Tools for Cplant

Ron Brightwell
Sandia National Laboratories
bright@cs.sandia.gov

Computational Plant

Outline

- Motivation
- Goals
- Approach
- Phase II Machine - Alaska
- Compile Environment
- Runtime Environment
- Sample Output and Performance
- Plans

Motivation

- A cluster run-time environment that is:
  - Scalable (ten thousand nodes)
  - High Performance (computation and communication)
  - Flexible (adapts to cutting edge technology)
  - Dynamic (supports adding & pruning nodes)
  - Configurable (supports partition model)
  - Familiar (allows users to adapt easily)
  - Multi-user (supports many simultaneous users)
  - General-purpose (supports a wide range of applications)

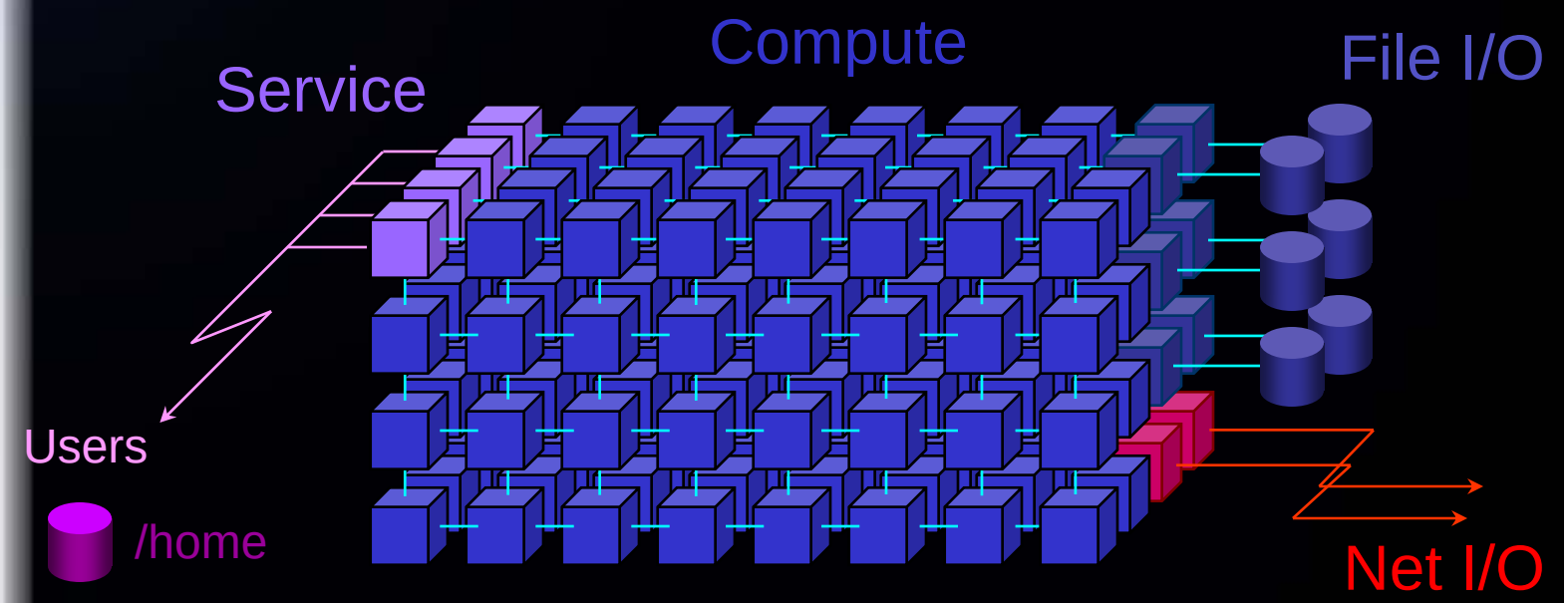
Goals

- Fast response from utilities which may query thousands of nodes, “distant” nodes or may transmit large messages
- Scalable dispersal of load data, acquisition of status data, and display
- Fault tolerant design for utilities and daemons
- Single system image
- Familiar look and feel for Tflops users

