Compiling and Running a Parallel Program on a First Generation Cray SuperCluster

Frank Chism

Product Specialist

Cray Inc.

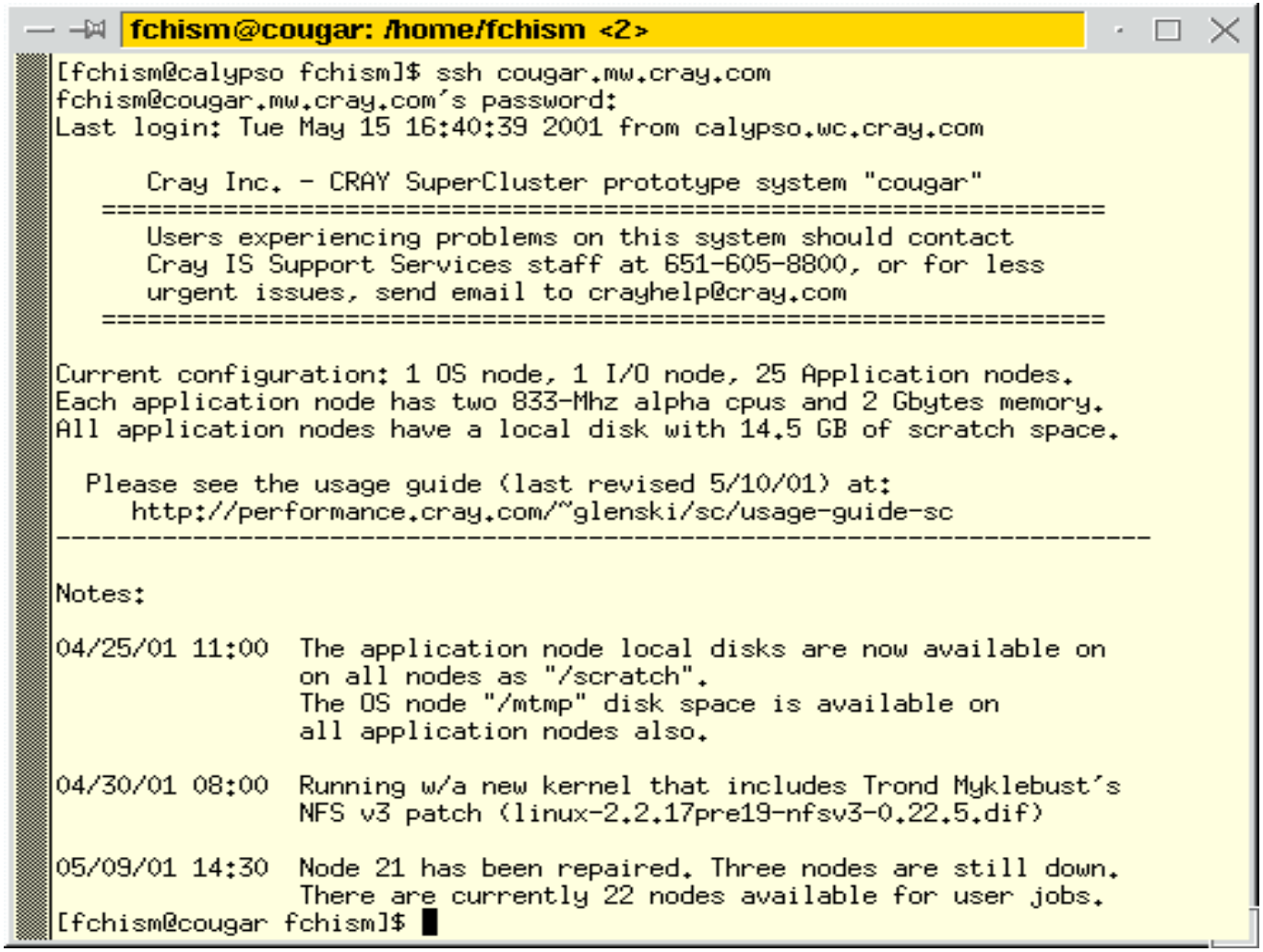
fchism@cray.com

CRAY

Logging On

- The Cray SuperCluster Presents a single identity (IP address) to the outside world
- Please don't telnet or ftp
- ssh is good
- MD5 passwords
- Remember security starts at the front door

