

TACC Overview & Lustre Experiences

Karl W. Schulz
HPC Group
The University of Texas

Lustre User Group Meeting
Miami, Florida
April 22, 2007

Outline

- TACC Overview
 - Technology Areas
 - Resources at TACC
- Lustre at TACC
 - Initial testing and evaluation
 - First production setup (*small*)
 - Planned production setup (*not so small*)

Focus Areas and Major Projects

- High Performance Computing (HPC)
 - Performance benchmarking, analysis, optimization
 - Linear algebra, solvers
 - CFD, computational chemistry, weather/ocean modeling, computational biomedicine
- Data & Information Analysis (DIA)
 - Scientific visualization
 - Data collections management, analysis & mining
- Distributed & Collaborative Computing (DCC)
 - Portals & gateways
 - Middleware for scheduling, workflow, orchestration
- **Some Example Projects**
 - TeraGrid (the US academic cyberinfrastructure for computational research)
 - NCSA, SDSC, PSC, Indiana, Purdue, Argonne, Oak Ridge, and NCAR
 - Dell Benchmarking Center
 - GotoBLAS
 - DoD application performance tuning (includes I/O)

TACC HPC & Storage Resources

LONESTAR

Dell Dual-Core Cluster
Recently Expanded
Infiniband Interconnect
Lustre File System



CHAMPION

IBM Power 5
96 CPUs, ~1 Teraflops
192 GB memory
Federation Interconnect
5 TB GPFS File System



ARCHIVE

STK PowderHorns (2)
2.8 PB max capacity
managed by Cray DMF



GLOBAL DISK

Sun SANs and
Data Direct Disk
> 50TB



TACC's Lonestar Cluster

- **Lonestar** is our main HPC workhorse
- Ranked as #12 on the November Top 500 list
- Q: Now wait a second, haven't y'all had Lonestar for a while now, how can it still be #12 on Top500?
- A: It's a bit confusing but there are many versions of Lonestar
 - **Late 90's**: Lonestar was a Cray T3E
 - **January 2004** : Lonestar was a Cray-Dell cluster (600 procs originally) interconnected via Myrinet
 - **October 2006**: Lonestar hardware swapped to blade based, dual-core Woodcrest nodes interconnected via Infiniband
 - **March 2007**: Lonestar expanded with more compute nodes and disk
 - **5840 Cores, ~62 Teraflops**
 - **11.7 TB memory**
 - **69 TB Lustre File System**

Initial Lustre Evaluation

- We began first evaluating Lustre sometime towards the end of 2005 (*we were complete newbies*)
- Attended training class in Feb. 2006
- Motivating factors:
 - seeing poor performance with other distributed file system on SATA drives
 - putting together a proposal with a fairly substantial Lustre file system (*more on that later*)
 - stability / failover support
 - allure of native high-speed interconnect support

Lustre Evaluation

- We performed comparison tests using DDN controller with SATA drives and exact same I/O node hardware (*March 2006*)
- GigE tests from client (big and small file ops):
 - **Test 1**: Write 16GB file from 1 client
 - **Test 2**: Write 4GB files from 4 clients (*parallel*)
 - **Test 3**: Write 10000 64K Files, 500 @ a time (*parallel*)
 - **Test 4**: Write 10000 64K Files (*serial*)

	Lustre	Other	Result
Test 1: Write Speed (MB/S)	94.11	43.32	Lustre 117% faster
Test 2: Write Speed (MB/S)	83.45	32.28	Lustre 159% faster
Test 3 : Average Time (secs)	84.25	90.38	Lustre 7% faster
Test 4 : Average Time (secs)	85.66	82.60	OTHER 4% faster

TACC's Lonestar Cluster - \$WORK

- The latest dual-core version of Lonestar is the first production resource at TACC to run Lustre
- Serves as our \$WORK parallel file system for staging of all jobs
- \$WORK is not quota'd
- \$WORK is purged on a 10-day policy