Approach

- Emulate the TFLOPS environment
  - Partition model (functional decomposition)
  - Space sharing (reduce turnaround time)
  - Scalable services (allocator, loader, launcher)
  - Ephemeral user environment
  - Complete resource dedication

Partition Model



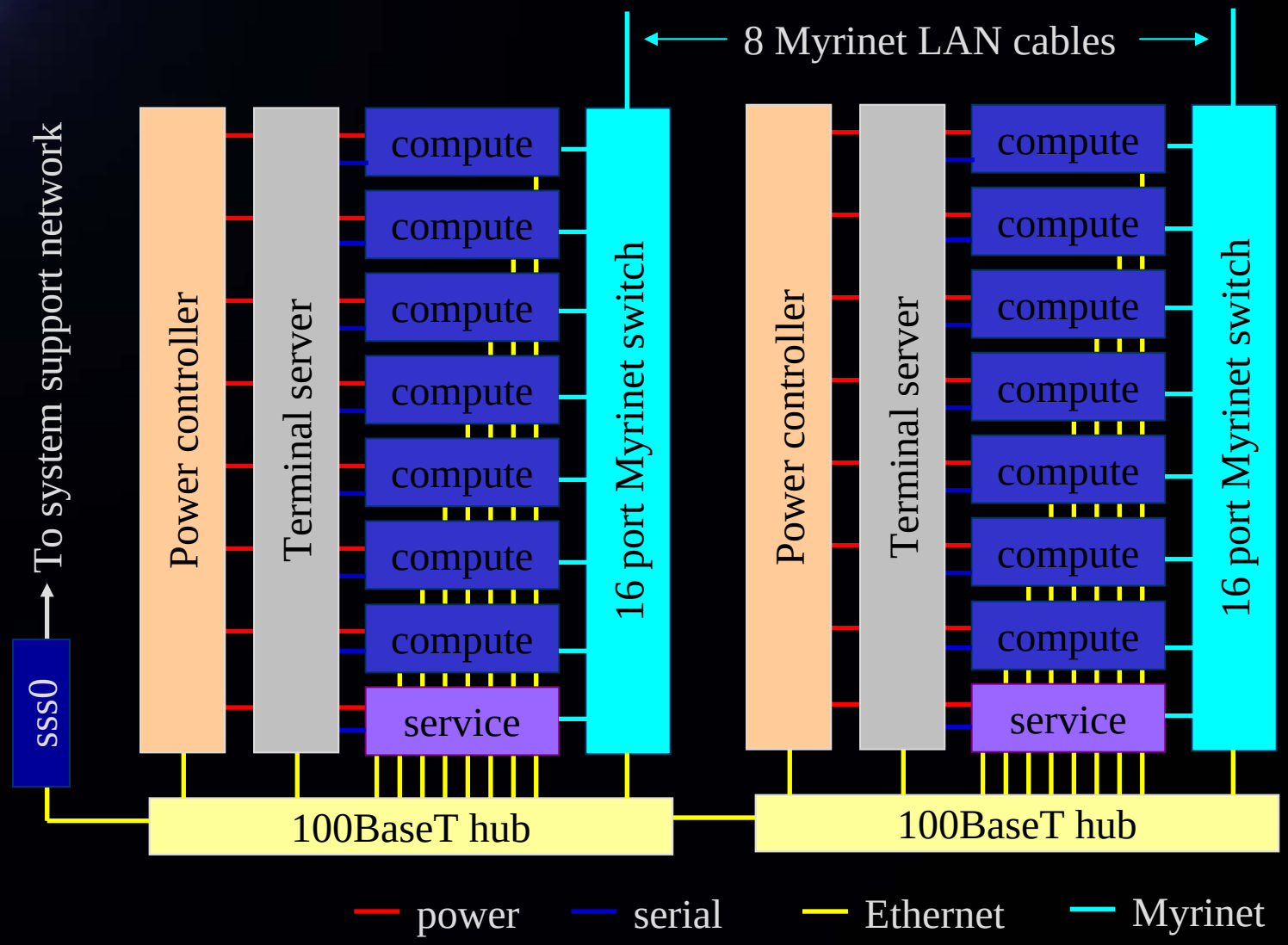
Alaska: DEC Alpha Cplant

- 427 DEC Alpha PWS 500a (Miata) 400 nodes visible to users
- No disks, CD ROM, kbd, floppy, or monitors in main partition
- 6 DEC AS1200, 12 RAID (.75 Tbyte) || file server
- 1 DEC AS4100 compile & user file server

