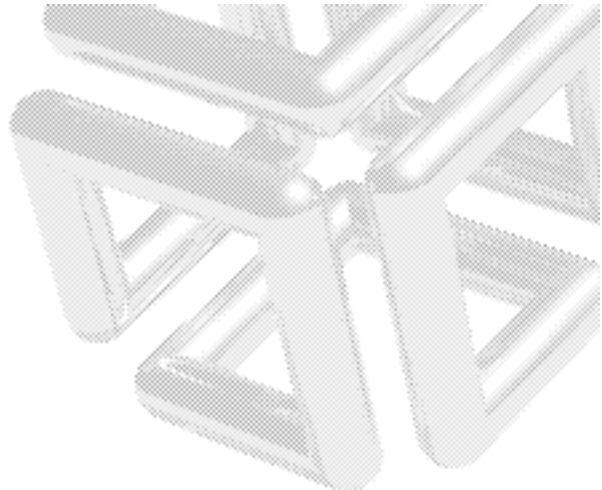




T3E Scalable I/O

Mark DuChene - Filesystem Engineer

Brian Gaffey - Engineering Manager

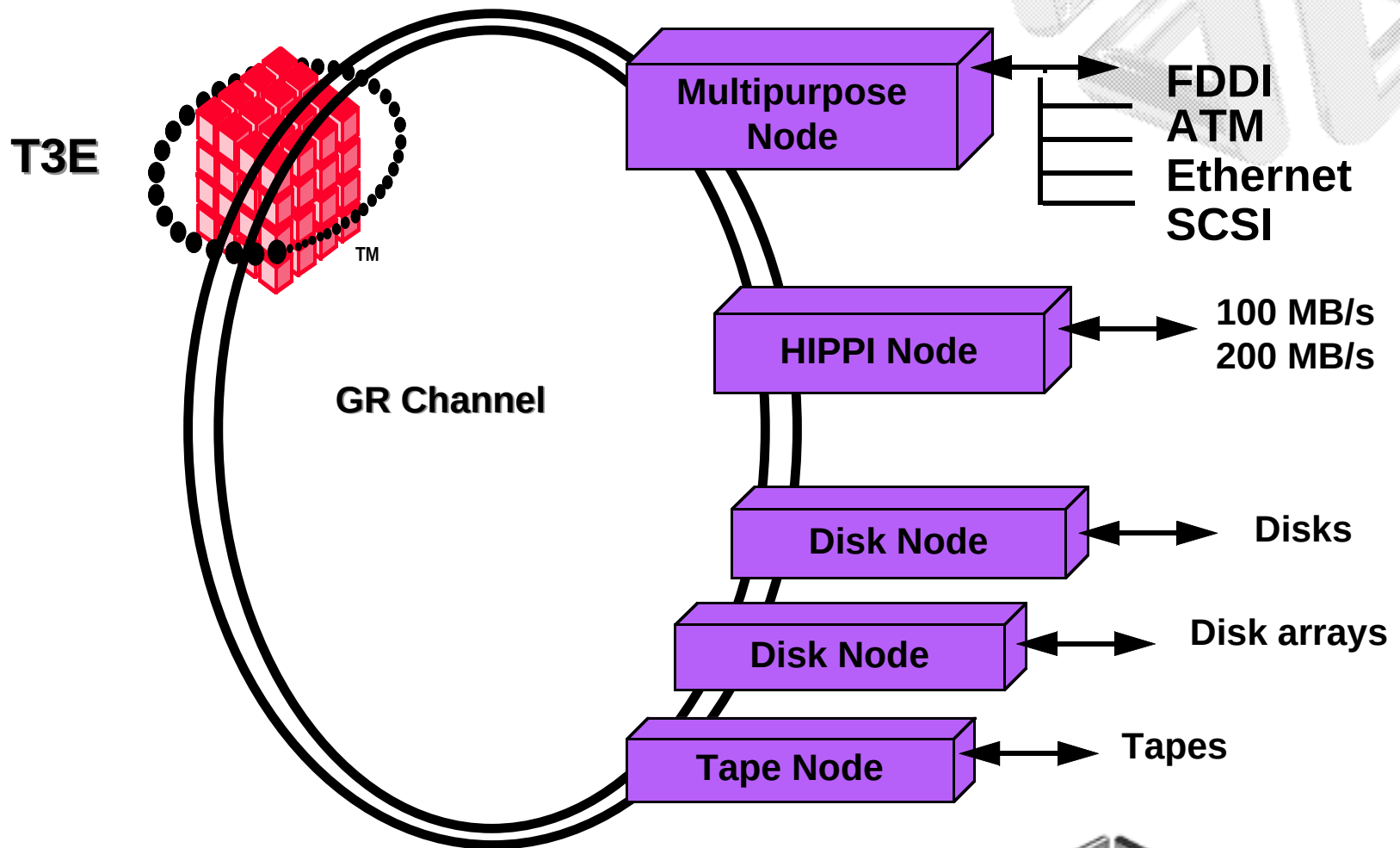
Topics

- *Hardware Overview*
- *Software Overview*
- *System I/O Flowtrace*
- *I/O Scalability Methods*
- *Example Programs*
- *Performance Data Analysis*

Purpose

- ***Our goal was to provide applications with sufficient file system bandwidth to allow scaling of applications from a handful of processors to thousands***
- ***Our approach is to allow the use of as many of the hardware components as desired***
- ***No programming changes required***

Hardware Overview



T3E Scalable I / O

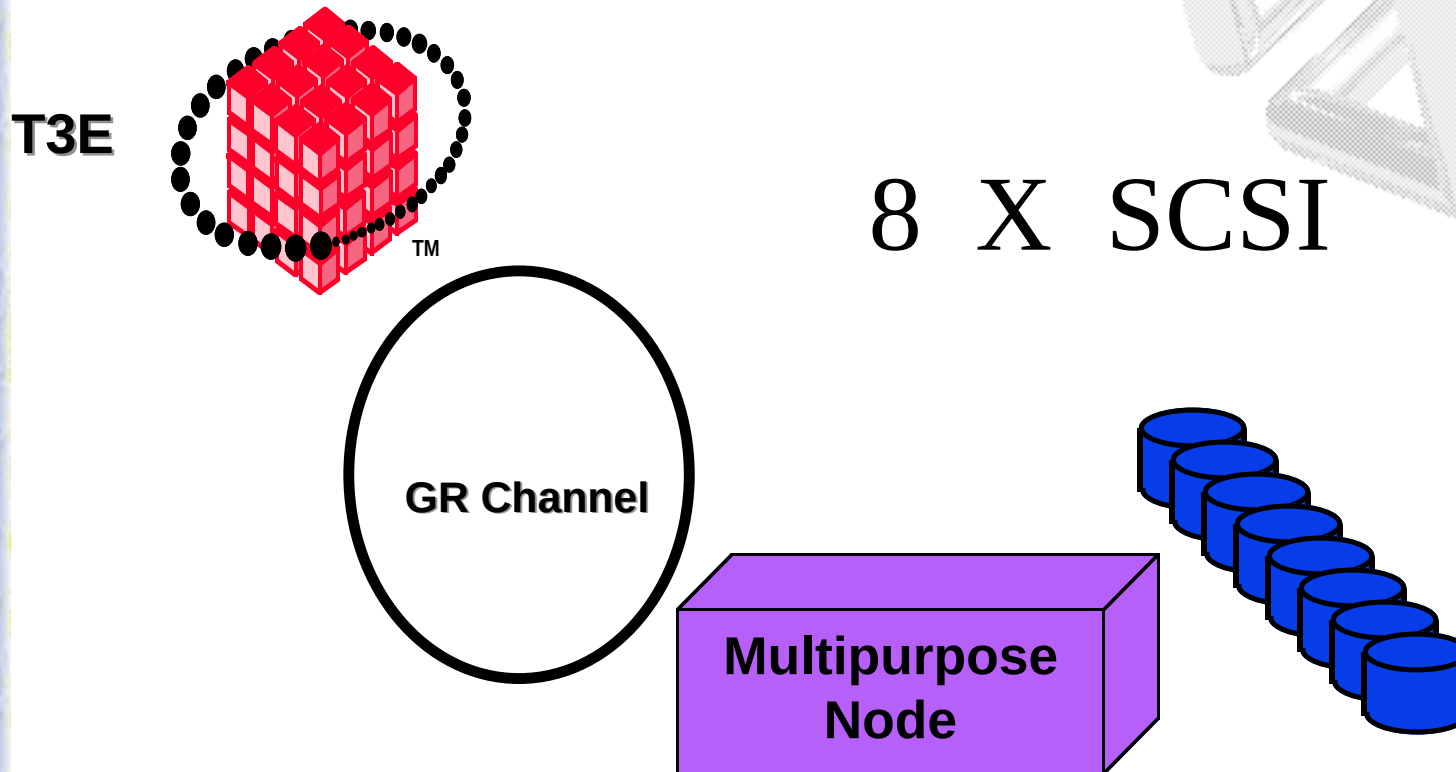
Cray Proprietary



CRAY
RESEARCH
A Silicon Graphics Company

Hardware Overview

Multiple Drives



T3E Scalable I / O

Cray Proprietary



Performance Information

- ***I/O bandwidth and scalability***
- ***Disk drive performance profiles***
- ***I/O capacity and scalability***

