Computer Organization FINAL EXAM

2012.12.08.

student number : **name :** **(signature)**

Please return this paper after exam !

Total point : 221

1) Match the memory hierarchy element on the left with the closest phrase on the right:[3]

1. L1 cache	a. A cache for a cache
2. L2 cache	b. A cache for disks
3. Main memory	c. A cache for a main memory
4. TLB	d. A cache for page table entries

2) Which of the following statements (if any) are generally true? [2] why?[2]
1. There is no way to reduce compulsory misses.
2. Fully associate caches have no conflict misses.
3. In reducing misses, associativity is more important than capacity.

3) Which of the following are true about dependability? [2] why?[2]
1. If a system is up, then all its components are accomplishing their expected service.
2. Availability is a quantitative measure of the percentage of time a system is accomplishing its expected service.
3. Reliability is a quantitative measure of continuous service accomplishment by a system.
4. The major source of outages today is software.

4) Which of the following are true about disk drives? [2] why?[2]
1. 3.5-inch disks perform more IOs per second than 2.5-inch disks.
2. 2.5-inch disks offer the highest gigabytes per watt.
3. It takes hours to read the contents of a high capacity disk sequentially.
4. It takes months to read the contents of a high capacity disk using random 512-byte sectors.

5) Which of the following are true about flash memory? [2] why?[2]
1. Like DRAM, flash is a semiconductor memory.
2. Like disks, flash does not lose information if it loses power.
3. The read access time of NOR flash is similar to DRAM.
4. The read bandwidth of NAND flash is similar to disk.

6) Both networks and buses connect components together. Which of the following are true about them? [2] why?[2]
1. I/O networks and I/O buses are almost always standardized.
2. I/O networks and I/O buses are almost always synchronous.

7) In ranking the three ways of doing I/O, which statements are true? [2] why?[2]
1. If we want the lowest latency for an I/O operation to a single I/O device, the order is polling, DMA, and interrupt driven.
2. In terms of lowest impact on processor utilization from a single I/O device, the order is DMA, interrupt driven, and polling.

8) Are the following true or false? Unlike processor benchmarks, I/O benchmarks. [2] why?[2]
1. concentrate on throughput rather than latency
2. can require that the data set scale in size or number of users to achieve performance milestones.
3. often report cost performance

9) Which of the following are true about RAID levels 1, 3, 4, 5, and 6? [2] why?[2]
   1. RAID systems rely on redundancy to achieve high availability.
   2. RAID 1 (mirroring) has the highest check disk overhead.
   3. For small writes, RAID 3 (bit-interleaved parity) has the worst throughput.
   4. For large writes, RAID 3, 4, and 5 have the same throughput.

10) True or false : To benefit from a multiprocessor, an application must be concurrent. [2]

11) True or false : Strong scaling is not bound by Amdahl's law.[2]

12) True or false : Shared memory multiprocessors cannot take advantage of job level parallelism. [2]

13) True or false : Like SMPs, message-passing computers rely on locks for synchronization. [2]

14) True or false : Unlike SMPs, message-passing computers need multiple copies of the parallel processing program and the operating system. [2]

15) True or false : Both multithreading and multicore rely on parallelism to get more efficiency from a chip. [2]

16) True or false : Simultaneous multithreading uses threads to improve resource utilization of a dynamically scheduled, out-of-order processor. [2]

17) True or false : As exemplified in the x86, multimedia extensions can be thought of as a vector architecture with short vectors that supports only sequential vector data transfers. [2]

18) True or false : GPUs rely on graphics DRAM chips to reduce memory latency and thereby increase performance on graphics applications. [2]

19) True or false : The main drawback with conventional approaches to bench-marks for parallel computers is that the rules that the rules that ensure fairness also suppress innovation. [2]

20) Unlike CPUs, there has been no support for double precision floating-point arithmetic.[3]

21) CUDA is an environment for writing parallel programs for only GPU.[3]

22) To hide memory latency, each streaming processor has Hardware-supported threads. Each group of 32 threads is called warp.[3]

23) A Streaming multiprocessor(SM) contains only one thread blocks.[3]

24) The Y-axis of the roofline model in our textbook means availability of floating-point performance[GFLOPS/second] and The X-axis means arithmetic intensity[FLOPS/DRAM bytes accessed].[3]

25) If the ridge point is far to the right, then only kernels with very low arithmetic intensity can achieve the max performance of the computer.[3]

26) To reduce computational bottleneck, a method which improve ILP and apply SIMD can help almost any kernel in a roofline model.[3]

27) To reduce memory bottleneck, software prefetch can help almost any kernel in a roofline model.[3]

28) Media applications that play audio or video files are part of a class of workloads called “streaming” workloads; i.e., they bring in large amounts of data but do not reuse much of it. Consider a video streaming workload that accesses a 512KB working set sequentially with the following address stream: [20]

0, 4, 8, 12, 16, 20, 24, 28, 32, …

① Assume a 64KB direct-mapped cache with a 32-byte line. What is the miss rate for the address stream above. How is this miss rate sensitive to the size of the cache or the working set? How would you categorize the misses this workload is experiencing, based on the 3C model.