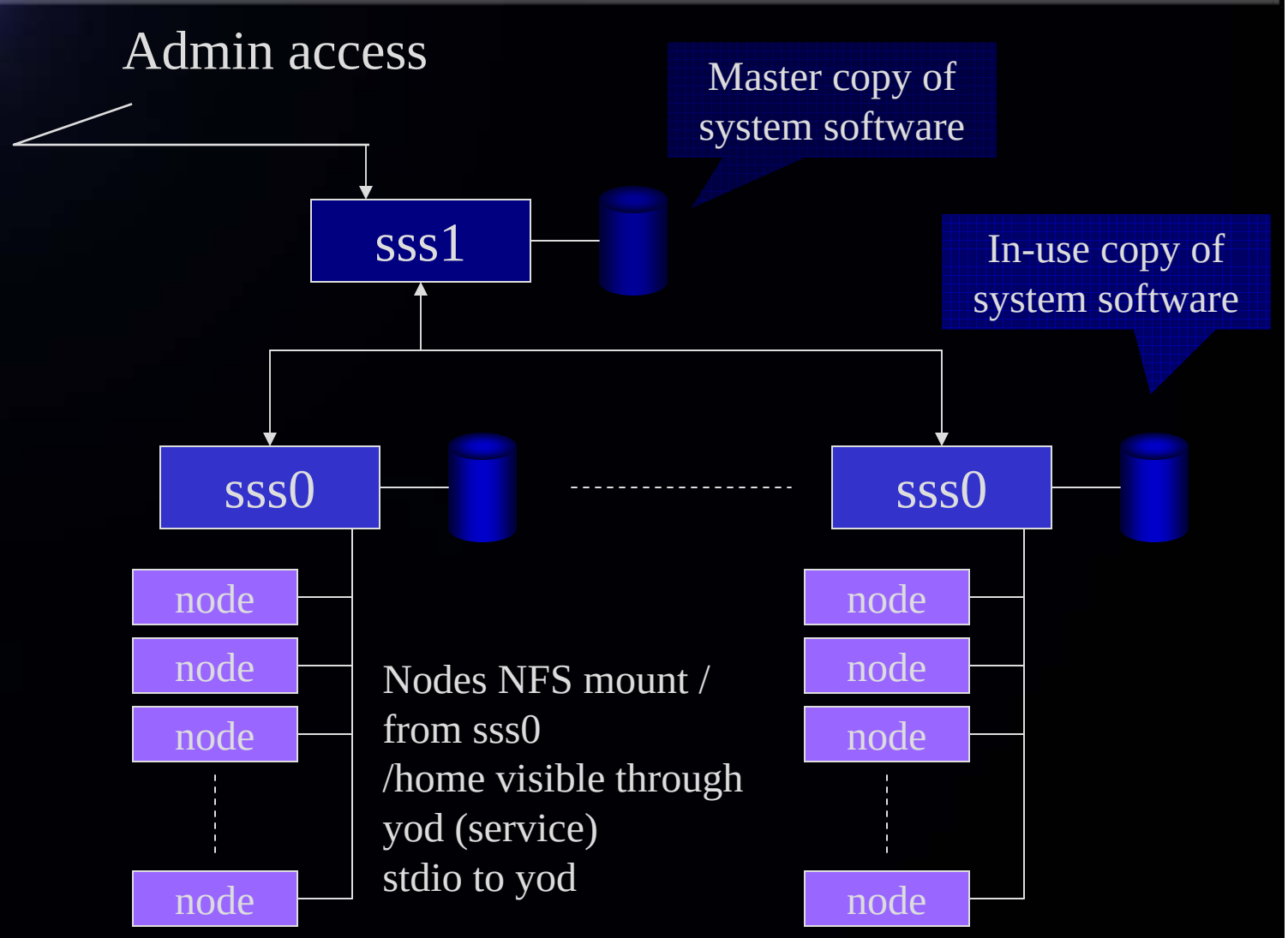
Alaska: DEC Alpha Cplant

- Networks:
  - High-performance message passing: Myrinet, 10GB/s bisection bandwidth (2-way)
  - System support: Localized Ethernet, 100BaseT
  - Diagnostic, control: Serial
- System support hierarchy: 27 DEC Alpha PWS 500a, 100BaseT network, terminal servers and serial lines for console access, remote power control of all components
- RedHat 5.2 - Linux 2.0.35

