

Cluster Scalability and Performance

Neil Gorsuch and Rob Pennington – NCSA

Jim Kasdorf - PSC



Session 3 Overview

Cluster Scalability Focus Group

- **Charter**
- **Current activities**
 - **NCSA view**
 - **Other presentations**
- **Wish-list**
- **Open discussion**



Cluster Scalability and Performance Charter

- **Determine cluster scalability and performance problem areas.**
- **Find existing and work-in-progress solutions.**
- **Gather and disseminate cluster scalability techniques.**
- **Generate scalability requirements.**
- **Determine common areas of interest, and collaborate on scalability solutions.**



Cluster Scalability Problems

- **I/O**
 - File systems
 - Bandwidth
- **Network services**
 - Single server failures
 - Privileged ports shortage
- **Job queuing and launching**
- **Interconnect mapping**
- **Software updates**
- **Management**
 - Grouping
 - Consolidation of commands output



Scalability Activities

- **OSCAR**
- **OSCAR Gold**
- **NCSA specific**
 - **VMI**
 - **KFS file system**
 - **Clumon**
 - **Large cluster practices and patches**
- **SCIDAC**



OSCAR

- **Framework/Installer for packaged open source cluster software**
- **Designed to support many Unix operating systems**
 - **Currently, Redhat Linux and Mandrake Linux**
 - **Soon to be released – other Linux distributions**
- **Supported and developed by:**
 - **NCSA**
 - **IBM**
 - **Dell**
 - **Intel**
 - **Oak Ridge Laboratories**
 - **Indiana University**
- **Popular open source cluster software package**



OSCAR and scalability

- **Framework for scalable software.**
- **Best/most used of each type of competing packages included.**
- **Allows multiple software packages for the same function.**
- **Best package for a function can be dynamically chosen.**
- **Upgraded packages do not have to be included in OSCAR.**
- **New packages can be downloaded during installation.**
- **Improved packages can be downloaded during installation.**
- **Packages can be upgraded to more scalable versions.**
- **SIS (node initializer/builder) now supports multi-cast.**
- **OSCAR has been used to install a 500 node cluster**



OSCAR Scalability Tweaks

- **PBS**
 - Home directory spooling (nfs instead of RSH)
 - Open file descriptor limit
 - Max server connections
 - Job basename length
 - Polling intervals
- **Maui**
 - Job attributes are limited to N nodes
- **SSH**
 - Non privileged ports (parallel SSH tasks)
 - User based keys



OSCAR Gold

- **Produced by NCSA and UIUC and Progeny and others**
- **Single CDROM includes:**
 - Linux Itanium operating system
 - Upgraded kernel
 - OSCAR for Itanium
- **When booted, it leads the operator through a Linux installation on the cluster head node**
- **The head node reboots from it's disk, then OSCAR:**
 - Installs the operating system on all compute nodes
 - Installs OSCAR compute clustering software stack
- **BETA testing now**
 - On Gelato portal for download



OSCAR / OSCAR Gold Components

- **Cluster Database**
- **SIS – Network Installer (IBM)**
- **C3 – Cluster Management Tools (ORNL)**
- **OpenSSH/OpenSSL – Secure Transactions**
- **Pfilter – Firewall System**
- **MPICH – Message Passing Interface**
- **LAM/MPI – Message Passing Interface (Indiana Univ)**
- **PVM – Parallel Virtual Machine (ORNL)**
- **Switcher – dynamic environment switching**
- **PBS – Job Queuing System**
- **Maui – Job Control System**



OSCAR / OSCAR Gold Add-on Components

- **Ganglia – cluster monitor**
- **Hdf5 – hierarchical data format system**
- **Clumon – cluster monitoring**
- **GM – Myrinet support**
- **PVFS – parallel file system**



Parallel File Systems

- **GPFS**
 - Used on various NCSA clusters
 - Some testing was done on NCSA clusters
- **Lustre**
 - Will be used on 17.7 Tflop NCSA cluster
- **PVFS**
 - Included in OSCAR



NCSA VMI and Scalability

- **Middleware communication layer**
- **Addresses availability, usability, and management for large-scale grids**
- **Stripes across heterogeneous networks**
- **Failover from one network onto a heterogeneous network**
- **Uses appropriate scaling process launch method**
 - **PBS**
 - **LFS**
 - **MPD**
- **Uses minimum memory resources**
- **Scalable process monitoring using trees**



