Ranger User Guide

General System Info Section

Important System and Programming Notes.

Login Nodes now have barcelona chips
: The 4-socket Login3 and Login4 (ranger.tacc.utexas.edu) nodes have been populated with barcelona chips.
*** These are 2.2 GHz chips, the compute nodes run at 2.0 GHz. (Please do not run codes on the login nodes.)

Compiling on Login Nodes
: When you login to ranger.tacc.utexas.edu you will be connected to either login3.ranger.tacc.utexas.edu or login4.ranger.tacc.utexas.edu (login1 and login2 are not available yet). Initially, login3 and login4 will be dual-core Opterons. Hence, you should not let the compilers automatically detect the hardware of the login nodes. Compile with the "barcelona" hardware options suggested in the Compiling Section.

MPI Support for Compilers
: Only the Intel and PGI compilers will support MPI. The mvapich2 libraries have been compiled with both compilers, and are automatically linked by the mpicc and mpif90 compiler drivers when correctly loaded through the module commands. (By default the MPI compiler drivers use the PGI-compiled mvapich2 libraries and the default compilers are PGI.)

Debugging and Profiling
: DDT is not available yet. Please use the idb (Intel) debugger, pgdbg and pgprof (PGI), and gdb and gprof (GNU) for debugging and profiling.

/tmp on Compute Nodes
: In the compute nodes, the only physical storage device is an 8GB compact flash, which stores the OS. Only 150MB are available in /tmp for user storage. Program developers should use $SCRATCH to store temporary files. (The /tmp directories on login nodes are 36G disk devices.)

Parallel Environment (using less than 16 cores/node)
: The Parallel Environment Section shows how to use less than 16 tasks per node, and how to run hybrid codes.

MPI (mvapich) Options for Scalable code
: See the mvapich1/2 User Guides.

Core Affinity and Memory Allocation Policy
: See Numa Section for controlling process/thread execution on sockets and cores; and memory allocation policy on sockets.

Core Count for Batch SGE Jobs
: See Numa Section (look for MY_NSLOTS) for core counts other than a multiple of 16.

Experienced Users
: Check out the Quick Start Notes.

Introduction

Ranger is one of the largest computational resources in the world, serving NSF TeraGrid researchers throughout the United States, academic institutions within Texas, and the components of The University of Texas System.

The Sun Constellation Linux Cluster, Ranger, is configured with 3,936 16-way SMP compute-nodes (blades), 123 TB of total memory and 1.73PB of global disk space. The theoretical peak performance is 504 TFLOPS. Nodes are interconnected with InfiniBand technology in a full-CLOS topology providing a 1GB/sec point-to-point bandwidth. Also, a 2.8 PB archive system and 5TB SAN network storage system are available through the login/development nodes.

Figure 1. One of 6 rows:Management/IO Racks (black), Compute Rack (silver), and In-row Cooler (black).

Figure 2. SunBlade x6420 motherboard.

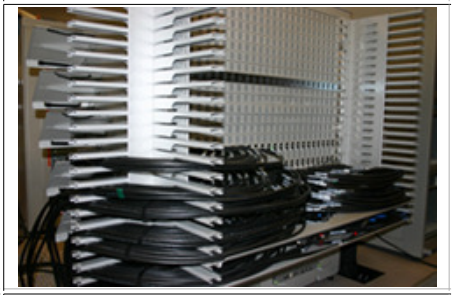

Figure 3. Constellation Switch (partially wired).

Architecture

The Ranger compute and login nodes run a Linux OS and are managed by the Rocks 4.1 cluster toolkit. Two 3456 port Constellation switches provide dual-plane access between NEMs (Network Element Modules) of each 12-blade chassis. Several global, parallel Lustre file systemss have been configured to target different storage needs. Each compute node contains 16 cores as a 4-socket, quad-core platform. The configuration and features for the compute nodes, interconnect and I/O systems are described below, and summarized in Tables 1-3.

Compute Nodes: Ranger is a blade-based system. Each node is a SunBlade x6420 blade running a 2.2 x86_64 Linux kernel from kernel.org. Each node contains four AMD Opteron Quad-Core 64-bit processors (16 cores in all) on a single board, as an SMP unit. The core frequency is 2.0GHz and supports 4 floating-point operations per clock period with a peak performance of 8 GFLOPS/core or 128 GFLOPS/node.

