Parallel Data Lab Research Overview

Garth A. Gibson, Carnegie Mellon Univ.
http://www.pdl.cs.cmu.edu

Reliable, Parallel Storage Subsystems (RAID)

- configurable architectures; rapid prototyping; completed

Discovering/Managing Storage Parallelism (TIP)

- cost-benefit exploitation of application disclosure

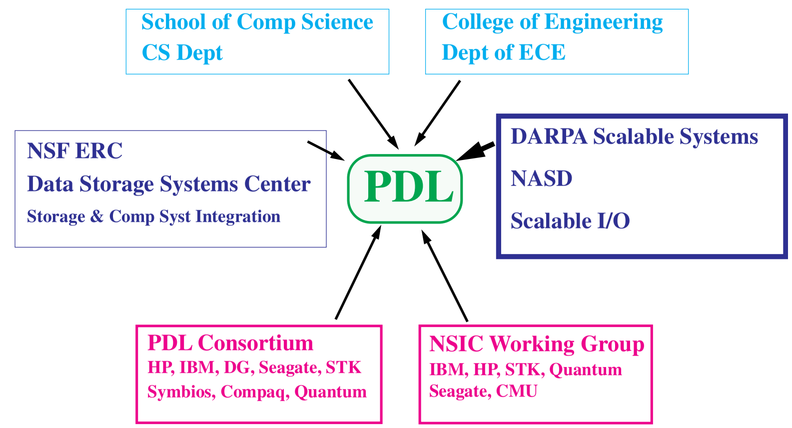
Parallel Filesystems for Parallel Programs (LLAPI)

- application “controls”: hints, cache directives, redundancy

New Interfaces for Network-Attached Disks (NASD)

- scalable, secure, extensible storage systems

Parallel Data Lab Organization



Relationship with DSSC

PDL work increasingly funded outside of DSSC

- almost all effort now in ARPA funded NASD project

PDC membership changed for NSIC NASD group

- NSIC NASD working group members bound by intellectual property sharing agreement
- PDC membership level 2 changed to accomodate PDL involvement in NSIC NASD
- DSSC membership rights inappropriate for NASD project

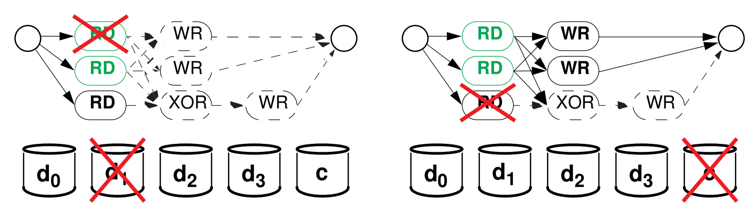
New view: PDL is independent partner of DSSC

- DSSC members need to actively seek PDC membership (ie. redirecting DSSC membership funds) if desired

Rapid Prototyping with RAIDFrame Simplifies Coding

RAID architecture as program

- separate policy & mechanism; RAID-unaware graph engine
- automate error recovery; retry uses different graph



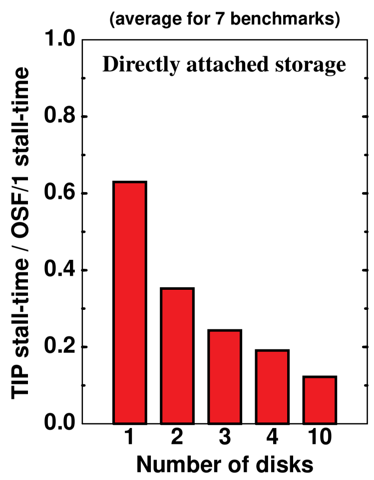
unsuccessful commit, roll back
retry using new method

successful commit, roll forward
operation is complete

Code/Docs released 9/96 - project complete

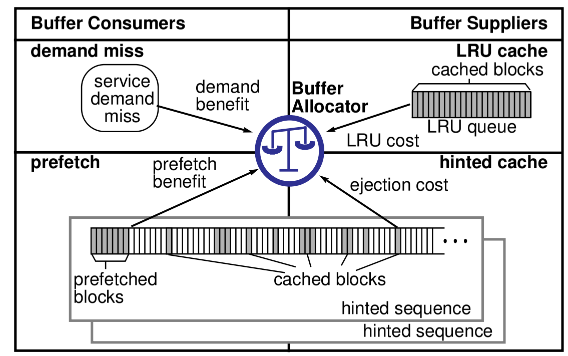
- http://www.pdl.cs.cmu.edu/RAIDframe

Informed Filesystems: "Parallelizing" Application I/O



- **Application discloses future accesses**
- **Exposes concurrency**
  - **overlap I/O and computation**
  - **overlap I/O and think time**
  - **overlap I/O and I/O !!!!**
  - **I/O optimization**
    - seek scheduling
    - batch processing
- **Cache management**
  - **balance buffers between prefetch and demand**

How does TIP work?