Original Configuration

- Raw Disk
 - DDN S2A 9500 couplet
 - 8 port FC-4 controller
 - 146 300GB FC drives
- I/O Servers
 - 16 Dell 1850 servers
 - 2S Xeon (3.2 GHz)
 - PCI-E Mellanox HCAs
 - Initially had 1 OST per OSS

Updated Configuration

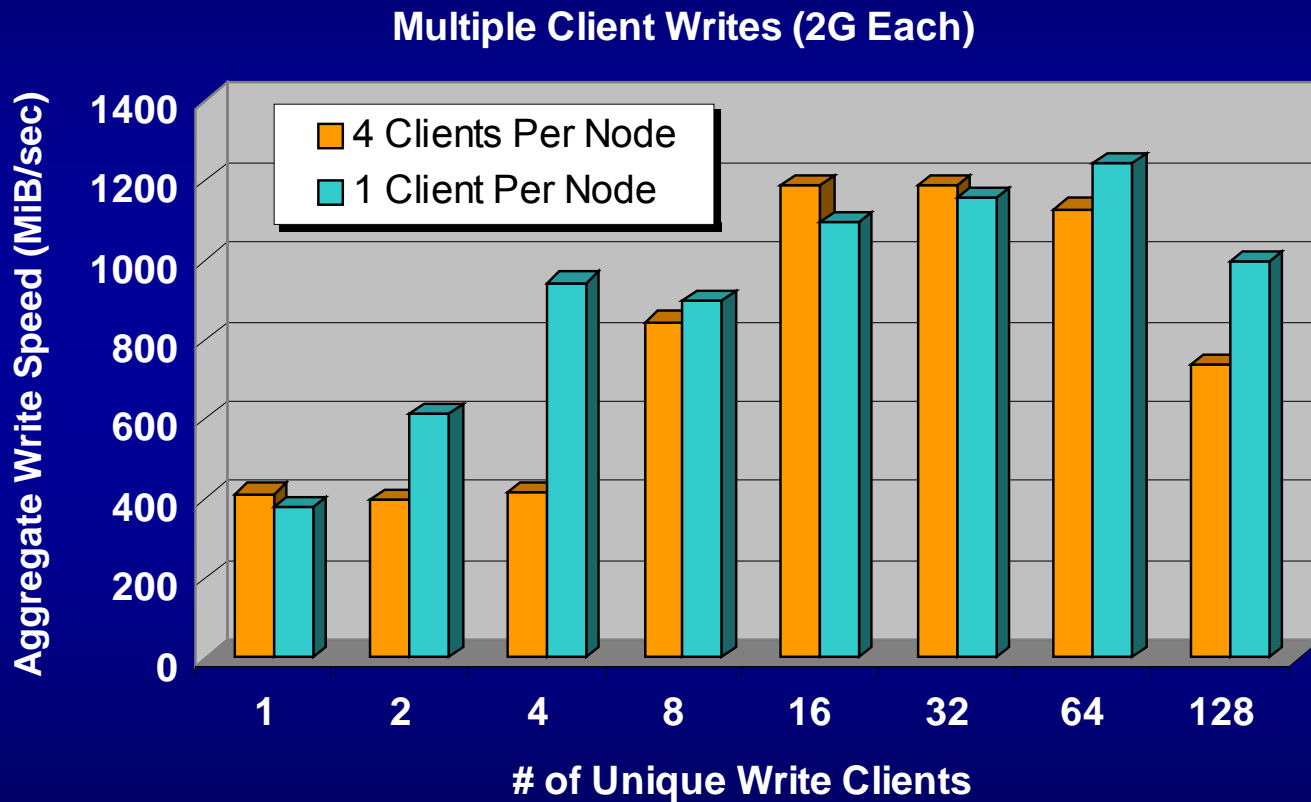
- Added another S2A 9500
- 2 OSTs per OSS

TACC's Lonestar Cluster - \$WORK

- Originally went into production with GigE only
- Intent was always to use raw Infiniband but this was problematic originally (more on that in a sec)
- To compromise, we switched to using a blend of GigE and IPoIB - *this seemed to work quite well* (eg. Large file writes from single client):
 - GigE (2-stripes): 97.9 MB/sec - *Initial Production*
 - GigE/IPoIB (2-stripes): 202 MB/sec
 - GigE/IPoIB (4-stripes): 380-400 MB/sec - *Amended Production*

