HPC Resource

DataStar : A 10TF Power4 + Federation Switch System

Amit Majumdar

(majumdar@sdsc.edu)

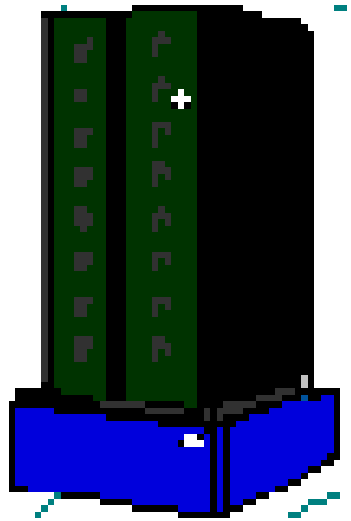
Scientific Computing Applications Group
San Diego Supercomputer Center

University of California San Diego

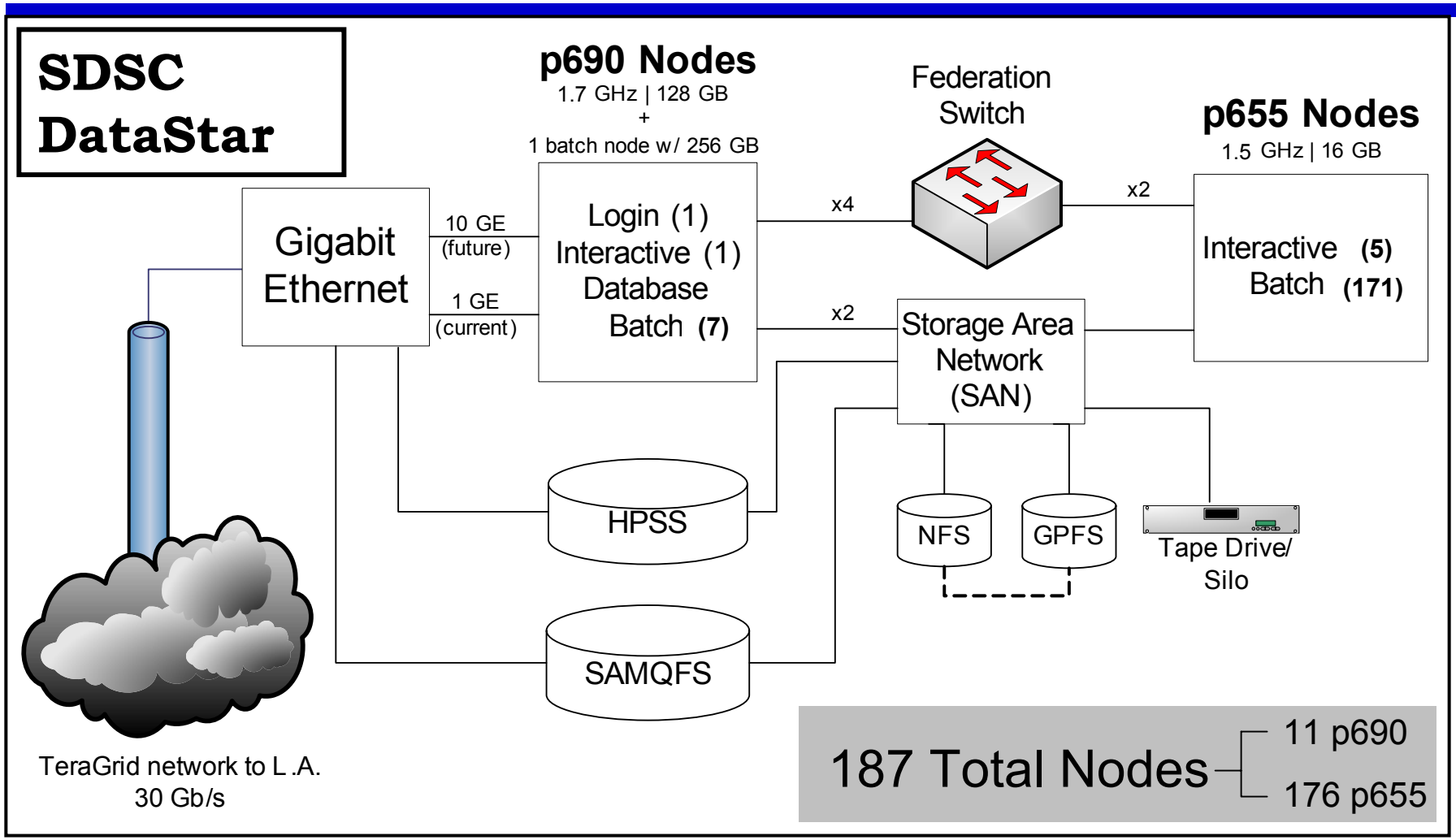
SDSC

- **A NSF center with compute and data resources allocated *freely* through peer review process**
- **One of the emphasis of SDSC for national Cyberinfrastructure initiative is Data intensive computing**
- **Acquired Datastar (DS) at the beginning of the year to replace Bluehorizon (BH) as the main compute engine**
- **Per NSF guidance we are planning on procuring our next big machine in FY06**