Test Configuration

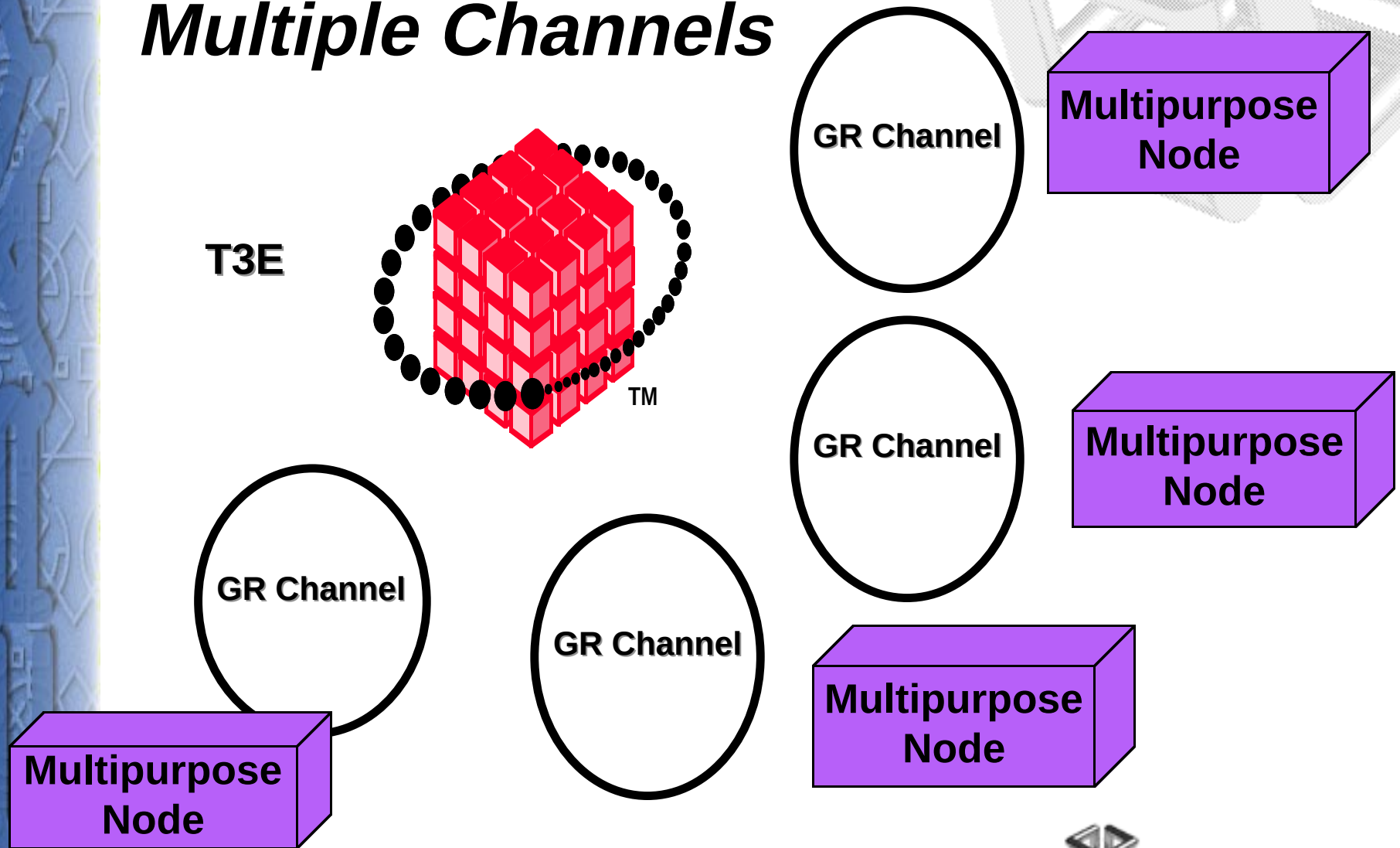
- ***Cray T3E AC64 (68 PEs) - sn6549 (acidrain)***
- ***16MWords of memory per PE***
- ***4 Gigaring***
- ***1 MPN-1 per ring***
- ***2 SBUSs per MPN-1 (except for one)***
- ***2 - 3 SCSI adapters per SBUS***

SDD (*Standard Data Disclaimer*)

- **Current development software at the time of testing**
- **The purpose of the data used during this presentation is to highlight the decisions and performance of the scalability features available on Unicos/MK platforms.**
- **For current device specific performance contact CRI**

Hardware Overview

Multiple Channels



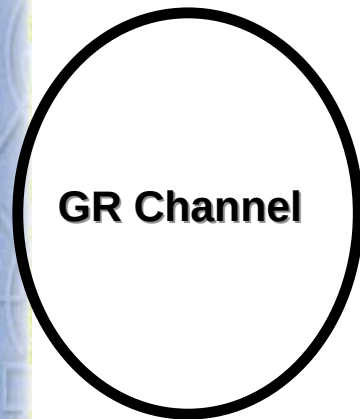
T3E Scalable I / O

Cray Proprietary

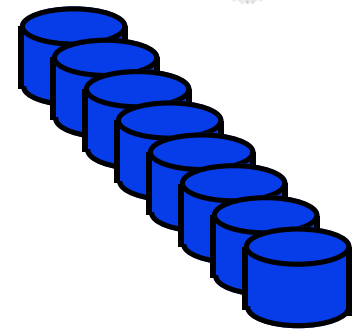
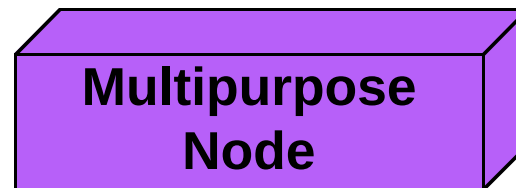


CRAY
RESEARCH
A Silicon Graphics Company

T3E I / O Rates



70 MB/ sec
(Goal 68Mb)

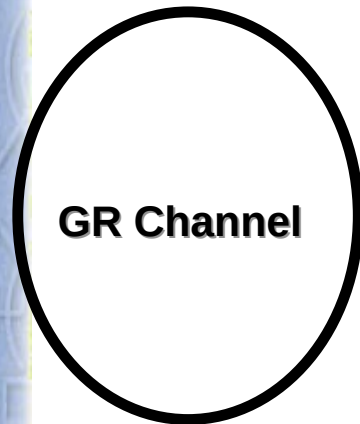


1200 MB/ sec

DD314 7.6 MB/ sec

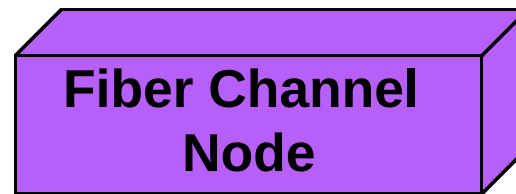
DD318 12 MB /sec

T3E I / O Rates

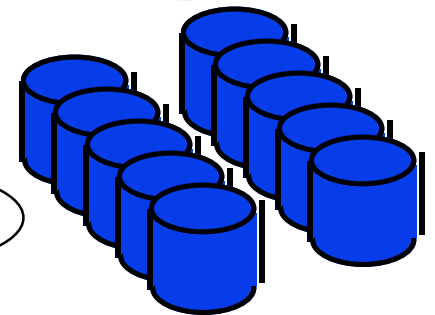
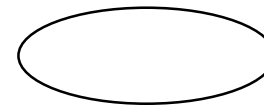


