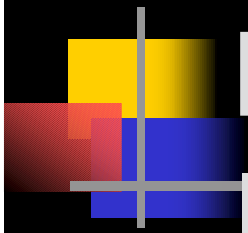
Evaluation of Sun Products as HPC Infrastructure



Presented by

Kevin Wohlever

Ohio Supercomputer Center

Special Thanks To

- Mark Juaire
- Steve Luzmoor
- Jeff Doak
- Paul Buerger
- Don Mengel
- Bob Rekieta
- Steve Johnson

Agenda

- Discuss OSC
- Describe Sun COE environment at OSC
- Review testing done on COE systems for HPC support
- Discuss problems found (and successes)
- Future Directions

Mission

The Ohio Supercomputer Center was established in 1987 to position Ohio universities and industries at the forefront of computationally intensive research, development, engineering and networking.

The Ohio Supercomputer Center is dedicated to demonstrating Ohio's leadership in and commitment to science and technology. OSC provides a reliable, high performance computing and ***c****ommunications* infrastructure for a diverse, statewide/regional community; including education, industry, and state government. In collaboration with this community; the Center evaluates, implements, and supports new and emerging *information technologies*. The Center, as a shared resource, accelerates the use of *information technologies* to strengthen the state's attractiveness and global competitiveness.

OSC Primary Areas

- High Performance Computing
- Statewide internet services -- OARnet
- Platform Lab
- Department of Defense Modernization Program
- Technology Policy research
  - International Privacy Conference
  - Regional studies of bandwidth services (Ohio, Kentucky, Michigan, Maryland, and Wisconsin)

Overall Hardware Directions

OSC

Athlon Cluster
64 nodes 128 GB

CRAY SV1ex
16p 32GB

Itanium Cluster
72 nodes 288 GB

Mass Storage
SGI Origin

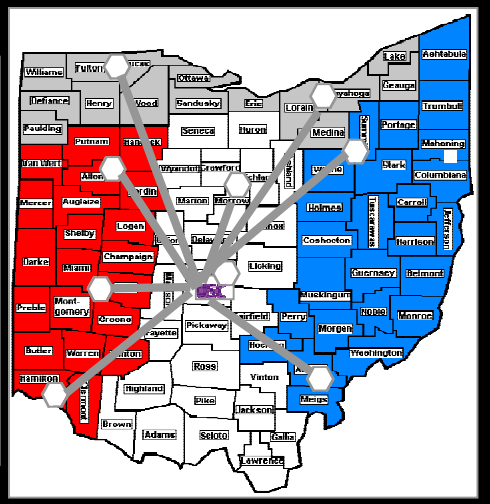
HIPPI /
gigabit
SWITCH

SGI Origin
32p 64GB

IBM3494
Tape Robot
120 TeraBytes

Sun HPCE
Systems
24p 48GB
24p 24GB
12p 12GB
12p 12GB
2.4 TB Disk

Internet



COE-HPCE Cray MOU Project Goals

- MOU announced at SC'2001
- IO / Cray SV2 testing
  - Testing network configurations
  - Testing possible system configurations
  - Disk I/O Issues
  - Network Storage
- Hierarchical Storage Testing
  - Comparing Cray DMF and Sun SAM-FS
  - Migration issues from DMF to SAM-FS

Overview of the Test Environment

COE1

- Sunfire 6800 with 12 900 MHz CPU with 12 GB Memory, 36GB UltraSCSI Disks, 327GB Sun StorEdge T3ES Rack. Two 10/100 Ethernet NICs. 2 Gigabit Ethernet NICs.
- Storage Area Network for systems in Columbus, 2620GB Sun StorEdge T3ES
  - 3 SAM-FS filesystems configured for testing using this test area.

Overview of the Test Environment

COE2 - Split into 2 domains

- Domain 1
  - Sunfire 6800 with 8 900 MHz CPU with 8 GB Memory, 36 GB of UltraSCSI Disk, 118 GB of StorEdge T3E disk. One 10/100 Ethernet NICs. 1 Gigabit Ethernet NICs.
- Domain 2
  - Sunfire 6800 with 4 900 MHz CPU with 4 GB Memory, 36 GB of UltraSCSI Disk, 118 GB of StorEdge T3E disk. One 10/100 Ethernet NICs. 1 Gigabit Ethernet NICs.
  - TimeLogic FPGA hardware for Bioinformatic processing.