Scalable Unit



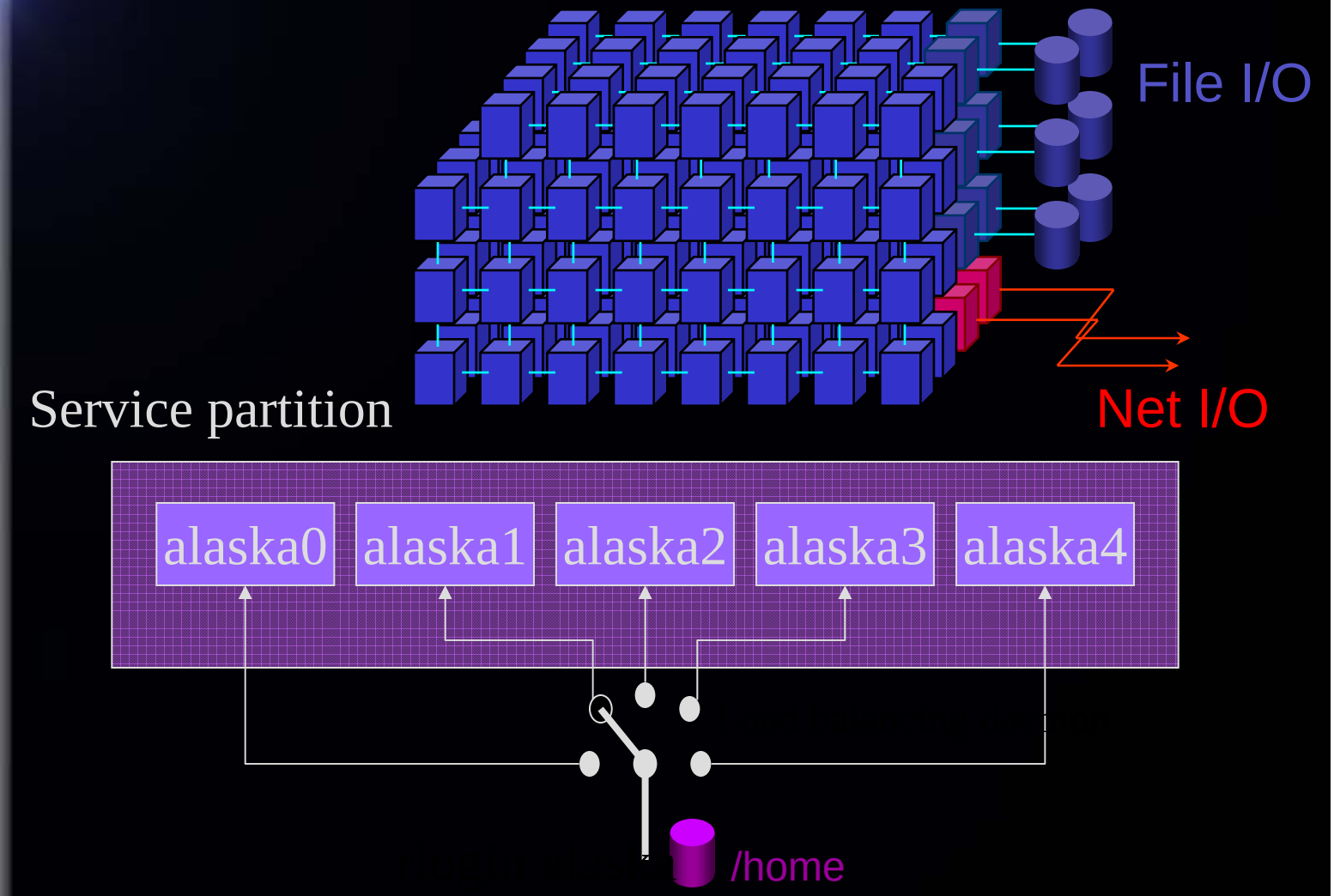
System Support Hierarchy



Compile Environment

- Shell scripts that call Digital UNIX compilers
- Support libraries - startup module, PUMA library, redirected I/O and C libraries, TFLOPS MPI library
- Centrally located for any DU machine
- Dedicated DU machine
- Can be tar'd up and moved to any DU machine
- Avoids Linux-specific configurations and immature GNU FORTRAN compiler

Logging In


File I/O
Net I/O
Service partition
alaska0
alaska1
alaska2
alaska3
alaska4
/home

Runtime Environment

- yod - service node parallel job launcher
- bebopd - compute node allocator
- PCT - process control thread, compute node daemon
- pingd - compute node status tool
- sfyod - file I/O striper

Runtime Environment

- Yod provides the following services:
  - contacts compute node allocator
  - launches the application into the compute partition
  - redirects all application I/O (stdio, file I/O)
  - makes any filesystem visible in the service partition visible to the application
  - redirects any UNIX signals to compute node processes
  - allows user to choose specific compute nodes
  - can launch multiple (up to 5) different binaries

Runtime Environment

- PCT provides the following services:
  - contacts bebopd to join compute partition
  - forms a spanning tree with other PCT's to fan out the executable, shell environment, signals, etc.
  - *fork()*'s, *exec()*'s, and monitors status of child process
  - cleans up a parallel job
  - provides a backtrace for process faults

Runtime Environment

- Bebopd provides the following services:
  - accepts requests from PCT's to join the compute partition
  - accepts requests from yod for compute nodes
  - accepts requests from pingd for status of compute nodes
  - allows for multiple compute partitions