② Recompute the miss rate when the cache line size is 16bytes, 64bytes, and 128bytes? What kind of locality is this workload exploiting?

③ “Prefetching” is a technique that leverages predictable address patterns to speculatively bring in additional cache lines when a particular cache line is accessed. One example of prefetching is a stream buffer that prefetches sequentially adjacent cache lines into a separate buffer when a particular cache line is brought in. If the data is found in the prefetch buffer, it is considered as a hit and moved into the cache and the next cache line is prefetched. Assume a two-entry stream buffer and assume that the cache latency is such that a cache line can be loaded before the computation on the previous cache line is completed. What is the miss rate for the address stream above?

Cache block size (B) can affect both miss rate and miss latency. Assuming the following miss rate table, assuming a 1-CPI machine with an average of 1.35 references(both instruction and data) per instruction, help find the optimal block size given the following miss rates for various block sizes.

	8	16	32	64	128
a.	8%	3%	1.8%	1.5%	2%
b.	4%	4%	3%	1.5%	2%

④ What’s the optimal block size for a miss latency of 20xBcycles?

⑤ What’s the optimal block size for a miss latency of 24+B cycles?

⑥ For constant miss latency, what’s the optimal block size?

29) Virtual memory uses a page table to track the mapping of virtual addresses to physical addresses. This exercise shows how this table must be updated as addresses are accessed. The following table is a stream of virtual addresses as seen on a system. Assume 4KB pages, a four-entry fully associative TLB, and true LRU replacement. If pages must be brought in from disk, increment the next largest page number. [15]

a.	4095, 31272, 15789, 15000, 7193, 4096, 8912
b.	9452, 30964, 19136, 46502, 38110, 16653, 48480

4095	0000 1111 1111 1111
31272	0111 1010 0010 1000
15789	0011 1101 1010 1101
15000	0011 1010 1001 1000
7193	0001 1100 0001 1001
4096	0001 0000 0000 0000
8912	0010 0010 1101 0000

9452	0010 0100 1110 1100
30964	0111 1000 1111 0100
19136	0100 1010 1100 0000
46502	1011 0101 1010 0110
38110	1001 0100 1101 1110
16653	0100 0001 0000 1101
48480	1011 1101 0110 0000

TLB

Valid	Tag	Physical Page Number
1	11	12
1	7	4
1	3	6
0	4	9

Page table

Valid	Physical page or in disk
1	5
0	Disk

0	Disk
1	6
1	9
1	11
0	Disk
1	4
0	Disk
0	Disk
1	3
1	12

① Given the address stream in the table, and the shown initial state of the TLB and page table, show the final state of the system. Also list for each reference if it is a hit in the TLB, a hit in the page table, or a page fault.

② Repeat problem ①, but this time use 16KB pages instead of 4KB pages. What would be some of the advantages of having a larger page size? What are some of the disadvantages?

③ Show the final contents of the TLB if it is two-way set-associative. Also show the contents of the TLB if it is direct-mapped? Discuss the importance of having a TLB to high performance. How would virtual memory accesses be handled if there were no TLB?

There are several parameters that impact the overall size of the page table. Listed below are several key page table parameters.

	Virtual address size	Page size	Page table entry size
a.	32 bits	4 KB	4 bytes
b.	64 bits	16 KB	8 bytes

④ Given the parameters in the table above, calculate the total page table size for a system running five applications that utilize half of the memory available.

⑤ Given the parameters in the table above, calculate the total page table size for a system running five applications that utilize half of the memory available, given a two-level page table approach with 256 entries. Assume each entry of the main page table is 6 bytes. Calculate the minimum and maximum amount of memory required.

⑥ A cache designer wants to increase the size of a 4KB virtually indexed, physically tagged cache. Given the page size listed in the table above, is it possible to make a 16KB direct-mapped cache, assuming two words per block? How would the designer increase the data size of the cache?

30) In this problem, we will examine how replacement policies impact miss rate. Assume a two-way set-associative cache with four blocks. There is example (below table) as demonstrated below on the address sequence “0, 1, 2, 3, 4”. [20]

Address of memory block accessed	Hit or miss	Evicted block	Contents of cache blocks after			
			Set0	Set0	Set1	Set1
0	Miss		Mem[0]			
1	Miss		Mem[0]		Mem[1]	
2	Miss		Mem[0]	Mem[2]	Mem[1]	
3	Miss		Mem[0]	Mem[2]	Mem[1]	Mem[3]
4	Miss	0	Mem[4]	Mem[2]	Mem[1]	Mem[3]
…						

The following table shows address sequences.