GR Channel

110 MB/ sec
(Goal 100Mb)



**Fiber Channel
Node**



1200 MB/ sec

DD308 11.4 MB/ sec

Scalable I/O Architecture

■ *Industry Leading I/O Performance*

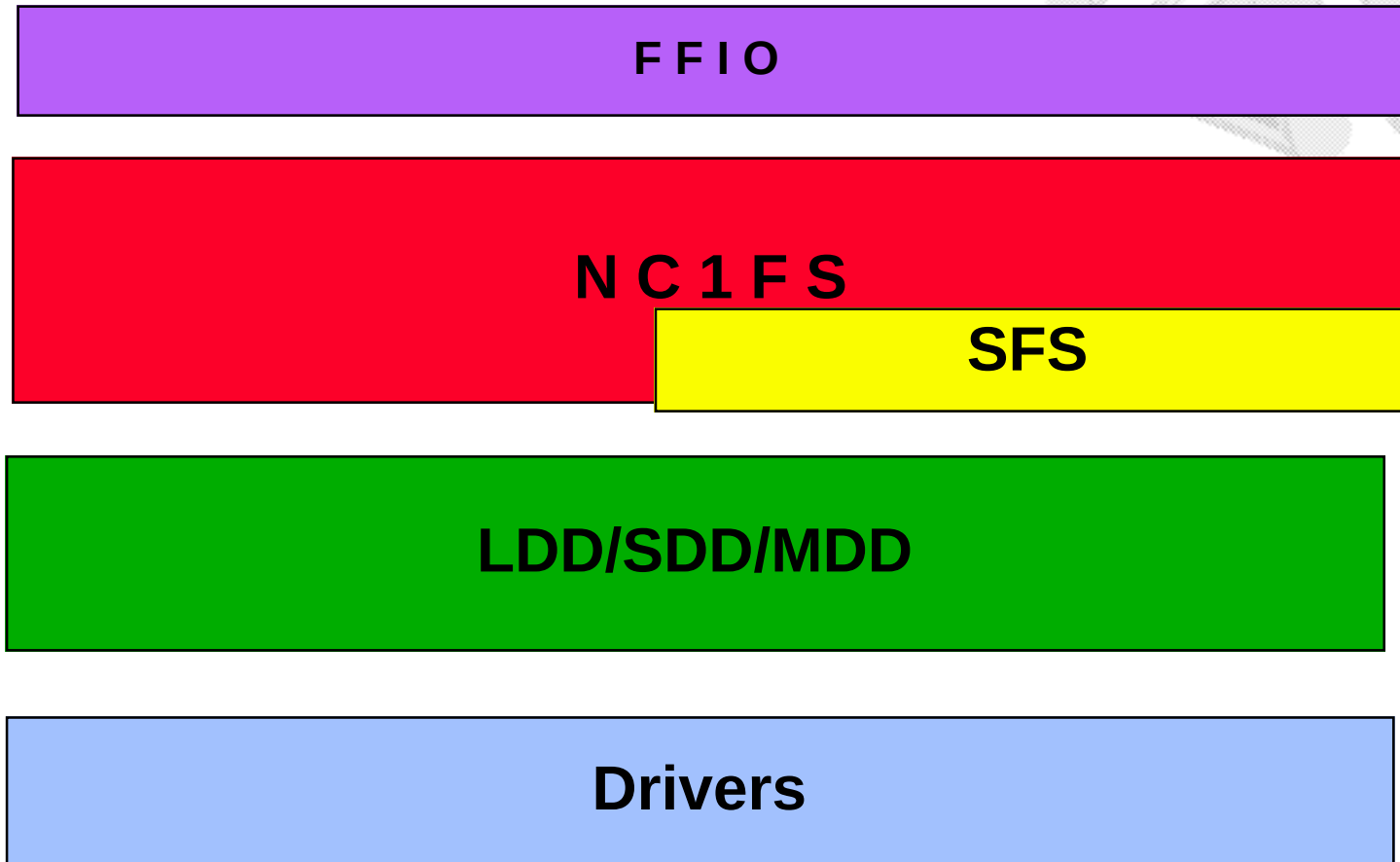
- 1200 MB/s GigaRing channel rate***
- ~800 MB/s data payload half duplex***

■ *Scalable configurations*

- Point-to-point or ring topology***
- Performance and/or capacity***



UNICOS Software Overview



UNICOS/mK Software Overview

File Server

servlib/wrapper

N C 1 F S

LDD/SDD/MDD

Disk Server

Packet Server

P S

SCSI

Fiber

IPI

HIPPI

T3E Scalable I / O

Cray Proprietary



CRAY
RESEARCH
A Silicon Graphics Company

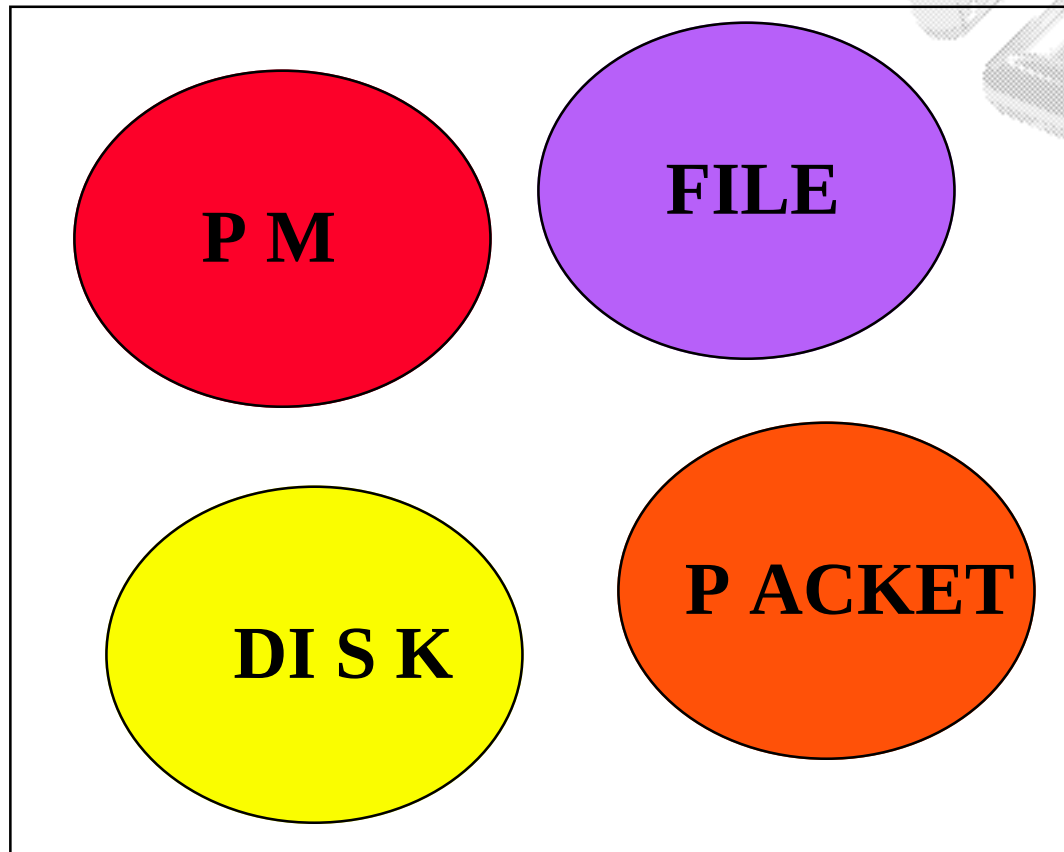
Parallel IO Programming

- ***No changes required for scaling***
- ***POSIX compliant***
- ***Administrator controlled scaling techniques***
- ***Programming optimizations***

Software Scalability Methods

- ***Multiple Disk & Packet Servers***
- ***Disk Striping***
- ***Disk Banding***
- ***Remote Mount***
- ***File Server Assistant (FSA)***
- ***Distio***
- ***Pcache***

