University of California, Berkeley
College of Engineering
Computer Science Division — EECS

Spring 1998 D.A. Patterson

Quiz 2
April 22, 1998
CS252 Graduate Computer Architecture

You are allowed to use a calculator and two 8.5" by 11" double-sided page of notes. Show your work on all problems. Good luck!

Your name:	
SID number:	
Email address:	

1	/ 15
2	/ 15
3	/ 30
Total	/ 60

Question 1: Bigger, Better, Faster?

A computer system has the following characteristics:

- Uses 10GB disks that rotate at 10000 RPM, have a data transfer rate of 10 MByte/s (for each disk), and have a 8 ms seek time
- Has an average I/O size of 32 KByte
- Is limited only by the disks
- Has a total of 20 disks

Each disk can handle only one request at a time, but each disk in the system can be handling a different request. The data is not striped (all I/O for each request has to go to one disk).
a) What is the average *service time* for a request?

b) Given the average I/O size from above and a random distribution of disk locations, what is the maximum number of I/Os per second (IOPS) for the system?

Question 1 (continued)

Someone suggests improving the system by using new, better disks. For the same total price as the original disks, you can get 11 disks that have 19 GB each, rotate at 12000 RPM, transfer at 12 MB/s, and have a 6 ms seek time.

c) What would be the average *service time* for a request in the new system?

d) What is the maximum number of IOPS in the new system?

Question 1 (continued)

e) Treat the entire system as a M/M/m queue (that is, a system with m servers rather than one), where each disk is a server. All requests are in a single queue. Requests may not overlap. Assume both systems receive an average of 950 I/O requests per second. Assume that any disk can service any request. What is the mean response time of the old system? The new one? You might find the following equation for an M/M/m queue useful:

$$\text{Server utilization} = \frac{\text{Arrival rate}}{\frac{1}{\text{Time}_{\text{server}}/\text{m}}} = \text{Arrival rate} \times \frac{\text{Time}_{\text{server}}}{\text{m}}$$

$$\text{Time}_{\text{system}} = \text{Time}_{\text{server}} \times \left(1 + \frac{\text{Server utilization}}{\text{m}(1 - \text{Server utilization})}\right)$$

f) Which system has a lower average response time? Why?

Question 2: A MESI Situation

Figure 1 below shows the three-phase write-back cache coherence protocol from the book.

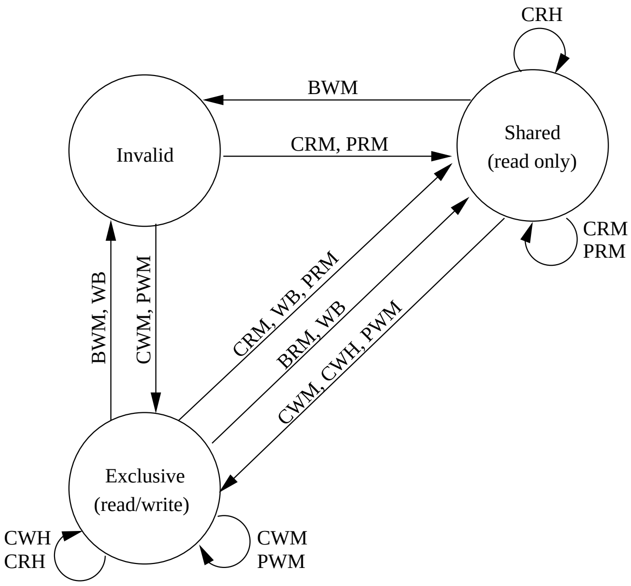

Figure 1: Three-Phase Protocol

The following terminology is used:

	Label	Stimulus or action
CPU stimulus causing transition	CRH	CPU read hit
	CRM	CPU read miss
	CWH	CPU write hit
	CWM	CPU write miss
Operation on bus causing transition	BRM	read miss for this block
	BWM	write miss for this block
CPU action on bus	PRM	place CPU read miss on bus
	PWM	place CPU write miss on bus
	WB	write back cache block

Question 2 (continued)

Figure 2 below shows a write-back MESI (Modified, Exclusive, Shared, Invalid) protocol. Assume that the processor is able to detect whether a read miss is a shared read miss or an exclusive read miss.