NCSA VMI More Scalability Features

- **An VMI-MPICH GRID job consists of one or more sub-jobs.**
- **A sub-job is launched on each site using individual mpirun commands.**
- **The higher performance SAN (Infiniband or Myrinet) is used for intra site communication. Cross site communication uses TCP automatically.**
 - **It's possible to have two clusters, one with Infiniband and other Myrinet and span a GRID job across them. Infiniband will be used for intra cluster communication within the first cluster and Myrinet for communication within the second. TCP will be used for communication between the clusters.**



NCSA KFS File System and Scalability

- **Design/testing phase**
- **Uses a dynamic server pool**
- **Support for foreign KFS file systems:**
 - **Grouping**
 - **Caching**
 - **Importing**
- **Interconnect support:**
 - **Myrinet**
 - **Infiniband**
 - **GigE**
- **Distributed metadata**



NCSA Clumon and Scalability

- **Used to monitor all NCSA clusters**
- **Part of OSCAR**
- **Scalability:**
 - **Single server will be used on 1,450 node cluster**
 - **Allows cascading of servers**

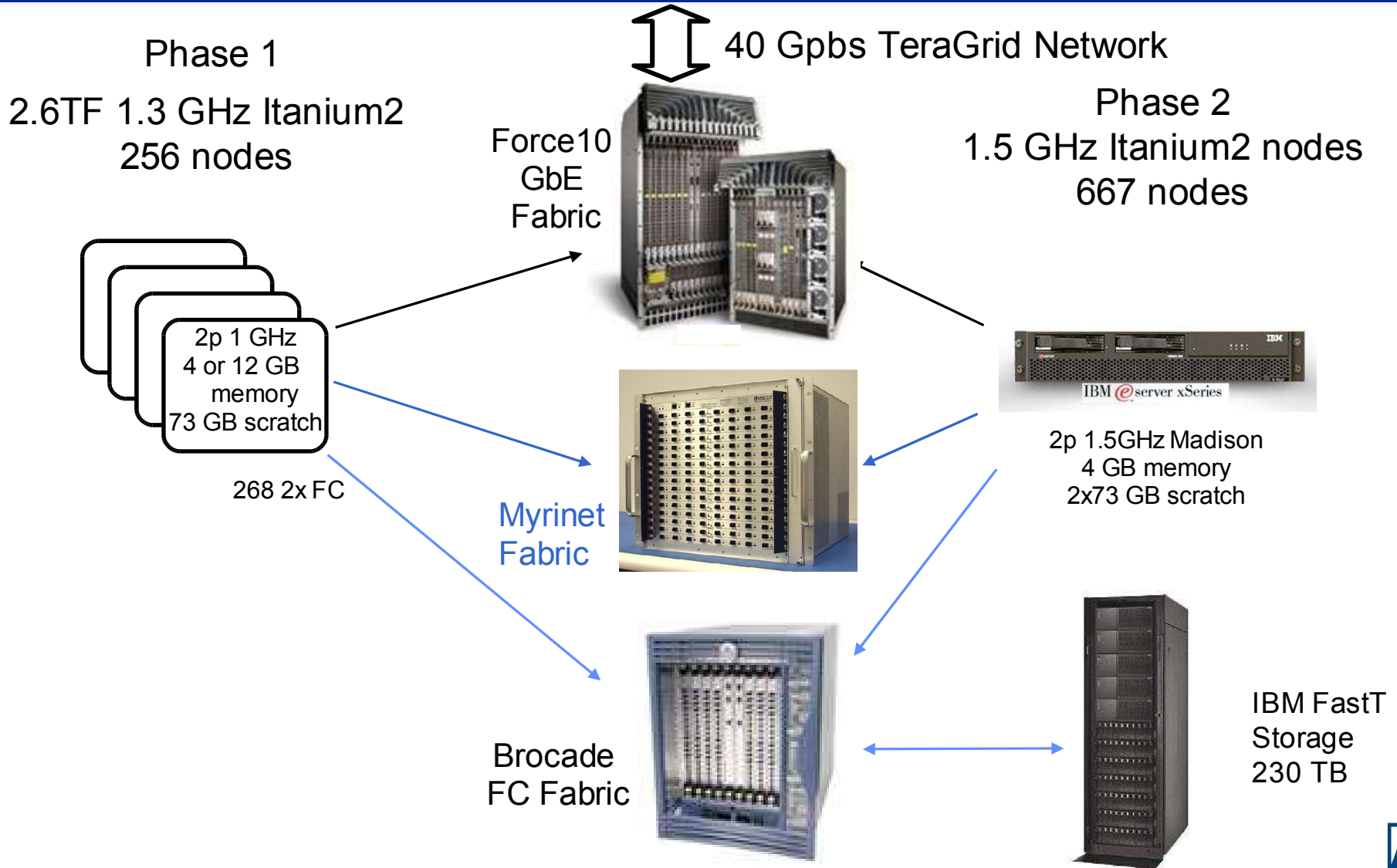


NCSA tweaks and patches

- **Non-privileged ports are used for ssh/scp**
- **Kernel patched for extreme network packet sizing**