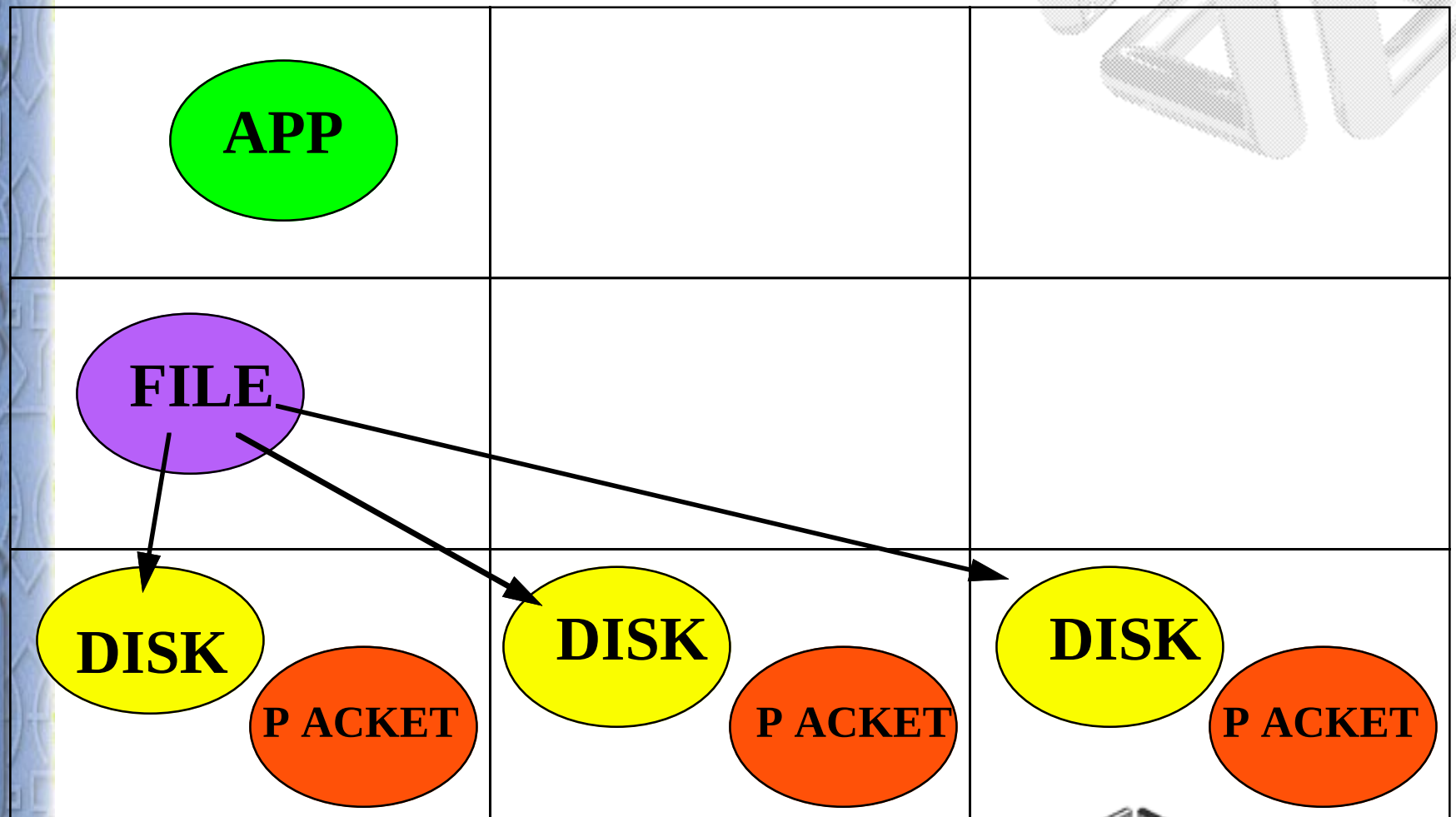
Default OS Node Configuration



Multiple Disk & Packet Servers

- ***Provides scalability by duplicating the function of the disk server.***
- ***Duplicate copies of the disk server code maintaining reliability and performance.***

Multiple Disk and Packet Servers



T3E Scalable I / O

Cray Proprietary



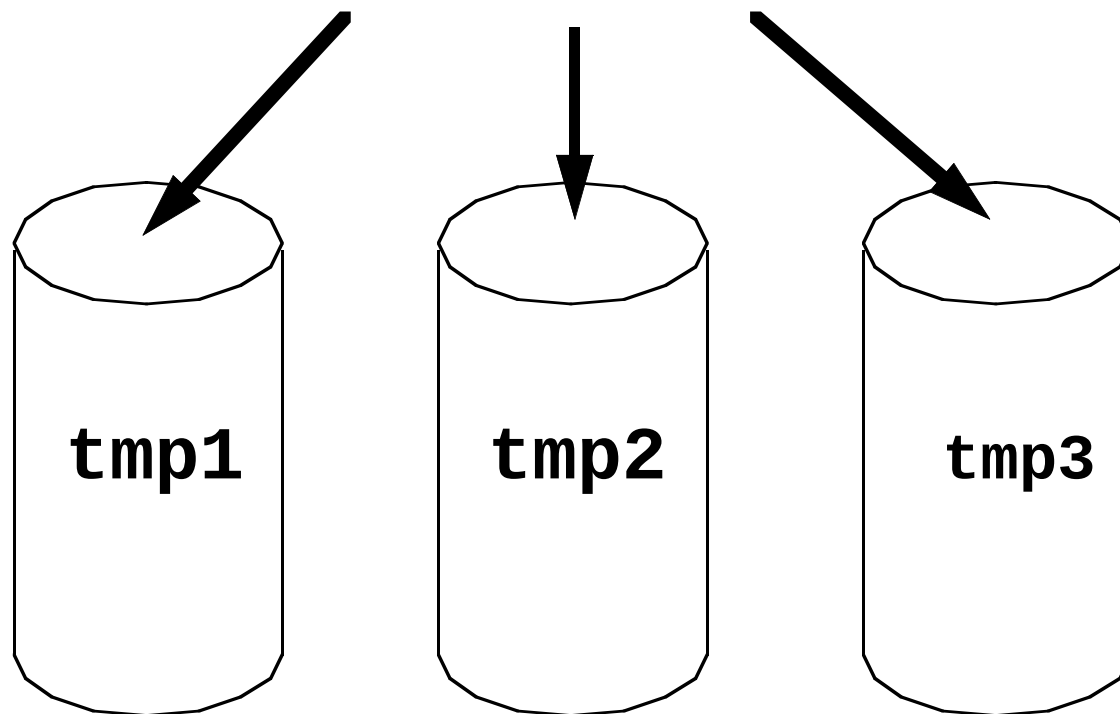
CRAY
RESEARCH
A Silicon Graphics Company

Disk Striping / Banding

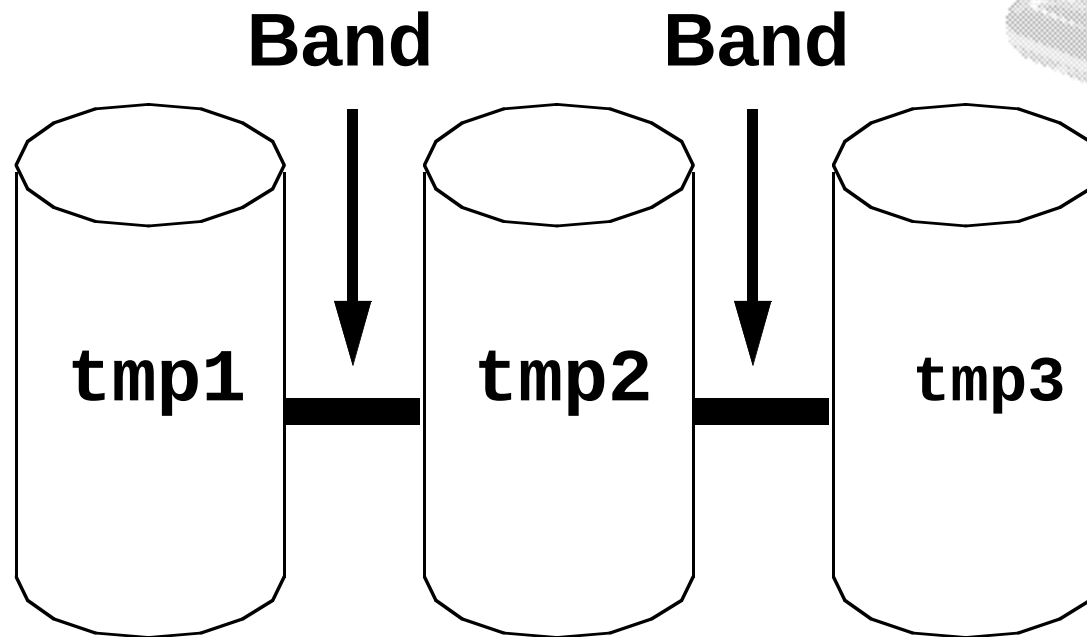
- ***Improved disk I/O performance***
- ***Improved disk I/O bandwidth***
- ***Known, well understood techniques***
- ***Admin controlled***

