# Ranger InfiniBand BOF

Tommy Minyard

Associate Director
Texas Advanced Computing Center
The University of Texas at Austin

SC08 Conference
November 18, 2008

# Ranger: Introduction

- Ranger is a unique instrument for computational scientific research housed at TACC's new machine room
- Results from over 2 ½ years of initial planning and deployment efforts
- Funded by the National Science Foundation as part of a unique program to reinvigorate High Performance Computing in the United States (*Office of Cyberinfrastructure)*

TACC

# How Much Did it Cost and Who's Involved?

- TACC selected for very first NSF 'Track2' HPC system
  - $30M system acquisition
  - Sun Microsystems is the vendor
  - All InfiniBand all the time
    - ~4100 endpoint hosts
    - >1350 MT47396 switches


- TACC, ICES, Cornell Theory Center, Arizona State HPCI are teamed to operate/support the system four 4 years ($29M)

TACC

# Ranger System Summary

- Compute power – 579.4 Teraflops
  - 3,936 Sun four-socket blades
  - 15,744 AMD “Barcelona” processors
    - Quad-core, four flops/clock cycle
- Memory - 123 Terabytes
  - 2 GB/core,  32 GB/node
  - 123 TB/s aggregate bandwidth
- Disk subsystem - 1.7 Petabytes
  - 72 Sun x4500 “Thumper” I/O servers, 24TB each
  - 50 GB/sec total aggregate I/O bandwidth
  - 1 PB raw capacity in largest filesystem
- Interconnect – 1 GB/s, 1.6-2.85 μsec latency, 7.8 TB/s backplane
  - Sun InfiniBand switches (2), up to 3456 4x ports each
  - Full non-blocking 7-stage fabric
  - Mellanox ConnectX InfiniBand

TACC

# InfiniBand Switch Configuration

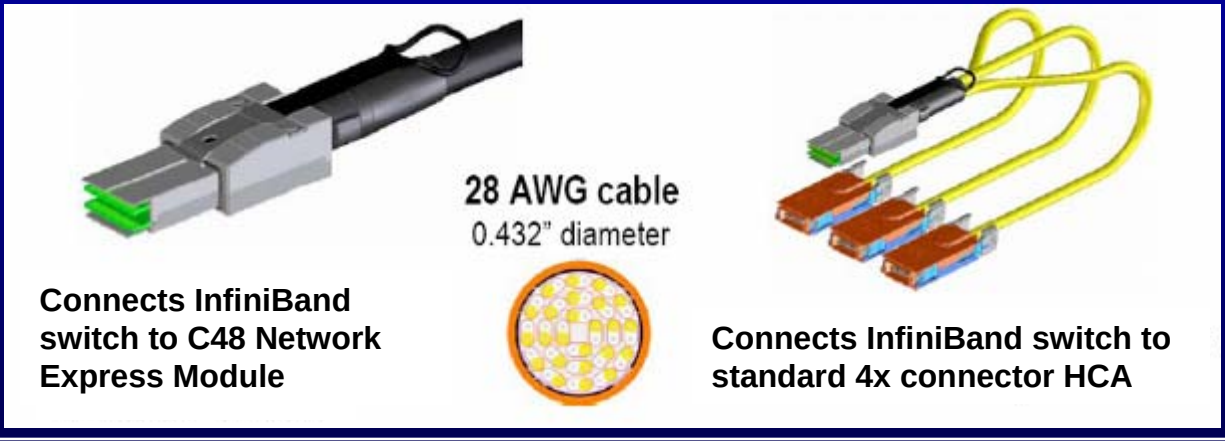
- 18 Fabric cards – 8 Infinihost3 switch chips
  - 144 Root switches
- 16 Line cards – 24 InifiniHost3 switch chips
  - 192 switches at first level
  - 192 switches at second level

# Ranger InfiniBand Configuration

- 288 Root switches
- 384 switches at rank 1
- 384 switches at rank 2
- 328 Network Express Modules (NEMs)
  - Integrated I3 switch chip
  - 12 InfiniBand Mellanox ConnectX HCAs in each

