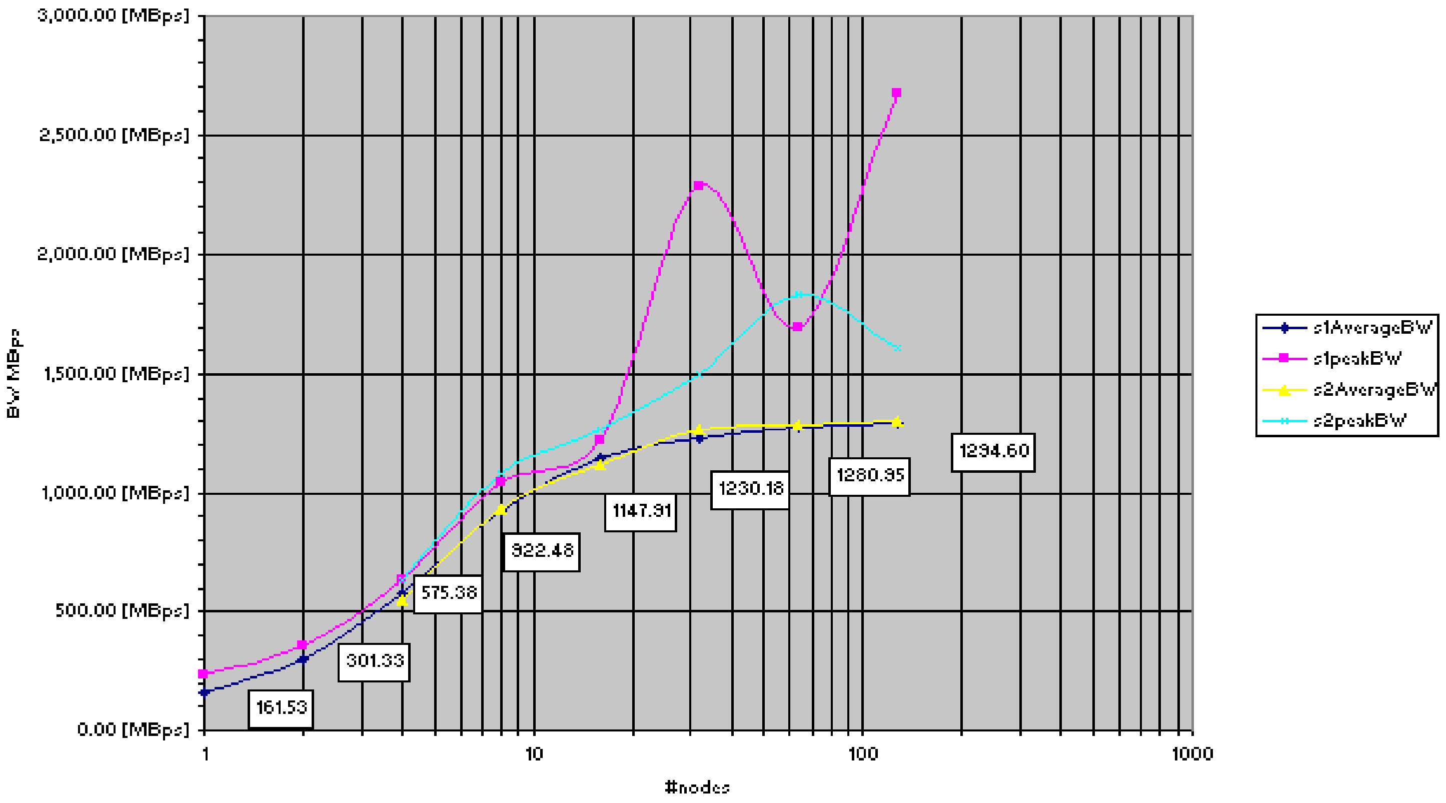
Reads from GPFS

- **Each file in GPFS is striped over every disk in the file system (256 KB block size)**
- **One 5.2 TB file system (to maximize spindle count)**
- **First runs at 8 threads per node, 4-128 nodes connected via Trailblazer switch (150 MB/s)**
- **Second runs were done with the Colony switch (450 MB/s), significantly better results**