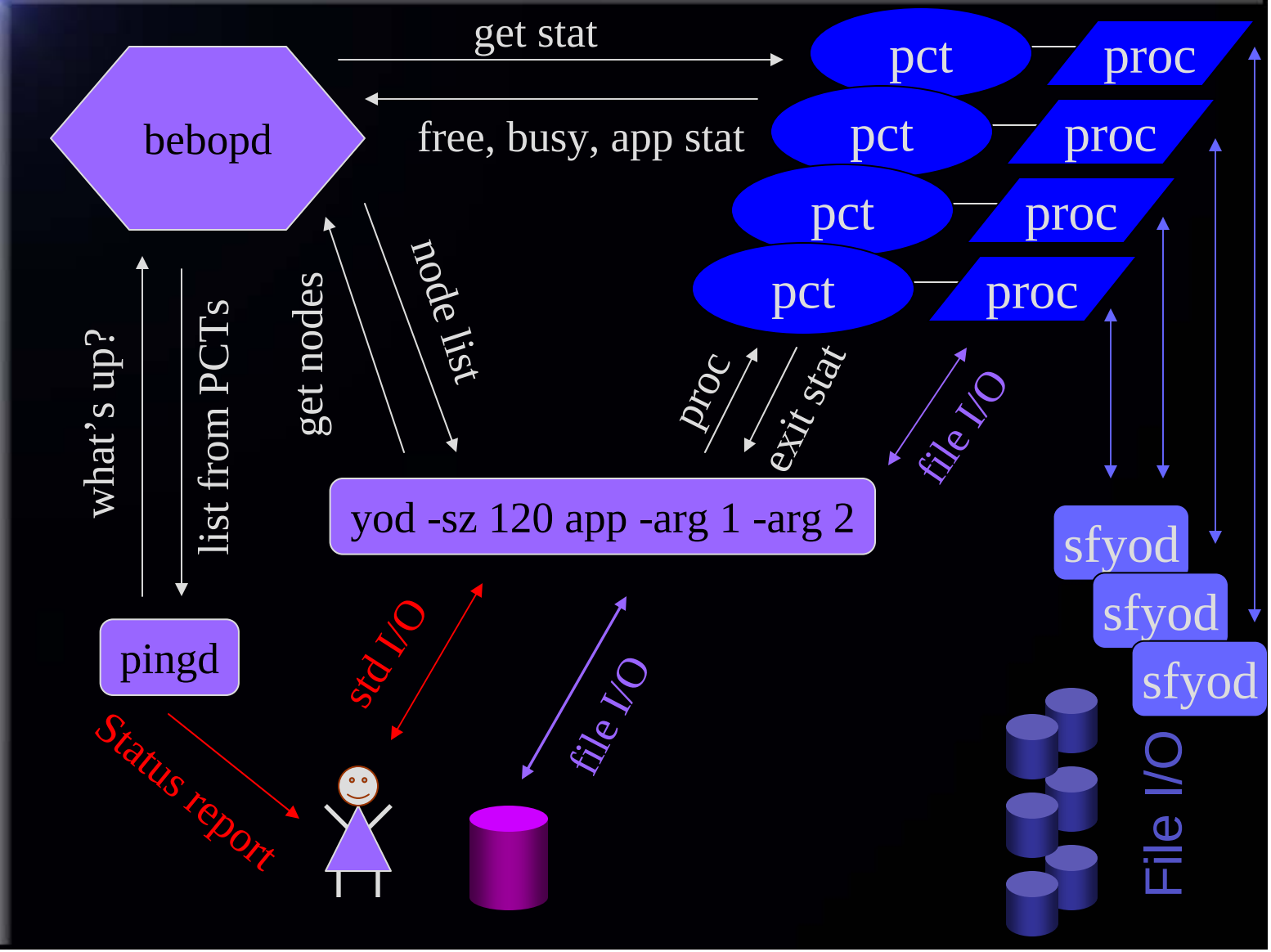
Runtime Environment

- Pingd provides the following services:
  - displays list of available compute nodes
  - displays list of compute nodes in use
  - displays owner, elapsed time of jobs
  - allows users to kill their jobs
  - allows administrators to kill jobs and free up specific nodes
  - allows administrators to remove nodes from the compute partition

Runtime Environment

- Sfyod:
  - runs on nodes in the file I/O partition
  - stripes files across multiple RAIDs
  - offloads I/O work from yod

Runtime Environment



Runtime Plans

- fyod - distributed file servers on IO nodes
- Totalview parallel debugger
- PBS (portable batch system) batch scheduling
- Threads (waiting for Compaq compilers)
- Thread-aware MPI (talking to MPI Software Technologies, Inc.)
- Performance analysis (Jumpshot,Vampir?)

Sample Output

```
% time pingd
Awaiting status from bebopd...
Awaiting pct list from bebopd

            node              Job ID    PID/rank      owner         elapsed
  -------------------------     ------     -----------   ------------      ----------
  0 (    SU-000 n-00    )
  1 (    SU-000 n-01    )
  2 (    SU-000 n-02    )
  3 (    SU-000 n-03    )
  4 (    SU-000 n-04    )
  5 (    SU-000 n-05    )
  6 (    SU-000 n-06    )
  7 (    SU-000 n-07    )  node is unavailable.
  8 (    SU-000 n-08    )
  9 (    SU-000 n-09    )
 10 (    SU-000 n-10    )
          .  .  .
349 (    SU-021 n-13    )
350 (    SU-021 n-14    )
351 (    SU-021 n-15    )

Total: 339
Total busy: 0
Total free: 336
Total nodes unavailable: 3
Command exited zero status
0.22user 0.25system 0:00.90elapsed 53%CPU (0avgtext+0avgdata 0maxresident)k
0inputs+0outputs (291major+369minor)pagefaults 0swaps
```