Lonestar Cluster - IOR

Measurements made in production



TACC's Lonestar Cluster - \$WORK

- Experiences with gen2 IB Lustre mount (o2iblnd)
 - we initially kept running into stability and performance problems when using native IB w/ OFED
 - Initial single-client tests were fine, but user application tests would act quite “squirrely”
 - Files appeared to write and flush quickly but had problem completing the close (*eg. might take 2-5 minutes to close individual files*)
 - More than a single user running jobs was generally sufficient to lockup a few OSS servers
- Consider the following trace - a beer is on me if you know what the problem is

MPI Application Write/Close Stalls

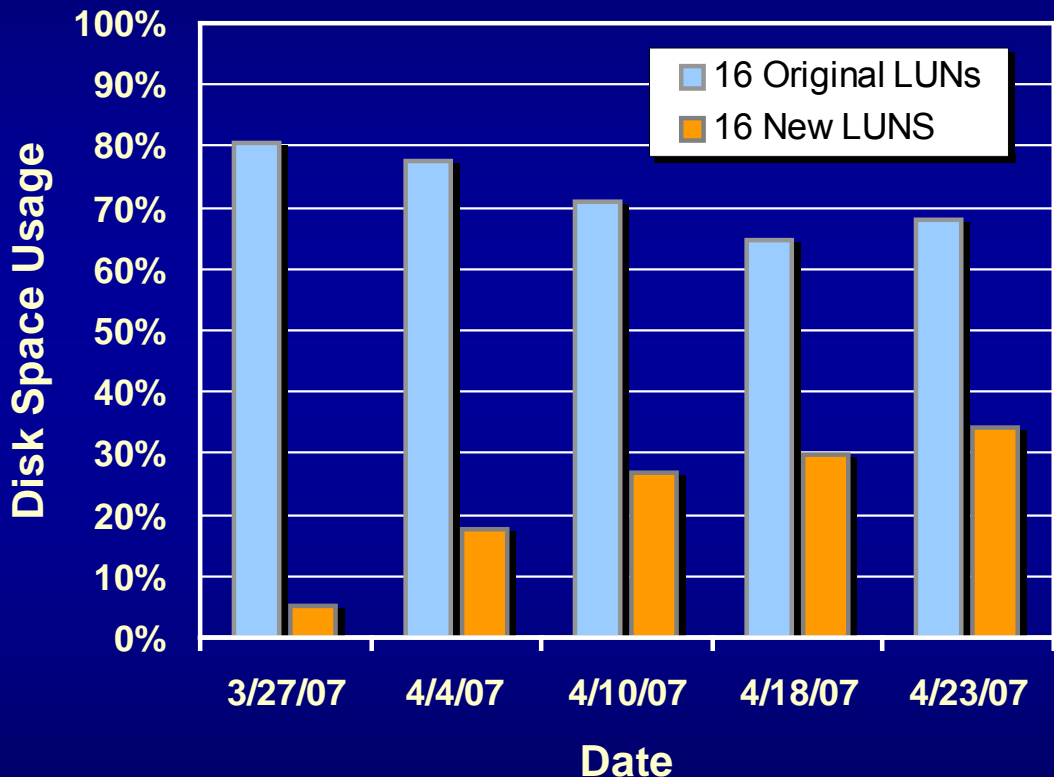
```
Program received signal SIGINT, Interrupt.
[Switching to Thread 46912510182240 (LWP 30376)]
0x00002aaaab1dbc01 in intra_RDMA_barrier () from ...mvapich-gen2/libmpich.so.1.0
(gdb) where
#0  0x00002aaaab1dbc01 in intra_RDMA_barrier () from ...mvapich-gen2/libmpich.so.1.0
#1  0x00002aaaab030ff7 in intra_Barrier () from ...mvapich-gen2/libmpich.so.1.0
#2  0x00002aaaab1b802b in MPI_Barrier () from ...mvapich-gen2/libmpich.so.1.0
#3  0x00002aaaab1b80d4 in mpi_barrier_ () from ...mvapich-gen2/libmpich.so.1.0
#4  0x00000000004323e6 in write_separate_restart_ ()
#5  0x000000000040f817 in MAIN__ ()
#6  0x000000000040507a in main ()
(gdb) cont
Continuing.
```

- Note: IB performance (b/w and latency checked out to all data servers); no problems with IPoIB mounts
- And the answer is.....?