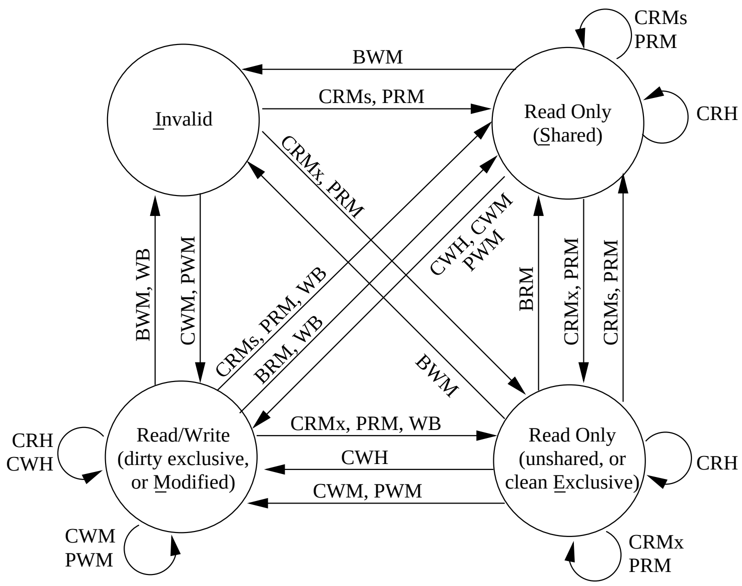

Figure 2: MESI Protocol

The following terminology is used:

	Label	Stimulus or Action
CPU stimulus causing transition	CRH	CPU read hit
	CRMs	CPU read miss (shared)
	CRMx	CPU read miss (exclusive)
	CWH	CPU write hit
	CWM	CPU write miss
Operation on bus causing transition	BRM	read miss for this block
	BWM	write miss for this block
CPU action on bus	PRM	place CPU read miss on bus
	PWM	place CPU write miss on bus
	WB	write back cache block

Question 2 (continued)

Here is a sequence of memory accesses. Assume only 2 processors, with the value 5 stored in address A1. All cache locations start out in the invalid state.

- P1 reads A1
- P1 writes 10 to A1
- P2 reads A1
- P2 writes 15 to A1

Below are the actions that occur for the above sequence on a group of machines using the three-phase protocol. Mark in the table any of the actions that change when the machines use the MESI protocol. *Only show the items that change.* Extra blank lines have been provided for you to show your changes. There may be more blank lines than you need. "Read" for bus state means that a processor is reading the value that is on the bus. In the table below, a bus action in one line affects processors and memory in the next line. For bus actions, denote a shared read miss as "RdMsS" and an exclusive read miss as "RdMsX". For states, use "Mod", "Excl", "Shar", or "Inv" to represent the read/write, read only unshared, read only shared, and invalid states. Use "(None)" to represent an item that exists in the three-phase protocol, but not in the MESI protocol.

	P1			P2			Bus				Memory	
Operation	State	Addr	Val	State	Addr	Val	Action	Proc	Addr	Val	Addr	Val
P1 Rd A1	Shar	A1					RdMs	P1	A1		A1	5
	Shar	A1	5				Read		A1	5	A1	5
P1 Wr 10 to A1	Ex	A1	10				WrMs	P1	A1		A1	5
P2 Rd A1	Ex	A1	10	Shar	A1		RdMs	P2	A1		A1	5
	Shar	A1	10	Shar	A1		WrBk	P1	A1	10	A1	5
	Shar	A1	10	Shar	A1	10					A1	10
P2 Wr 15 to A1	Shar	A1	10	Excl	A1	15	WrMs	P2	A1		A1	10
	Inv	A1		Excl	A1	15					A1	10

Question 3: Cluster vs SMP*

Evaluate the resource utilization while performing streaming I/O on the following three architectures:

- A single workstation
- A cluster of workstations
- A symmetric multiprocessor (SMP)

The basis for the first two architectures is shown in Figure 3. The cluster is built of 8 copies of the single workstation and is shown in Figure 4. The workstation contains a 167 MHz processor (with 512 KB of L2 cache) and 128 Mbyte of memory. The memory bus is 128 bits wide and operates at 83.3 MHz. The workstation contains one 32-bit, 25 MHz I/O bus (called the S-Bus). Attached to this I/O bus are two fast-wide (16-bit, 10 MHz) SCSI controllers. In the cluster, a Myrinet network interface, which is a switch based network that can support 1280 Mbit/s in each direction, is also installed in each machine; the machines are all connected to a single eight-port switch.