Estimate:

- ***benefit* of giving a buffer to a *consumer***
- ***cost* of taking a buffer from a *supplier***

Reallocate when *benefit > cost*

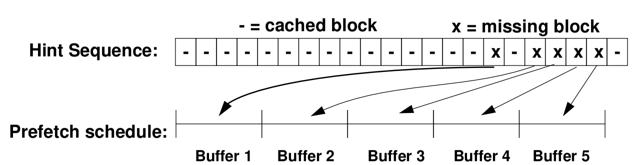
Recent Results: TIPTOE

New TIP estimator for unbalanced disk loads

- **unfortunate strided storage access; insufficient disks**
- **prefetch more deeply to avoid I/O stall provided savings per buffer per access is large enough**

Compared to LRU-SP (Princeton/Washington)

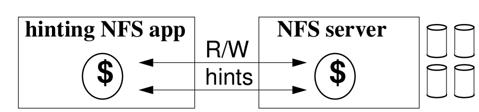
- **LRU-SP allocates buffers without regard to locality**
- **largest impact when multiple I/O-bound applications**



Recent Results: Remote TIP

Prefetching hides latency in remote filesystem

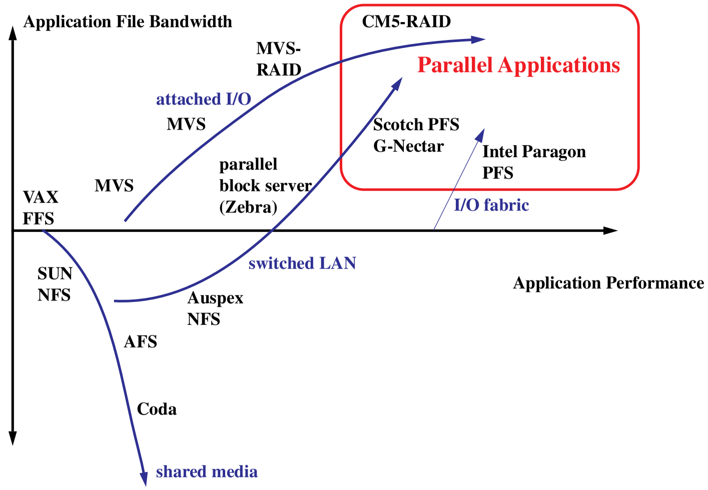
- **remote access has more overhead, congestion = delay**
- **informed prefetching effective given efficient networking**
- **three organizations: TIP in client only, TIP in server only, TIP in both (for smart use of distributed buffering)**
- **prototypes built for first two organizations**



- **ie., RTIP in client only**
- **Gnuld 361 modules over ATM**
- **data striped on 3 disks**

	NFS	Local
Hints	43s	43s
Nohints	126s	104s

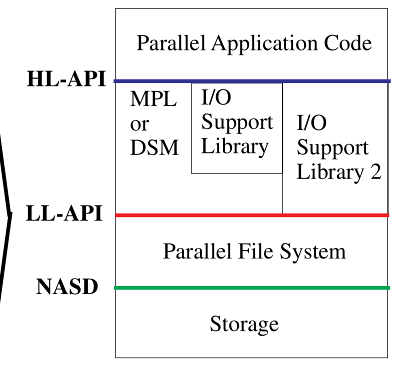
Support for I/O-intensive Multicomputer Apps



Efficient, scalable file access in heterogenous multicomputers

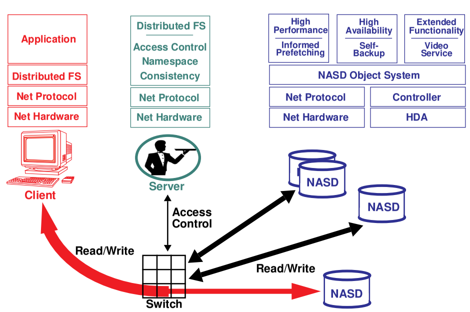
SIO Parallel File System Low Level API

- Scatter/Gather
- Asynch
- Collective Transfer
- Copy-on-write
- Client cache control
- Hints to/from storage



MPL = Message Passing Library
DSM = Distributed Shared Memory

Network Attached Secure Disks - next talk



Attach drives to network

- fewer copies, more band-width
- "huge" addressability

Port filesystems to NASD

- "traffic cop" DMA management

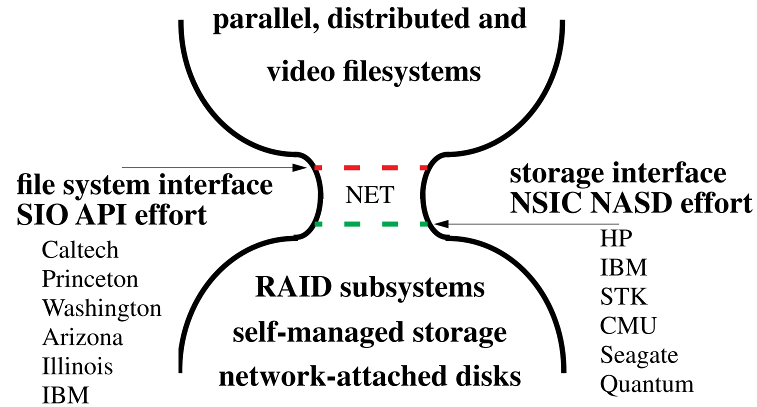
Raise drive functional interface to file system level

- < file, offset, length > for better readahead, remapping, ...

Integrate drive into LAN security protocol

- message digest or encryption for authentication check

Overall Strategy



Summary: Evolving Parallel Storage Requires ...

Rapid prototyping for RAID: RAIDframe

- flexible, architecture-rich, automated recovery

File system support for storage parallelism (TIP)

- informed prefetching and caching

Parallel file systems for parallel applications (LLAPI)

- highly available, highly scalable, global resource management

Network-Attached, Secure Disks (NASD)

- eliminate workstation as DMA device and raise interface level

Industrial interaction and support

- HP, Symbios, IBM, Seagate, Quantum, DG, Compaq, STK