...raw HCA firmware (doh)

Lonestar File System Expansion

- \$WORK file system doubled to 69TB ~ 1 month ago (16 new LUNS driven by same hardware)
- Original usage shown in blue
- Usage on new Luns shown in orange
- Demand remains strong, hope to add another 35 TB this summer



Transition to a Larger Lustre Deployment

baby steps to the elevator...

First NSF Track2 System: 500+ Tflops!

- TACC selected for first NSF 'Track2' HPC system
 - \$30M system acquisition
 - Sun is the vendor
 - Competed against almost every open science HPC center
- TACC, ICES, Cornell, ASU operating/supporting system four 4 years (\$29M)



Team Partners & Roles

- **TACC / UT Austin:** project leadership, system hosting & ops, user support, apps collaborations, tech evaluation & dev
- **ICES / UT Austin:** apps collaborations, algorithm/technique dev & transfer
- **Cornell Theory Center:** large-scale data management & analysis, user support
- **Arizona State HPCI:** tech evaluation & dev, user support

Ranger System Configuration

- **Compute Power** - 529 Teraflops Peak Performance
 - 3,936 Sun four-socket, quad-core compute nodes
 - 15,744 AMD Opteron “Barcelona” processors
 - Quad-core, four flops/cycle (dual pipelines)
- **Memory**
 - 2 GB/core, 32 GB/node, 125 TB total
 - 132 GB/s aggregate bandwidth
- **Infiniband interconnect**
 - Full non-blocking 7-stage Clos fabric
 - Low latency ($\sim 2.3 \mu\text{sec}$), high-bandwidth ($\sim 950 \text{ MB/s}$)

Impact in NSF TeraGrid

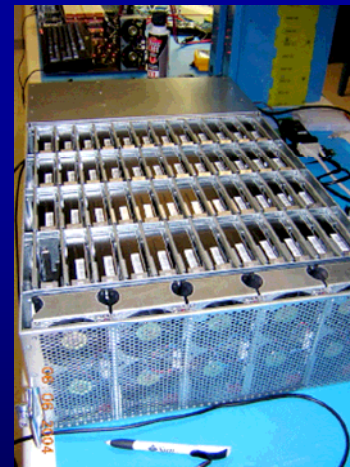
- 460M CPU hours to TeraGrid per year
 - more than double current total capacity of all TG HPC systems
 - 1.8 Billion CPU hours over operational life
- 529 Teraflops peak
 - 2x total performance of all TG HPC systems
 - 8x top TG HPC system in performance, memory, disk
- Balanced, general-purpose capability system
 - More than 60,000 cores available
 - Unprecedented scaling opportunities for computational science and research
- Re-establish NSF as a leader in HPC
- *Jumpstarts progress to petascale for entire US academic research community*

Ranger User Environment

- The overall look and feel of *Ranger* from the user perspective will be very similar to our current Linux cluster
 - Full Linux OS w/ hardware counter patches on login and compute nodes (2.6.12.6 is starting working kernel)
 - Lustre File System
 - \$HOME, and multiple \$WORKS will be available
 - Largest \$WORK will be ~1PB total
 - Standard 3rd party packages
 - Infiniband using next generation of Open Fabrics
 - MVAPICH and OpenMPI (MPI1 and MPI2)
- Suite of compilers
 - Portland Group PGI
 - Sun Studio
 - PathScale
 - *Possibly the Intel compiler*