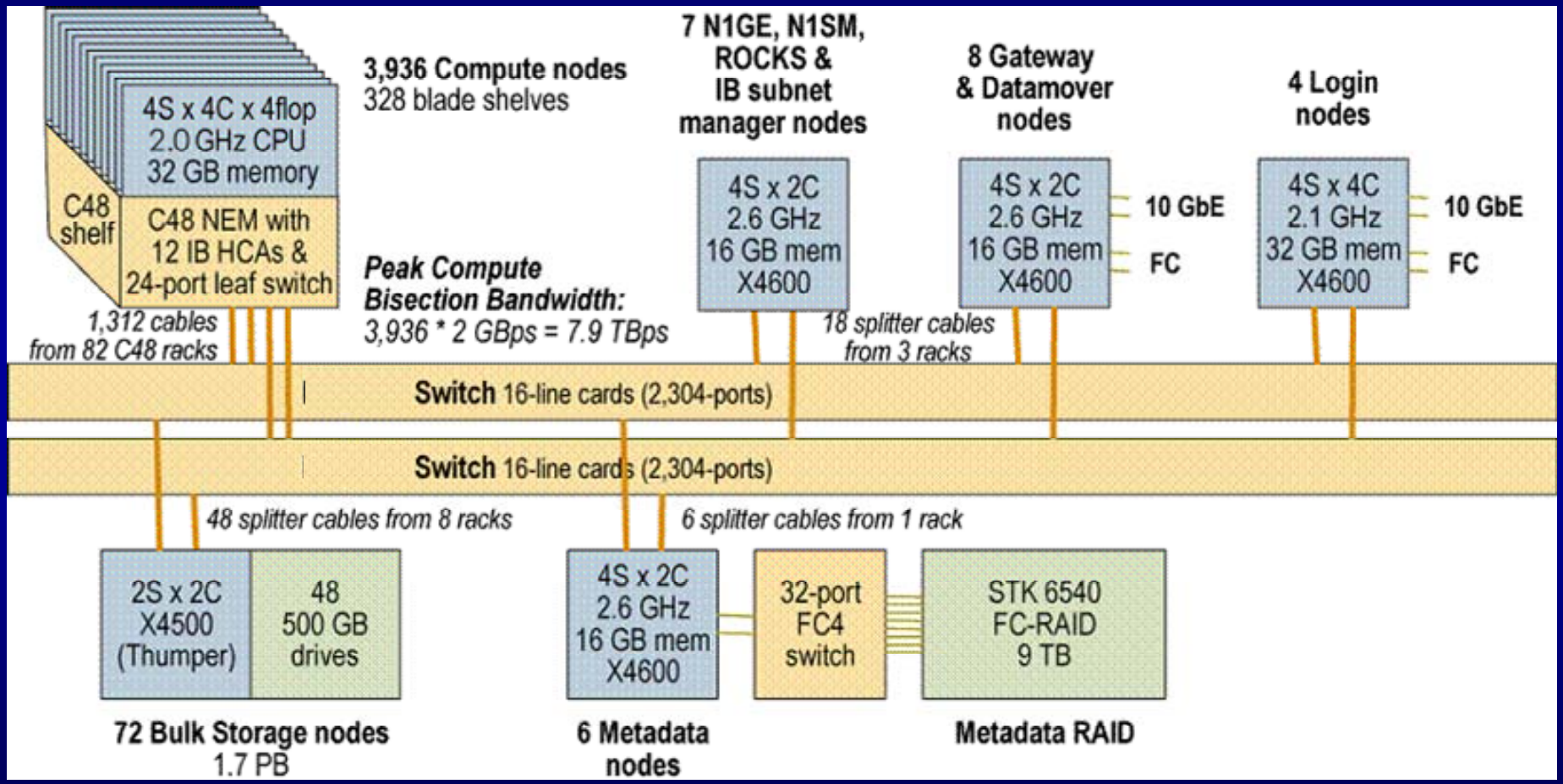
# InfiniBand Cabling for Ranger

- Sun switch design with reduced cable count, manageable, but still a challenge to cable
  - 1312 InfiniBand 12x to 12x cables
  - 78 InfiniBand 12x to three 4x splitter cables
  - Cable lengths range from 7-16m, average 11m
- 15.4 km total InfiniBand cable length