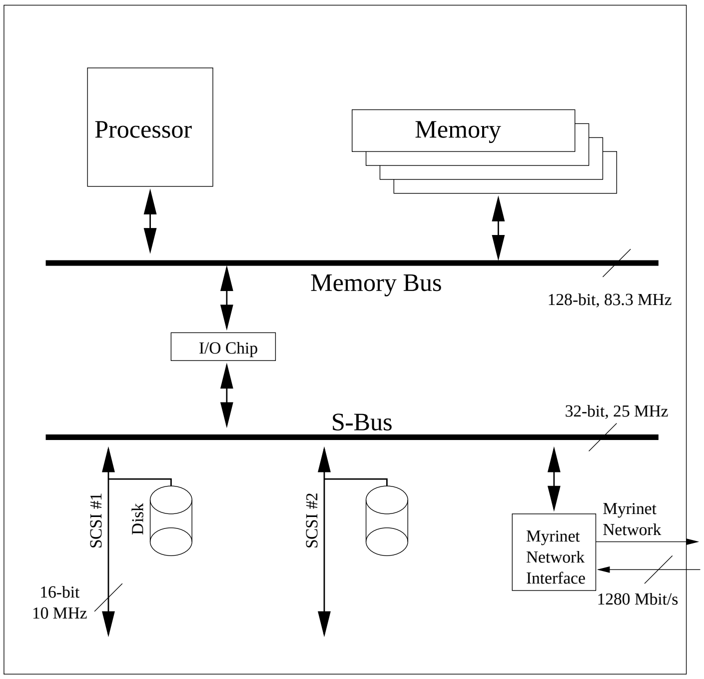

Figure 3: The Workstation

*This problem is based on a simplified version of the study "The Architectural Costs of Streaming I/O: A Comparison of Workstations, Clusters, and SMPs" by Remzi H. Arpaci-Dusseau, Andrea C. Arpaci-Dusseau, David E. Culler, Joseph M. Hellerstein, and David A. Patterson from the Fourth International Symposium on High-Performance Computer Architecture. The paper is available at `http://www.cs.berkeley.edu/~remzi/Postscript/hpca98.ps.gz`

Question 3 (continued)

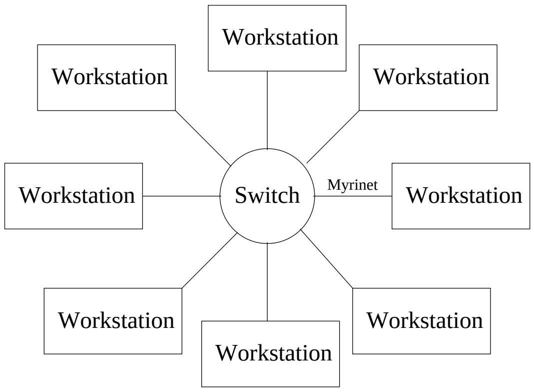

Figure 4: The Cluster

The SMP is shown in Figure 5. The system consists of four CPU/Memory and four S-Bus I/O boards connected via the GigaPlane memory bus. The GigaPlane is a 256-bit wide 83.3 MHz bus. Each CPU/Memory board contains two 167 MHz processors (each with 512 KB of L2 cache) and 256 Mbyte of memory. Each I/O board contains two S-Busses. Each S-Bus has one fast-wide (16-bit, 10 MHz) SCSI controller. All communication is performed via loads and stores to shared memory. All memory accesses are uniform access time.

