



Florida Site Report

US CMS Tier-2 Facilities Workshop

April 7, 2014

Bockjoo Kim
University of Florida



Outline

- **Site Overview**
- **Computing Resources**
- **Site Status**
- **Future Plans**
- **Summary**



Florida Tier-2

- Paul Avery (PI)
- Bockjoo Kim (Florida T2 Manager)
- Yu Fu (T2 and HPC Liaison, T2 Admin)
- Dimitri Bourilkov (Lustre/Analysis Support)
- Florida Tier-2 Facility/Hardware Management:
U of Florida IT and Research Computing

UF HiperGator Supercomputing



- 21,000 cores, 4GB RAM/ core
- 100Gbps WAN
- 55Gbps FDR Infiniband Intercon.
- 40Gbps HNAT for workers
- 25k ft² Data Center
- Battery-backed central UPS
- 72 HR emergency power+backup
- 600 kW PDUs (Will add 800kW)
- 600 Ton Cooling



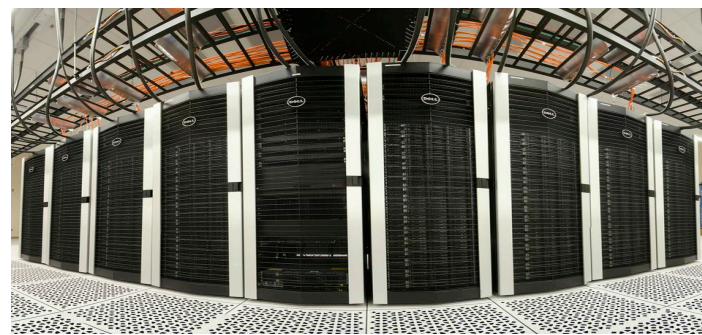
Plenty of Room for Growth

4/7/2014



Computing Power

- **U of Florida HiperGator : 21044 cores**
 - Majority is Opteron 6378 @ 2.40GHz (64 cores per node)
 - Torque-Moab Batch System
 - HSA06 : 184889



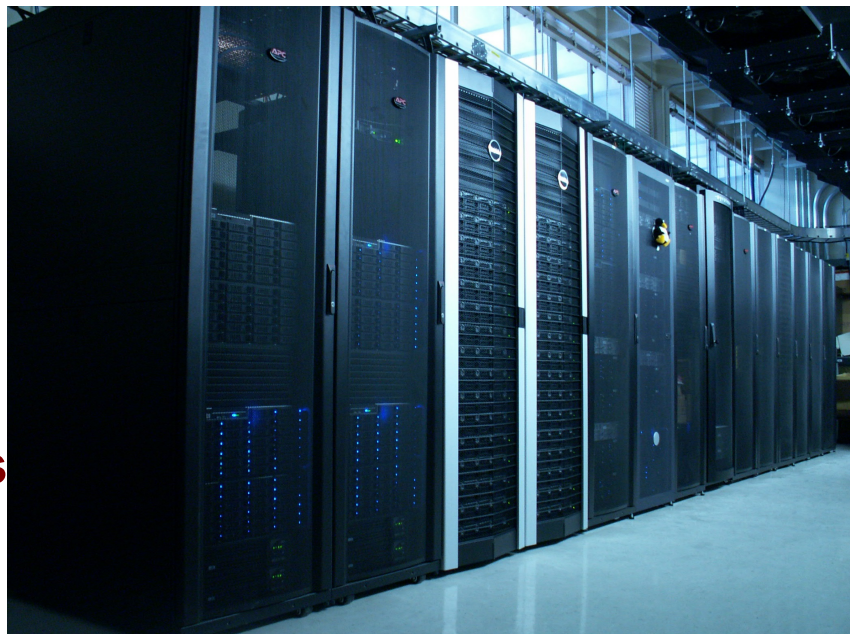
- **Tier2 Dedicated: 4126 cores**
 - Under OS RHEL 6.4, OSG 3.2.4 CE with patch
 - 4GB RAM / core typically
 - 1TB of total space for OSG_WN_TMP
 - HSA06 by scaling : 36647
- **Opportunistic: 25% total cores available : CMS can get 5250 cores maximum**