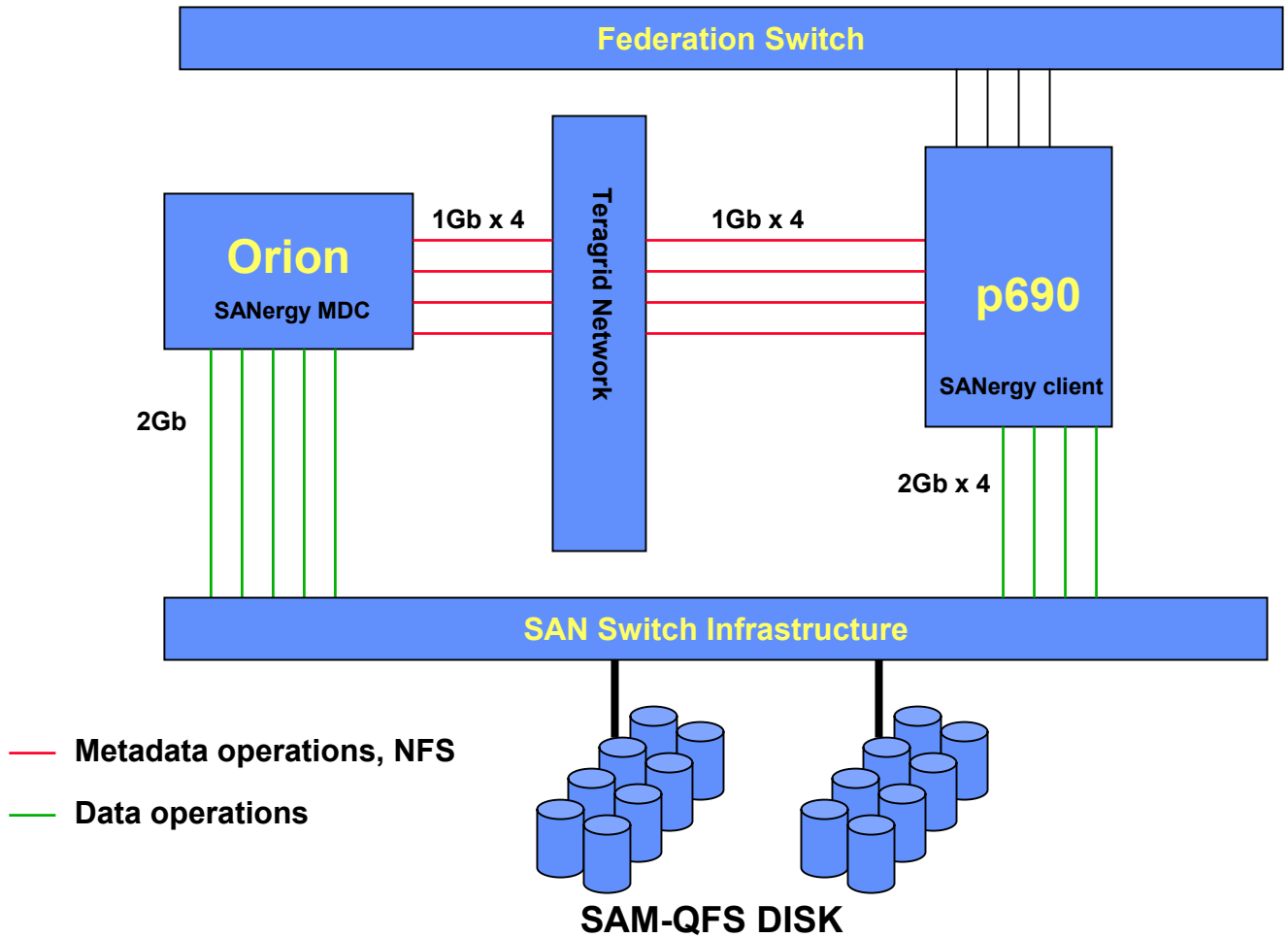
DataStar

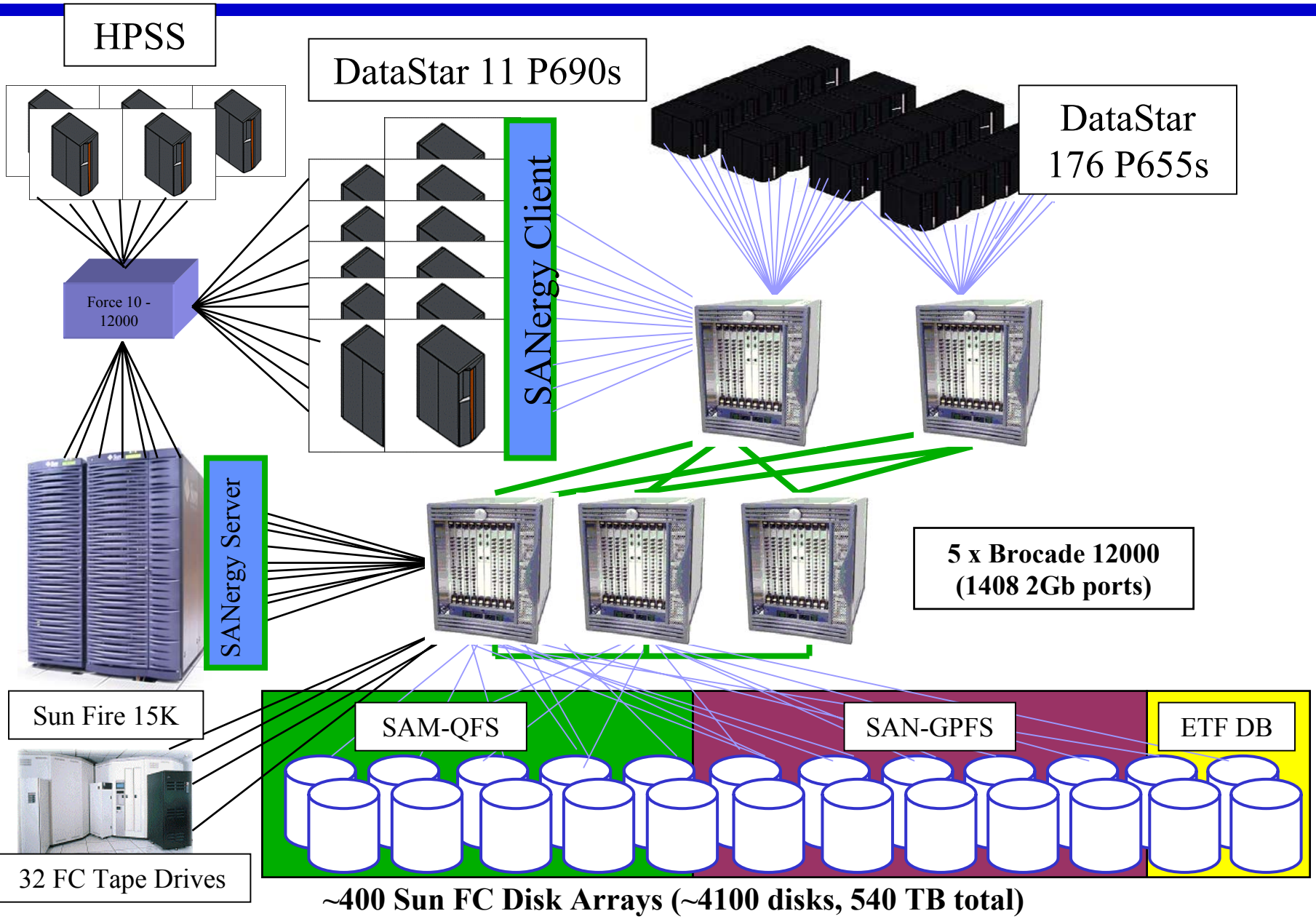


- **10.1 TF, 1760 processors total**
- **11 32-way 1.7 GHz IBM p690s**
  - 2 nodes 64 GB memory for login and interactive use
  - 6 nodes 128 GB memory for scientific computation
  - 2 nodes 128 GB memory for database, DiscoveryLink
  - 1 node 256 GB memory for batch scientific computation
  - All p690s connected to Gigabit Ethernet with 10 GE coming soon
- **176 8-way 1.5 GHz IBM p655**
  - 16 GB memory
  - Batch scientific computation
- **All nodes Federation switch attached**
- **All nodes SAN attached**
- **Currently 66 TB GPFS – will be increased in the near future**


SDSC DataStar
p690 Nodes
1.7 GHz | 128 GB
+
1 batch node w/ 256 GB
Federation Switch
p655 Nodes
1.5 GHz | 16 GB
Gigabit Ethernet
10 GE (future)
1 GE (current)
Login (1)
Interactive (1)
Database
Batch (7)
x4
x2
Interactive (5)
Batch (171)
x2
Storage Area Network (SAN)
HPSS
NFS
GPFS
Tape Drive/ Silo
SAMQFS
TeraGrid network to L .A.
30 Gb/s
187 Total Nodes
11 p690
176 p655

