Lecture 25: I/O—UNIX File System Performance and Benchmarking

Professor Randy H. Katz
Computer Science 252
Spring 1996

Review: Storage System Issues

- Historical Context of Storage I/O
- Storage I/O Performance Measures
- Secondary and Tertiary Storage Devices
- A Little Queuing Theory
- Processor Interface Issues
- I/O & Memory Buses
- Show and Tell
- ABCs of UNIX File Systems
- RAID
- I/O Benchmarks
- Comparing UNIX File System Performance
- Tertiary Storage Possibilities

Review: I/O Benchmarks

- **Scaling to track technological change**
- **TPC: price performance as normalizing configuration feature**
- **Auditing to ensure no foul play**
- **Throughput with restricted response time is normal measure**

Review—I/O Benchmarks

- **Alternative: self-scaling benchmark; automatically and dynamically increase aspects of workload to match characteristics of system measured**
  - **Measures wide range of current & future**
- **Describe 3 self-scaling benchmarks**
  - **Transaction Processing: TPC-A, TPC-B, TPC-C**
  - **NFS: SPEC SFS (LADDIS)**
  - **Unix I/O: Willy**

Review—TPC Results

TPC-A

Machine	*tpsA-local*	*K$/tps*	*OS/DB*	*Date*
HP 852S	43	24	HPUX 7/Infmx 4	12/90
VAX 4000	41	23	VMS 5.4/Dec 6	7/90
IBM RS6/550	32	20	Aix 3.1/infmx 4	1/91
Compaq SysPro	172	5	??	1/93
SPARCserve41	108	7	??	1/93
HP 9000 890/4	710	8	??	1/93

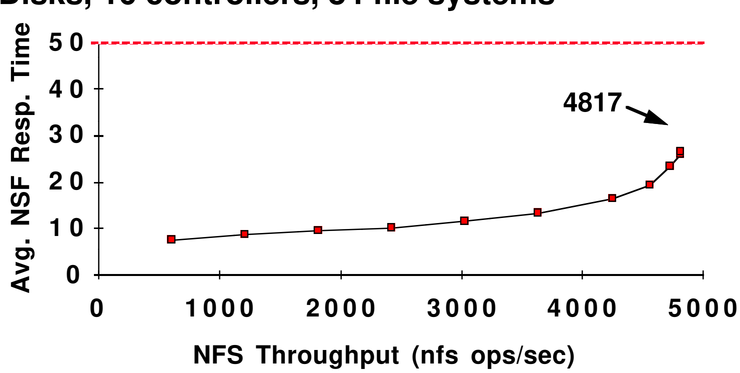
TPC-B

Machine	*tpsB*	*K$/tps*	*OS/DB*	*Date*
HP 852S	90	5	HPUX 7/Infmx 4	12/90
IBM RS6/550	58	5	Aix 3.1/infmx 4	1/91
Sun SS 490	57	8	Sun4.1/Sybase 4	10/90
Sun SS 2	52	4	Sun4.1/Sybase 4	10/90
Sun SC2000/10	1400	?	Solaris2/Sybase ?	9/94

RHK.S96 5

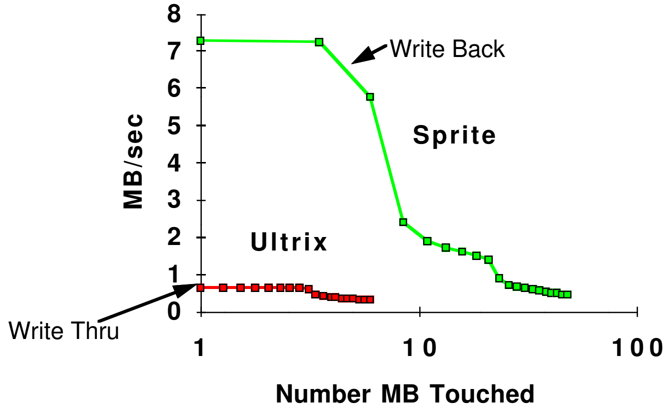
Review—Example SPEC SFS Result: DEC Alpha