Sample Output

```
% pingd
Awaiting status from bebopd...
Awaiting pct list from bebopd

            node              Job ID   PID/rank     owner        elapsed
 -------------------------    ------   -----------  ------------  ----------
 0 (   SU-000 n-00   )        101      191/0        bright            1:14
 1 (   SU-000 n-01   )        101      191/1        bright            1:14
 2 (   SU-000 n-02   )        101      191/2        bright            1:15
 3 (   SU-000 n-03   )        101      191/3        bright            1:14
 4 (   SU-000 n-04   )
 5 (   SU-000 n-05   )
 6 (   SU-000 n-06   )        102      197/0        lafisk             :30
 7 (   SU-000 n-07   )        102      197/1        lafisk             :29
 8 (   SU-000 n-08   )        102      197/2        lafisk             :30
 9 (   SU-000 n-09   )        102      197/3        lafisk             :30
10 (   SU-000 n-10   )        102      197/4        lafisk             :30
11 (   SU-000 n-11   )        102      197/5        lafisk             :30
12 (   SU-000 n-12   )        102      197/6        lafisk             :30
13 (   SU-000 n-13   )        102      197/7        lafisk             :30
14 (   SU-000 n-14   )        104      201/0        bright        1:01:44

Total: 15
Total busy: 13
Total free: 2
Total nodes unavailable: 0
```

Sample Output

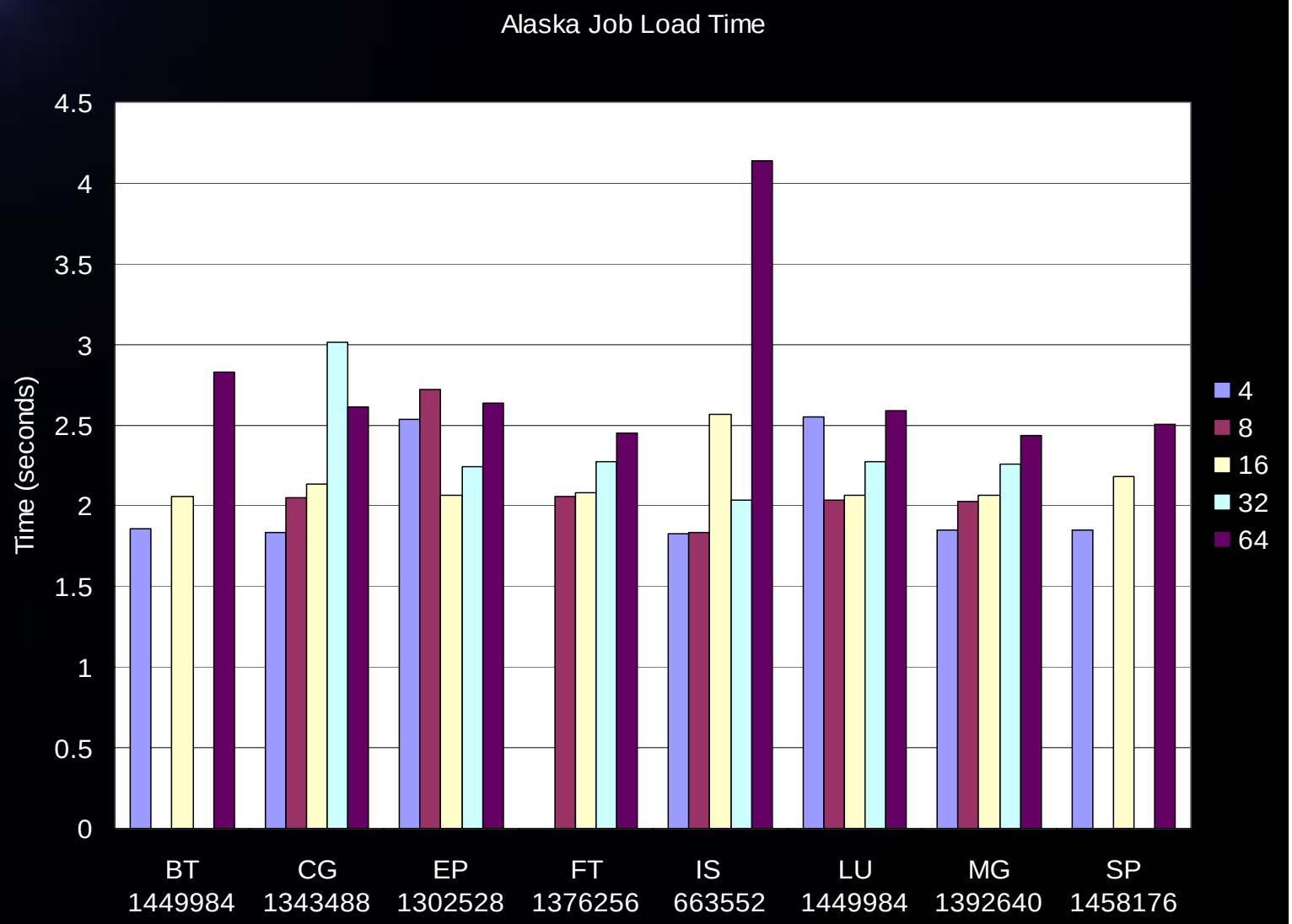
```
[root@c-15 lafisk]# yod -sz 5 -g tshello
Contacting node allocation daemon...
5 nodes are allocated to your job ID 109
Compute partition synchronized
Application processes begin user code.

>>> Terminating signal SIGFPE on node 21 <<<
>>> further messages suppressed

   Node name     Rank  Node   PID   Elapsed  Exit Code  Terminating Signal
---------------- ----  ----  -----  -------  ---------  ------------------
  SU-001 n-04       0    20    204  0:00:06       0       none
  SU-001 n-05       1    21    205  0:00:07
```

