HPC Committee (RCAC-HPC)

Minutes February 27, 2012 (taken by Erik Deumens)

Present: Paul Avery, Peter Barnes, Erik Deumens, Eric Ford, Charles Taylor, Craig Prescott

Reports

Eastside Campus Data Center (ECDC)

The work on ECDC is progressing according to plans. The foundations are being laid. The 100% drawings have been submitted to UF and haven been reviewed. The construction will be concrete walls poured on site and then raised up. The target date for having systems operational in the ECDC is December 2012.

Acquisitions of new hardware

During the fall of 2011, several new systems were purchased to address a range of specific needs and provide services that were purchased by researchers and research groups on campus within the Research Computing Matching program. (See the minutes of October 17, 2011 for detailed report on the matching program.) The needs are so diverse that there is no single large system acquisition associated with this Phase 5 acquisition as there has been in the past with Phase 1 through and Phase 4.

1. In May 2011, 512 GB of RAM was purchased and installed into an existing AMD-based Dell server. The purchase was split with Brad Barbazuk who needed the large physical address space for genome assembly. This purchase provided the genome assembly community on campus with an essential capability that was previously lacking in the HPC Center. The cost was $22K.
2. In June 2011, a 115 TB useable space was added to the parallel Lustre file system to bring the total capacity to 230 TB. This system is fully redundant and has fail-over mechanism built in so that no single component is a single point of failure. The cost was $104K.
3. In November 2011, a test cluster was installed targeted to the computational biology community while the configuration of the “biocluster” is being worked out. This cluster has 16 nodes with 96 GB of RAM in each node and two Intel X5575 processors running at 3.0 GHz, providing a total of 192 compute cores. The cost was $95K.
4. In December 2011, a GPU test cluster was installed to determine the optimal configuration for applications of interest to researchers at UF, mostly in chemistry and astronomy. It has one node with 8 NVIDIA M2070 GPU cards and two 4-core Intel CPUs in it, and 4 nodes each with 2 NVIDIA M2070 GPU cards, 1 6-core Intel x5675 CPU, and 32GB of DDR3 RAM. The 4 nodes are connected to the InfiniBand fabric so that experiments can be carried where the GPUs can send messages directly to each other. The cost was $56K.

5. Two NAS systems (network attached storage) from Nexenta were purchased each with a useable capacity of 96 TB to serve the needs of our long term storage investors as well as the investor who bought replicated long term storage. For the latter investors the replication of the data from one system is done on the other system. The file system on these systems is ZFS. The storage is also accessible via CIFS to Windows, Mac and Linux desktops and laptops, so that users can see and manipulate their research data directly as well as run high-performance computations on them. The systems became operational in January and February 2012. The cost was $130K.
6. In February 2012, a rack was bought with 16 nodes, optimized for AMBER simulations, that each have two NVIDIA M2090 GPUs and two 6-core 2.8 GHz 4284 AMD processors and 32 GB of RAM. It also has 8 nodes for large RAM electronic structure computations with 6-core 3.06 GHz Intel X5675 processors and 96 GB RAM. The cost was $187K.
7. In February 2012, a rack with 1088 AMD Interlagos cores was added to replace some 6 years-old nodes with 8 times more cores for the same amount of electrical power and cooling capacity. The rack sits in NPB 1114 and the nodes are connected by Gigabit Ethernet only. This makes these nodes ideal for serial jobs and jobs with up to 16-way parallelism that do not require high input-output bandwidth to the scratch files system. The cost was $200K.

Purchasing unit costs

With every set of acquisitions we adjust the rates for basic units to reflect the actual cost that the HPC Center pays to the vendors for completely functional systems.

1. **NCU (normalized compute unit, i.e. essentially a compute core with the necessary environment to make it useful for 5 years)** The rate of $400 per NCU remains the same.
2. **OSU (optional storage unit in TB per year)** The rate stays at $125 per TB per year for regular storage and $250 per TB per year for replicate storage. Note that this storage is in addition to the use of the 230 TB parallel Lustre file system for scratch data; it is for long term storage of research data.
3. **NGU (normalized graphics processor unit)** This is a new option with Phase 5. We require that any investor who buys an NGU also buys one NCU to go with it, since a GPU cannot be used without the participation of a CPU core. From the nodes in the test cluster under item4 above that have 1 6-core CPU, 32 GB of RAM and 2 GPUs at $8,000 each, we compute that the cost of two NGUs = $8,000 – (4 x $400) = $6400. We subtracted the 4 cores that are in the node, but are not essential to use the GPU. Thus the rate is $3,200 per NGU.

If we apply the same formula to the higher-density node with 8 GPUs, we get a cost estimate for 8 NGUs = $24,000 or $3000 per NGU because the 8 GPU system only has 8 cores. Therefore purchasing GPU's in systems with 8 GPU's per chassis only saves about 7% over the cost of 2 GPU systems. We conclude that there is not much to be gained from purchasing GPUs in this format.

Discussion

Multiple computer architecture support

The cluster always has a mixture of Intel and AMD processors. Initially we supported multiple compilers: Intel, Portland Group, and PathScale. For several years we have only supported Intel compilers, because the benefit of the compilers that can generate better code for AMD did not outweigh the complexity and cause for confusion with the users, and the extra effort for HPC staff to maintain multiple versions of all supported software.

1. The Intel compiler has options to generate code on the submit node, which has an older, less capable CPU, that will run on all nodes, including the newer ones and the AMD nodes. This is done by generating extra sections in the executable that are conditionally executed depending on the hardware capability of the CPU cores the code is running on. The flags are –xSSE2 –axSSE3,SSE4.2,SSSE. The Intel compiler still produces quite good, if not completely optimal, code for AMD processors. The OMP directives in the Intel compiler do not generate multithreaded code for AMD.
2. The GNU compilers also produce code that is quite good on Intel and AMD CPUs.
3. The rights to the PathScale compiler have now been acquired by AMD, which has released it as an open source compiler called Open64.

At this time there are no plans to support multiple compilers again for the reasons stated above.

Evaluating job schedulers

For years, we have been using the Moab scheduler and Torque resource manager from Adaptive Computing to manage the scheduling of jobs on the clusters. Although the software is stable, there are many features that are advertised and that would be useful for us to put into production, but they turn out not to work. This has resulted in a lot of time wasted trying to make them work and pursuing the company to work on the bugs.

1. Last summer we took a very detailed look at LSF from Platform Computing, now bought by IBM. That software is also complex and would imply a lot of changes for the users, if we decide to take that into production.
2. We are now also evaluating SGE (Sun Grid Engine), which is supported by Univa after SUN was bought by Oracle.
3. A third candidate being evaluated is PBSpro from Altair. Our initial assessment is that that software will not work for our needs.
4. Finally we are considering the open source combination of Maui as scheduler and Torque as resource manager.

In the next few months we will decide on a strategy to minimize maintenance cost in terms of dollars and staff time.

Next meeting will be of the RCAC committee on March 12 at 1:30 pm in NPB 2205.