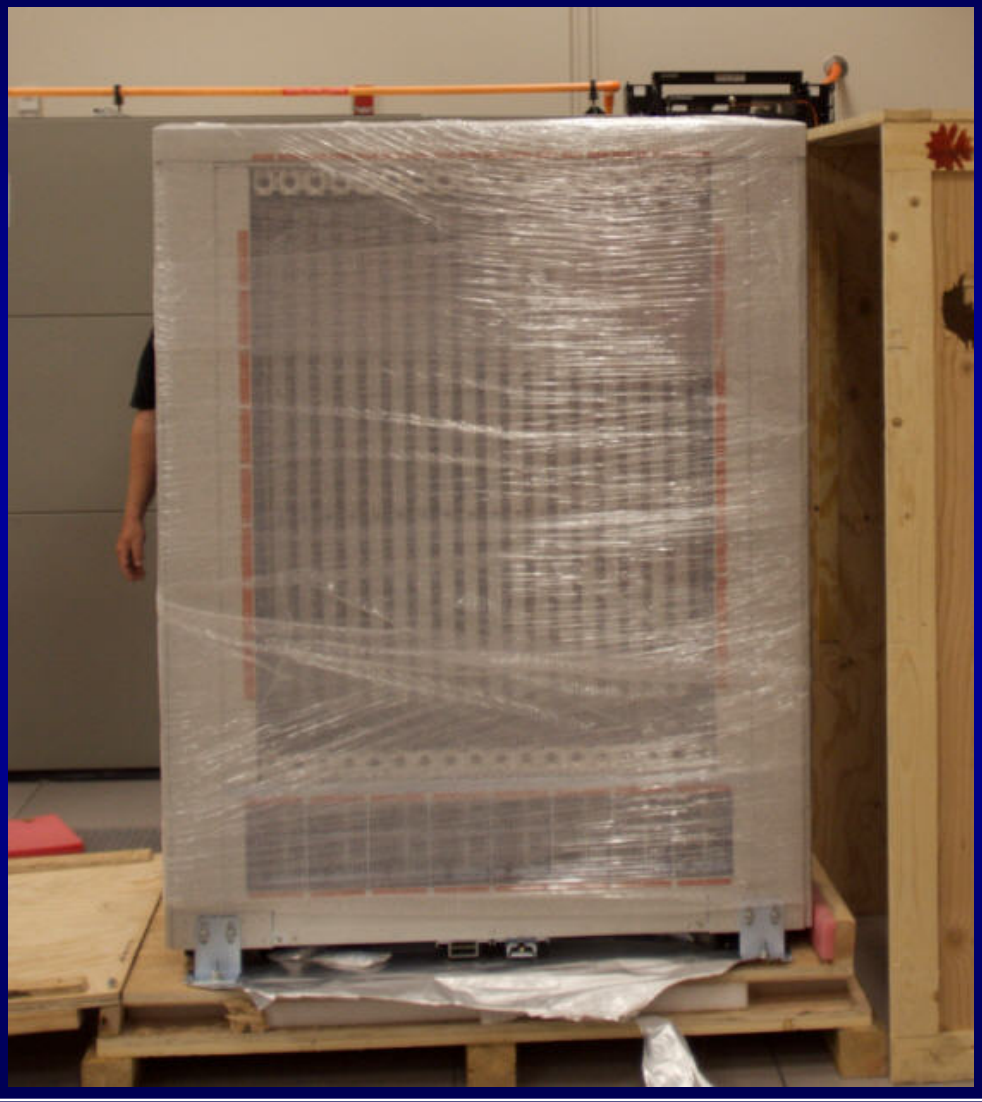



TACC

# InfiniBand Fabric Connectivity



# The switch arrives



TACC

# Switches in place


TACC

# InfiniBand Cabling in Progress


TACC

# Hot aisles enclosed


TACC

# InfiniBand Cabling Complete


TACC

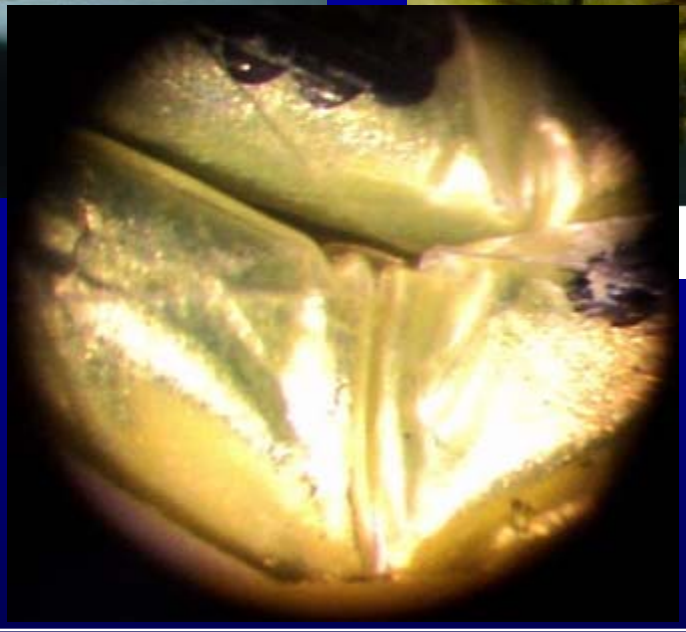
# InifinBand Hardware Deployment Challenges

- Shear quantity of components, logistics nightmare
- InfiniBand cable quality
- New hardware, firmware updates
- Scale, scale, scale