NCSA TeraGrid Scale: 10.6 TF Itanium2/Madison



SciDAC

- **(Scientific Discovery through Advanced Computing)**
- **DOE Office of Science 5 year program launched 2001**
- **Develop Hardware and Software Infrastructure needed to use terascale computers in energy related areas including fusion and nuclear physics.**
- **Focus areas include:**
 - ...
 - **Scientific Computing Software Infrastructure (including Scalable System Software for Terascale Computer Centers)**



SciDAC Participating Organizations

- **DOE Labs, NSF Supercomputer Centers, Vendors**

ORNL

SNL

PSC

Cray

SGI

ANL

LANL

SDSC

Intel

HP

LBNL

Ames

IBM

Unlimited Scale

PNNL

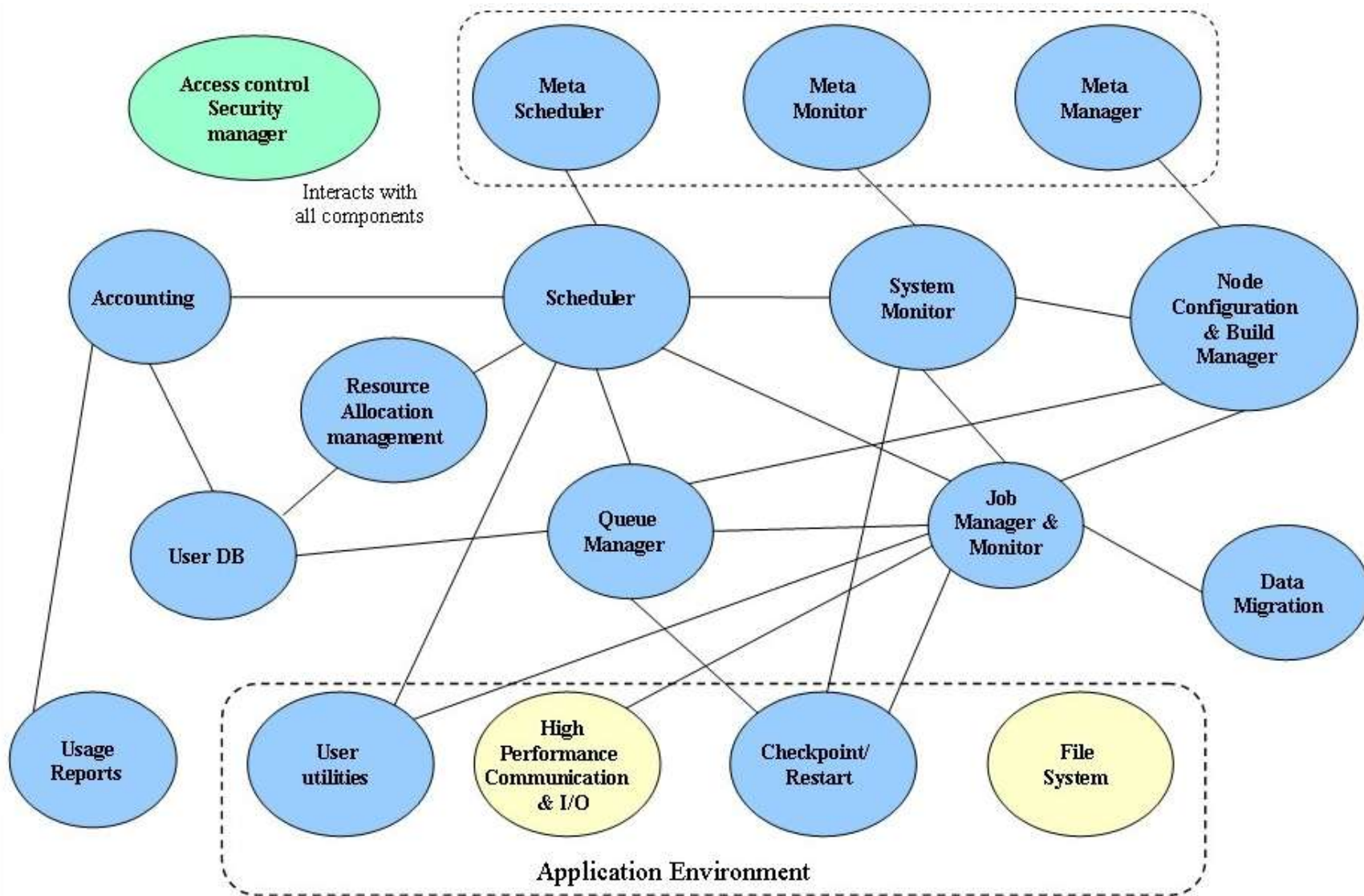
NCSA

SciDAC Scalable Systems Software Goals

- **Agree on and specify standardized interfaces between system components**
 - MPI-like process to promote interoperability, portability, and long term usability
- **Produce a fully integrated suite of systems software and tools**
 - Reference implementation for the management and utilization of terascale computational resources
- **Research and development of more advanced versions of the components**
 - To support the scalability, fault tolerance, and performance requirements of large science applications. Up to 10,000 nodes.



SciDAC Scope and Components



SciDAC Progress

- **PBS replacement being tested**
 - **Uses MPD for scalability**
- **Other components being designed and tested**
- **NCSA contributing cluster monitoring portion**
- **SciDAC components are being ported into OSCAR as OSCAR packages**