SANergy Data Movement




HPSS
DataStar 11 P690s
DataStar
176 P655s
SANergy Client
Force 10 -
12000
SANergy Server
5 x Brocade 12000
(1408 2Gb ports)
Sun Fire 15K
SAM-QFS
SAN-GPFS
ETF DB
32 FC Tape Drives
~400 Sun FC Disk Arrays (~4100 disks, 540 TB total)

Porting Issues

- **We moved from default 32bit to default 64bit**
- **Fairly easy to port from BH to DS**
- **Mixed Fortran + C/C++ codes encountered some trouble**

Initial Setup Difficulties

- **Substantial jump in weight, power and cooling load for DS compared to BH**
- **Memory and performance leak**
  - Fixed through NFS automounts
  - Removing unnecessary daemons
- **Problems related to GPFS over SAN**
  - Upgrade of IBM FC adapters, Brocade switches and SUN disks
- **Loss of processors, memory, cache**
- **HMC issues**
- **Federation issues**

Large Scale Pre-production Computing

- 100k to 200k hours per project
- Onuchic, UCSD: Study the folding kinetics of a beta hairpin at room temperature in explicit water
- Yeung - 2048**3 turbulence run
- Goodrich, BU - 3D calculation of the shearing and eruption of solar active regions on 201 x 251 x 251 mesh
- Richard Klein, UCB: dynamical evolution, gravitational collapse and fragmentation of large turbulent molecular cloud in the galaxy
- NREL, Cornell: Cellulose project, 1 million atom CHARMM simulation of protein interactions

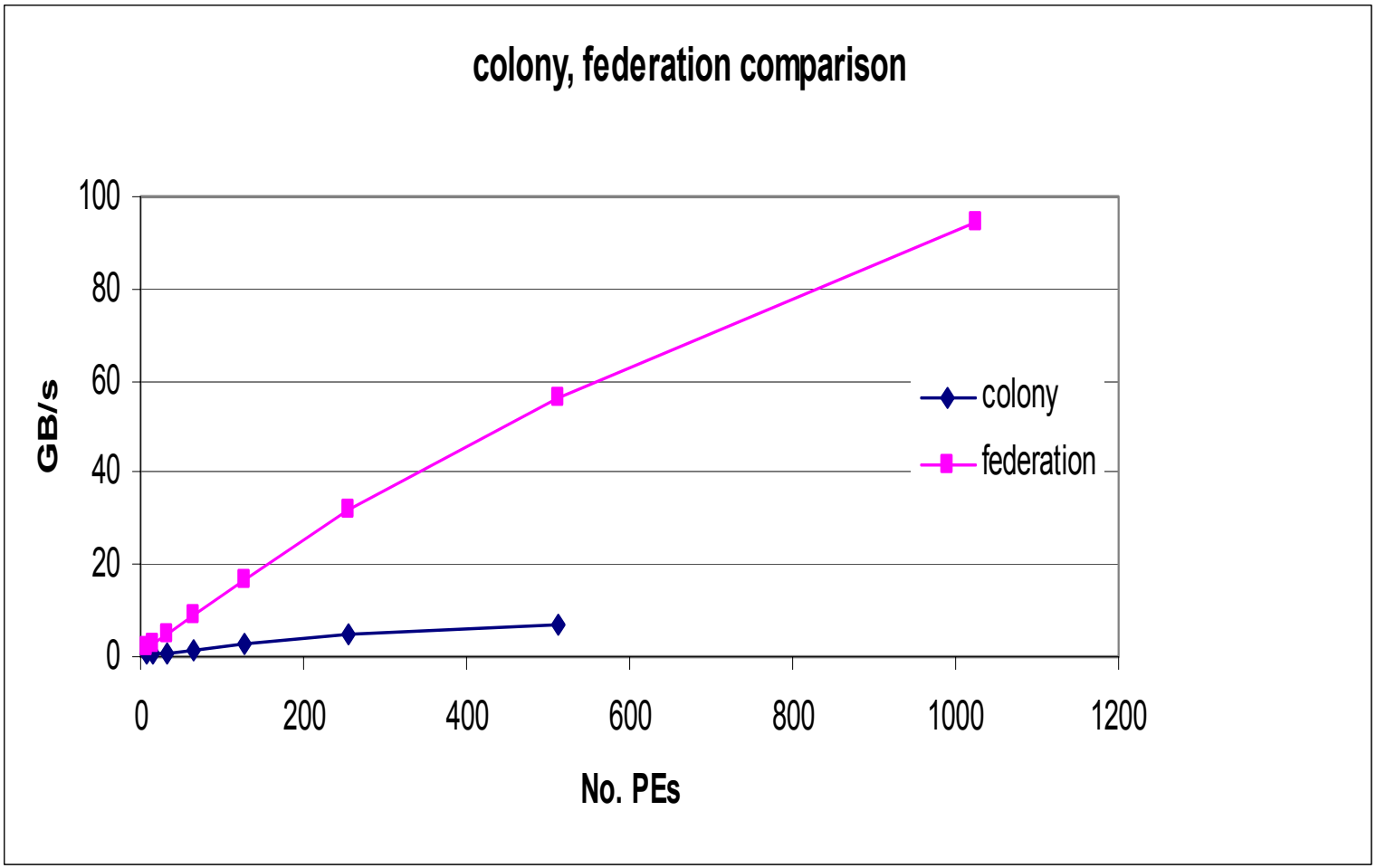
Performance Analysis

- **All codes are 64bit compiled**
- **Still some problems, so the results may not be the best**
- **All the runs are done on 1.5GHz P655s**
- **NAS benchmarks CG, FT, LU and MG**
- **Applications ENZO**

