Achieving Maximum Disk Performance on Large T3Es: Hardware, Software and Users

John Urbanic, Chad Vizino

Pittsburgh Supercomputing Center

Context

Actual Scientific Applications on Large (at least 256 PE) T3E's.

Motivation

Scientific codes have similar I/O patterns. Most save approximately 5 – 10 % of their data frequently (every so many 10's or 100's of time-steps) for Visualization purposes. And most save 20 – 40 % of their data less frequently, as well as at start-up and finish, for Checkpointing purposes.

Motivation

On a “production” run this can translate to
64 GB x 10% = 6.4 GB max, let’s say
3 GB for a typical visualization file.
64 GB x 40% = 25 GB max, or around
12 GB for a typical checkpoint file.

Motivation

If an application can run at 600 time-steps/hour, and it saves a visualization file every 10 time-steps at an all-too-typical bandwidth of 100 MB/sec across a 3 GB file, then:

0.5 hr / (1 hr + 0.5 hr) =

33% of runtime is I/O

Motivation

If the code is checkpointed every hour, add another 3% to the runtime for IO overhead. Now 36% of runtime is IO.

Motivation

If we can get bandwidths to ~1 GB/s then we improve to:

5% overhead for visualization I/O

and

<1% for checkpointing I/O

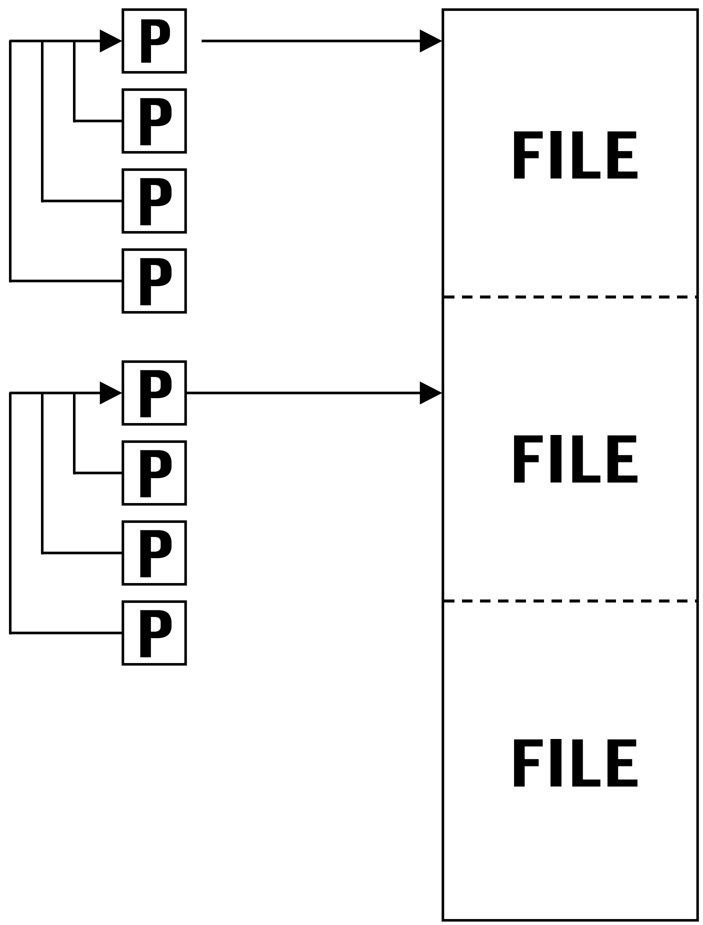
GFLOPS go way up.

Motivation

$\text{Performance}_{\text{Nature}} = \text{Performance}_{\text{Kernel}}\ (1 - 10\%)$

Ex: 100 GFLOPS code (during development) ends up running at 64 GFLOPS with real dataset. With tuned IO system, this could be 95 GFLOPS.

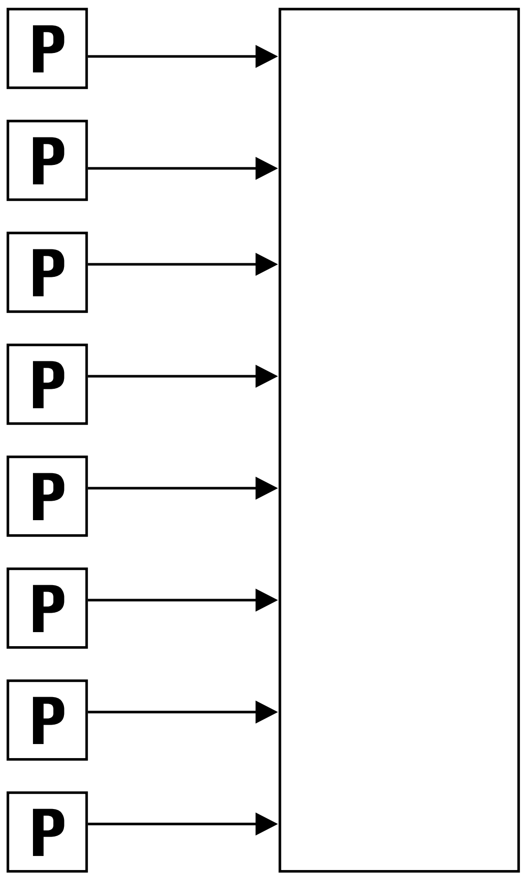
Knowledgeable Users



May be one file.

May be four or eight or...files.

Naive Users



One big file.

User Mix

We must accommodate both at PSC while we strive to educate users.