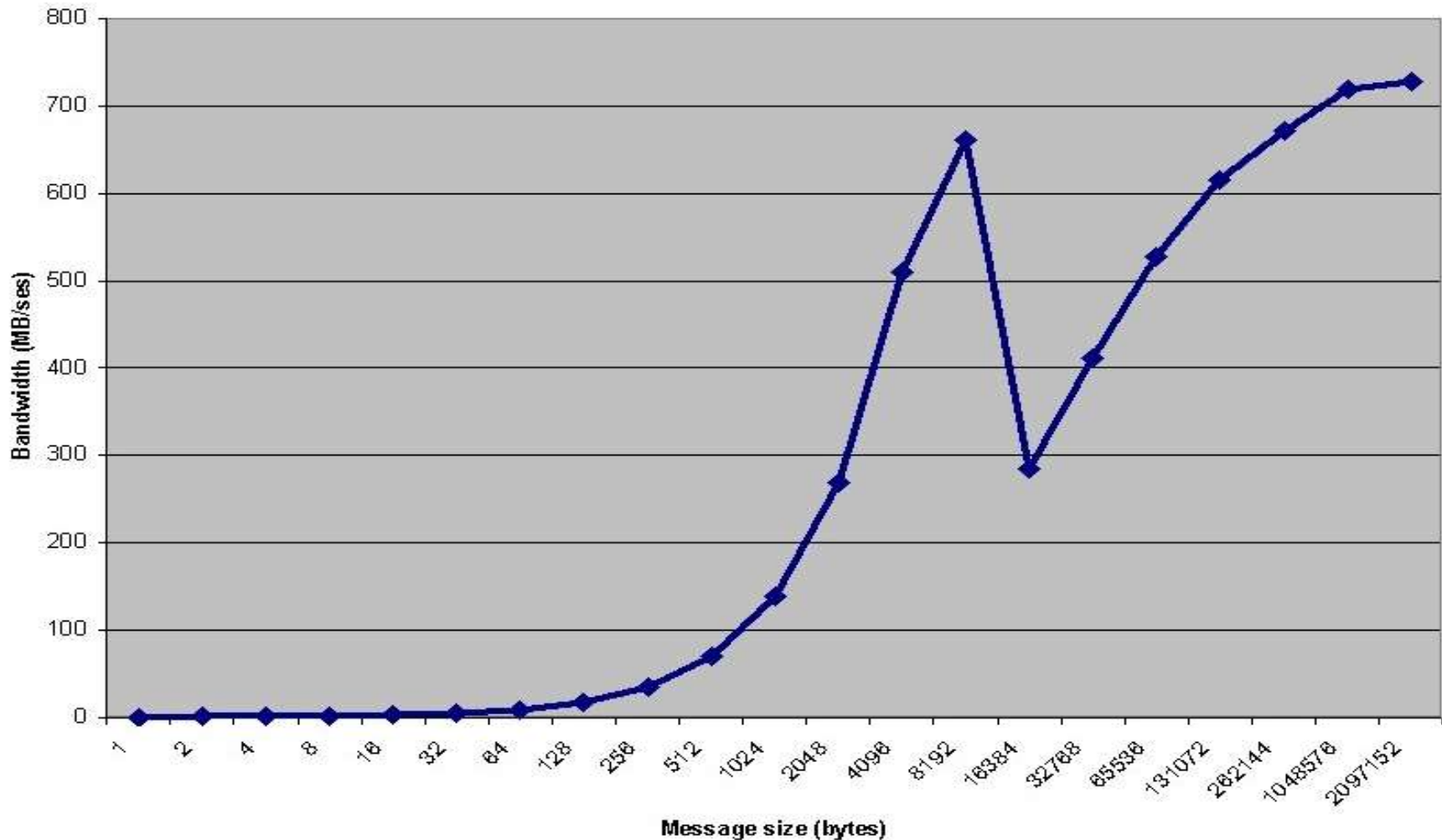
NCSA Itanium Infiniband Test Cluster

- **64 compute + 1 head node HP rx2600:**
 - dual 900 Mhz Itanium 2 processors
 - 2 GB ram
 - 36 GB disk
 - separate management and control processor with a serial interface for initial setup and a 100 BaseT network interface
 - Serial console output connected to a Cyclades concentrator
 - 100 BaseT and GigE interfaces
- **OSCAR Gold installation**