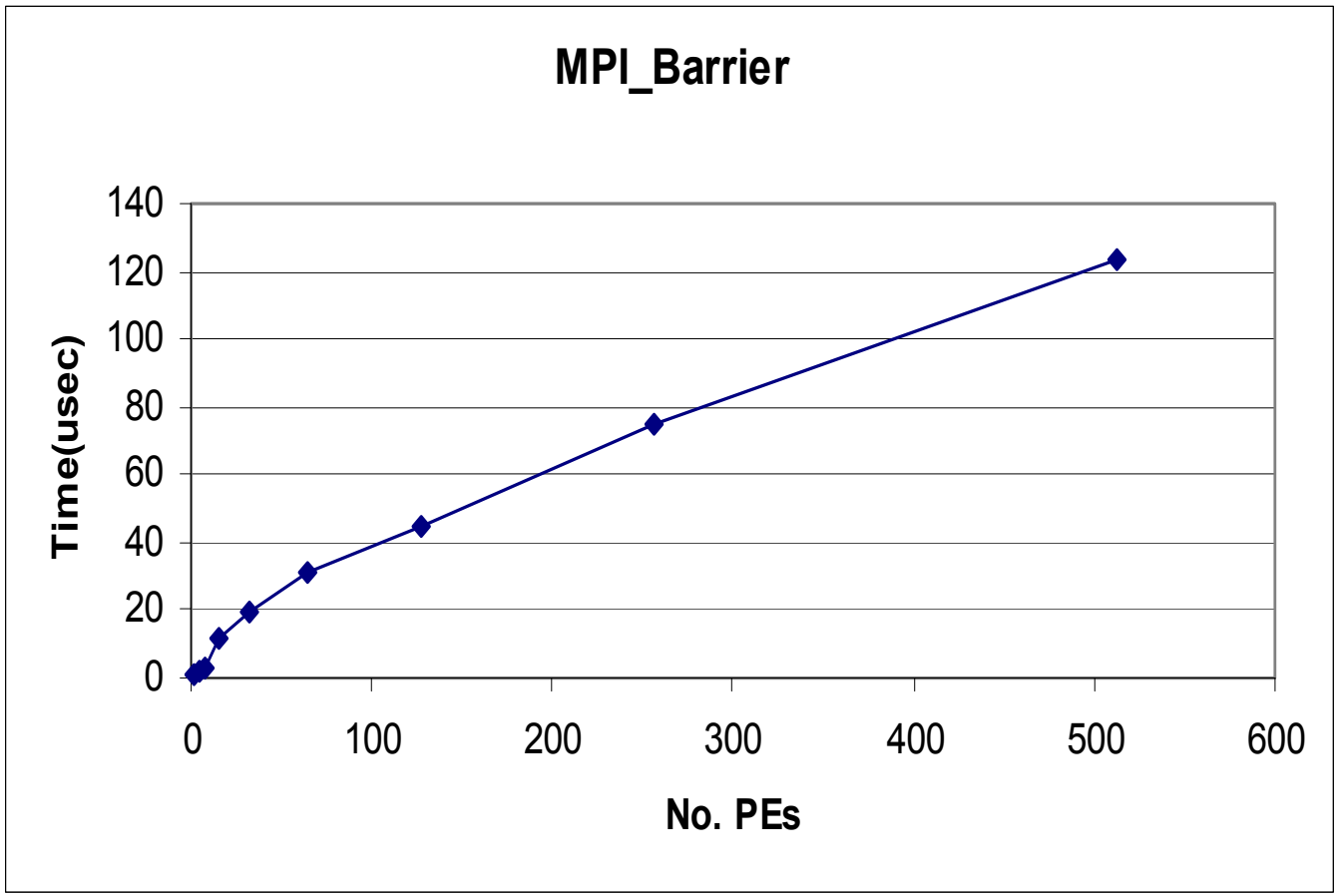
BH Versus DS Latency and Bandwidth

MPI	Latencies (usec)		Bandwidth (MB/s)	
	BH	DS	BH	DS
• Intra-node	12.68	3.9	512.2	3120.4
• Inter-node	18.6	7.56	353.6	1379.1