- 200 MHz 21064: 8KI + 8KD + 2MB L2; 512 MB; 1 Gigaswitch
- DEC OSF/1 v2.0
- 4 FDDI networks; 32 NFS Daemons, 24 GB file size
- 88 Disks, 16 controllers, 84 file systems



RHK.S96 6

Review—Willy: DS 5000 Number Bytes Touched



- **Log Structured File System: effective write cache of LFS much smaller (5-8 MB) than read cache (20 MB) => reads cached while writes not => 3 plateaus**

UNIX File System Performance

- **9 Machines & OS**

	Machine	***OS***	***Year***	***Price***	***Memory***
Desktop	**Alpha AXP 3000/400**	**OSF/1**	**1993**	**$30,000**	**64 MB**
	DECstation 5000/200	**Sprite LFS**	**1990**	**$20,000**	**32 MB**
	DECstation 5000/200	**Ultrix 4.2**	**1990**	**$20,000**	**32 MB**
	HP 730	**HP/UX 8 & 9**	**1991**	**$35,000**	**64 MB**
	IBM RS/6000/550	**AIX 3.1.5**	**1991**	**$30,000**	**64 MB**
	SparcStation 1+	**SunOS 4.1**	**1989**	**$30,000**	**28 MB**
	SparcStation 10/30	**Solaris 2.1**	**1992**	**$20,000**	**128 MB**
Mini/Mainframe	**Convex C2/240**	**Convex OS**	**1988**	**$750,000**	**1024 MB**
	IBM 3090/600J VF	**AIX/ESA**	**1990**	**$1,000,000**	**128 MB**

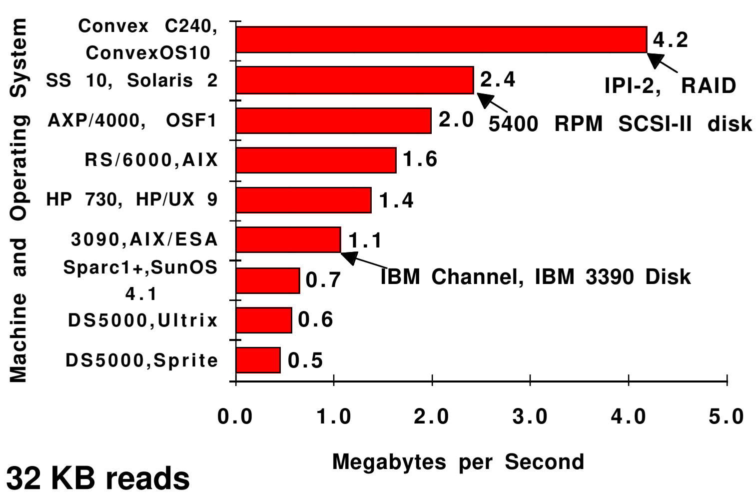
Disk Performance

- I/O limited by weakest link in chain from processor to disk
- What is a fair comparison: disks, disk controller, I/O bus, CPU/Memory bus, CPU, OS?
- Common across machines?

Machine	*OS*	*I/O bus*	*Disk*
Alpha AXP 3000/400	OSF/1	TurboChannel	SCSI RZ26
DECstation 5000/200	Sprite LFS	SCSI-I	3 CDC Wren
DECstation 5000/200	Ultrix 4.2	SCSI-I	DEC RZ56
HP 730	HP/UX 8 & 9	Fast SCSI-II	HP 1350SX
IBM RS/6000/550	AIX 3.1.5	SCSI-I	IBM 2355
SparcStation 1+	SunOS 4.1	SCSI-I	CDC Wren IV
SparcStation 10/30	Solaris 2.1	SCSI-I	Seagate Elite
Convex C2/240	Convex OS	IPI-2	4 DKD-502
IBM 3090/600J VF	AIX/ESA	Channel	IBM 3390