Overview of the Test Environment

COE3

- A Sunfire 6800 with 24 900 MHz CPU with 48 GB Memory, 36 GB of UltraSCSI disk and 655 GB of T3 Disk Storage. Two 10/100 Ethernet NICs. 2 Gigabit Ethernet NICs.

COE4

- A Sunfire 6800 24 900 MHz CPU with 24 GB Memory, 36 GB of UltraSCSI disk and 655 GB of T3 Disk Storage. Two 10/100 Ethernet NICs. 2 Gigabit Ethernet NICs.

Software for the COE Environment

- Optimized BioInformatic software for use on the TimeLogic hardware.
- SUN/LSC Software
- SAM 250 for one system (Tape Library Support)
- QFS and FS for each system
- Sun Grid Engine Software
- SUN HPC Cluster Tools
- Sun FORTE development tools

The HPCE SUN COE Plans and Projects

- Project overview
  - GRID Computing
  - Portals
  - Storage
  - SANs / NAS
  - Domaining
  - WAN
  - Bioinformatic support (TimeLogic)

Sun as the HPC Environment Infrastructure

- Issues with this testing
  - Common MAC address impact on VPN
- SGE configuration expectations
  - Filesystems
  - Accounts
- Technical compute portal issues

Cray Tests Run

- SAM-FS I/O tests run 750 MHz chips.
  - Filesystem composed of 5 T3 arrays on 1/30-31
    - 1 T3 array for metadata
    - 4 T3 arrays for user files (thru 2 Qlogic HBAs)
  - 100 GBs of I/O, varied # streams and block sizes
  - Independent files in one directory
    - Around 50 MB/sec for one stream (file), able to aggregate I/O at 170 to 180 MB/sec.
  - All processors (12) read one file
    - 56 MB/sec to (1 stream) 1761 MB/sec (32 streams)
    - Slightly slower with 64 and 128 streams

SAM-FS Testing

- Interleaved test, processes read/write to non-overlapping sections of a single shared file.
  - Tested with a range of buffer sizes and varying number of streams.
  - Test results with 64 Kbyte buffers (aggregate rates):
    - Writes from 46 MB/sec. (1 stream) down to 6 MB/sec (128 streams)
    - Reads from 40 MB/sec. (1 stream) down to 20 MB/sec (128 streams)

SAM-FS Testing

- Suspected possible serialization of parallel I/O to shared file
- Later mounting with with SAM-QFS “-o qwrite” did not appear to help
- 900 MHz CPU Upgrage
  - Read reates improved 10 to 25% for larger buffers with many streams
  - Impact on write rates and with smaller buffers was minor

NFS Testing

- 32 K read / write sizes using UDP
- Other mount options used Solaris defaults
- To SAM-FS filesystem
- 15 GB file size (more than system memory)

NFS Testing Results

- Single stream rates over Fiberchannel
  - Write 20.7 MB/sec, Read 15.3 MB/sec
  - Initial results of 2 GB file show write rate of > 100 MB/sec.
  - Aggregate rates with independent files up to 32 MB/sec write, 50 MB/sec read
  - Interleaved I/O very slow with smaller block sizes, especially for writes
  - All streams reading single file saw aggregate rates up to 900 MB/sec
    - Implies smart cacheing on client side

NFS Testing

- Used a single T3 array
- Journaling was on for the UFS NFS tests.
- Single stream rates over Gigabit Ethernet
  - Write 15.2 MB/sec, Read 14.6 MB/sec
- Aggregate rates for independent file I/O with varying numbers of streams and a 1 Mbyte block size
  - 9 MB/sec write (1 stream) down to 2 MB/sec write (128 streams)
  - 23 MB/sec read (1 stream) up to 28 MB/sec read (128 streams)

UFS Local File System Testing

- These UFS rates are with the previously mentioned single T3 array with UFS logging on
- Aggregate rates for independent file I/O with varying numbers of streams and a 1 Mbyte block size.
- 30 MB/sec write (1 stream) up to 34 MB/sec write (128 streams)
- 67 MB/sec read (1 stream) down to 35 MB/sec read (128 streams)

UFS Local File System Testing

- Interleaved read rates were very good
  - ~100 MB/sec with 128 streams and 1 MB buffers.
- All streams reading a single file saw aggregates up to 1350 MB/sec
- With UFS loggin on, write rates for man concurrent streams substantially degraded.
- UFS logging had a relatively small impact on reads with 1 MB buffers, but had a significant impact with 64 Kbyte buffers.