Cray T3E at PSC

- sn6301 -- first T3E delivered
- 544 Pes (512 app, 10 os, 22 cmd)
- 9 GigaRings
- 6 FCNs
- 140 DD-308s configured as 27 DA-308s
- /tmp file system: 25 DA-308s + 5 DD-308 = 1TB

The DD-308

- Seagate Barracuda 9FC
- Fibre Channel Interface
- Capacity: 9.5 GB
- Performance: 7-12MB/s

The DA–308

- Disk array of DD–308s (4 data + 1 parity)
- Capacity: 38GB
- Performance: 48MB/s

The GigaRing

- **One GigaRing is two counter-rotating 500MB/s rings**

The FCN

- Fibre Channel Node
- Allows attachment of fibre disks to GigaRing and ultimately to IO controllers on T3E
- Bandwidth: 240MB/s
- Supports up to 5 fibre loops (100MB/s each) with up to 125 disks on each loop

What We Had

- Lots of DD-308s
- Some DD-308s configured as DA-308s
- Poor documentation
- Default configuration
  - IO controller PE/GigaRing connections on nearby PEs

Before Configuration

```
   RING 4           RING 5           RING 6           RING 7           RING 8           RING 9
   FCN  0           FCN  1           FCN  2           FCN  3           FCN  4           FCN  5
0  1  2  3  4    0  1  2  3  4    0  1  2  3  4    0  1  2  3  4    0  1  2  3  4    0  1  2  3  4
-------------    -------------    -------------    -------------    -------------    -------------
R  R  R  O  O    R  R  R  R  O    R  R  R  O  O    R  R  R  O  X    R  R  R  O  X    R  R  R  O  X
A  A  A  |  |    A  A  A  A  |    A  A  A  |  |    A  A  A  |       A  A  A  |       A  A  A  |
I  I  I  O  O    I  I  I  I  O    I  I  I  O  O    I  I  I  O       I  I  I  O       I  I  I  O
D  D  D  |  |    D  D  D  D  |    D  D  D  |  |    D  D  D  |       D  D  D  |       D  D  D  |
         O  O                O             O  O             O                O                O
         |  |                |             |  |             |                |                |
         O  O                O             O  O             O                O                O
         |  |                |             |  |             |                |                |
         O  O                O             O  O             O                O                O
            |
            O
            |                        Key: RAID = DA-308
            O                             O    = DD-308
            |                             X    = no connection
            O
            |
            O
            |
            O
```

HW Steps

- Relocated GigaRing connections throughout the torus
- Rebalanced DD-308s evenly across all FCNs
- Reconfigured 125 DD-308s into 25 DA-308s
- Put each FCN on dedicated GigaRing

After Configuration

```
RING 4         RING 5         RING 6         RING 7         RING 8         RING 9
   FCN  0         FCN  1         FCN  2         FCN  3         FCN  4         FCN  5
0  1  2  3  4  0  1  2  3  4  0  1  2  3  4  0  1  2  3  4  0  1  2  3  4  0  1  2  3  4
-------------  -------------  -------------  -------------  -------------  -------------
R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  O  X  X
A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  A  |
I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  O
D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  D  |
                                                                                   O
    Ts0            Ts1            Ts2            Ts3            Ts4                |
                                                                                   O
                                                                                   |
                                                                                   O
                                                                                  Tp


Key: RAID = DA-308
     O    = DD-308
     X    = no connections

/tmp = {Tp0, Tp1, Tp2, Tp3, Tp4} + {Ts0, Ts1, Ts2, Ts3, Ts4}
```

Software Steps

- Set up 5 stripe sets of 5 DA-308s each (on each FCN)
- Setup 5 OS Pes, each running file, disk and packet servers to handle each FCN/GigaRing pair to reduce inter-processor communication among OS PEs

File System Layout

- Each software stripe set of 5 DA-308s forms one secondary partition of /tmp
- 5 DD-308s form primary partitions of /tmp
- /tmp: 5 secondary areas delivering up to 240MB/s each, 5 primary areas delivering up to 7-12 MB/s each

File System Layout Continued

- Used file size distribution to help determine optimal cutoff for “big” files (primary/secondary threshold)
- Found 95% of files on /tmp taking up < 1% of allocated space
- Selected 1MB primary cutoff

Application Steps

- Use setf (or similar) to preallocate file(s) precisely on secondary /tmp partitions
- Can use fck to make sure you got what you asked for
- Again: make sure representative PE for a PE group performs IO instead of every PE

System Performance Benefits

- Checkpoint/Restarts much faster since /tmp used for checkpoint files
- NQS shutdown time nearly cut in half from over 18 minutes down to about 10 minutes
- These were daily benefits due to nightly scheduled dedicated runs
- Live dumps much faster

Application Performance Benefits

- Naïve large-file users can benefit from the fast secondary partitions (up to 240 MB/s)
- Knowledgeable users can exploit the file system layout and spread file(s) over 5 secondary areas of /tmp, achieving over 1 GB/s aggregate bandwith (5 x 240 MB/s)

Application Performance Benefits

- Cut naïve dedicated users' IO wait time in half
- Improved knowledgeable users' bandwidth by ~10x

Conclusion (Programmers)

- Realize your system administrators may be able to make considerable IO system improvements
- This means bigger FLOPS

Conclusion (Sys Admins)

- **Know that IO system has direct impact on GFLOPS code performance for programmers**

Reference

- **Kent Koeninger's *GigaRing System View of IO* paper presented at CUG '97**

Request to Cray/SGI

- Would like to see up-to-date Performance Tuning Guide (similar to last one published for UNICOS 8.0)
- More documentation/specs on specific device/component limits