- **Mellanox 72 port Infiniband switch**
- **Mpich-vmi2 beta 3**
- **Mellanox VAPI (1.0 release)**



NCSA Measured Infiniband Bandwidth

Infiniband Bandwidth Results



NCSA Measured Infiniband Bandwidth

Size	MB/sec
1	0.142
2	0.279
4	0.568
8	1.134
16	2.249
32	4.510
64	8.894

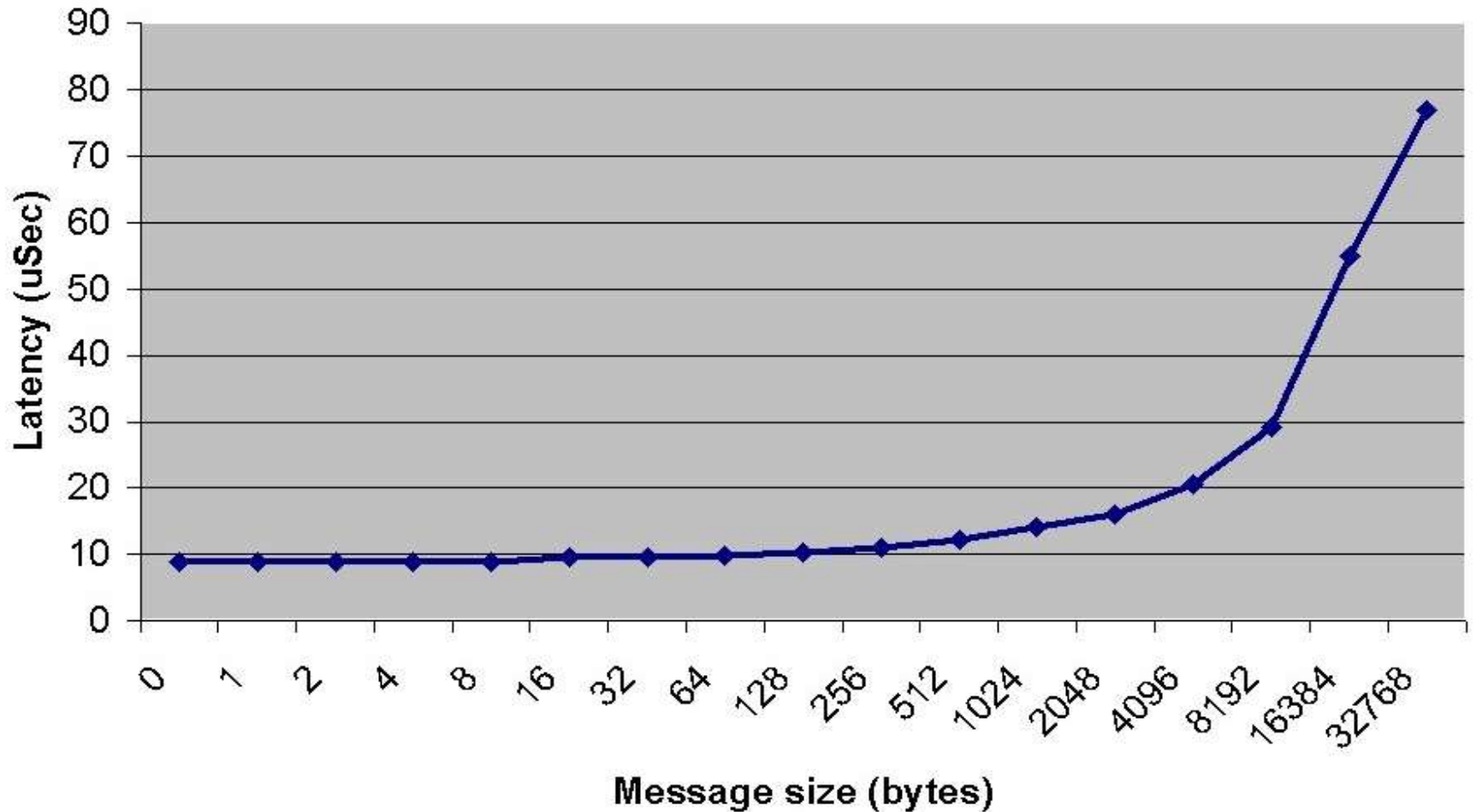
Size	MB/sec
128	17.770
256	35.228
512	69.481
1,024	139.03
2,048	268.31
4,096	509.26
8,192	661.92