# InfiniBand Cables, Connectors and Switch Firmware

- Initial InfiniBand cables poorly manufactured
  - Attachment of inner wires to connectors created pinch points during manufacturing
  - Any bending of cable exacerbated the damage
- Early 12x IPASS connectors generating too many errors, required additional capacitor
- First IB switch chip firmware had loopback setting incorrectly enabled, caused havoc with multicast traffic and difficult to find the cause

# Damage Analysis on Initial Cables



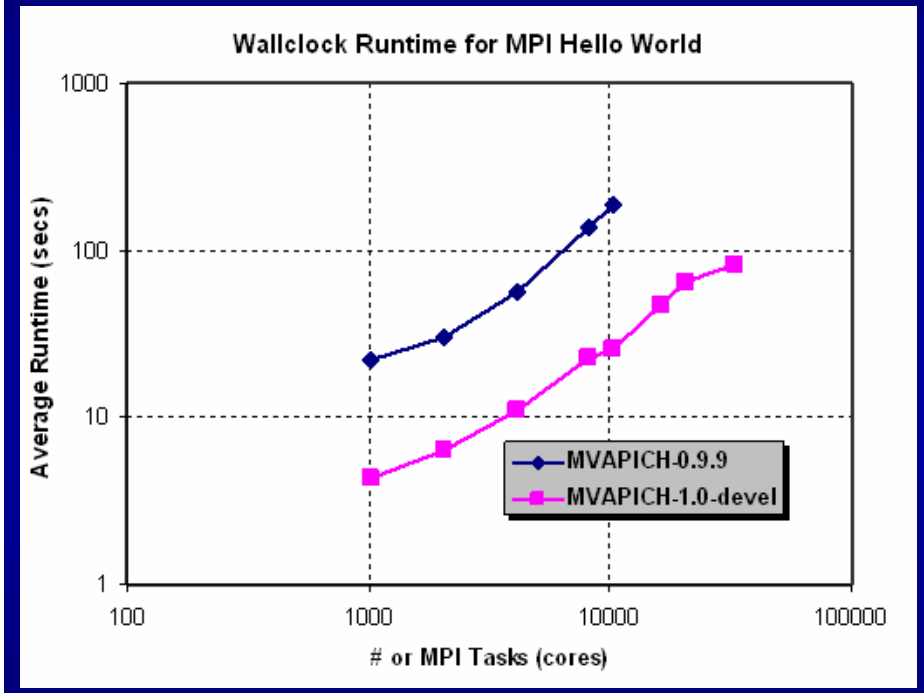
# Software Deployment Challenges

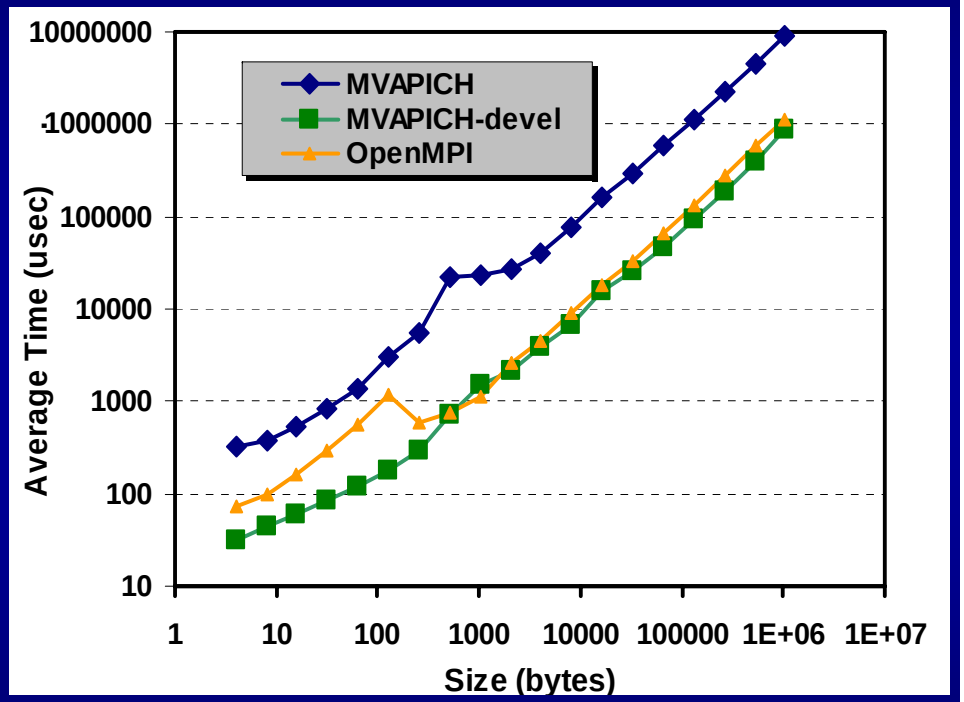
- InfiniBand Subnet manager and routing algorithms
- OFED driver stack and IPoIB
- MPI collective tuning and large job startup