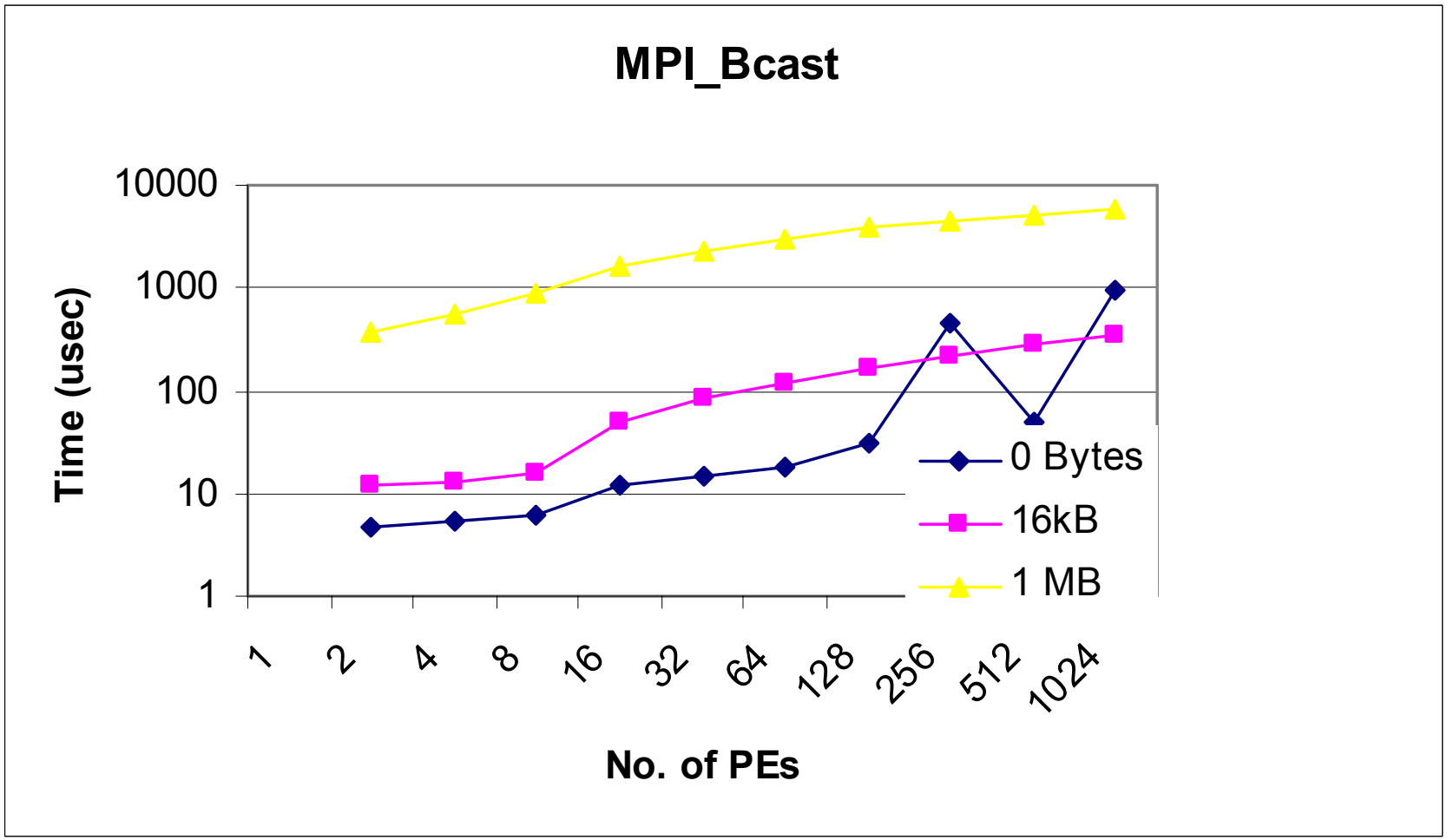
Bisection Bandwidth



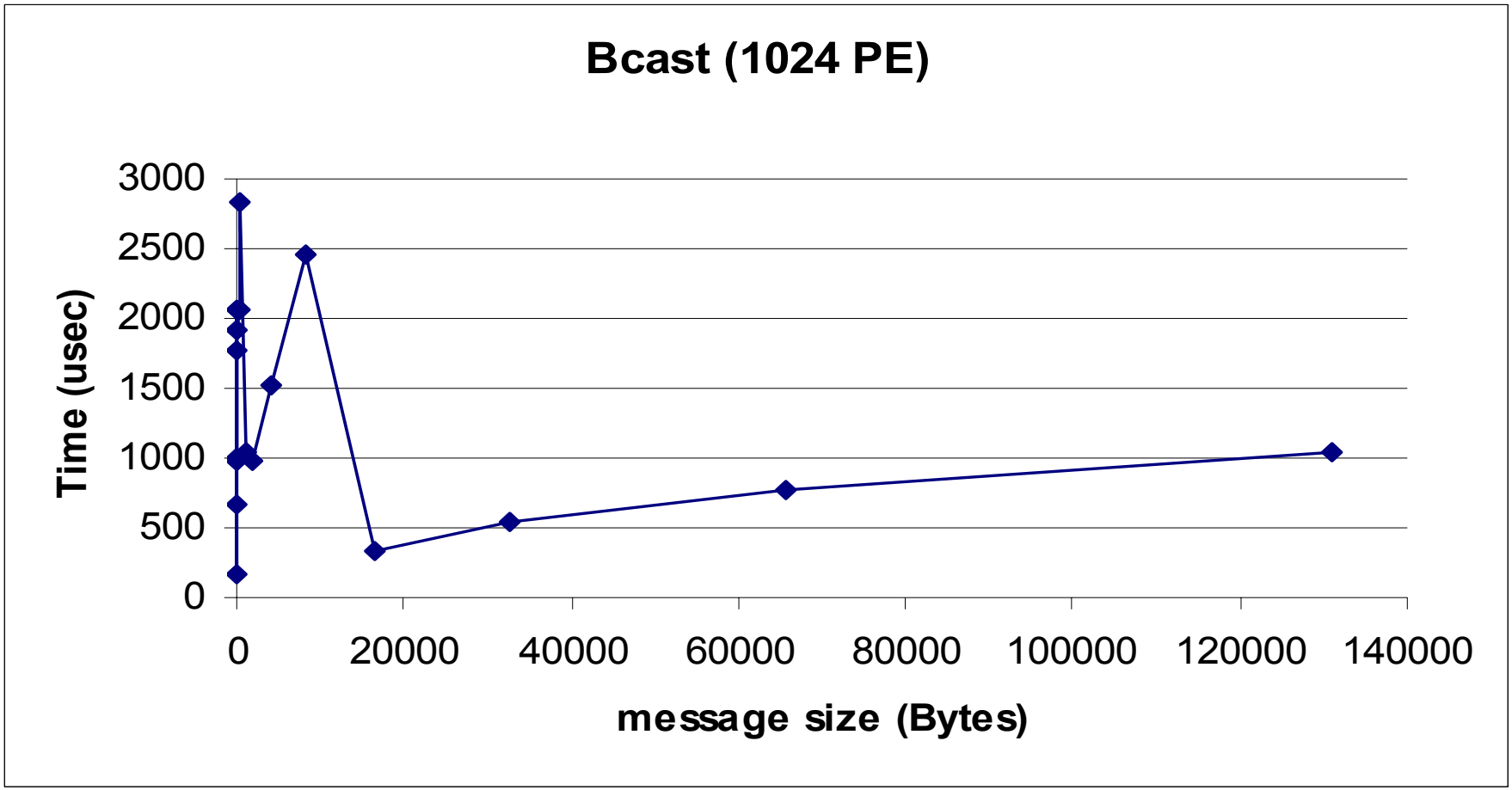
MPI Barrier Performance (Federation)