Disk Striping

Logical Partition



Disk Banding



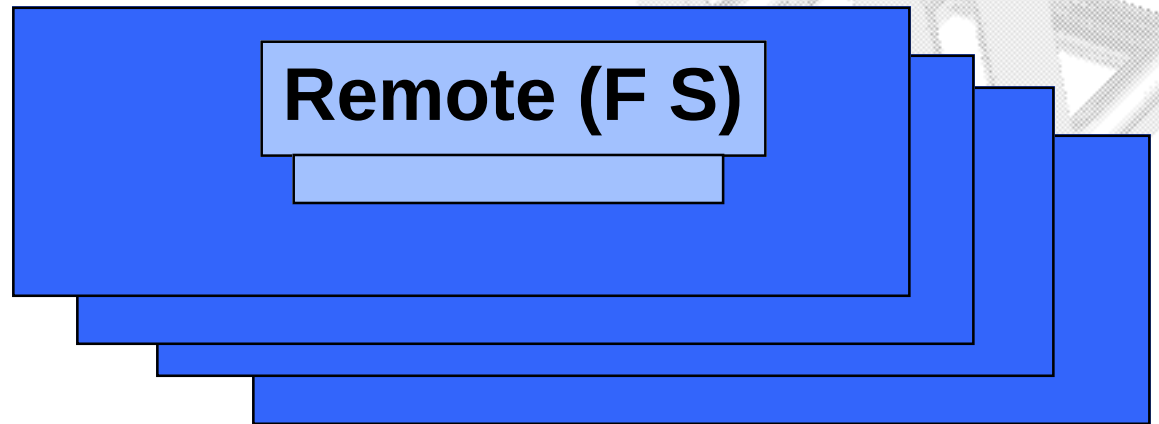
Remote Mount

- ***Provides the capability for multiple file servers running on OS PE's.***
- ***Greatly improves the system I/O scalability on a per filesystem basis.***
- ***Duplicate file server maintaining reliability.***
- ***Simple configuration (under admin control).***