Ranger Disk Subsystem - *Lustre*

- Disk system (OSS) is based on Sun x4500 “Thumper” servers - similar to TiTech installation
 - Each server has 48 SATA 500 GB drives (24TB total) - running internal software RAID
 - Dual Socket/Dual-Core Opterons @ 2.6 GHz
 - Downside is that these nodes have PCI-X - raw I/O bandwidth can exceed a single PCI-X 4X Infiniband HCA
 - **72 Servers Total: 1.7 PB raw storage**
- Metadata Servers (MDS) based on Sun Fire x4600s
- MDS is Fibre-channel connected to 9TB Flexline Storage
- Target Performance
 - Aggregate bandwidth: 40 GB/sec



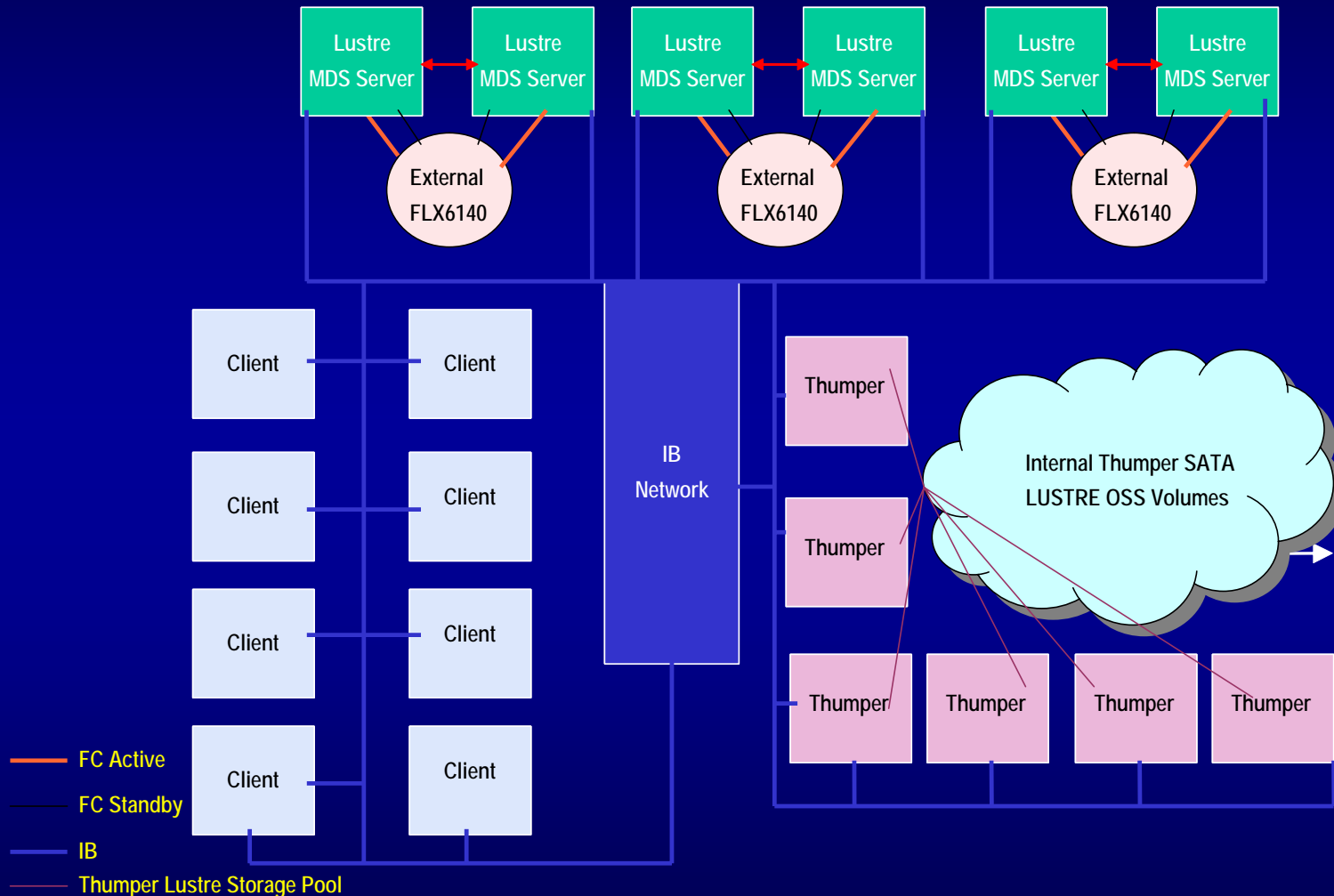
Design:

- Top loading Disks
- Front to rear airflow
- Redundant fans
- Passive Backplane
- No wires in box

Reliability/Availability

- Enterprise class SATA disks
- 1M hours MTBF
- RAID 0, 1, 5, 10
- Redundant Power
- Hot-swap FRUs

Ranger Disk Subsystem - *Lustre*



Ranger System Configuration

*At this scale, parallel file systems are universally required
Lustre and Sun X4500's are used for all volumes*

Logical Volume Name	Estimated Raw Capacity	Target Usage
<i>WORK1</i>	1 PB	Large temporary storage; not backed up, purged periodically
<i>WORK2</i>	~500 TB	Large allocated storage; not backed up, quota enforced
<i>PROJECTS</i>	2 TB	Repository for TeraGrid Community Software
<i>HOME1</i>	2 TB	Permanent user storage; automatically backed up, quota enforced
<i>HOME2</i>	2 TB	Permanent user storage; automatically backed up, quota enforced
<i>HOME3</i>	2 TB	Permanent user storage; automatically backed up, quota enforced

Ranger Space, Power and Cooling

- System Power: 3.0 MW total
- System: 2.4 MW
 - ~90 racks, in 6 row arrangement
 - ~100 in-row cooling units
 - ~4000 ft² total footprint
- Cooling: ~0.6 MW
 - In-row units fed by three 400-ton chillers
 - Enclosed hot-aisles
 - Supplemental 280-tons of cooling from CRAC units
- Observations:
 - Space less an issue than power
 - Cooling > 25kW per rack difficult
 - Power distribution a challenge, more than 1200 circuits

Ranger Project Timeline

Sep06	award, press, relief, beers
1Q07	equipment begins arriving
2Q07	facilities upgrades complete
3Q07	very friendly users
4Q07	more early users
Dec07	production, many beers
Jan08	allocations begin

Note: all US academics are eligible to apply for a TeraGrid/Ranger allocation:

<https://pops-submit.ci-partnership.org/>

Ranger: PDUs and In-Row Coolers



- APC In-row coolers are installed and plumbed
- Compute Racks will slide in between the coolers which are heat exchangers drawing from the hot aisles, exhausting ambient into the cold aisles

Overall Impressions

- ✓ Performance on SATA was much improved over our previous clustering file system
- ✓ Failover support has been very good and quite helpful in production operation
- ✓ Greatly appreciate the open-source availability to patch into our local kernels - we would not have been able to deploy our most recent cluster without it
- ✓ We've been in full production with Lustre for ~ 6months - users are generally quite happy - only outstanding complaints exist for small file reads/writes and a new problem involving direct access Fortran reads
- ✓ Now running Lustre with native o2iblnd with OFED 1.1 (~480-520 MB/sec client write for 8-stripes)- *hooray!*

Overall Impressions (cont)

- Logging issues - as relatively new users to Lustre, one of the things we struggle with is the interpretation of syslog:
 - ie. what errors should we really care about vs. which can be summarily ignored
- ✘ **Quotas:**
 - we normally endeavor to provide a large quota on \$WORK to avoid a single user from filling it up
 - we were a wee bit surprised when we turned on quotas for a user with a few hundred Gigs of existing storage :-)
- ☺ **Of Keen Interest (Ranger related):**
 - Network raid for our Thumper-based storage to handle failover
 - RAID 6 improvements
 - Multiple HCA support for o2ib (remove some network limitations)

Thanks for your time.

Questions?

karl@tacc.utexas.edu

Shameless Plug:

IEEE Cluster 2007 is in Austin in September (www.cluster2007.org)

Parallel File Systems and I/O Libraries always a good topic!

Papers Due: May 11, 2007