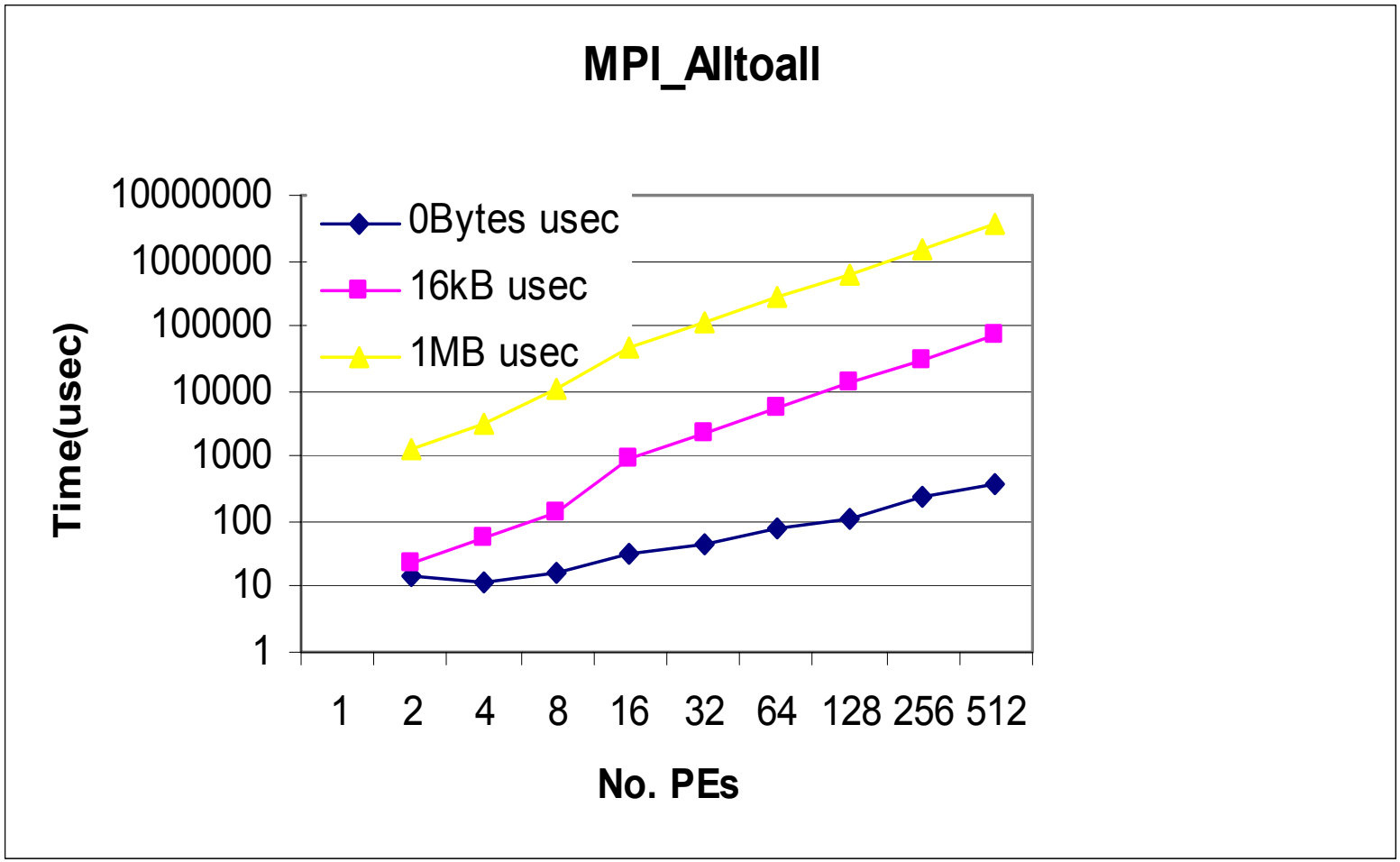
MPI Broadcast Performance



Unusual Bcast Behavior



MPI_Alltoall



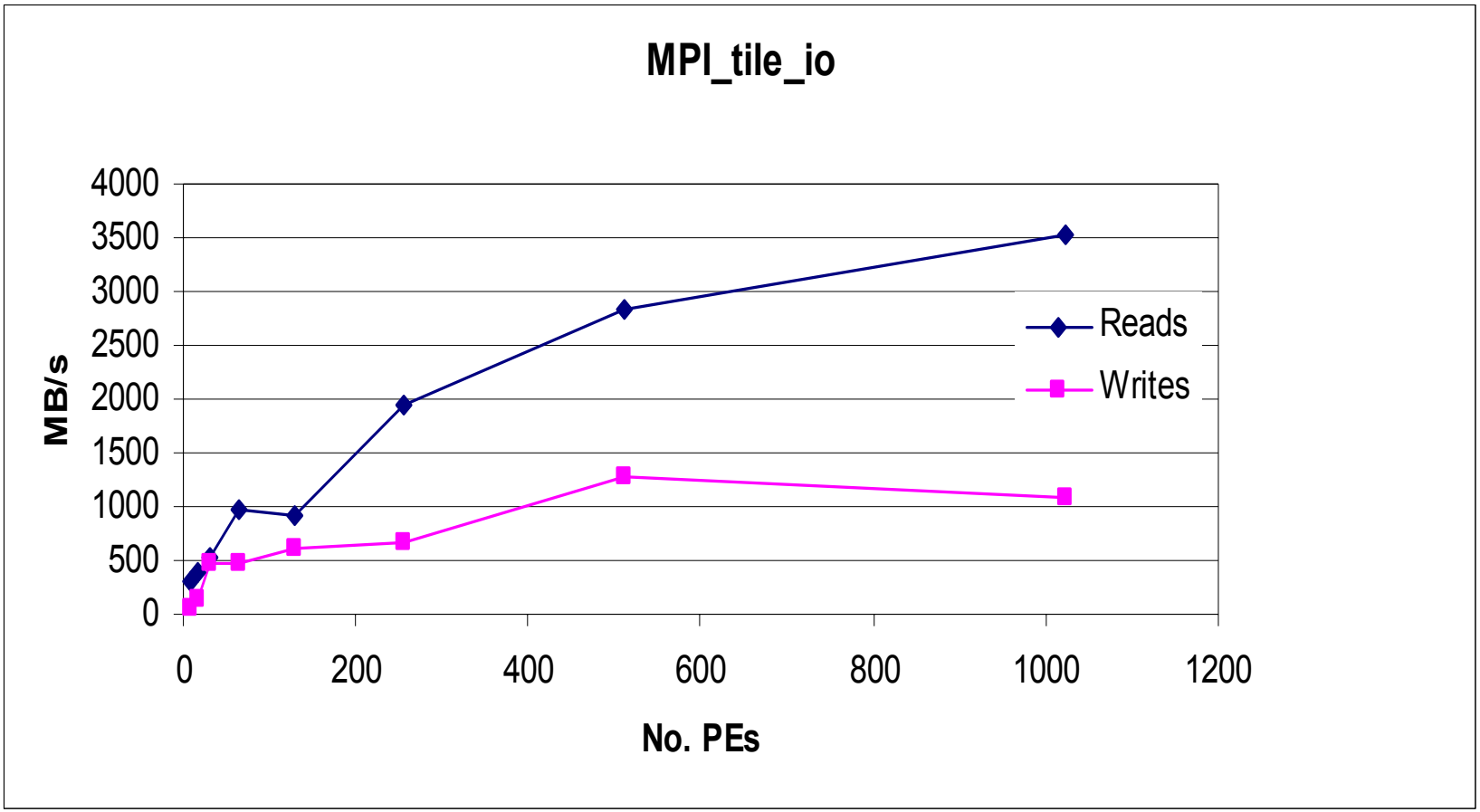
Basic MPI Performance

- **Understanding the raw MPI call performance (message size, No. of PEs) is useful to interpret real applications using MPI traces**