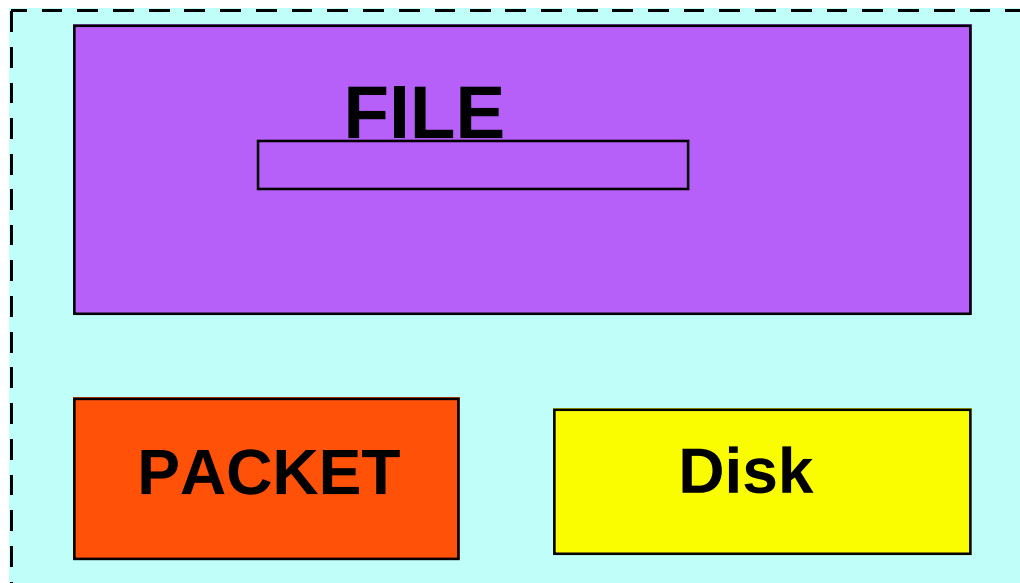


Remote Mount

Using a PE
per mounted
filesystem



P E 6-9



PE 5

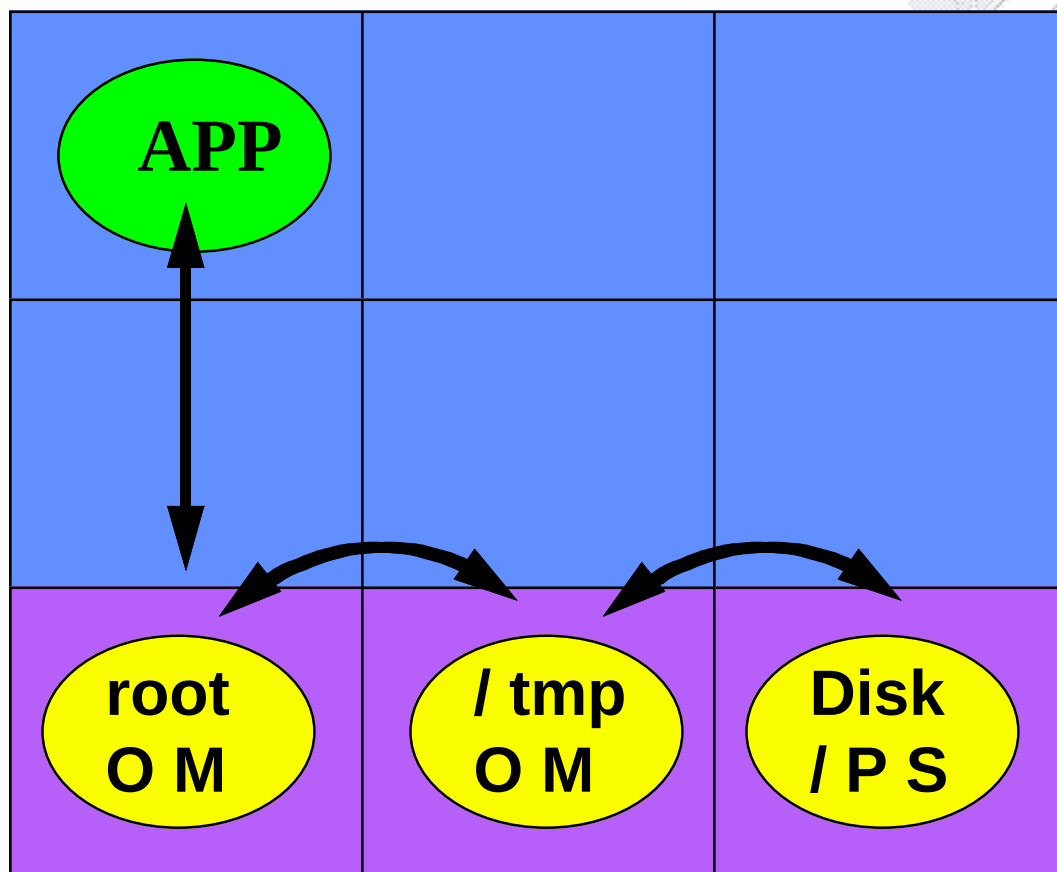
T3E Scalable I / O

Cray Proprietary

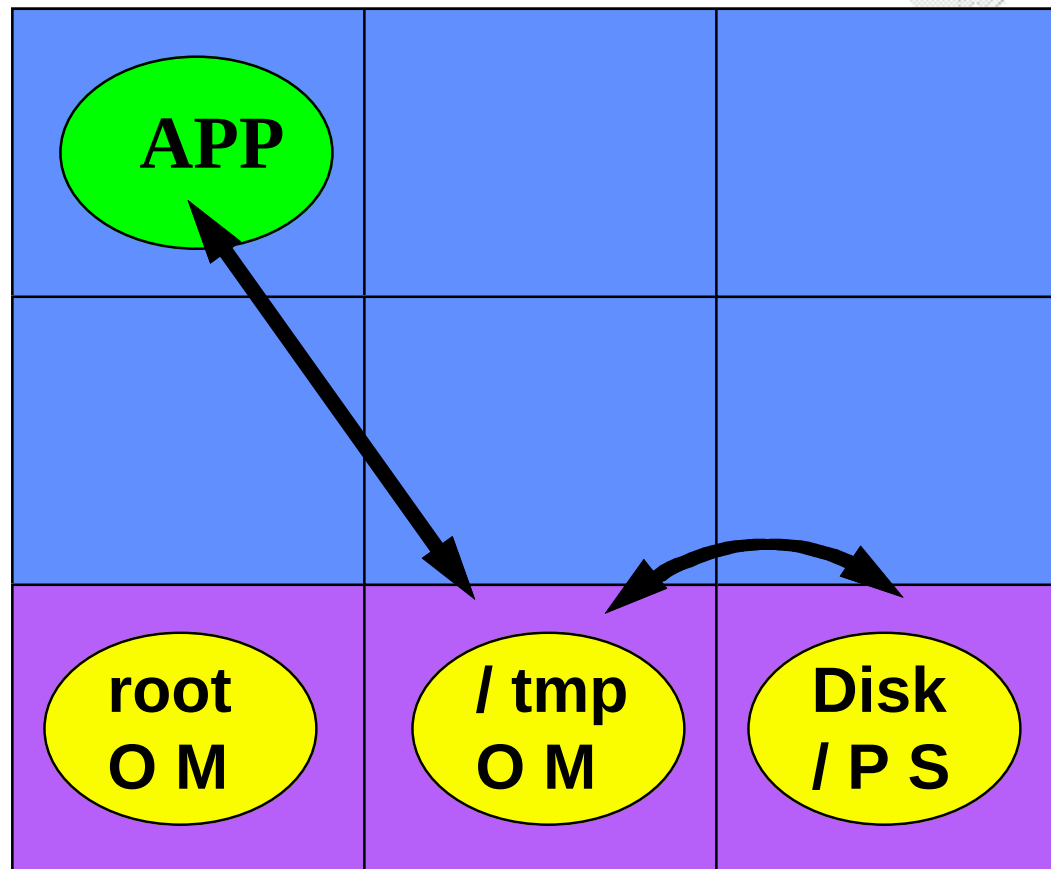


CRAY
RESEARCH
A Silicon Graphics Company

Remote Mount - Open



Remote Mount - Data



Remote Mount Configuration

All filesystems not assigned a file server
default to the root server

ldd[0].name = ***“workfs”***;

ldd[0].minor = ***2***;

ldd[0].file_pe = ***“ospe_c”***;

Remote Mount Configuration (Cont)

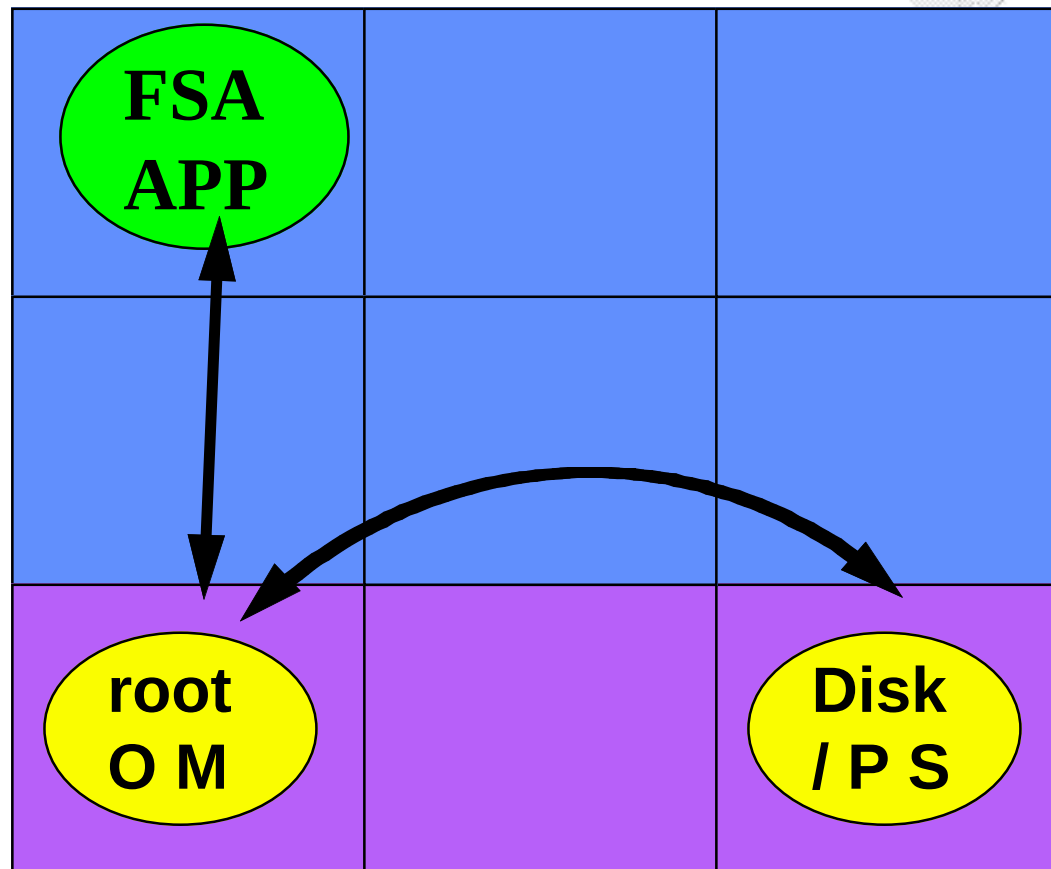
***Adding the additional file server(s)**

```
mf[0].mpp.pe_actors[5].num_actors = 6;  
mf[0].mpp.pe_actors[5].actor_name[0] = "kernel";  
mf[0].mpp.pe_actors[5].actor_name[1] = "em";  
mf[0].mpp.pe_actors[5].actor_name[2] = "PM";  
mf[0].mpp.pe_actors[5].actor_name[3] = "packet";  
mf[0].mpp.pe_actors[5].actor_name[4] = "disk";  
mf[0].mpp.pe_actors[5].actor_name[5] = "file";
```

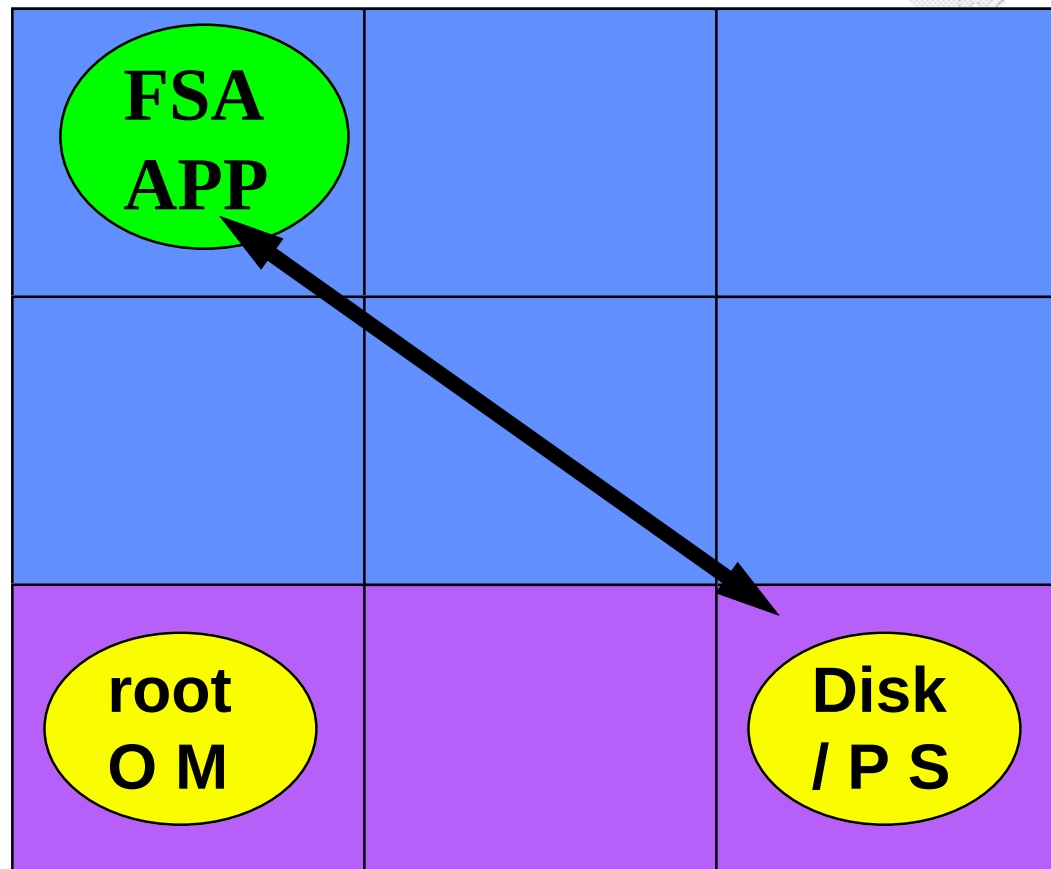
File Server Assistant (FSA)