RHK.S96 9

Disk Performance

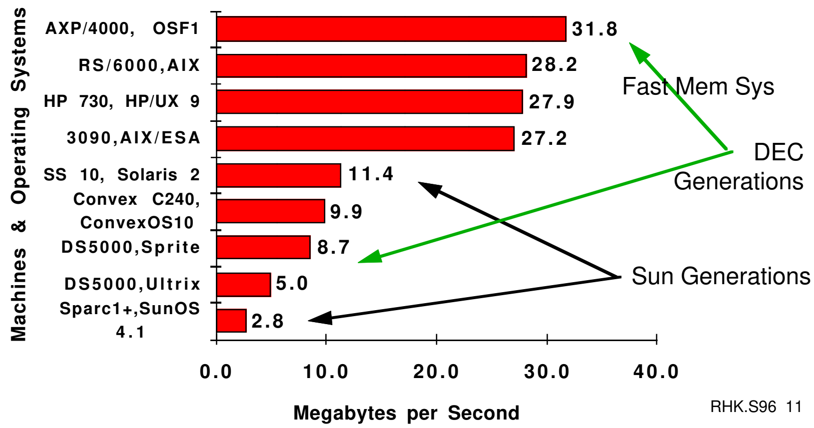


- 32 KB reads
- SS 10 disk spins 5400 RPM; 4 IPI disks on Convex

RHK.S96 10

File Cache Performance

- UNIX File System Performance: not how fast disk, but *whether* disk is used (32 KB reads; 7X speedup)
- 4X speedup between generations; DEC & Sparc



RHK.S96 11

OS Policies and Performance

- Thus far determined by HW: DIsk, bus, memory system
- Policies on machines/OS aimed at same market

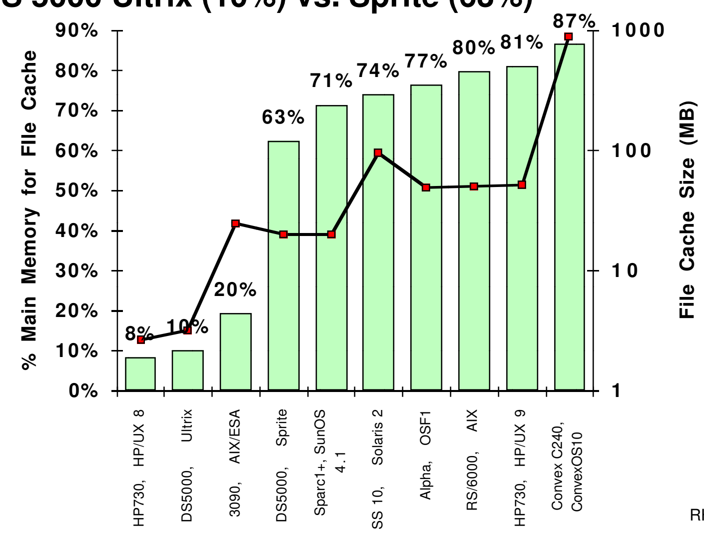
  1) How much main memory allocated for file cache?

  2) Can boundary change dynamically?