IP Over Fibre Channel

- All tests were done using nettest
- All tests were TCP based
- Between coe1 and coe2b using emulex light point fibre channel cards
- Using MTU of 1500
- Using a buffer size of 32768

IP Over Fiber Channel Single Stream Results

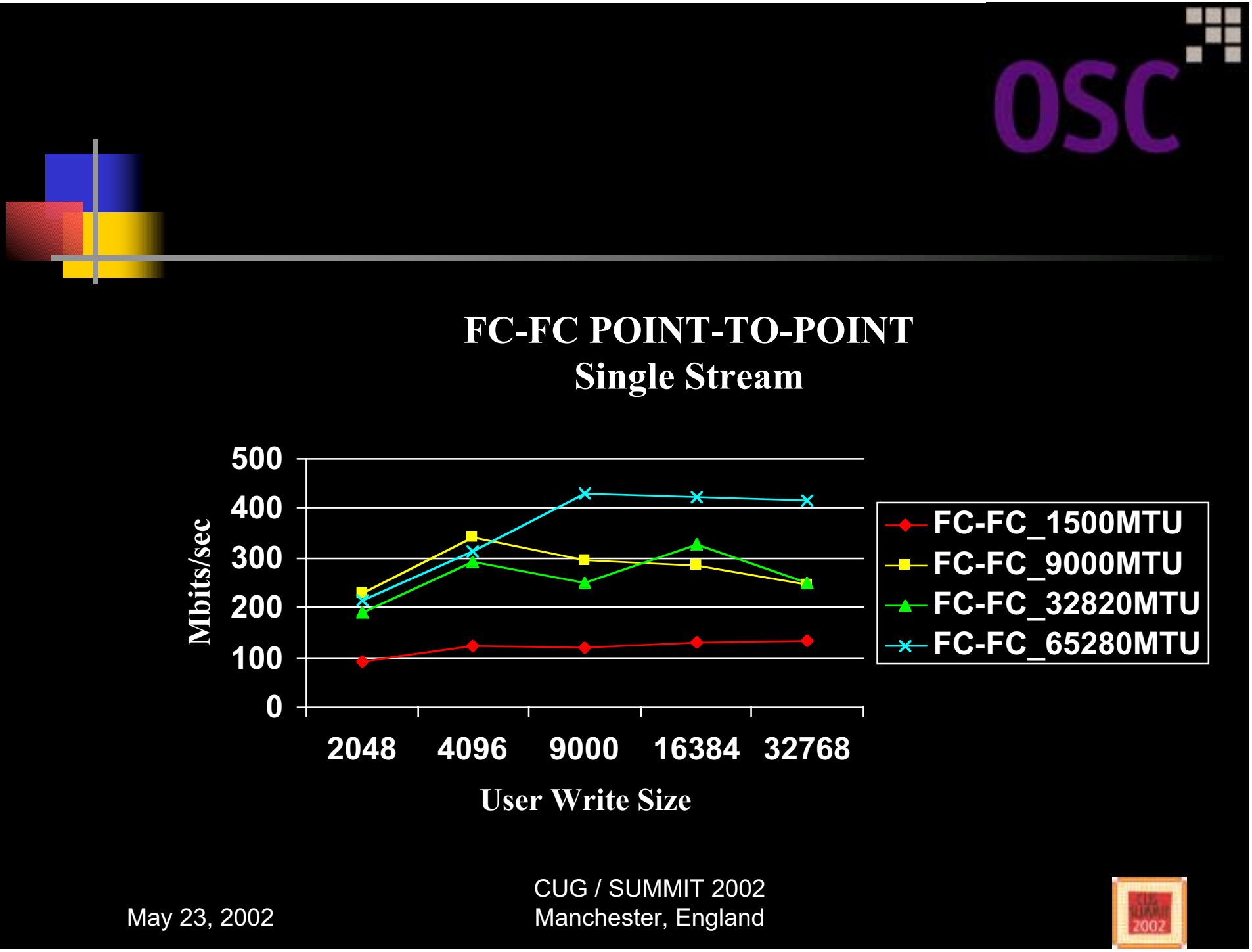
- Varying write sizes
  - 2048 write had a RW avg of 90.02 mbit/sec
  - 4096 write had a RW avg of 124.93 mbit/sec
  - 9000 write had a RW avg or 119.90 mbit/sec
  - 16384 write had a RW avg of 131.74 mbit/sec
  - 32768 write had a RW avg of 135.48 mbit/sec
- Multiple Stream Results
  - Write size of 16384
    - 2 streams Aggregate had a RW avg of 143.04 mbit/sec
    - 4 streams Aggregate had a RW avg of 188.79 mbits/sec
    - 8 streams Aggregate had a RW avg of 203.47 mbit/sec