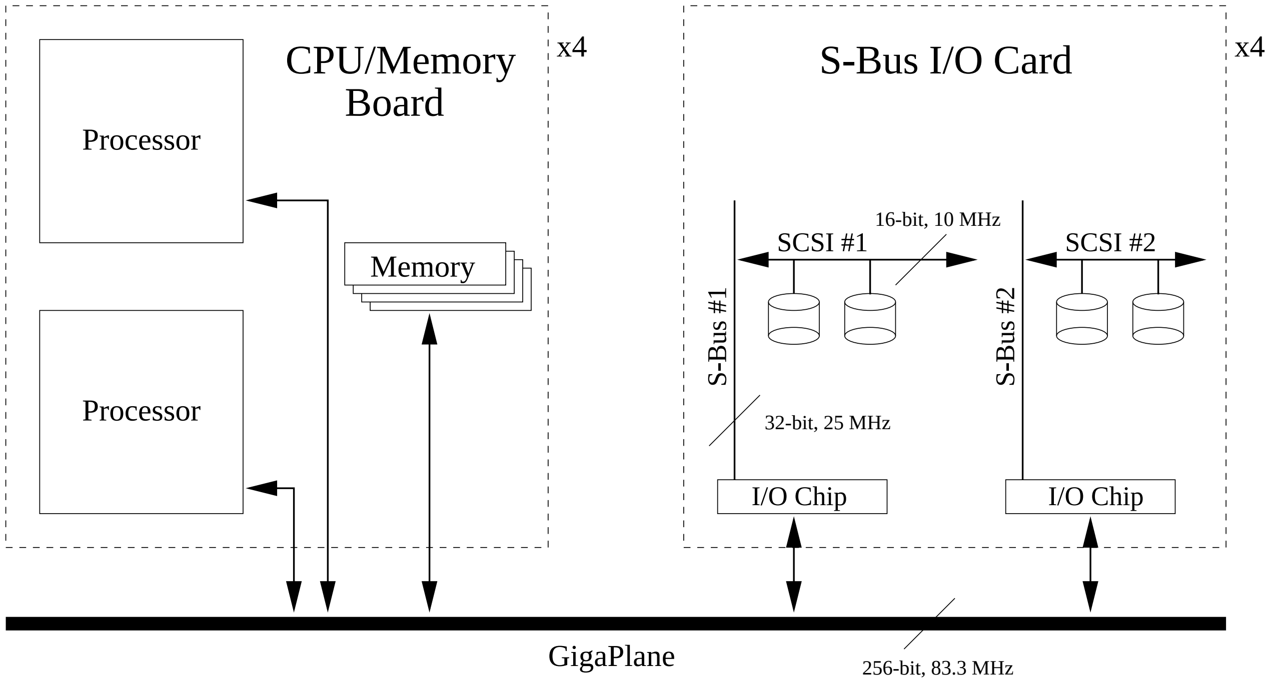

Figure 5: The SMP

Question 3 (continued)

The streaming I/O benchmark we will use is a sorting benchmark. The benchmark processes 100-byte records that include 10-byte keys. The basic algorithm is the same on all three platforms. In the first step, the records must be converted from the layout on disk to a format more suitable for efficient sorting. As records are read from disk, the key (which is part of the record) and a pointer to the full record are placed into buckets based on the top few bits of the key; this improves the cache behavior of the sort in two ways. First, the sort operates on only <partial key, pointer> pairs, thus copying only 8-bytes rather than 100-byte records as keys are compared and swapped. Second, the number of keys in each bucket matches the size of the second-level cache. The next step sorts the keys in each bucket.

Assume that the data is initially randomly placed over all disks. The basic algorithm has been slightly tailored for best performance on each platform.

Figures 6, 7, and 8 show a graphical representation of the read phase for each platform. The arrows show the order and direction of data that moves across busses, but does not show the relative sizes of each transfer. The following paragraphs refer to the numbers in those figures.

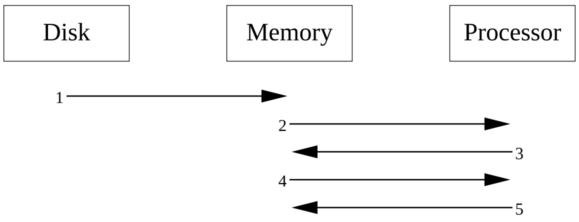

Figure 6: Workstation Sort Read Phase

In the workstation read phase, the input file is read into the user's address space (1). These records are then copied to an input buffer (2, 3). Each key is examined (4), and a <partial key, pointer> pair is written into the correct bucket (5).

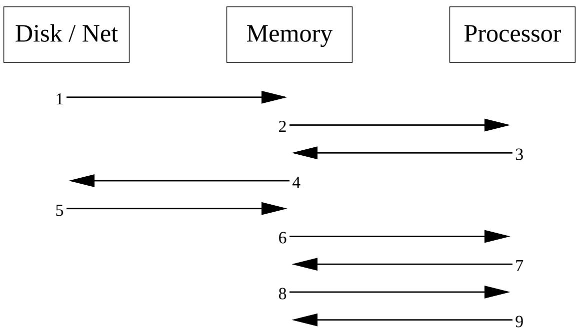

Figure 7: Cluster Sort Read Phase

In the cluster read phase, the input file is read into the user's address space (1). Records are then copied into one of 8 send buffers (2, 3); as each buffer fills, it is sent to the appropriate destination processor (4). Upon receipt of records from other processors (5), records are copied into a record buffer (6, 7). Then, each key is examined, and a <partial key, pointer> pair is written into the bucket array (8, 9), as in the single workstation sort.

