XT9? Integrating and Operating a Conjoined XT4+XT5 System

presented by
Don Maxwell
HPC Systems
ORNL
What is a Conjoined XT4+XT5?

Jaguar XT4

Jaguar XT5
# What is a Conjoined XT4+XT5?

<table>
<thead>
<tr>
<th></th>
<th>Jaguar XT5</th>
<th>Jaguar XT4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabinets</td>
<td>200</td>
<td>84</td>
</tr>
<tr>
<td>Processors</td>
<td>AMD Opteron 2.3 GHz quad-core</td>
<td>AMD Opteron 2.1 GHz quad-core</td>
</tr>
<tr>
<td>Compute Cores</td>
<td>149,504</td>
<td>31,328</td>
</tr>
<tr>
<td>Memory (TB)</td>
<td>300</td>
<td>62</td>
</tr>
<tr>
<td>Links</td>
<td>115,200</td>
<td>48,384</td>
</tr>
<tr>
<td>Theoretical Peak</td>
<td>1,375</td>
<td>263</td>
</tr>
<tr>
<td>Performance (TFLOPS/s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I/O Capacity (TB)</td>
<td>4,100*</td>
<td>700</td>
</tr>
<tr>
<td>I/O Bandwidth (GB/s)</td>
<td>100*</td>
<td>40</td>
</tr>
<tr>
<td>Service Nodes</td>
<td>256</td>
<td>116</td>
</tr>
</tbody>
</table>

* The current filesystem on Jaguar XT5 is an Infiniband direct-attached configuration using roughly half of the available storage capacity available. The other half is being used for development of a Lustre routed filesystem called Spider. The two halves will be merged into a Spider configuration which will be mounted center wide during the next few months.
What is a Conjoined XT4+XT5?

Combining two resources into one
SION
External Logins
Need a platform for access to both machines

Jaguar XT4

Cisco IB Core 1

Cisco IB Aggregation

Cisco IB 2nd Floor

External Logins

Jaguar XT5

Cisco IB Core 2
Routing XT Computes

XT Compute Node Routes
192 IB nodes XT5
48 IB nodes XT4
IB Router <-> IB Router
Selection based on IB switch
Compute node router selection based on distance
External Login Nodes

- **Motivation**
  - Single platform for accessing both XTs
  - To provide a much more capable platform for software development than the current service nodes directly attached to the XTs

- **Prototype Hardware**
  - Quad socket AMD Opteron 2.0 GHz quad-core
  - 32 GB memory
  - SLES 10.2
  - Autoyast
  - Cfengine
  - Conserver
External Login Nodes

- XT Software
  - Batch Systems
  - Filesystems
  - Cray XT Stack
Batch

- **Moab/TORQUE**
  - History dating back to 2005
  - First port to XT platform on ORNL development system
  - Requirements discussion in December for conjoined project
    - Two potential development paths
      - Modify existing XT native resource manager
      - Use grid model
    - Modifying existing RM seemed to be the easiest path
Moab features support NCCS mission

- Job templates to categorize job sizes
  - Large jobs favored to support capability mission
  - DOE metrics requirement for Capability Usage
    - In the first year following general availability of a new or upgraded system, 35% of the CPU time used on the system will be accumulated by jobs using 20% or more of the available processors
    - In subsequent years, 30% of the CPU time used on the system will be accumulated by jobs using 30% or more of the available processors
  - Supported through use of Moab job templates/fairshare/priorities

- Identity manager to import project priorities
  - RATS maintains project information
  - Priorities changed dynamically via import from ASCII file

- Size 0 jobs eliminate need for user cron jobs
  - Cron can causes issues with filesystem unmounts
    - Batch control more desirable
  - Accounting method same as traditional batch jobs

- LENS Visualization cluster job pre-emption
  - 32 nodes with each node containing four quad-core 2.3 GHz AMD Opteron processors with 64 GB of memory, and 2 NVIDIA 8800 GTX GPUs
  - Computational jobs allowed unless an analysis job appears
Batch

What’s the model?

ALPS only has knowledge of one XT/domain

Passwordless ssh using sudo for communication

External Moab allows each XT to operate independently
Batch

- Features
  - Target a particular resource
    - qsub
    - `msub -l partition=(xt4|xt5)`
  - No specific resource
    - msub
  - Load balancer
    - Simple algorithm based purely on availability of resources at the time of job launch
    - Open to more sophisticated algorithm
      - Delay choice until runtime
      - Queue depth
      - Historical utilization
  - Restrict each partition based on user
  - Direct jobs based on size using job templates
Filesystems

- Production
  - 3 Fibre-channel Lustre filesystems on XT4
    - 150TB spans first half of DDN 9550s
    - 150TB spans second half of DDN 9550s
    - 300TB spans all DDN 9550s
  - 1 Infiniband direct-attached 4.5PB Lustre filesystem on XT5

- How do I mount these filesystems on external login nodes?
  
  Answer: Not easily
Filesystems

- **Method**
  - LNET routing via SION

- **Advantages**
  - Users have same filesystems available to them on external login nodes

- **However...**
  - Using XTs as Lustre file servers is a bad idea
    - Hangs for users accessing filesystems
    - Users have to compile for multiple filesystems if allowing the system to choose the partition

- **LMON**
  - Script to monitor health of filesystems
  - Lctl ping mds to detect state
  - umount problems
    - /etc/mtab locking issues
Job Execution

Jaguar XT4

Jaguar XT5

SION

External Logins

Login Services

Filesystem
**Cray XT software**

- Same versions of XT software must be available on external logins

- Method
  - xt-rpm utility
    - External NFS Sharedroot for Cray XT software
    - /opt/xt* links back to External NFS Sharedroot
    - Separate RPM database

- Default programming environment for both XTs same
  - Software packages per machine can vary
XT Modules

- Module named XT4 or XT5 will be loaded as a key to determine which machine is being addressed
- XT-specific commands such as apstat, xtnodestats, etc. will be wrapped based on XT module
- Lustre scratch directory /tmp/work/$USER changes based on XT module
- Provides TORQUE environment
Status

- Prototype up and working
  - External login node up with SLES 10.2
  - Using XT5 TDS/XT4 TDS for XTs
  - Cray software installed and communication working with XTs using XT[45] modules
  - Local Lustre filesystems from each XT mounted
  - Single scheduler running on external server

- 4 External Logins in testing for Jaguar with SLES 10.2
  - Local Lustre filesystems from XT4/XT5 mounted
  - LMON hardening
  - Moab policy review for final configuration underway
• Futures
  – Filesystems
    • Spiders everywhere
  – More sophisticated Moab load-balancing algorithm
  – Moab priorities based on fairshare force Grid model?
  – Cray software is multi-XT aware
  – Spanning machines
    • Moab can span partitions using a QOS with SPAN feature
    • Requires OpenMPI or another MPI derivate
Questions?