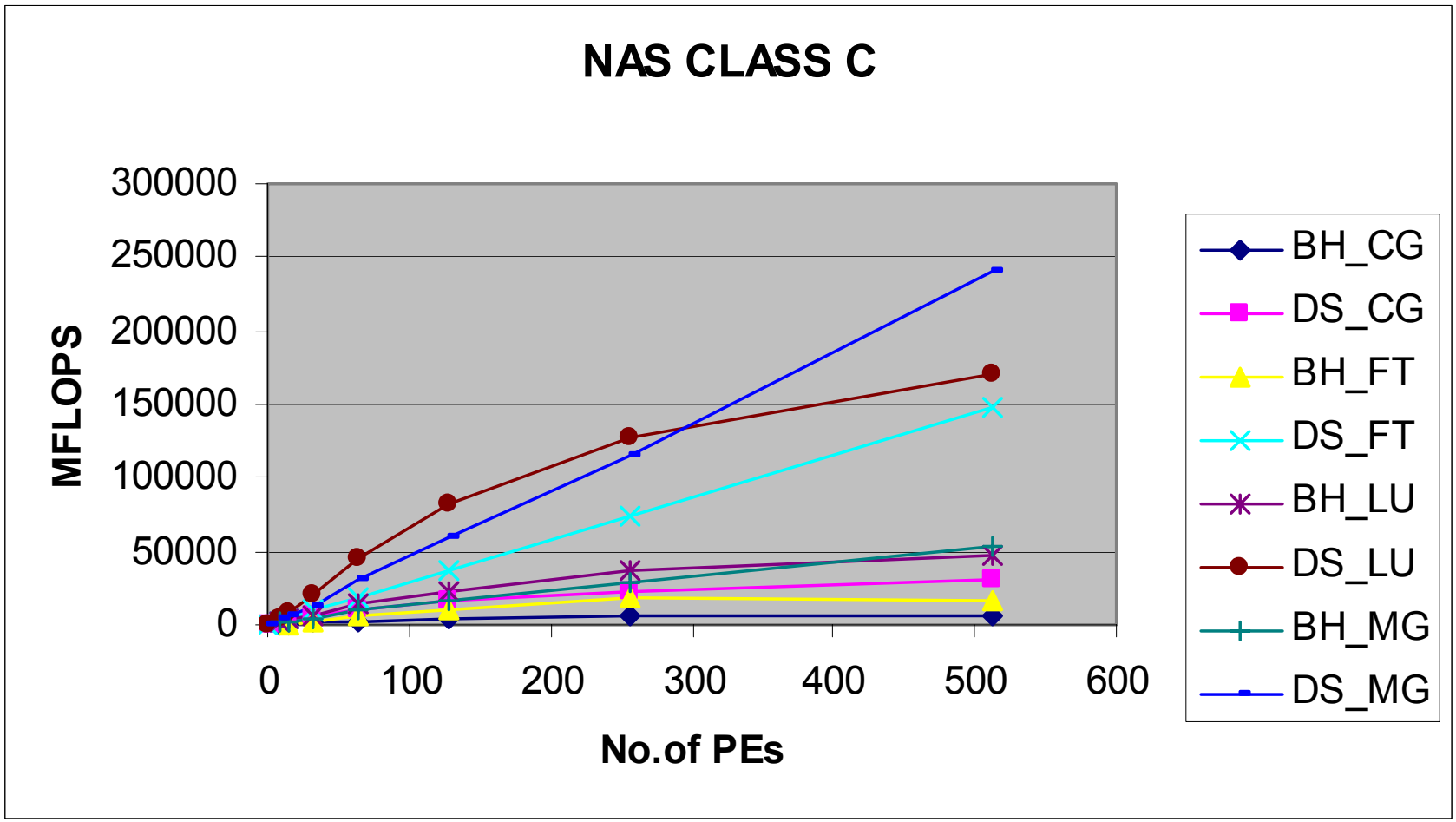
GPFS Performance

- **Metadata performance**
- **File read, write performance**
- **MPI_tile_io (IOR, Pallas MPI I/O, MDTEST)**
- **Each processor writes one rectangular tile**
  - 16 bytes per element
  - each tile is 1000 x 1000 elements
- **1024PE run writes 32X32 tiles: 16GB file**

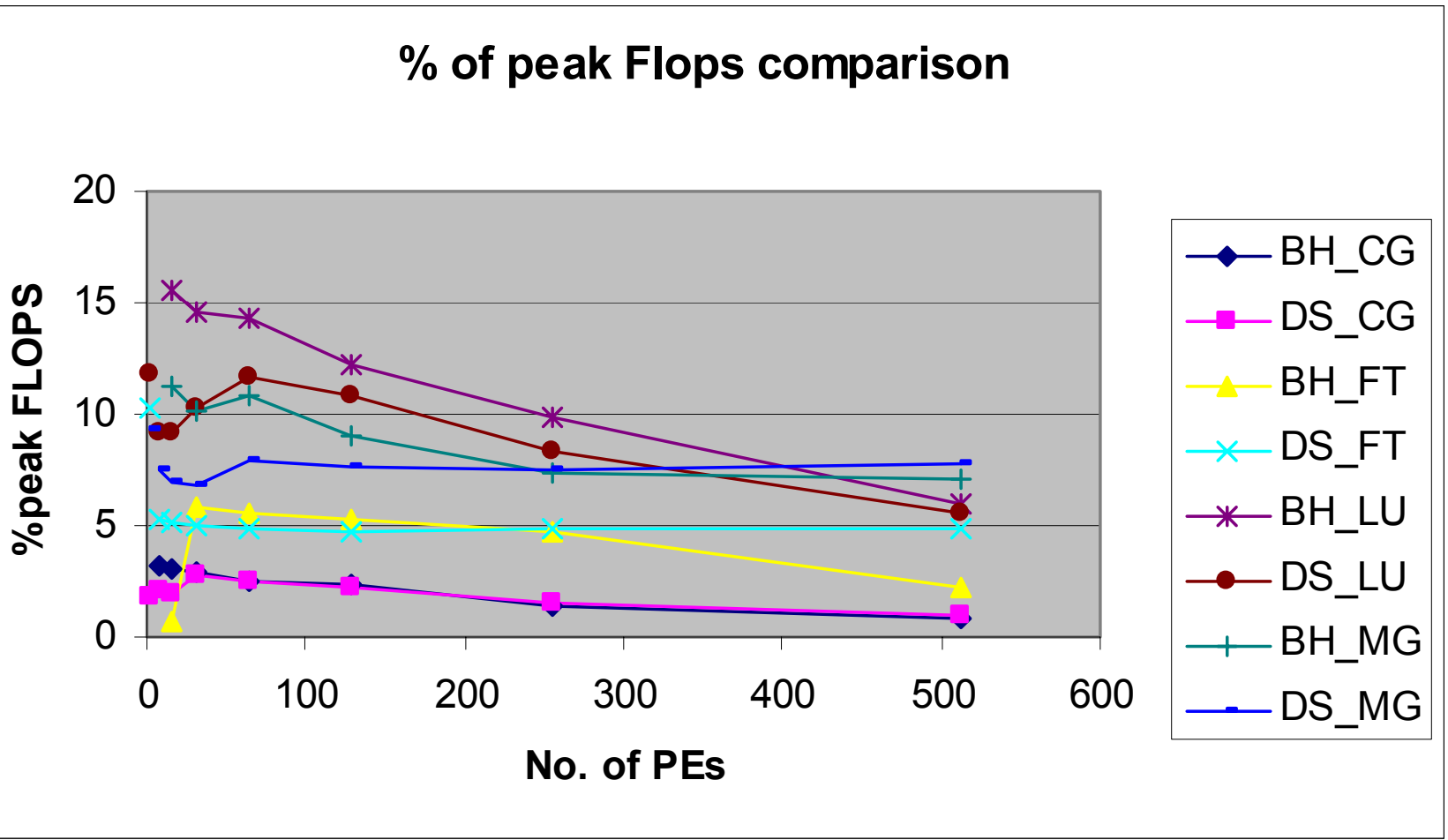
GPFS Performance (MPI I/O)