Each node contains 32GB of memory. The memory subsystem has an 800MHz Hypertransport system bus, and 2 channels with 667MHz Fully Buffered DIMMS. Each socket possesses an independent memory controller connected directly to L3 cache.

Interconnect: The interconnect topology is a full-CLOS fat tree. Each of the 328 12-node compute chassis is connected directly to the 2 core switches. 12 additional frames are also connected directly to the core switches and provide file systems, administration and login capabilities.

File systems: Ranger's file systems are built on 72 Sun x4500 disk servers, each containing 48 SATA drives, and two Sun x4600 metadata servers. From this aggregate space of 1.73PB, several file systems will be partitioned (see Table 5).

Table 1. System Configuration & Performance

Component	Technology	Performance/Size

Peak Floating Point Operations		504 TFLOPS (Theoretical)
Nodes(blades)	Four Quad-Core AMD Opteron processors	3,936 Nodes / 62,976 Cores
Memory	Distributed	123TB (Aggregate)
Shared Disk	Lustre, parallel File System	1.73PB
Local Disk	Compact Flash	31.4TB (Aggregate)
Interconnect	InfiniBand Switch	1 GB/s P-2-P Bandwidth

Table 2. SunBlade x6420 Compute Node

Component	Technology
Sockets per Node/Cores per Socket	4/4 (Barcelona)
Clock Speed	2.0GHz
Memory Per Node	32GB memory
System Bus	HyperTransport, 6.4GB bidirectional
Memory	2GB DDR2/667, PC2-5300 ECC-registered DIMMs
PCI Express	x8
Compact Flash	8GB

Table 3. Sun x4600 Login Nodes

Component	Technology
4 login nodes	ranger.tacc.utexas.edu (login1.tacc.utexas.edu Not Available) (login2.tacc.utexas.edu Not Available) (login3.tacc.utexas.edu) (login4.tacc.utexas.edu)
Sockets per Node/Cores per Socket	4/4 (Barcelona).
Clock Speed	2.2GHz
Memory Per Node	32GB

Table 4. AMD Barcelona Processor

Technology	64-bit
Clock Speed	2.0GHz
FP Results/Clock Period	4
Peak Performance/core	8GFLOPS/core
L3 Cache	2MB on-die (shared)
L2 Cache	4 x 512KB
L1 Cache	64KB

Table 5. Storage Systems

Storage Class	Size	Architecture	Features
Local	8GB/node	Compact Flash	not available to users (O/S only)
Parallel	1.73PB	Lustre, Sun x4500 disk servers	72 Sun x4500 I/O data servers, 2 Sun x4600 Metadata servers (See Table 6 for breakdown of the parallel file systems)
SAN	15TB	Synergy FS, SUN Storage Tek	QLogic switch, SUN V880 Server, mnt on /san/hpc/<project>
Ranch (Tape Storage)	2.8PB	SAMFS (Storage Archive Manager)	10Gb/s connection through 8 GridFTP Servers

Table 6. Parallel File systems

Storage Class	Size	Quota (per User)	Features
HOME	~100TB	6GB	Backed up nightly; Not purged
WORK	~200TB	350GB	Not backed up; Not purged
SCRATCH	~800TB	400TB	Not backed up; Purged every 10 days

System Access

SSH

To ensure a secure login session, users must connect to machines using the secure shell, **ssh** program. **Telnet** is no longer allowed because of the security vulnerabilities associated with it. The *"r"* commands **rlogin, rsh**, and **rcp**, as well as **ftp**, are also disabled on this machine for similar reasons. These commands are replaced by the more secure alternatives included in SSH --- **ssh**, **scp**, and **sftp**.

Before any login sessions can be initiated using **ssh**, a working SSH client needs to be present in the local machine. Go to the TACC introduction to SSH for information on downloading and installing SSH.

To initiate an ssh connection to a ranger login node, execute the following command on your local workstation

ssh <login-name> @ ranger.tacc.utexas.edu

Note <login-name> is needed only if the user name on the local machine and the TACC machine differ.

Additionally, each of the login nodes can be accessed directly, to allow users to move data to/from local disk space on the login nodes. These nodes are directly accessible by using the node name:

ssh <login-name> @ <login{3|4}>.ranger.tacc.utexas.edu

Password changes (with the *passwd* command) are forced to adhere to "strength checking" rules, and users are asked to comply with practices presented in the TACC password guide.