# OpenSM performance

- Initial fat-tree routing on Ranger fabric taking 20 minutes to sweep and computing routing
- Mellanox support team retuned OpenSM and implemented GUID endpoint caching
- Improved fat-tree routing now 3 minutes when doing a full remap
- Endpoint caching incurs on 20-second sweep of fabric

# MPI Scalability and Collective Tuning

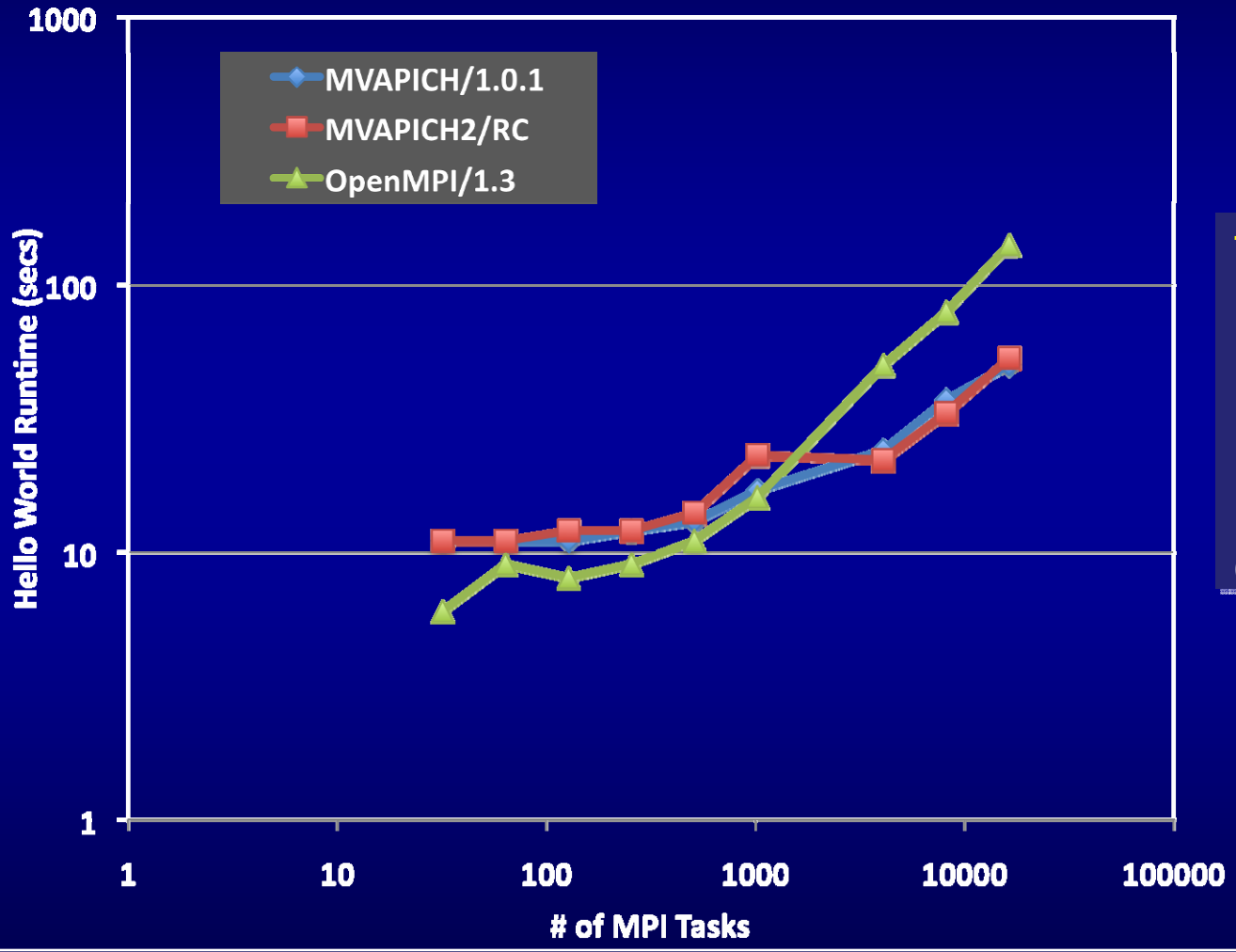



TACC

*Initial performance in Feb. 08*

# Software Challenges: Large MPI Jobs




Time to run 16K hello world:

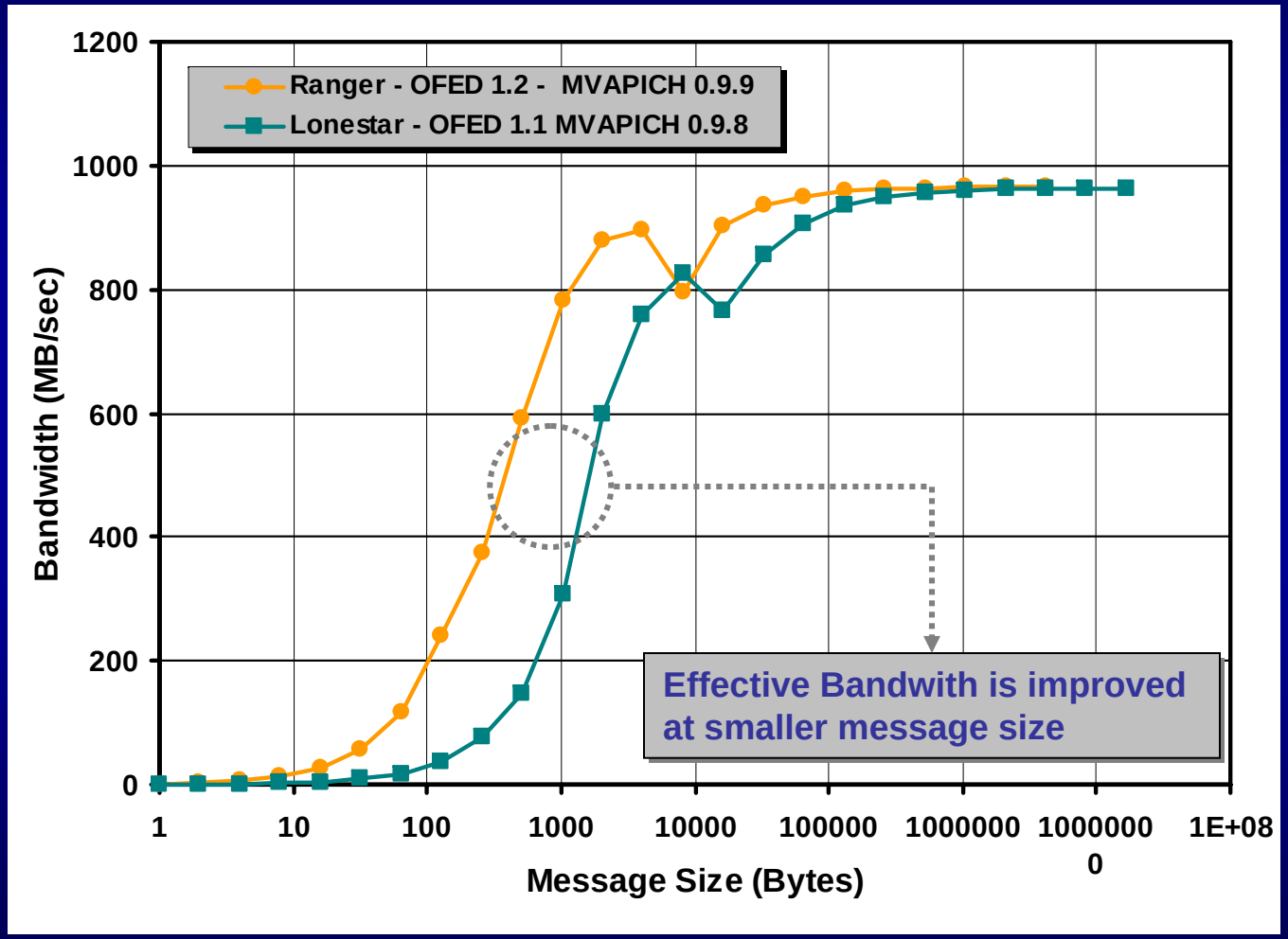
MVAPICH: 50 secs

OpenMPI: 140 secs

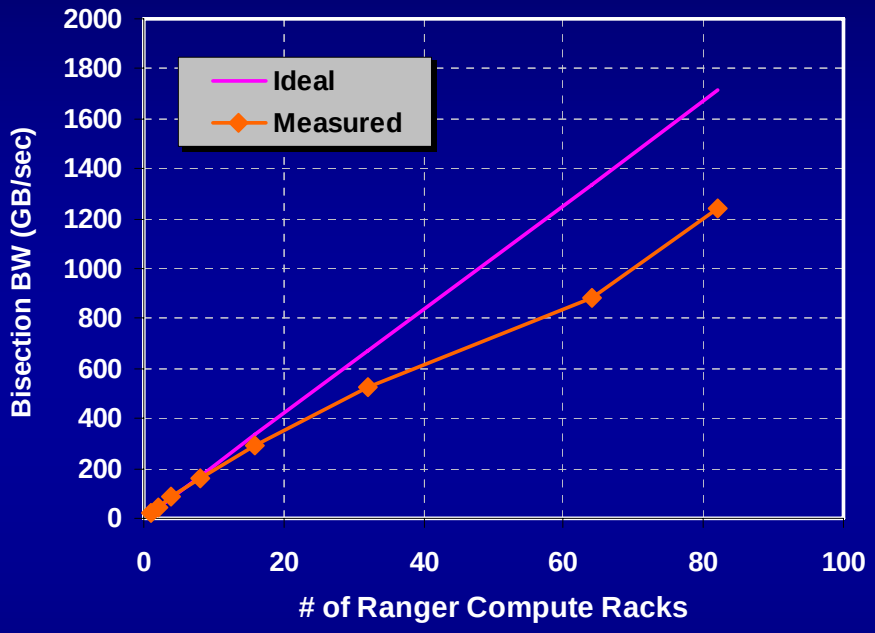
TACC

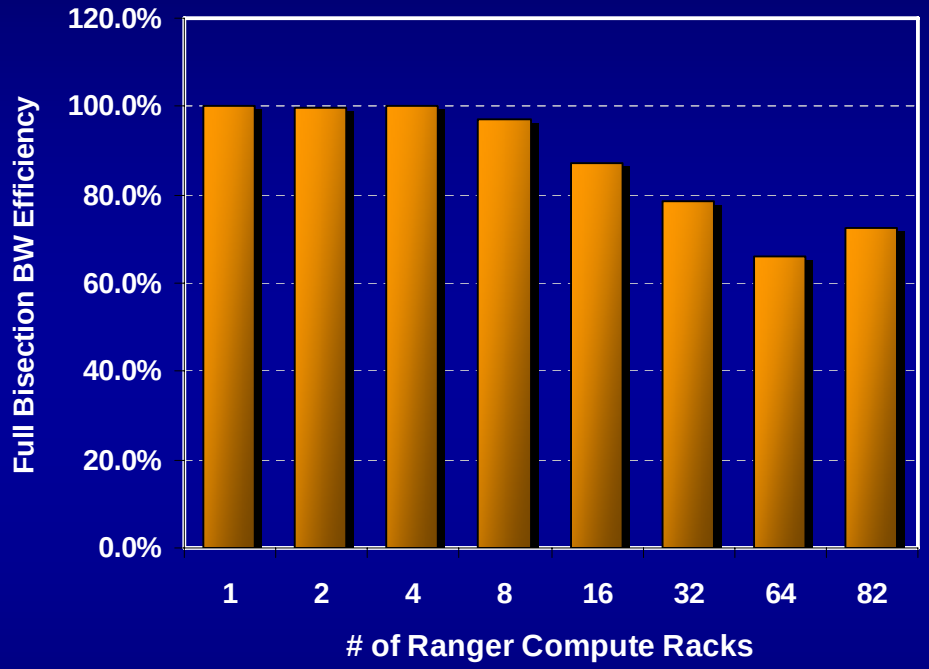
*Upgraded performance in Oct. 08*

# MPI Tests: P2P Bandwidth



# Ranger: Bisection BW Across 2 Magnums




- **Able to sustain ~73% bisection bandwidth efficiency with all nodes communicating (82 racks)**

- **Subnet routing is key! – Using special fat-tree routing from OFED 1.3 which has cached routing to minimize the overhead of remaps**