- ***Provides the capability for a small FSA server to reside on APP PE's.***
- ***FSA improves scalability on a per file basis for wellformed raw I/O.***
- ***Requires only a minor open flag alteration to the applications open statement.***
- ***Server is composed of a subset of the file server code for reliability & resilience.***

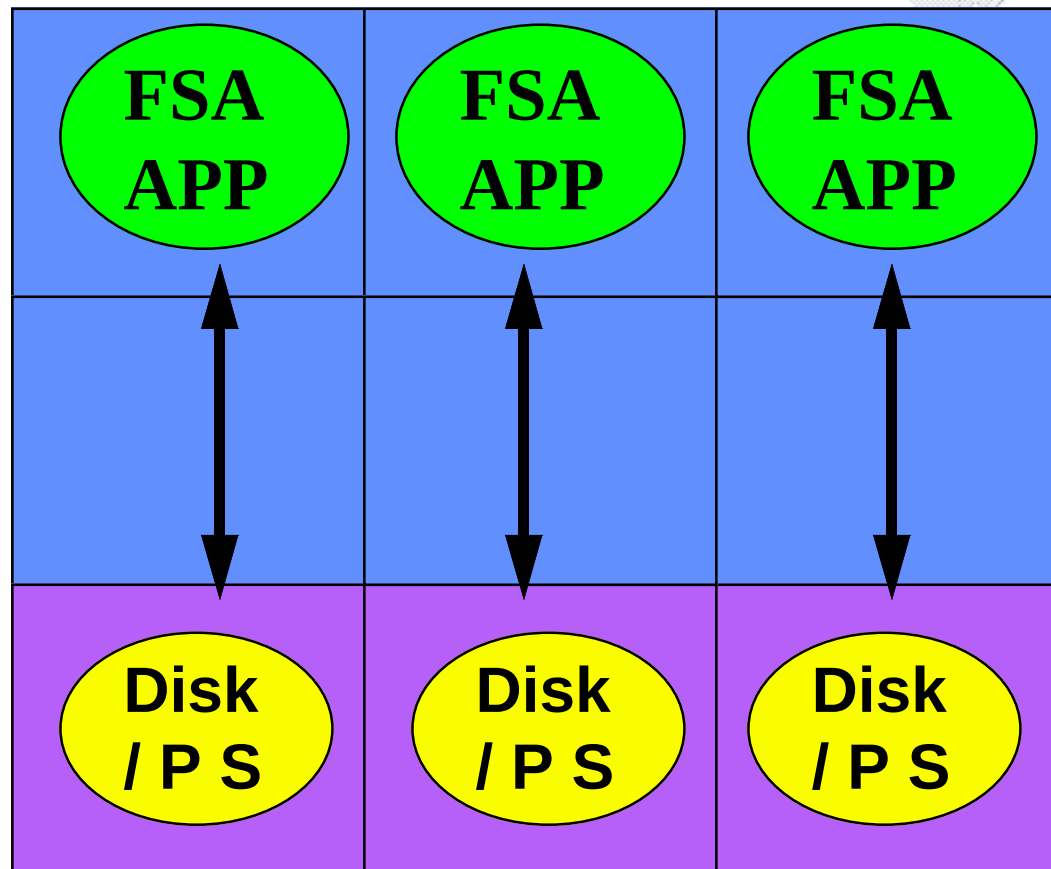
File System Assistant - Open



File System Assistant - I/O path



File System Assistant - Parallel



T3E Scalable I / O

Cray Proprietary



FSA Programming example

```
#pragma _CRI cache_align buffer
```

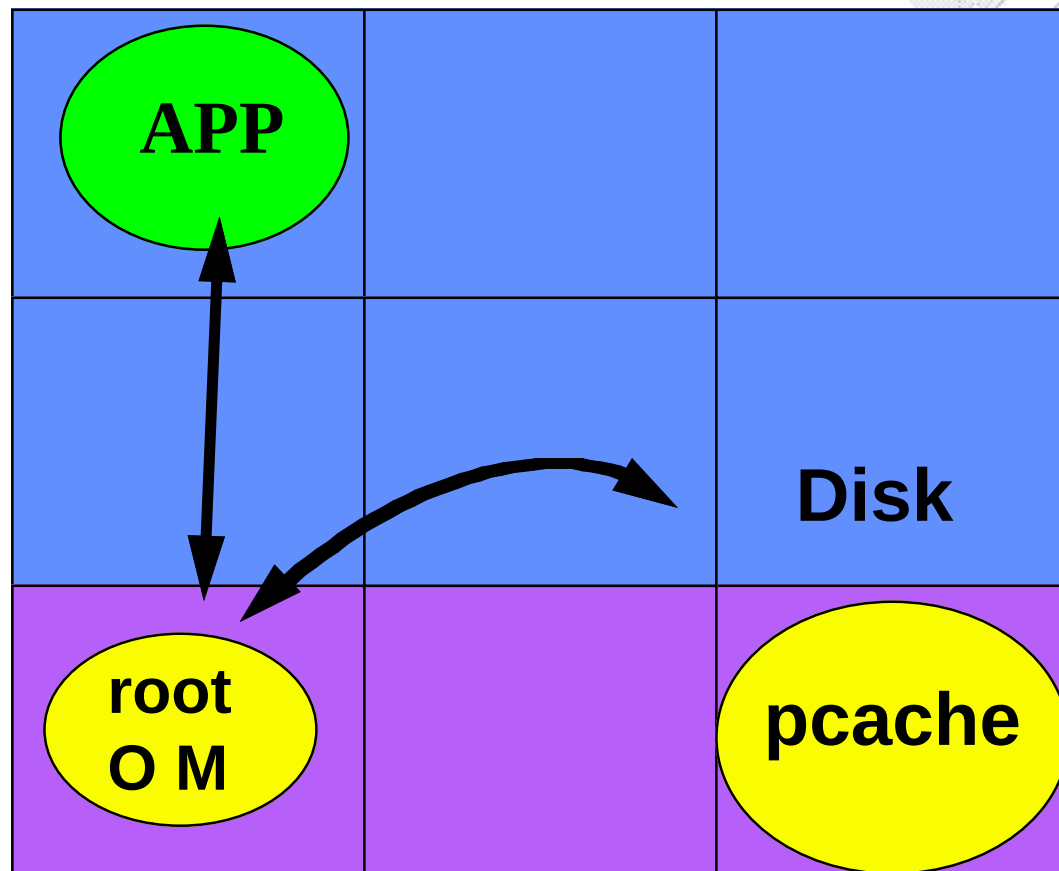
```
char  buffer[BSIZE * 64];
```

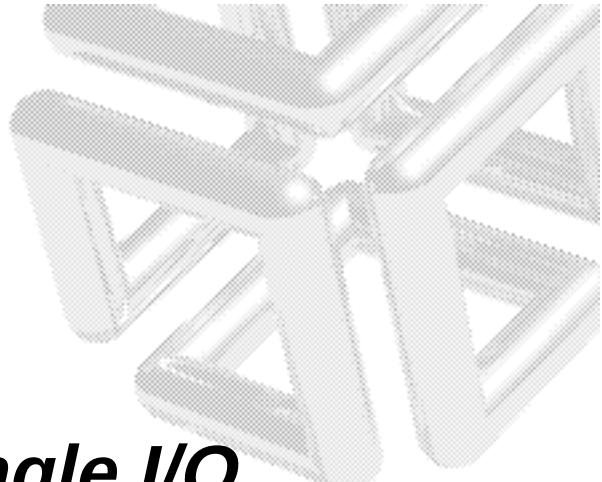
```
fd = open ("test.fsa", O_RDWR | O_RAW |  
O_WELLFORMED | O_PARALLEL);
```