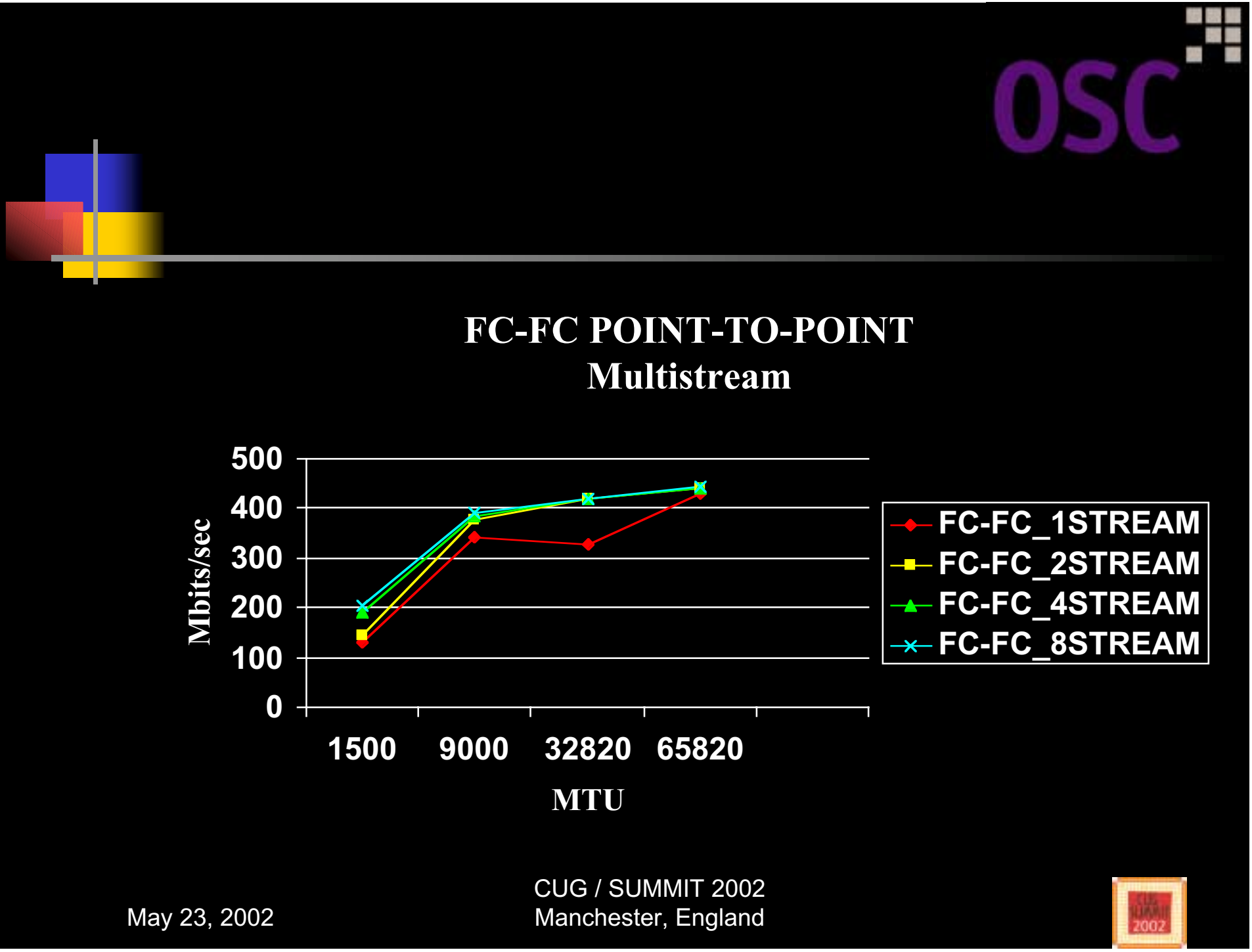
IP Over Fibre Channel

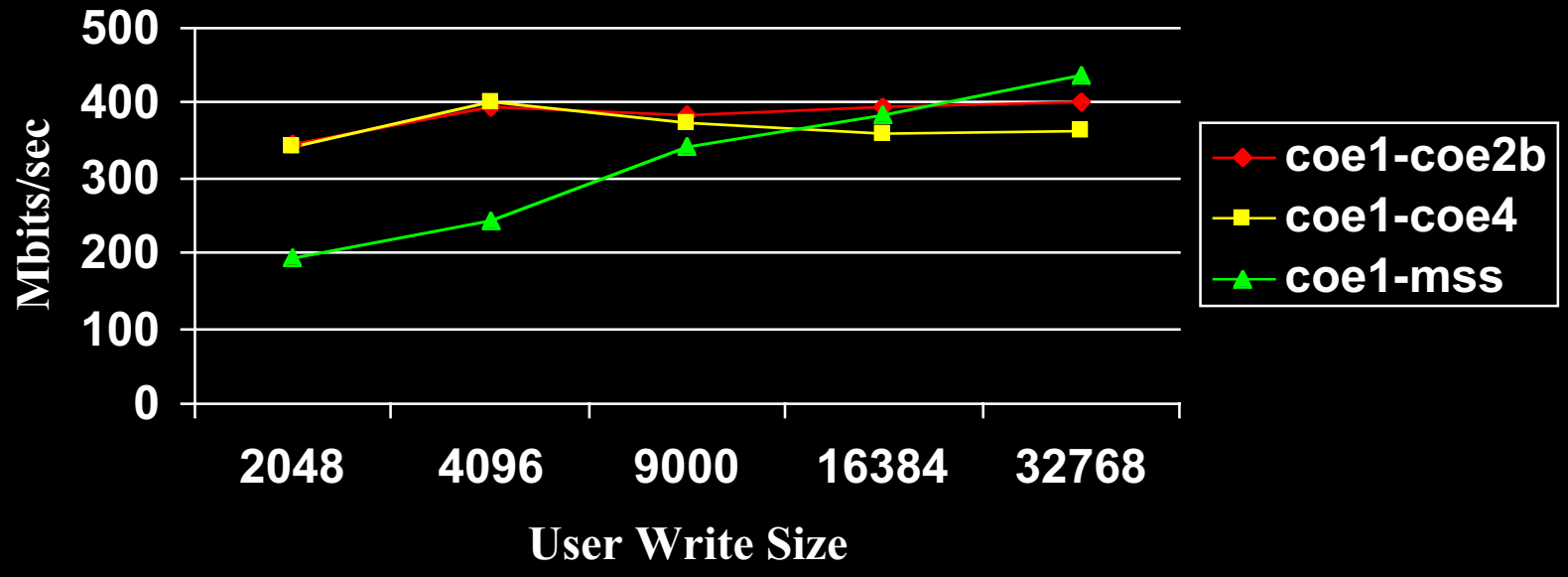
- Increasing buffer size to 65336
  - Single stream
    - RW Avg = 131.56 mbits/sec
- Using MTU of 9000 and buffer size of 4500
- Single Stream Results
  - Varying Write Sizes
    - 2048 Write had a RW avg of 228.79 mbit/sec
    - 4096 Write had a RW avg of 340.47 mbit/sec
    - 9000 Write had a RW avg or 294.87 mbit/sec
    - 16384 Write had a RW avg of 285.90 mbit/sec
    - 32768 Write had a RW avg of 245.53 mbit/sec