RHK.S96 12

File Cache Size

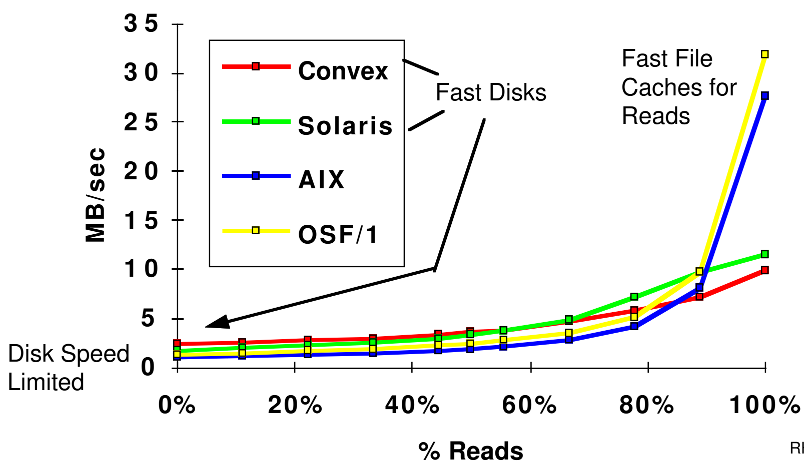
- **HP v8 (8%) vs. v9 (81%);**
  DS 5000 Ultrix (10%) vs. Sprite (63%)



RHK.S96 13

File System Write Policies

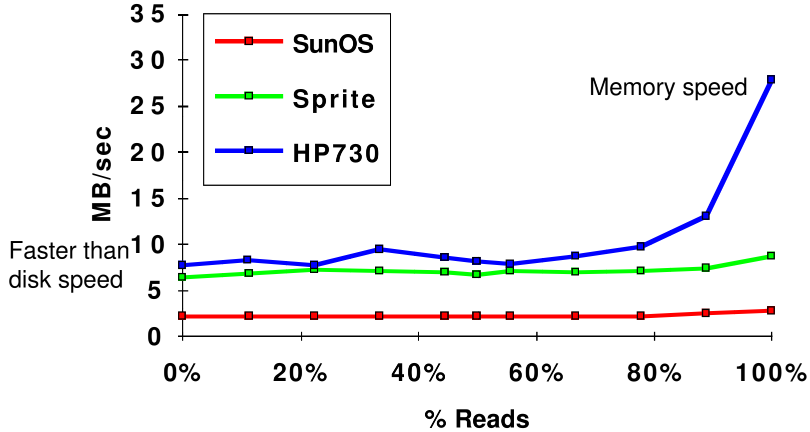
- **Write Through with Write Buffer (Asynchronous): AIX, Convex, OSF/1 w.t., Solaris, Ultrix**