Question 3 (continued)

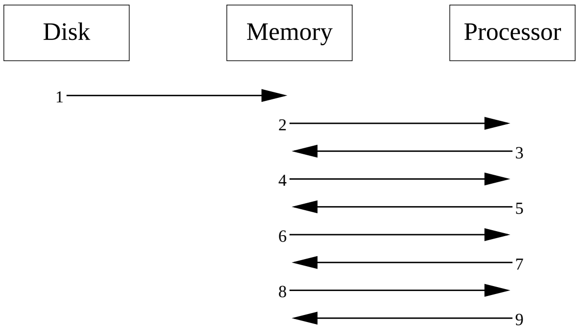

Figure 8: SMP Sort Read Phase

In the SMP read phase, the input file is read into the user's address space (1). Records are then copied into an input buffer (2, 3). Each key is examined (4), and a <partial key, pointer> pair is written into the correct bucket buffer (5). When a bucket buffer filles, the processor copies the <partial key, pointer> pair (6, 7) and records (8, 9) into a global array.

The GigaPlane bus can sustain 94% of its theoretical maximum transfer rate. The SCSI bus can sustain 80% of its theoretical maximum transfer rate. The workstation and cluster memory bus can sustain 75% of its theoretical maximum transfer rate. The S-Bus can sustain 55% of its theoretical maximum transfer rate.

The table below shows the number of millions of instructions required to processes each megabyte of data on the disk for the different platforms. The differences are mainly from overhead of sending and receiving network messages, and from slightly different ways of zeroing pages on the different platforms.

	Workstation	**Cluster**	**SMP**
Read Phase	4.1	5.5	4.6

The table below shows the measured CPI for each platform while running the benchmark.

	Workstation	**Cluster**	**SMP**
Read Phase	2.0	2.2	2.2

Question 3 (continued)

a) Determine how much of each resource (I/O bus and memory bus) is used *during the read phase* of the sort for each platform. First, write a general equation for how much of each resource is used in terms of the rate data is read from disk (D_r), the number of processors in the cluster or SMP (P), and the sizes of the records (rec), keys (key), and <partial key, pointer> pairs ($bucket$). D_r is the total rate that data is read from disks (the sum of all the individual disk rates). Give the combined bandwidth required for *all* the busses. Then, fill in the table on the next page with the summary. Provide a *short* justification for these equations. The resource usage for the workstation sort has been completed as an example.

Workstation

Memory Bus: During the read phase, data is read from disk (D_r), then copied into memory ($2D_r$). The keys are read ($\frac{key}{rec} \times D_r$), and <partial key, pointer> pairs written to the right bucket ($\frac{bucket}{rec} \times D_r$).

I/O Bus: Data is read from the disk (D_r).

Cluster

SMP

Question 3 (continued)

	Resource Usage
Workstation Memory Bus	$D_r + (2D_r) + (\frac{key}{rec})D_r + (\frac{bucket}{rec})D_r$
Workstation I/O Bus	D_r
Cluster Memory Bus	
Cluster I/O Bus	
SMP Memory Bus	
SMP I/O Bus	

b) Fill in the values of the general equations from part (a), using the following values: 8 processors in the cluster and SMP (P), 100-byte records (rec), 10-byte keys (key), and 8-byte <partial key, pointer> pairs ($bucket$). Leave the term D_r in your equations. The read phase of the workstation sort has been completed as an example.

	Memory Bus Usage	**I/O Bus Usage**
Workstation	$3.18D_r$	D_r
Cluster		
SMP		

Question 3 (continued)

c) Each disk can read data at 5.5 Mbyte/s. Assume disks are organized the most efficient way possible (the disks are equally spread over all the busses available). If we use 2 disks per processor, what is the utilization of each resource (SCSI bus, I/O bus, memory bus, processor) during the read phase of the sort for *only* the cluster and SMP platforms? (Determine utilization as a percent of maximum sustainable transfer rate for each bus.)

SCSI Bus

Cluster:

SMP:

I/O Bus

Cluster:

SMP:

Question 3 (continued)

Memory Bus

Cluster:

SMP:

CPU

Cluster:

SMP:

e) Explain briefly which system scales the best (in terms of adding more disks) for this benchmark.