Size	MB/sec
16,384	284.52
32,768	410.69
65,536	528.16
131,072	615.80
262,144	671.23
1,048,576	720.24
2,097,152	727.96



NCSA Measured Infiniband Latency

Infiniband Latency Results



NCSA Measured Infiniband Latency

Size	uSec
0	8.87
1	8.87
2	8.83
4	8.85
8	8.90
16	9.67
32	9.71
64	9.96

Size	uSec
128	10.34
256	10.99
512	12.22
1,024	14.21
2,048	16.15
4,096	20.56
8,192	29.28

Size	uSec
16,384	54.99
32,768	76.95
65,536	121.33
131,072	210.04
262,144	387.46
1,048,576	1454.54
2,097,152	2881.06



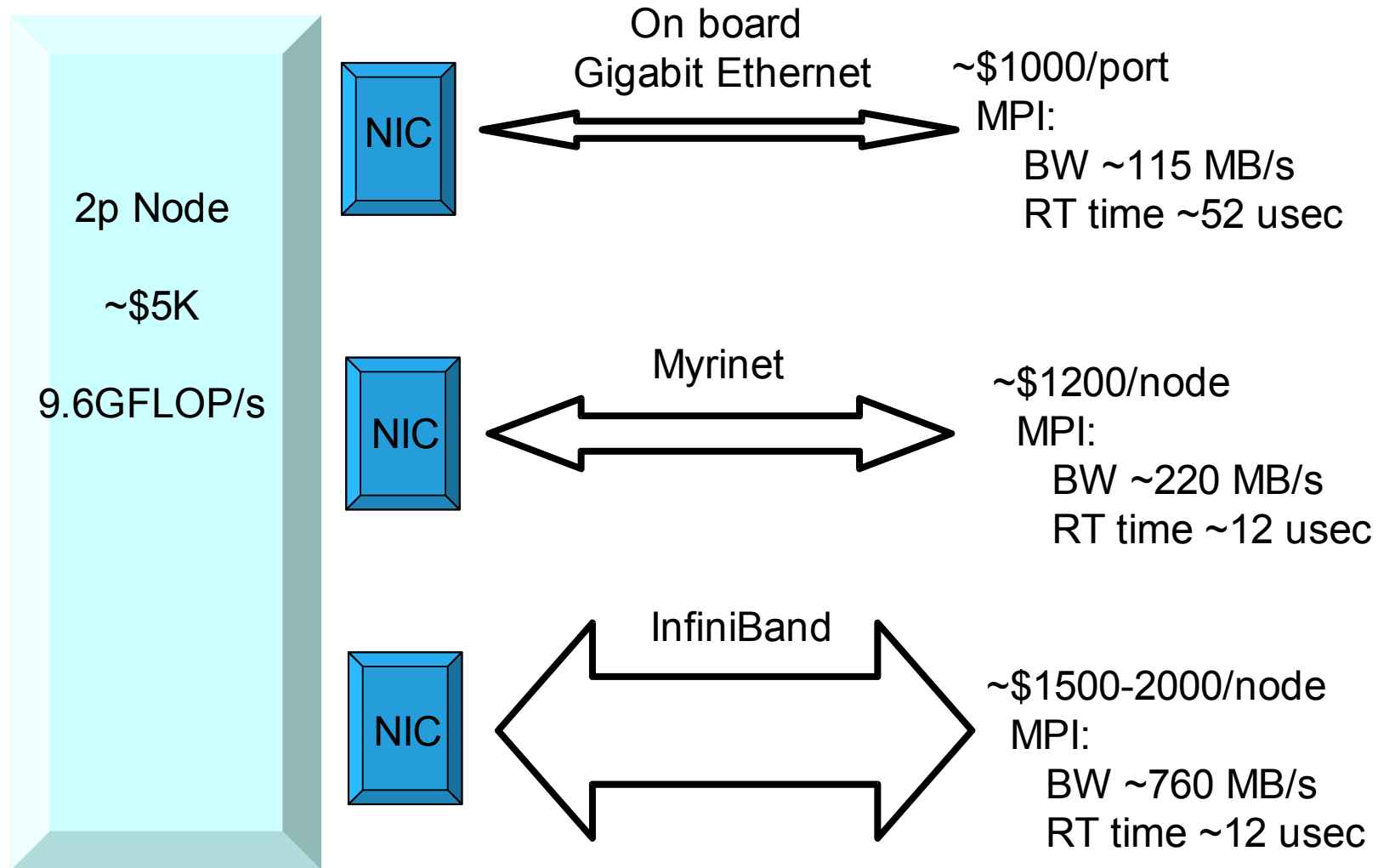
NCSA Infiniband Cluster Efficiency

- **Measured:** 392.6 Gflops
- **Theoretical:** 460.8 Gflops
- **Efficiency:** 85.2 %

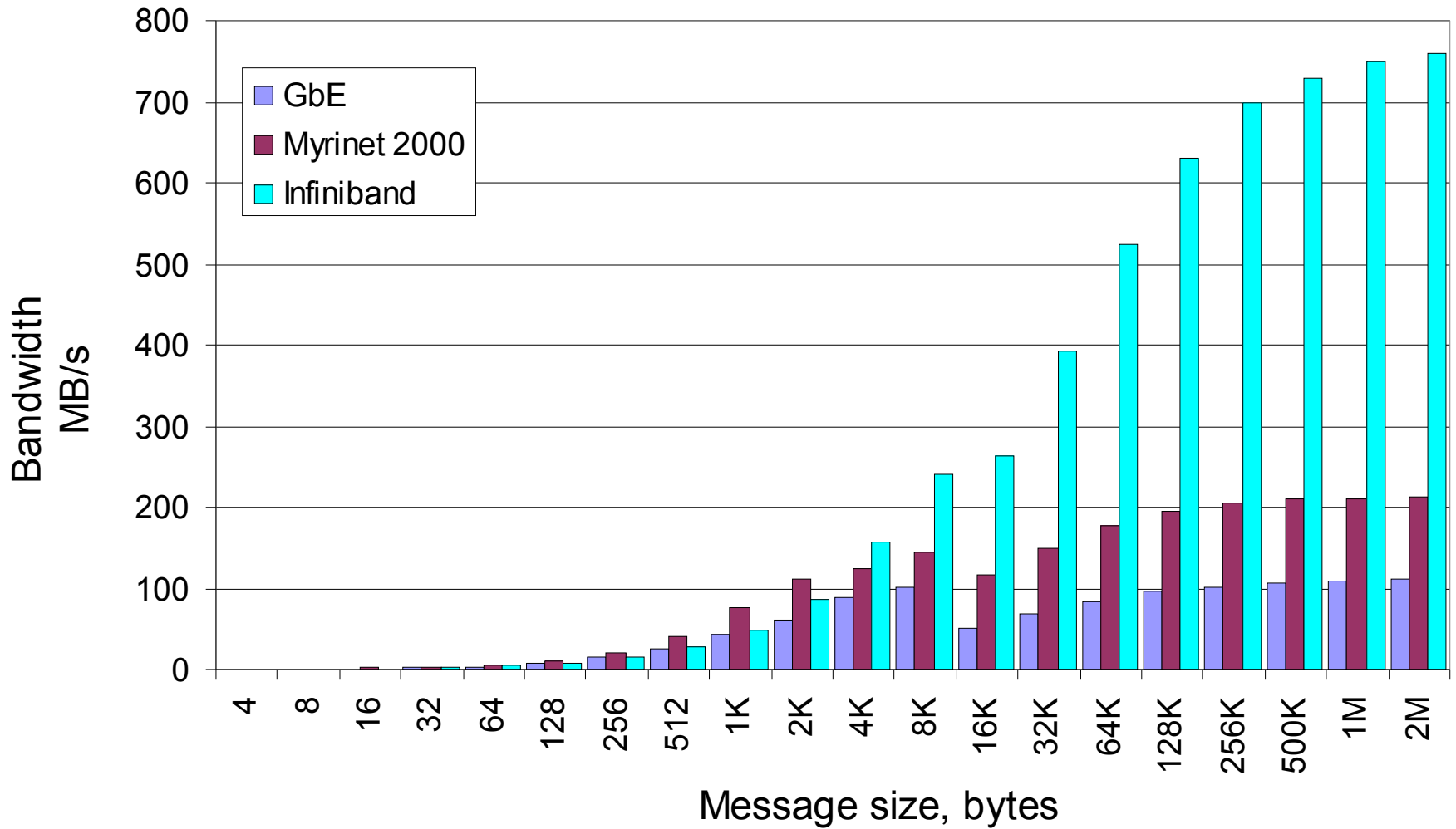
- **Top 500 list clusters:**
 - Average efficiency 49.7%
 - Best efficiency 82.6%