CRAY


fchism@cougar: /home/fchism <2>
[fchism@calypso fchism]$ ssh cougar.mw.cray.com
fchism@cougar.mw.cray.com's password:
Last login: Tue May 15 16:40:39 2001 from calypso.wc.cray.com
Cray Inc. - CRAY SuperCluster prototype system "cougar"
Users experiencing problems on this system should contact
Cray IS Support Services staff at 651-605-8800, or for less
urgent issues, send email to crayhelp@cray.com
Current configuration: 1 OS node, 1 I/O node, 25 Application nodes.
Each application node has two 833-Mhz alpha cpus and 2 Gbytes memory.
All application nodes have a local disk with 14.5 GB of scratch space.
Please see the usage guide (last revised 5/10/01) at:
http://performance.cray.com/~glenski/sc/usage-guide-sc
Notes:
04/25/01 11:00 The application node local disks are now available on
on all nodes as "/scratch".
The OS node "/mtmp" disk space is available on
all application nodes also.
04/30/01 08:00 Running w/a new kernel that includes Trond Myklebust's
NFS v3 patch (linux-2.2.17pre19-nfsv3-0.22.5.dif)
05/09/01 14:30 Node 21 has been repaired. Three nodes are still down.
There are currently 22 nodes available for user jobs.
[fchism@cougar fchism]$

Look Familiar? It's Linux.

- GNU/Linux is like UNIX (or UNICOS) only geekier
- All major shells
  - Bash (Bourne or sh)
  - tcsh (csh)
  - ash, zsh
  - ksh (Korn shell via pdksh)
- Perl, Python, and Wish are first class objects, that is they live in /usr/bin not some 'local' getto
- Editors up the wazoo
  - vi is vim
  - emacs
  - ed, Jed, joe, pico, xedit, and others

CRAY

I can. GNU?

- The full GNU tool set is available in the default log in environment
- make is gmake
- more is less

CRAY

Compaq Knows Alpha

- Compaq Alpha compilers are significantly faster than GNU compilers
- Fortran 77 and 90, C, C++
- Compaq Extended Math Library (CXML)
- Compaq Portable Math Library (CPML)
- Use: -O5 -fast -arch ev6 -tune ev6 (for you Guy)
- These are the compilers we chose to ship and support

CRAY

Sample Compaq MPI Job

```
[fchism@cougar scring]$ head ring.f
      program ring
c Measure the MPI latency using a  pass the token
c    around a ring approach
      include "mpif.h"
      integer nwords, nrepeat
      real*8 t0, t1
      integer target, source
      integer myrank, numprocs, ierr, status(MPI_STATUS_SIZE)
[fchism@cougar scring]$ more makefile
# Simple Compaq MPI makefile for Cray SuperCluster
#
ring: ring.f makefile
       mpif90 -o ring ring.f /home/bench/libtimfunc.a

clean:
       rm -f ring
       rm -f ring.o
[fchism@cougar scring]$ make ring
mpif90 -o ring ring.f /home/bench/libtimfunc.a
[fchism@cougar scring]$
```

```
[fchism@cougar scring]$ mpirun -np 7 ./ring

=====================================
=
  Begin Fortran based MPI latency test
        100000  repeats
 Processor           0  finished with test.
 Processor           1  finished with test.
 Processor           2  finished with test.
 Processor           3  finished with test.
 Processor           6  finished with test.
 Processor           4  finished with test.
 Processor           5  finished with test.
    total seconds=     10.29
         latency=     14.70  microseconds
 End of Fortran based MPI latency test

=====================================
=
[fchism@cougar scring]$
```

CRAY

Job Submission

- PBS is the batch system
  - We use a third party version
  - Looks like NQE to me, only the names have changed to confuse the innocent
  - More standard than NQE, so long live the new standard
- Interactive jobs as seen in the example
- Simple system scheduling now, but better in the future

CRAY

Where am I?

- **Use bpstat to see the neighborhood**

```
[fchism@puma fchism]$ bpstat
Node    Address         Status  User    Group
0       10.14.1.8       up      any     any
1       10.14.1.9       up      any     any
2       10.14.1.10      up      any     any
3       10.14.1.11      up      any     any
[fchism@puma fchism]$
```