GPFS reads with TBX switch



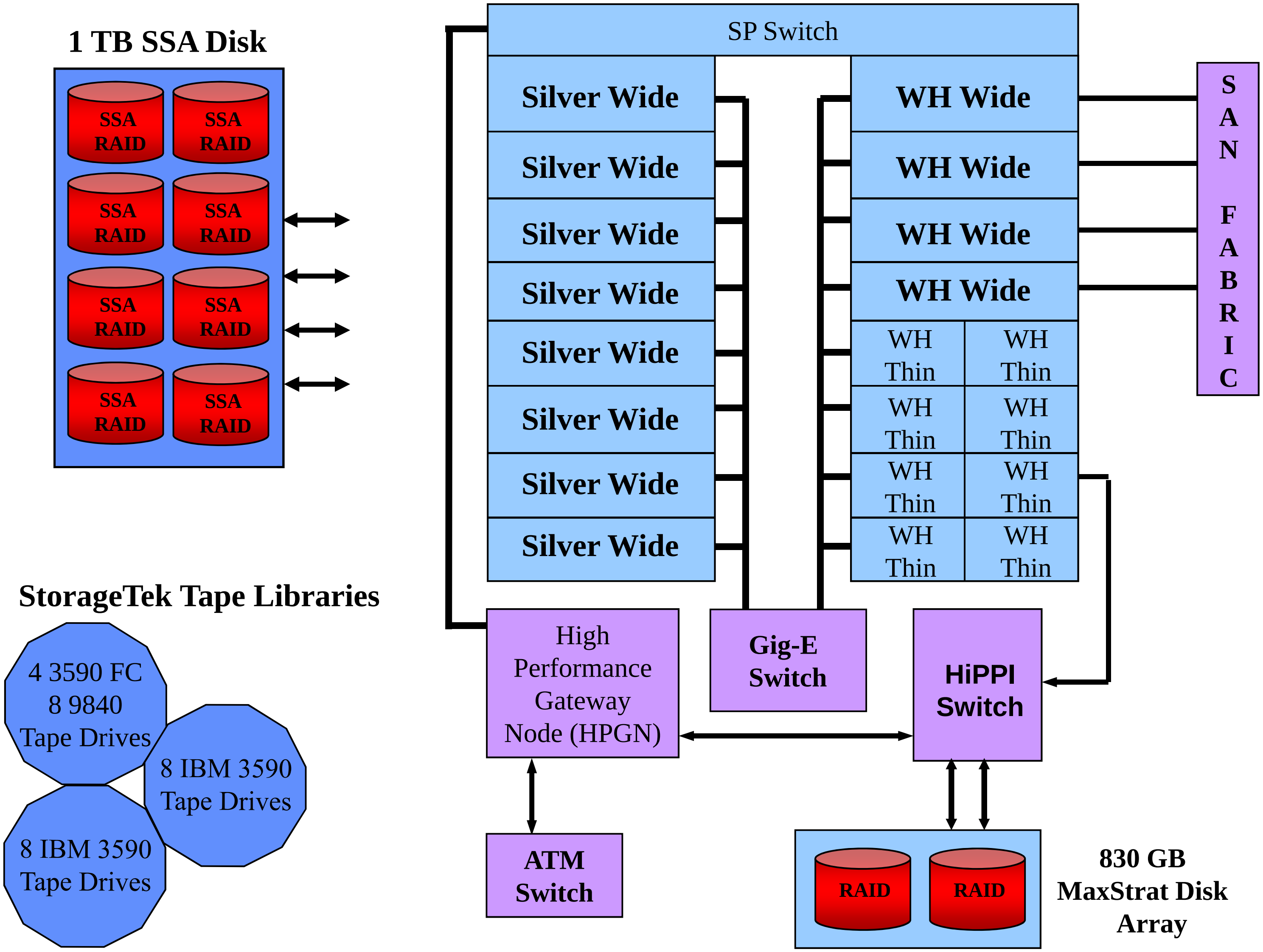
GPFS reads with Colony switch



HPSS Configuration

- SP system with 20 nodes; 8 silver wides, 4 Winterhawk wides, and 8 Winterhawk thins
- Direct network connectivity through Gigabit Ethernet, and HiPPI. ATM via HPGN.
- Close to 2 TB of disk cache as SSA raid or MaxStrat raid, 3TB of Fibrechannel T3 cache
- 28 tape drives; 20 IBM 3590E and 8 STK 9840
- Striping essential for required performance