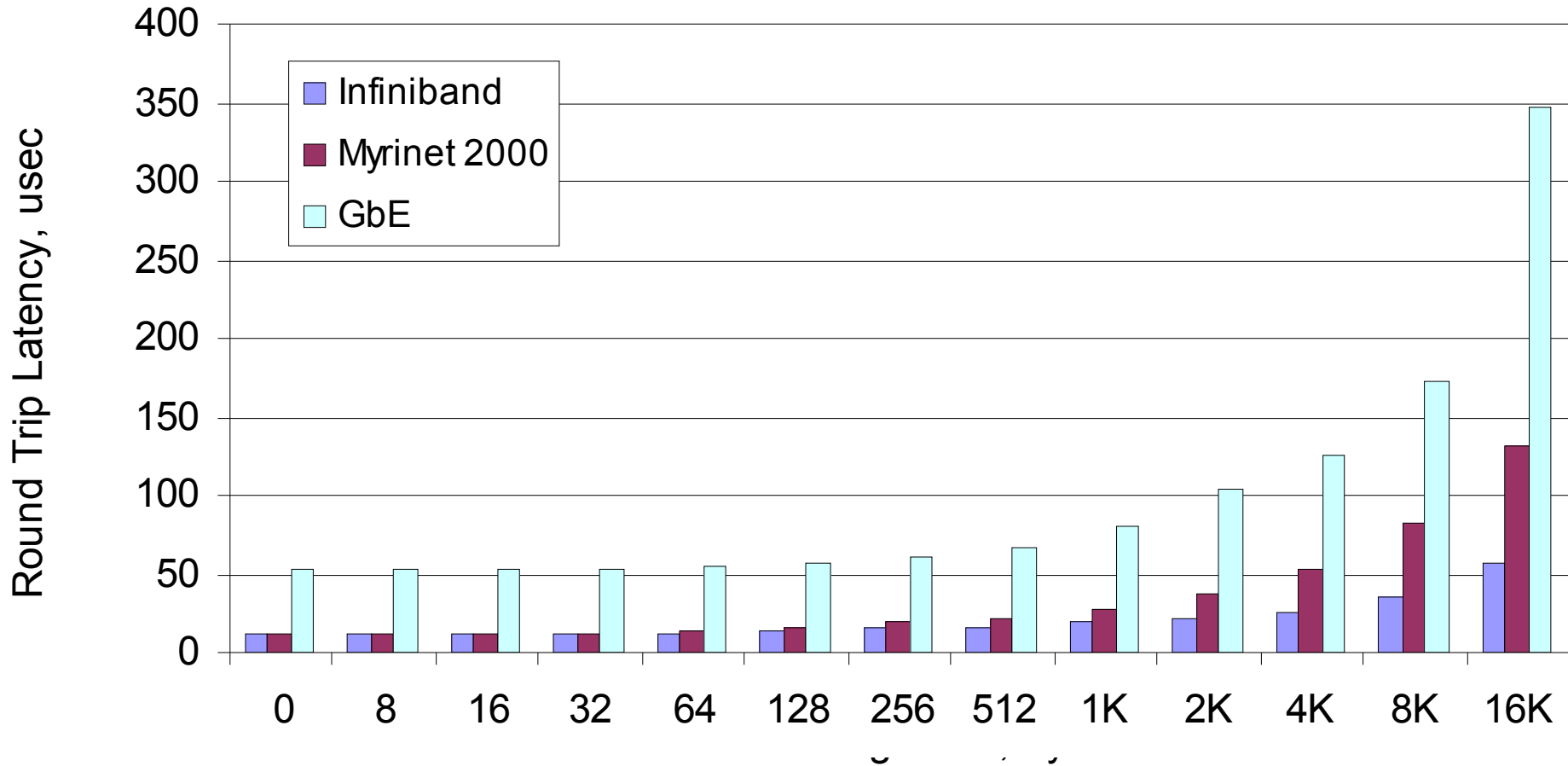
Node Interconnects, 2003



MPI Bandwidth – Delivered (Preliminary)



MPI Latency – Round trip time (Preliminary)



For Further Cluster Scalability Information

- **Contact the mailing list at:**
gelato-cluster-scal@gelato.unsw.edu.au
- **Contact me at:**
ngorsuch@ncsa.uiuc.edu
- **NCSA:** <http://www.ncsa.uiuc.edu>
- **OSCAR:** <http://sourceforge.net/projects/oscar/>
- **SciDAC:** <http://www.scidac.org/>

