Minnesota Supercomputing Institute
Introduction to MSI Systems

Andrew Gustafson
The Machines at MSI
Machine Type: Cluster

Machine Type: Cluster

- Clusters at MSI
  - Itasca
    - 1,091 HP ProLiant blade servers, each with two quad-core 2.8 GHz Intel Xeon X5560 "Nehalem EP" processors sharing 24 GiB of system memory. 40-gigabit QDR InfiniBand (IB) interconnect. In total, Itasca consists of 8,728 compute cores and 24 TiB of main memory.
    - itasca.msi.umn.edu
  - Calhoun
    - SGI Altix XE 1300 cluster with 180 compute nodes, each with 16 GB of memory. The nodes are interconnected with a 20-gigabit DDR InfiniBand (IB) network.
    - calhoun.msi.umn.edu
Machine Type: Shared Memory

Source: http://en.wikipedia.org/wiki/Shared_memory
Machine Type: Shared Memory

- Shared memory systems at MSI
  - Koronis
    - Altix UV 1000 node with 1,140 compute cores (190 6-core Intel Xeon X7542 Westmere processors at 2.66 GHz). 2.96 TiB of globally addressable shared memory in a single system image.
    - Dedicated to research related to human health.
    - koronis.msi.umn.edu
Machine Type: Heterogeneous

Source: http://electronicdesign.com/digital-ics/gpu-architecture-improves-embedded-application-support
Machine Type: Heterogeneous

- Heterogeneous systems at MSI
  - Cascade
    - 32 M2070 Telsa GPUs (4 per node, 8 nodes), 8 K20 Kepler GPUs (2 per node, 4 nodes), 2 Xeon Phi co-processors (1 per node, 2 nodes)
    - cascade.msi.umn.edu
Machine Type: Heterogeneous

Many Core NVIDIA GPGPU
Programmed with CUDA

http://www.nvidia.com/object/what-is-gpu-computing.html
Machine Type: Heterogeneous

Many Core Intel Co-processor
Programmed in C/C++/Fortran with special statements.

Figure 2. Intel’s Xeon Phi is a heterogeneous array of more than 50 x86 cores linked by a two-way racetrack interconnect scheme.

Source: http://www.altera.com/technology/system-design/articles/2012/multicore-many-core.html
Machine Types: Programming Complexity

General View:

Simplest: Shared Memory (OpenMP)
Medium: Cluster (MPI)
Complex: Heterogeneous (CUDA, special statements)
MSI Labs and Lab Servers

https://www.msi.umn.edu/labs

- Labs at 5 locations
  - Labs machines use the MSI software environment
  - Special high definition and 3D visualization equipment

- Labs Servers for small jobs
  - lab.msi.umn.edu
  - Many software packages.
  - Some special hardware such as liquid cooled overclocked nodes.
Interacting with MSI Systems
Connecting to MSI

https://www.msi.umn.edu/remote-access

SSH is the most reliable method

Some programs with SSH functionality in Windows include PuTTy and Cygwin

Other connection methods such as NX are available.

Most systems require you to first ssh into: login.msi.umn.edu
MSI Computing Environment

https://www.msi.umn.edu/resources/computing-environment

MSI systems are mostly Linux compute clusters running CentOS

Software is managed via a module system

Jobs are scheduled via a queueing system
Home directories are unified across all Linux systems (except for Koronis).

Each group has a disk quota which can be viewed with the command: `groupquota`

Panasas ActivStor 14: 1.28PB storage, capable of 30 GB/sec read/write, and 270,000 IOPS
Software Management

Software is managed via a module system

- `module load modulename` - loads a software module
- `module show modulename` - shows the actions of a module
- `module avail` - shows all available modules
- `module list` - shows a list of currently loaded modules
- `module unload modulename` - unloads a software module
- `module purge` - unloads all software modules
Job Scheduling

Jobs are scheduled using the Portable Batch System (PBS) queueing system

To schedule a job first make a PBS job script:

```
#!/bin/bash -l
#PBS -l walltime=8:00:00,nodes=3:ppn=8,pmem=1000mb
#PBS -m abe
#PBS -M sample_email@umn.edu

cd ~/program_directory
module load intel
module load ompi/intel
mpirun -np 24 program_name < inputfile > outputfile
```

https://www.msi.umn.edu/resources/job-submission-and-scheduling-pbs-scripts
https://www.msi.umn.edu/resources/job-queues
Job Scheduling

To submit a job script use the command:
qsub -q queuename scriptname

A list of queues available on different systems can be found here:
https://www.msi.umn.edu/resources/job-queues

Submit jobs to a queue which is appropriate for the resources needed.

To view queued jobs use the commands:
qstat -u username
showq -w user=username

To cancel a submitted job use the command:
qdel jobnumber
Interactive Jobs

Nodes may be requested for interactive use using the command:

```bash
qsub -I -l walltime=1:00:00, nodes=1:ppn=8
```

The job waits in the queue like all jobs, and when it begins the terminal returns control.
Service Units (SUs)

Jobs on the high performance computing (HPC) systems consume Service Units (SUs), which roughly correspond to processor time.

Each research group is given a service unit allocation at the beginning of the year. To view the number of service units remaining use the command: acctinfo

If a group is using service units faster than the "fairshare target", then the group's jobs will have lower queue priority.

https://www.msi.umn.edu/hpc/fairshare
Simple Parallelization: pbsdsh

#!/bin/bash
#PBS -l walltime=8:00:00,nodes=1:ppn=8,
#PBS -m abe
#PBS -M sample_email@umn.edu

cd ~/job_directory
pbsdsh -n 0 ~/jobdirectory/script0.sh
pbsdsh -n 1 ~/jobdirectory/script1.sh
pbsdsh -n 2 ~/jobdirectory/script2.sh
pbsdsh -n 3 ~/jobdirectory/script3.sh
pbsdsh -n 4 ~/jobdirectory/script4.sh
pbsdsh -n 5 ~/jobdirectory/script5.sh
pbsdsh -n 6 ~/jobdirectory/script6.sh
pbsdsh -n 7 ~/jobdirectory/script7.sh
wait
Job Problem Solving

MSI has resources available to help solve job problems.

https://www.msi.umn.edu/services/user-support

https://www.msi.umn.edu/resources/job-problem-solving

https://www.msi.umn.edu/support/faq

https://www.msi.umn.edu/resources/debugging

Help can be requested by emailing: help@msi.umn.edu
MSI Services

- At-cost Consulting

Complementing its core mission of providing supercomputing resources at no charge to the research communities of the University of Minnesota and other institutions of higher education in the state, MSI charges other units of the University at cost for the efforts of its staff when one of the following conditions obtains: Effort is @ 10% of an FTE or greater for at least one month; Effort is @ 5% of an FTE or greater for at least three months; or Effort is at any level and is grant funded.
MSI Services

● Data Storage

For groups that have unusually large storage requirements or groups that require guaranteed space at all times, MSI offers high performance, dedicated space for a fee. Currently, the fee is $261 / (TB-year). MSI offers this storage in increments of 1 TB, and for periods of time that are multiples of 1 year. This fee guarantees the contracted space availability, backups, and general management and maintenance. Dedicated storage can be accessed from all MSI computing systems.
MSI Services

- Tape Storage

For those users whose large data storage needs are largely of archival nature, MSI is also providing the capability of tape storage at cost. We provide high density tapes (LTO-5) for $100/tape, plus an associated labor charge of $109 (two hours of staff time) to move data to tape/s and to give tape/s to the researcher. Each tape has native capacity of 1.5 TB (3 TB compressed). Since this is off-line storage, data volume cannot exceed the capacity of the tape.