- **Use qstat -q to see the queues**

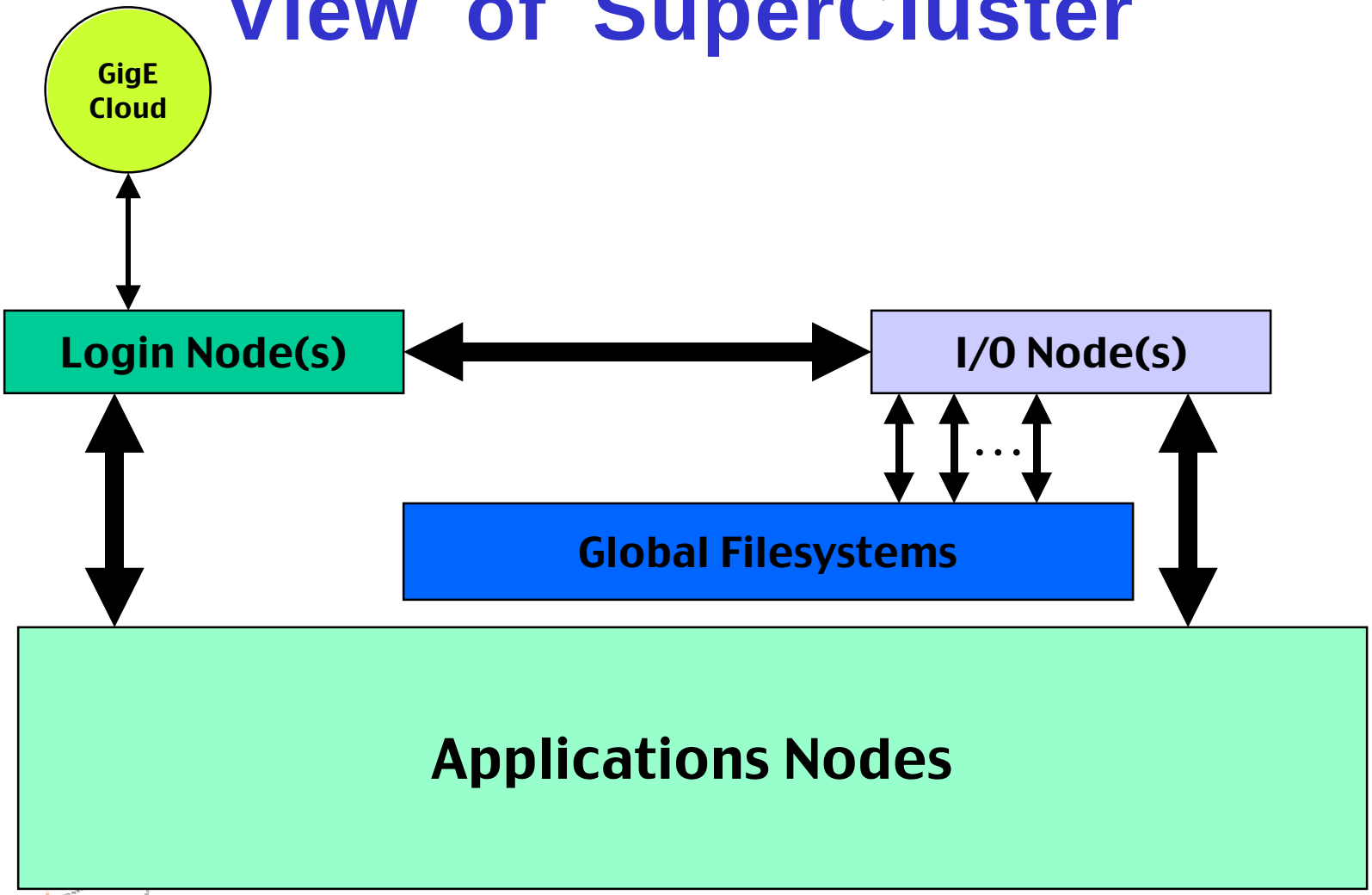
```
bash$ qstat -q
server: lynx

Queue            Memory CPU Time Walltime Node Run Que Lm  State
---------------- ------ -------- -------- ---- --- --- --  -----
workq              --      --       --     --    0   0 --   E R
                                                --- ---
                                                  0   0
```

The BIG Picture

- Outside world sees a single system via GigE
- Bproc provides a global process space for all nodes in the machine
- PS is your friend because of Bproc
- Your /home is everywhere
- Run scripts in the log in PE
- Run parallel jobs on the applications nodes
- I/O servers provide global filesystem
- Local node disks are an option for scratch I/O

CRAY

Cray Proprietary High Level View of SuperCluster



Thank You

- Joe Glenski (To whom all praise should be directed)
- Geir Johansen
- Frithjov Iverson
- Linus Thorvalds
- Richard M. Stallman
- CUG board and members
- Anne Lindsay
- Management with that Super gleam in their eyes

CRAY