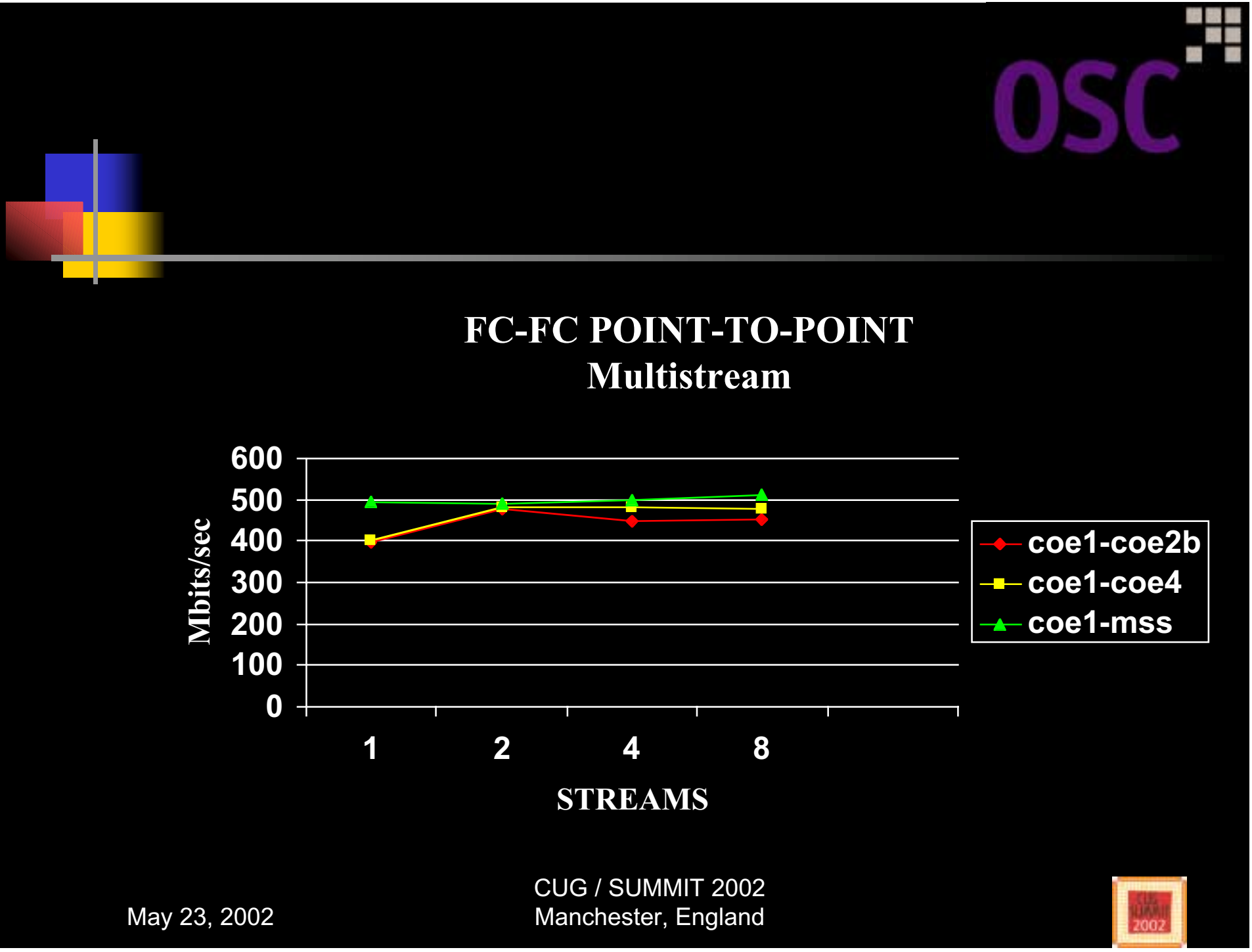
IP Over Fibre Channel

- Multiple Stream Results
  - Write size of 4096
    - 2 streams Aggregate RW avg of 378.58 mbit/sec
    - 4 streams Aggregate RW avg of 384.68 mbits/sec
    - 8 streams Aggregate RW avg of 389.25 mbit/sec


FC-FC POINT-TO-POINT
Single Stream
Mbits/sec
500
400
300
200
100
0
2048
4096
9000
16384
32768
User Write Size
FC-FC_1500MTU
FC-FC_9000MTU
FC-FC_32820MTU
FC-FC_65280MTU


FC-FC POINT-TO-POINT
Multistream
Mbits/sec
500
400
300
200
100
0
1500
9000
32820
65820
MTU
FC-FC_1STREAM
FC-FC_2STREAM
FC-FC_4STREAM
FC-FC_8STREAM


GIGABIT ETHERNET
Single Stream
Mbits/sec
500
400
300
200
100
0
2048
4096
9000
16384
32768
User Write Size
coe1-coe2b
coe1-coe4
coe1-mss