4/7/2014

Storage

- **Bestman SE with**
 - **2 10Gbps gridftp servers**
- **Lustre Parallel Filesystem:**
 - **1.7 PB for production**
 - **85 TB under older hardware**
 - **19 OSSes with 135 RAID OSTs**
- **Very high performance**
- **Outstanding reliability**
- **Excellent scalability/extendibility**
- **44 TB hardware-RAID-based NFS server for \$DATA etc.**





Network

- 100Gbps WAN
- 200Gbps within Campus
- 55Gps FDR Infiniband interconnection
- 40Gbps IB-IP bridges
- New 40Gbps high-performance NAT on 100G WAN. The University network team reconfigured it for the CMS jobs



Site Status

- 500 ~ 5000 pilot jobs at a single time
- Increased WLCG CPU contribution from this year
- SAM : ok except for frequent xrootd test errors
- HC: Usually better than 90%
- PhEDEx:
 - Hosts 3225 datasets
 - Supports 4 PAG/PDG groups
 - NLR to Internet2 switch. FNAL peering shows surge in transfer volume
 - New transfer rate measurement with T1s and T2s is needed and improvement is needed





Current Focus and Future Plans

- **Number of running glidein jobs:**
 - not enough, glidein team knows the issue
 - Heavy hacking of g-j-m perl modules for stageout issue
 - NAT and WN network improvement
 - Add one more CE to relieve stress on the single CE
 - Multicore jobs: will work with the glidein team
- **Popular dataset placement and Glidein FE:**
 - To attract more jobs and meet more than the WLCG pledge
 - More core, more storage space, more jobs
- **PhEDEx: transfer rate improvement**
 - Perfsonar in full 10Gbps for the BW test
 - Conversation with the University network team
 - Exchange ideas with sites for transfer rate improvement



P1: Improving Transfer

- **100G WAN**
 - **We have the capability**
 - **We need to test and evaluate by working with other sites**
- **IPv6: in progress**

We will start to test IPv6 on some public servers this fall, other nodes will be considered thereafter.
- **Improving PhEDEx Transfer Rate**
 - **Reference rate collection**
 - **Utilize perfsonar network aspect**





P2: Cluster Problem Solving

- **Monitor glidein jobs**
 - **Access FE condor status and logs**
 - **Analyze dashboard more proactively**
 - **Submit analysis jobs to feel how users feel about the site**
 - **Start migrating to using CRAB3 client**

- **Transfer Improvement**
 - **Use perfsonar**
 - **Use grid tools for transfer rate comparisons**



Automated Resource Management

- System provisioning (imaging) system handles most installs and post-install configs.
- Ganglia and Nagios monitor system.
- perfSONAR for network monitor.
- OSSEC for security scanning and logging.
- Customized monitor and automation systems:
 - Lots of tools to monitor CE, SE, glidein/dashboard have been developed and are evolving as the CMS system changes. They all start with ftool*.sh or cms* (All shell)
 - PhEDEx transfer monitoring/approval/deletion automated/cronized



Site Evolution

- **Three hardware locations connected via 200Gbps campus network.**
- **The new Data Center for the Cluster**
- **Storage and HA failover equipment will remain in the second location (Larson Hall)**
- **All equipment in the first location (Oldest) will be moved to two locations**
- **All equipment will be in two locations**
- **Ironically, they are furthest from our offices**





Site Policy

- All Florida Tier2 hardware will be under UF Research Computing Center's management and maintenance
- UF Research Computing has a 5-year expiration policy for any investments. This is better than a typical worker node lifespan
- For best reliability, older than 3 years (typical warranty period) CMS Tier2 Lustre OSTs will move from production pool to non-production pool and keep serving there until retirement.





Summary

- 4.2K dedicated slots contribute to the daily CMS computing needs
- 1.8PB usable space is provided currently
- We utilize local supercomputing power and management efficiently
- Highly customized tools to monitor health of the site are constantly developed
- Job monitoring in glidein FE and dashboard will be developed
- Need new measurement/improvement in the PhEDEx transfer