RHK.S96 14

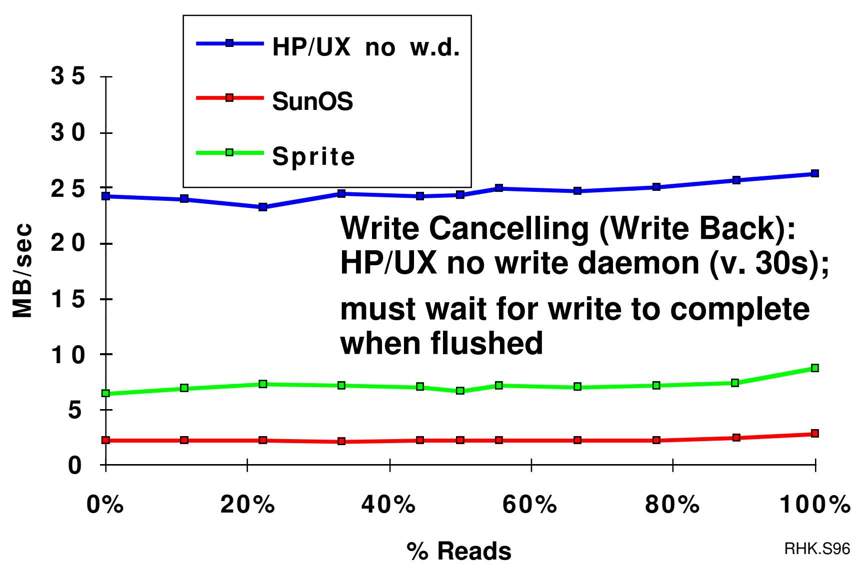
File System Write Policies

- **Write Cancelling (Write Back): HP/UX, Sprite, OSF/1 w.c., Sun OS**



- **Why HP/UX goes up with reads?**

File System Write Policies



**Write Cancelling (Write Back):
HP/UX no write daemon (v. 30s);**

**must wait for write to complete
when flushed**

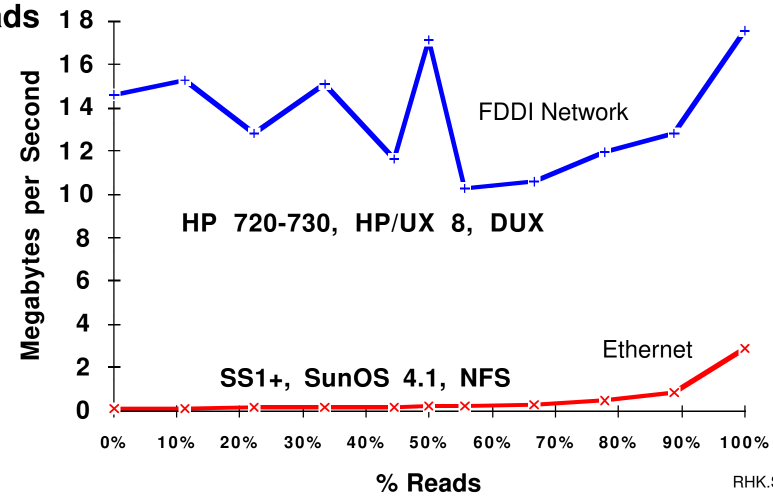
Will File Cache Blocks be Rewritten?

- 36% to 63% of all bytes do not survive 30 second window
- 60% to 95% do not survive 1000 second window
- Short file lifetimes => blocks will be rewritten

RHK.S96 17

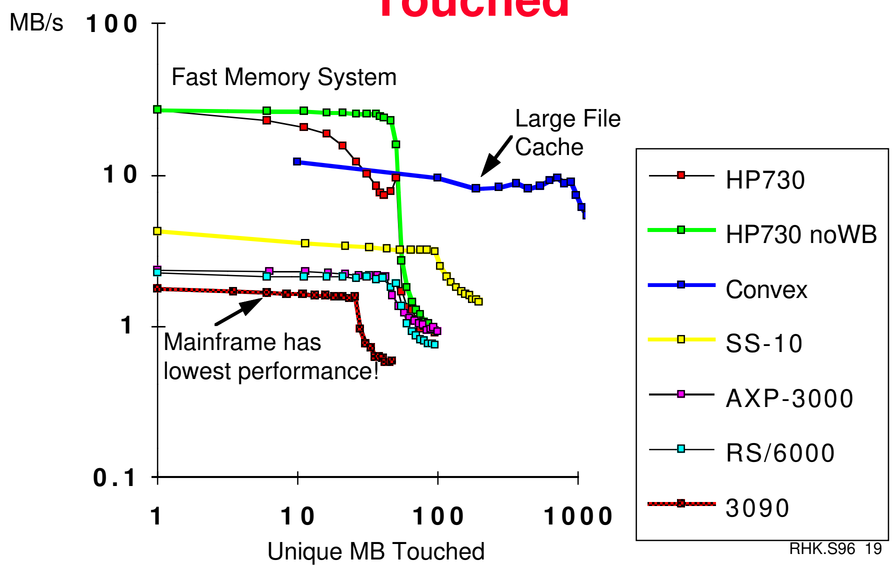
Client Server

- NFS: write through on close (no buffers)
- HPUX: client-level caching of writes; 25X faster @ 80% reads



RHK.S96 18

Overall: Bandwidth vs. Bytes Touched



RHK.S96 19

Summary: UNIX I/O

- HW determines potential performance, OS policies determine how much potential delivered
- File cache performance improving rapidly: 4X in 3 years
- File cache performance on mainframes & minisupercomputers ≤ workstations
- Write cancelation (write buffer) improves file cache performance
- File cache policy (for machines aimed at same market) determine performance; 1st place to start

RHK.S96 20