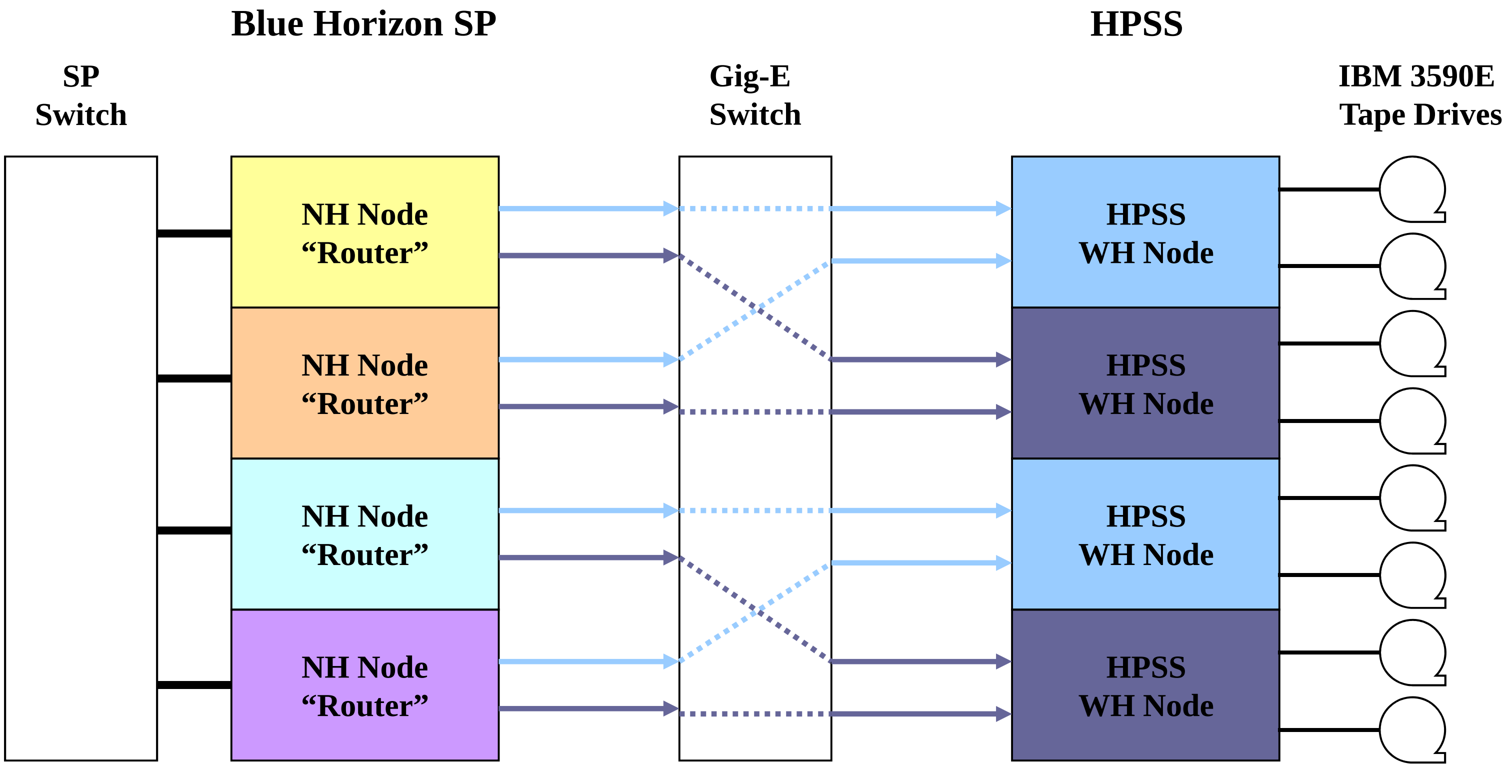
HPSS configuration



Striped Data Transfers to HPSS

- Machines are logically sub-netted at the router
- “Blue Horizon” SP is organized in 4 network ‘quadrants’
- HPSS servers divided across 2 networks
- “Router” nodes do network I/O to HPSS on behalf of remaining nodes in a quadrant.

8-way striped GbE transfers to HPSS



Transfers directly to HPSS tape

- **Used up to 8 IBM 3590E tape drives (14 MB/s nominal, compression always on)**
- **All striping of single data sources was by HPSS software (no parity)**
- **Transfers were by the HSI software interface which allows multiple parallel data streams**

Write performance to HPSS tape

Transfer type	Sparse file	Uncompressible file	Scientific data
One one-way	16.4 MB/s	11.4 MB/s	16.3 MB/s
One two-way	29 MB/s	23.7 MB/s	25.5 MB/s
Two two-ways	52.1 MB/s	45.5 MB/s	50.4 MB/s
Four two-ways	108 MB/s	89.6 MB/s	106.3 MB/s
One eight-way	36.6 MB/s	30.8 MB/s	31.7 MB/s