Sample Output

```
GNU gdb 4.17
Copyright 1998 Free Software Foundation, Inc.
Program received signal SIGFPE, Arithmetic exception.
__divqu () at ../sysdeps/alpha/divrem.h:165
#0  __divqu () at ../sysdeps/alpha/divrem.h:165
#1  0x1200028f0 in fpe (x=0x11ffff9f8) at ../tshello.c:32
#2  0x120002894 in main () at ../tshello.c:24
- - - - - - - - - - - - - - - - - - - - - -
  SU-001 n-06       2    22    196  0:00:06       0       none
  SU-001 n-07       3    23    196  0:00:06
- - - - - - - - - - - - - - - - - - - - - -
GNU gdb 4.17
Copyright 1998 Free Software Foundation, Inc.
Program received signal SIGFPE, Arithmetic exception.
__divqu () at ../sysdeps/alpha/divrem.h:165
#0  __divqu () at ../sysdeps/alpha/divrem.h:165
#1  0x1200028f0 in fpe (x=0x11ffff9f8) at ../tshello.c:32
#2  0x120002894 in main () at ../tshello.c:24
- - - - - - - - - - - - - - - - - - - - - -
  SU-001 n-08       4    24    196  0:00:06       0       none
```

Sample Output

```
% yod -sz 5 tshello
Contacting node allocation daemon…
5 nodes are allocated to your job ID 109
Compute partition synchronized
Application processes begin user code.
hello 0/203/0/0
hello 1/203/1/0
hello 2/203/2/0
hello 3/203/3/0
hello 4/203/4/0
hello 5/203/5/0
    Node name     Rank  Node   PID   Elapsed  Exit Code  Term Sig
 ---------------- ----  ----  -----  -------  ---------  --------
  SU-001 n-04       0    20    203  0:01:02       0       none
  SU-001 n-05       1    21    203  0:01:02       0       none
  SU-001 n-06       2    22    203  0:01:01       0       none
  SU-001 n-07       3    23    203  0:01:02       0       none
  SU-001 n-08       4    24    203  0:01:02       0       none
```


Alaska Job Load Time
Time (seconds)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
BT 1449984
CG 1343488
EP 1302528
FT 1376256
IS 663552
LU 1449984
MG 1392640
SP 1458176
4
8
16
32
64

Current Status

- 256-node MPLinpack job loaded in 2.6 seconds and achieved 87.1 GFLOPS
- Usage logs
- 400-node machine boots in 3 minutes
- Sys SW upgrade of all nodes in < 1 min.
- Machine in production mode

Plans

- Linux
  - physically contiguous memory
  - in-memory *exec()*
  - replace redirected libraries with Linux filesystem
- Intelligent allocator
- Parallel filesystem
- Portals 3.0
- Compaq Linux compilers

http://www.cs.sandia.gov/cplant

- Bill Davidson
- Lee Ann Fisk
- Tramm Hudson
- Mike Levenhagen
- Jim Otto
- Rolf Riesen
- John Van Dyke
- Lee Ward