Pcache

- *Provides a disk partition cache at the device driver level within the disk server.*
- *Effectively uses OS PE memory to increase I/O performance and scalability.*
- *Administrator controlled within the configuration file and admin commands.*

Partition Cache (pcache)

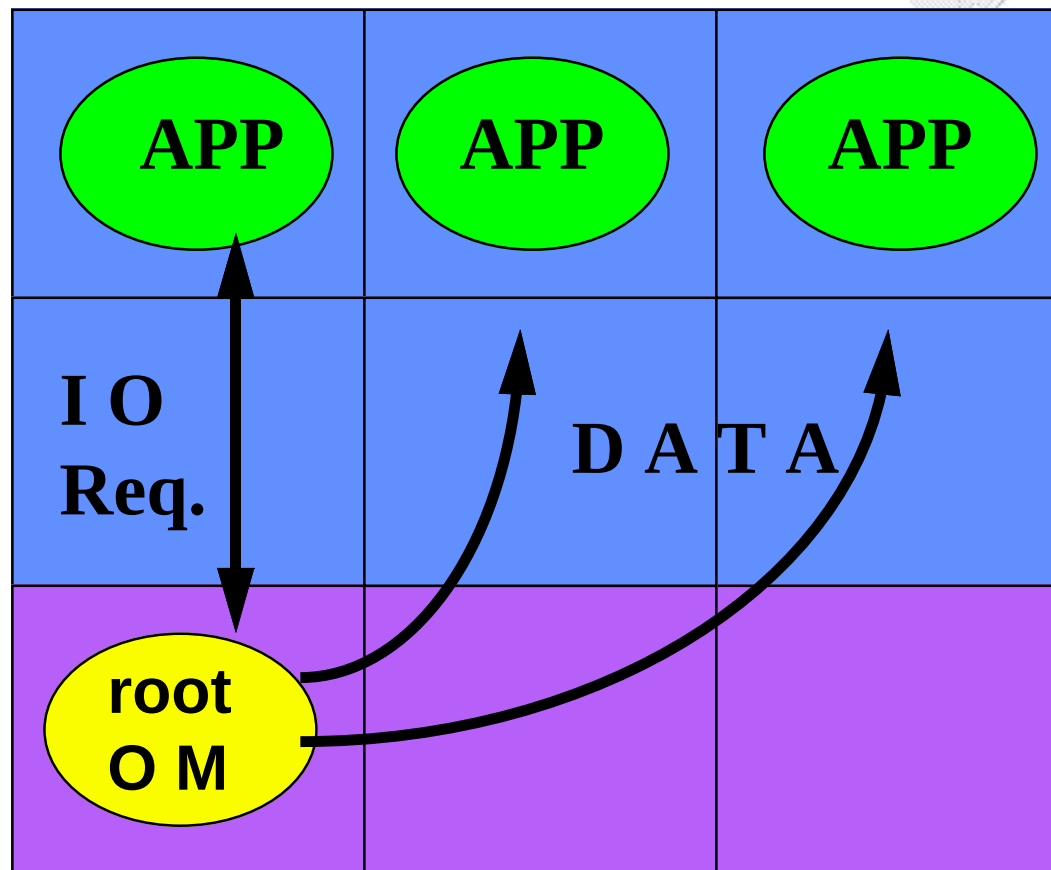




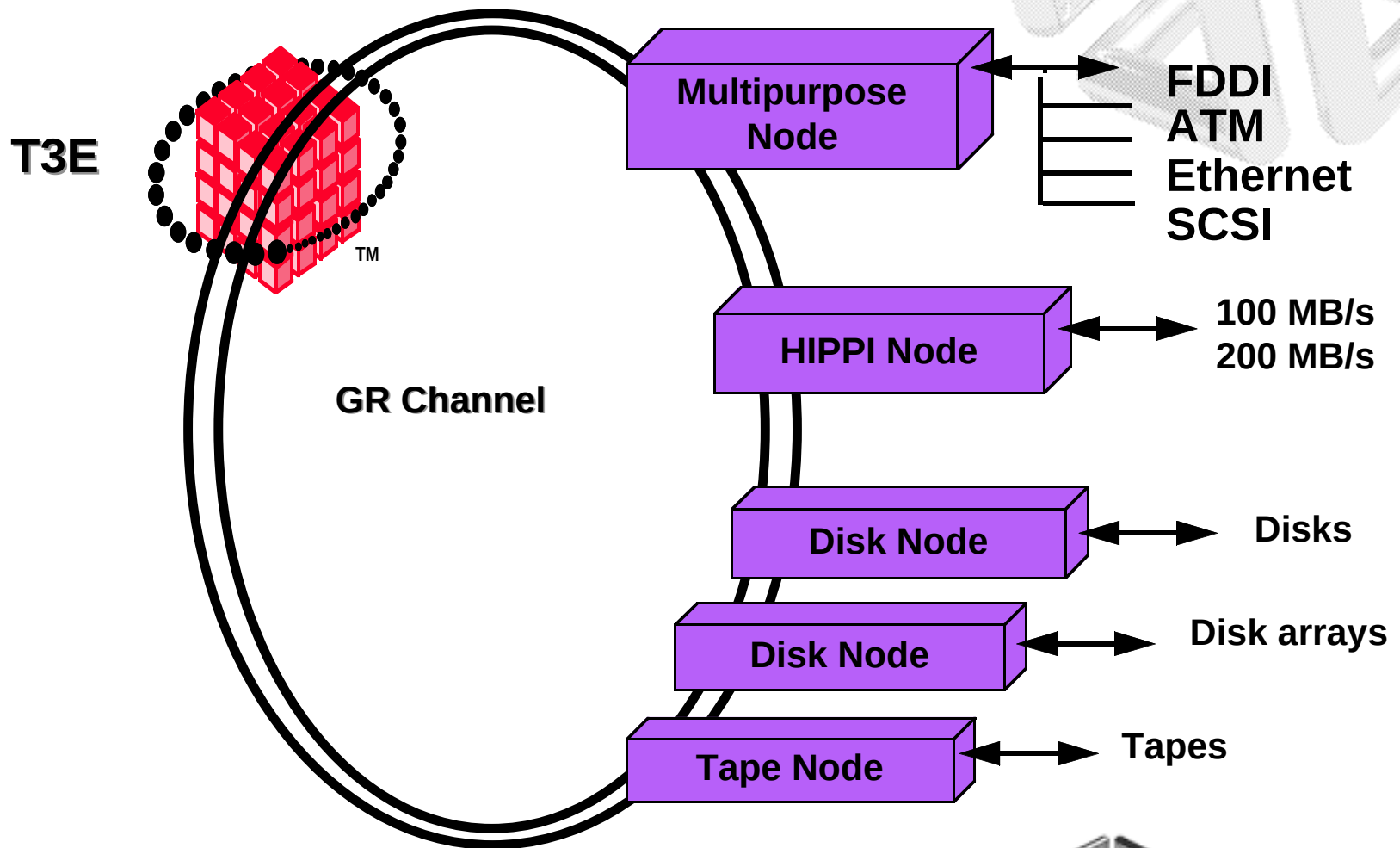
Distio

- ***Provides for the data of a single I/O request to be dispersed across user application PE's.***
- ***Implemented via the existing listio system call.***

DISTIO Data Flow



Questions ?



T3E Scalable I / O

Cray Proprietary



CRAY
RESEARCH
A Silicon Graphics Company