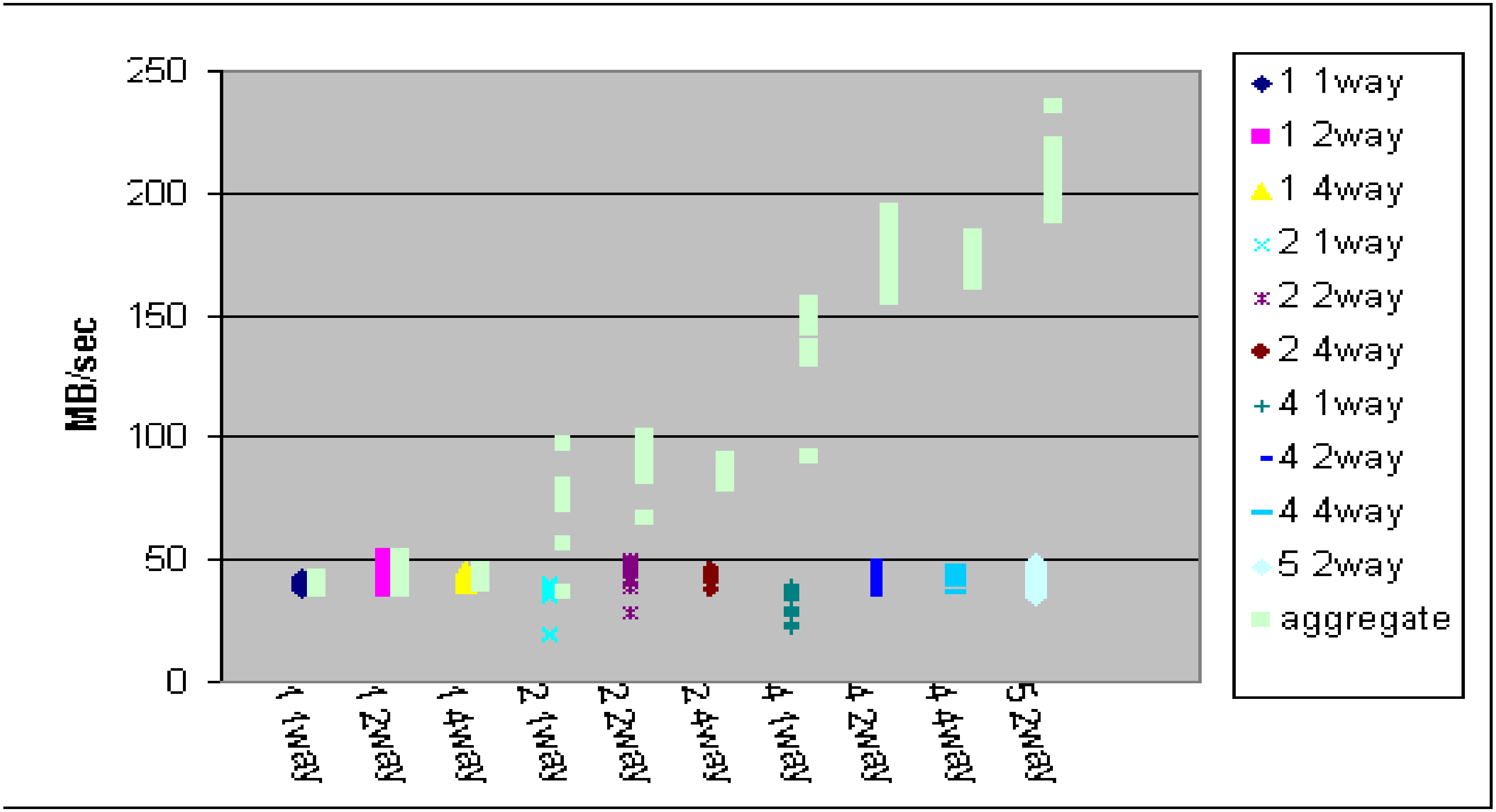
HPSS Disk Cache Configuration

- **IBM 9GB SSA drives in a 6+P RAID 5, 25 MB/s per RAID stripe**
- **MaxStrat HiPPI attached disk, (HiPPI limited into box)**
- **Sun T3 Fibre Channel disk, 72GB drives in an 8+P RAID 5, 55 MB/s per brick, attached via Storage Area Network (4 x 16 port Brocade switches)**

Transfers to HPSS Disk Cache

- Used up to 6 Sun T3 FC disk “bricks”
- Used GbE “jumbo” packets (actually 9000 bytes)
- Believe limited by GbE at ~55 MB/s
- Best performance, 197 MB/s using 4 physical GbE connections, 235 MB/s, 5 nodes, two streams per node (5 physical connections)

BH transfers to HPSS disk cache



Futures

- Heavily into SAN operations, looking at direct disk-tape transfers using “extended copy” commands. See Poster later today
- Hope to use true “RAIT” (Redundant Array of Independent Tape-drives) in combination with FC SAN access
- Expect to see GbE -> 10 GbE and FC/SAN to go 1->2->10 Gb/sec