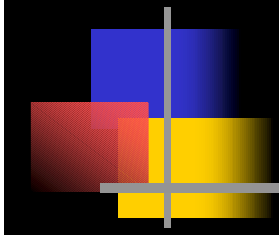

FC-FC POINT-TO-POINT
Multistream
600
500
400
300
200
100
0
Mbits/sec
1
2
4
8
STREAMS
coe1-coe2b
coe1-coe4
coe1-mss

IP Forwarding Testing

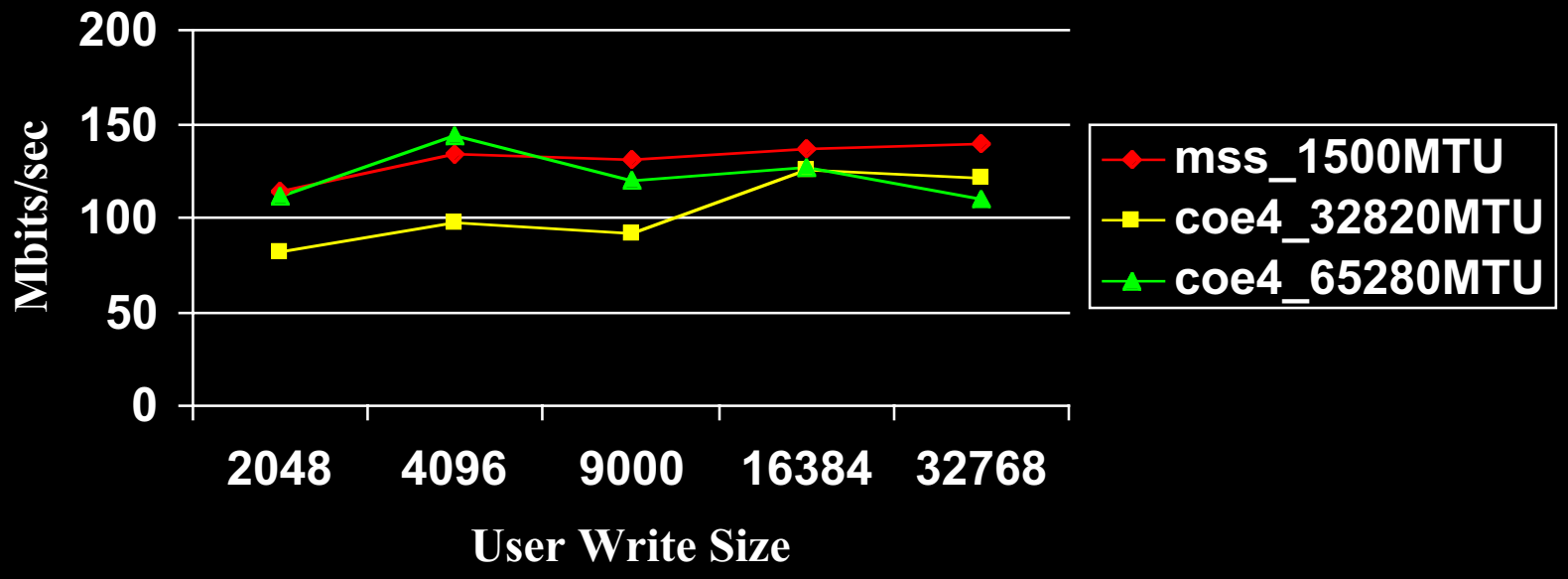
- All tests were done using nettest
- Over VLAN Gig-E, Sun 6800 <-> Sun 6800 <-> Origin 2000
- Unable to get MTU to sync up at higher packets sizes
  - 135 mbits/sec avg. (read & write) single stream
  - ~ 155 mbits/sec avg. multiple streams

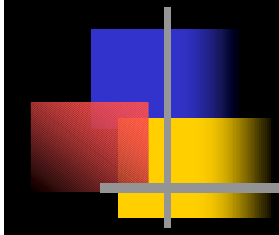
IP Forwarding Testing

- Over VLAN Gig-E, Sun 6800 <-> Sun 6800 <-> Sun 6800
- Unable to get MTU to sync up at higher packet sizes.
- MTU Able to Sync up proper
  - 147 mbits/sec avg single stream
  - 143 mb/sec avg multiple stream
- Conclusion: unknown problems with Extreme Switch that is prohibiting MTU sizes greater than 1500.