NAS Benchmarks: Strong scaling



NAS Benchmarks: strong scaling



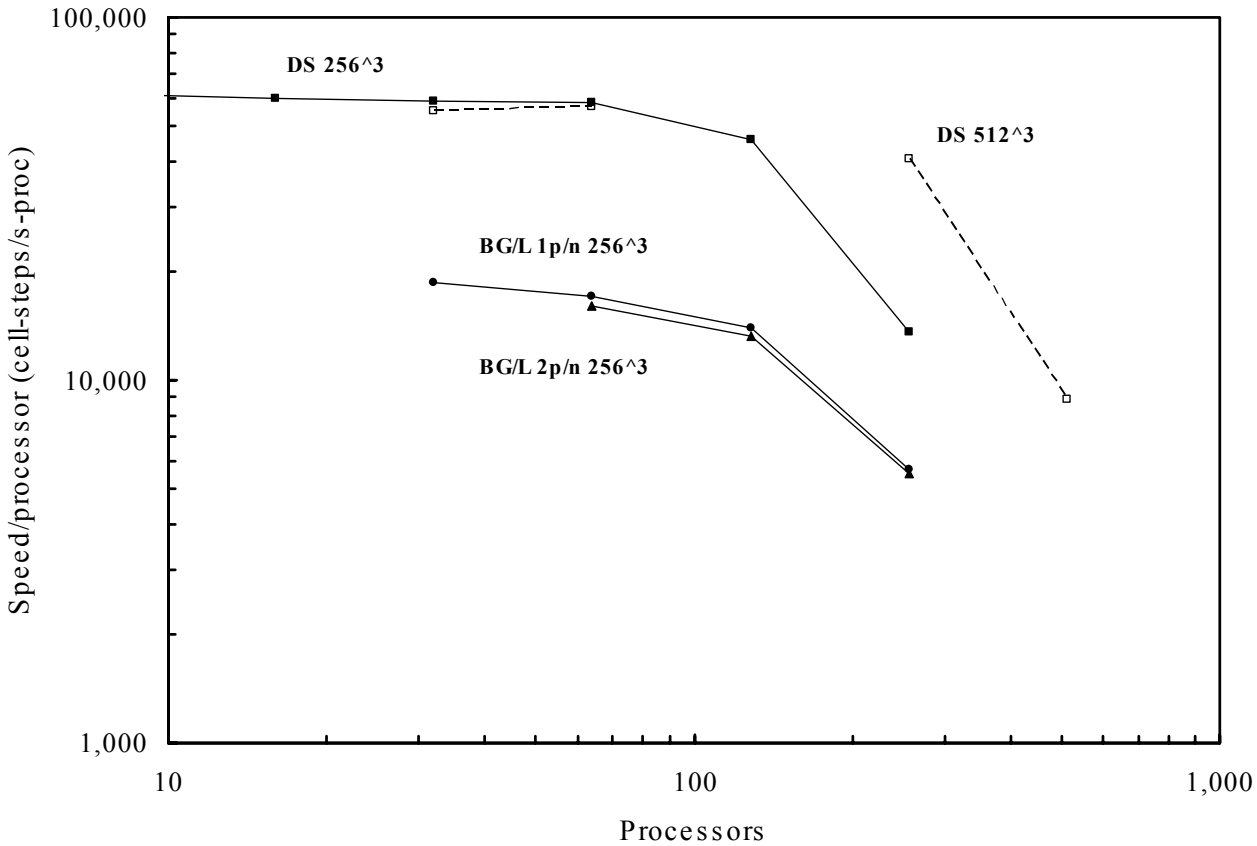
NAS Benchmarks: strong scaling

- DS %peak is more flat than BH due to the better bandwidth
- Initially %peak increases due to cache effects and then drops due to communication over heads
- Sparse matrix codes give the worst %peak and dense matrix codes gives the best %peak

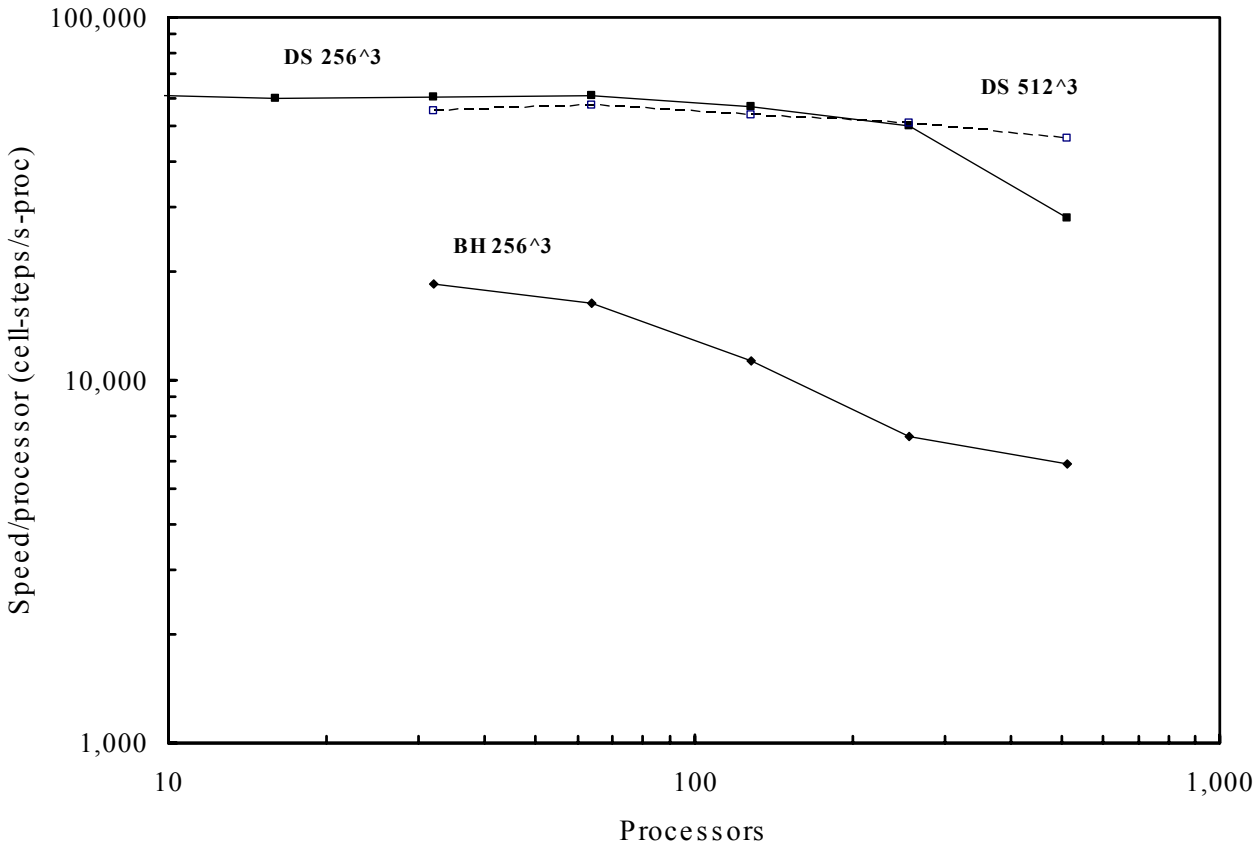
Application Benchmarks

- **ENZO: Astrophysics code**
  - Consumes 1 Million hours on Datastar

Original Enzo Performance on DS,BG/L



Improved Enzo Performance on BH and DS



Datastar Summary

- Datastar has become stable with good performance
- DS Federation switch and GPFS performance is significantly improved compared to BH
- P690s can be used for pre/post processing of large memory runs – no need to switch to different machine

Blue Gene/L

- We are looking into getting a 1-rack Blue Gene/L system – this is not finalized, not official yet
- If we do get this 1-rack system, a possible configuration would be 1024 nodes, 2.7/5.4 TFLOP peak speed, 0.5 GB memory/node, and 128 I/O nodes making compute to I/O node ratio 8:1 (LLNL BG/L this ratio is 64:1)
- The machine will be used initially for benchmarking performance of various applications from SDSC and ASCI alliance