TACC

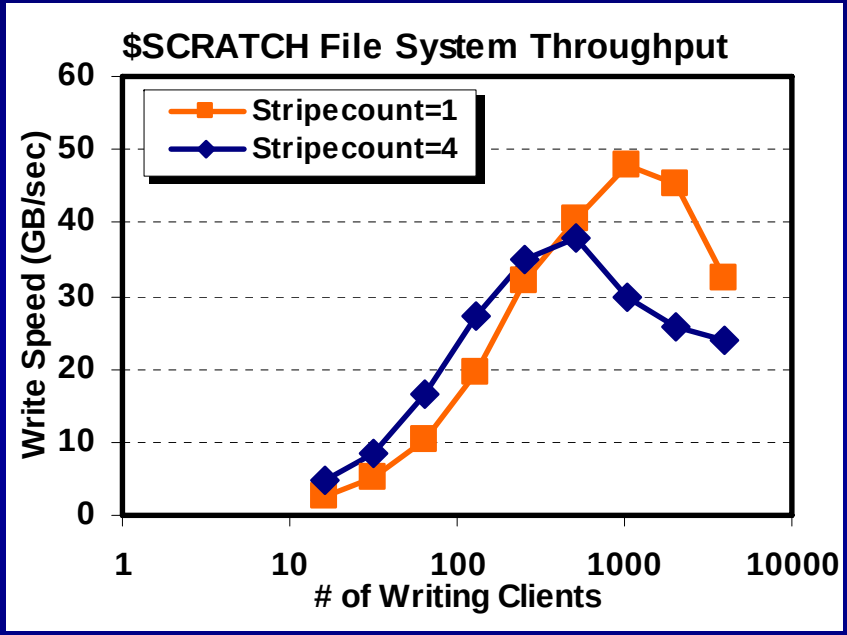
# Ranger Usage

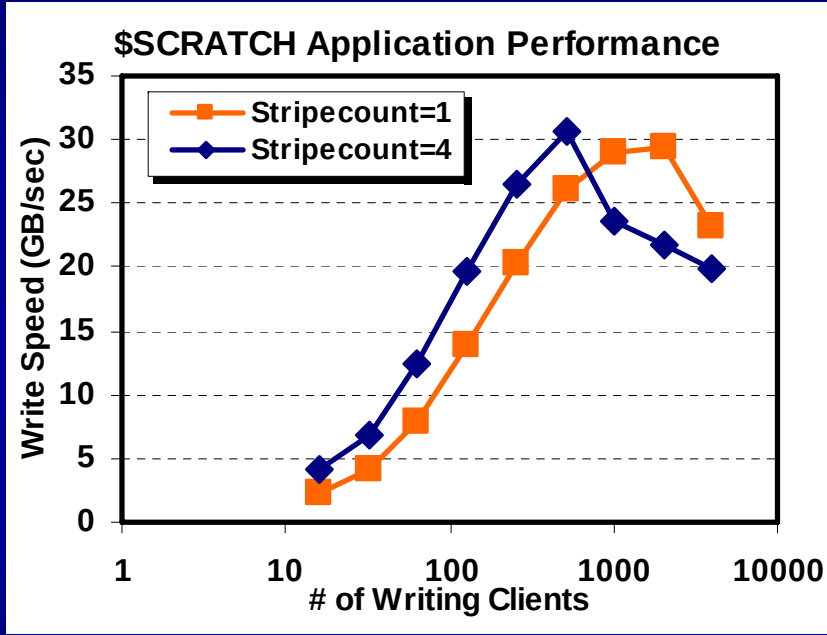
- Who uses Ranger?
  - a community of researchers from around the country (along with international collaborators)
  - more than 1500 allocated users as of Nov 2008
  - ~400 individual research projects
- Usage to Date?
  - ~300,000 jobs have been run through the queues
  - ~220 million CPU hours consumed
- How long did it take to fill up the big Lustre file system?
  - We were able to go ~6months prior to turning on the file purging mechanism
  - Steady state usage allows us to retain data for about 30 days
  - Generate ~5-20 TB a day

# Initial Production Experiences

- Demand for system exceeding expectations
- Applications scaling better than predicted
  - Jobs with 16K MPI tasks routine on system now
  - Several groups scaling to 62K processors
- Filesystem performance very good

# Parallel Filesystem Performance




**More than 600TB of data generated in 4 months of production**

**Some applications measuring 35GB/s of performance**

TACC

# Application Performance Variability Problem

- User code running and performing consistently per iteration at 8K and 16K tasks
- Intermittently during a run, iterations would slow down for a while, then resume
- Impact was tracked to be system wide
- Monitoring InfiniBand error counters isolated problem to single node HCA causing congestion
- Users don’t have access to the IB switch counters, hard to diagnose in application

# Ongoing Challenges

- Supporting multiple groups running at 32K+ cores
- Continued application and library porting and tuning
- Continued tuning of MPI
- Ensuring filesystem and IB fabric stability

# Latest Updates

- OFED 1.3.1 installed in early October
- MVAPICH 1.0.1 and MVAPICH2 1.1 installed
- Much improved MPI job startup performance and reliability
- Still resolving a few minor IB issues with NEMs and HCAs

# Summary

- Significant challenges deploying system at the size of Ranger, especially with new hardware and software
- Application scalability and system usage exceeding expectations
- Collaborative effort with many groups successfully overcoming the challenges posed by system of this scale