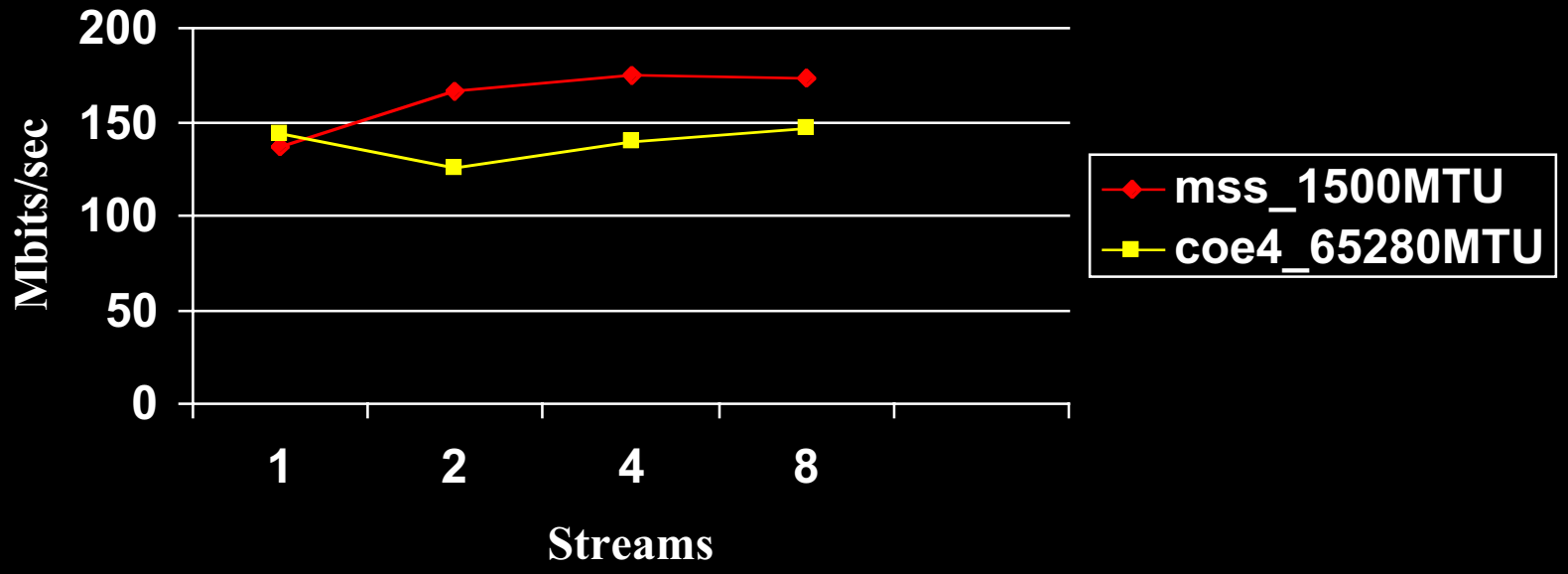


IP Forwarding GIGE<-->FC
Single Stream





IP Forwarding GIGE<-->FC
Multistream



Problems Encountered

- LSC testing
  - 1,000,000 files in a directory
  - Arfind problem
- Fiber channel testing
  - Not able to run more than one light point card at once on a system.
- Disk configuration
- Sun upgrades
- Veritas upgrades
- OS upgrades

Administrative Issues

- Technical compute portal overview
- SGE Issues

Benefits of the Testing

- Network configurations
  - VPN
  - Gig-E
  - IP over FC
- NFS Testing Results
  - Tests run

HSM Testing Issues

- Cray will present at CUG
- SAM-QFS arfind command consuming most CPU resources

Network Testing

- GIG-E jumbo frames not supported
- Ttcp test results
- Problem with Multiple NIC sharing one MAC address (Default for Sun system (RAS))

Future Directions

- Understanding the TCP performance issues with 64k MTUs
- Resolving the IP routing issues and MTU discovery problem(s)
- Exploring some larger SAM-QFS configurations to see if I/O rates scale
- Seeing if Sun/LSC can do anything to improve rates for interleaved I/O
- Trying various SAM-QFS capabilities (Direct I/O, striping, etc.) we didn't get to.
- f) Trying SAM-QFS in a SAN environment, hopefully with the multiple writer capability.