	Address sequence
a.	0, 2, 4, 0, 2, 4, 0, 2, 4
b,	0, 2, 4, 2, 0, 2, 4, 0, 2

① Assuming an LRU replacement policy, how many hits does this address sequence exhibit?

② Assuming an MRU(most recently used) replacement policy, how many hits does this address sequence exhibit?

③ Simulate a random replacement policy by flipping a coin. For example, “heads” means to evict the first block in a set and “tails” means to evict the second block in a set. How hits does this address sequence exhibit?

④ Which address should be evicted at each replacement to maximize the number of hits? How many hits does this address sequence exhibit if you follow this “optimal” policy?

⑤ Describe why it is difficult to implement a cache replacement policy that is optimal for all address sequences.

⑥ Assume you could make a decision upon each memory reference whether or not you want the requested address to be cached. What impact could this have on miss rate?

31) Explore the nature of Flash memory by answering the questions related to performance for Flash memories with the following characteristics. [12]

	Data Transfer Rate	Controller Transfer Rate
a.	34 MB/s	480 MB/s
b.	30 MB/s	500 MB/s

① Calculate the average time to read or write a 1024-byte sector for each Flash memory listed in the table.

② Calculate the minimum time to read or write a 512-byte sector for each Flash memory listed in the table.

③ Figure 6.6 shows that Flash memory read and write access times increase as Flash memory gets larger. Is this unexpected? What factors cause this?

Characteristics	Kingston SecureDigital (SD) SD4/8 GB	Transend Type I CompactFlash TS16GCF133	RiDATA Solid State Disk 2.5 inch SATA
Formatted data capacity (GB)	8	16	32
Bytes per sector	512	512	512
Data transfer rate (read/write MB/sec)	4	20/18	68/50
Power operating/standby (W)	0.66/0.15	0.66/0.15	2.1/—
Size: height × width × depth (inches)	0.94 × 1.26 × 0.08	1.43 × 1.68 × 0.13	0.35 × 2.75 × 4.00
Weight in grams (454 grams/pound)	2.5	11.4	52
Mean time between failures (hours)	> 1,000,000	> 1,000,000	> 4,000,000
GB/cu. in., GB/watt	84 GB/cu.in., 12 GB/W	51 GB/cu.in., 24 GB/W	8 GB/cu.in., 16 GB/W
Best price (2008)	~ $30	~ $70	~ $300

FIGURE 6.6 Characteristics of three flash storage products. The CompactFlash standard package was proposed by Sandisk Corporation in 1994 for the PCMCIA-ATA cards of portable PCs. Because it follows the ATA interface, it simulates a disk interface, including seek commands, logical tracks, and so on. The RiDATA product imitates an SATA 2.5-inch disk interface.

32) Measurements and statistics provided by storage vendors must be carefully interpreted to gain meaningful predictions about their system behavior. The following table provides data for various disk drives. [20]

	# of Drives	Hours/Drive	Hours/Failure
a.	1000	8,760	1,000,000
b.	1000	10,512	1,500,000

① Calculate annual failure rate (AFR) for disks in the table.

② Assume that annual failure rate varies over the lifetime of disks in the previous table. Specifically, assume that AFR is three times as high in the first month of operation and doubles every year starting in the fifth year. How many disks would be replaced after 7 years of operation? What about 10 years?

③ Assume that disks with lower failure rates are more expensive. Specifically, disks are available at a higher cost that will start doubling their failure rate in year 8 rather than year 5. How much more would you pay for disks if your intent is to keep them for 7 years? What about 10 years?

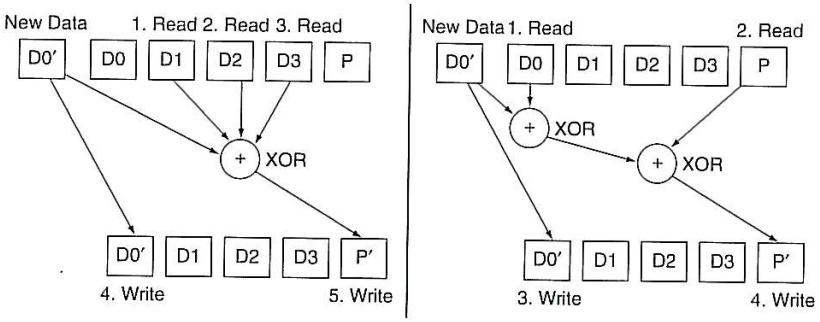
For disks in the above table, assume that your vendor offers a RAID 0 configuration that will increase storage system throughput by 70% and a RAID 1 configuration that will drop AFR of disk pairs by 2. Assume that the cost of each solution is 1.6 times the original solution cost.

④ Given only the original problem parameters, would you recommend upgrading to either RAID 0 or RAID 1 assuming individual disk parameters remain the same in the previous table?

⑤ Given that your company operates a global search engine with a large disk farm, does upgrading to either RAID 0 or RAID 1 make economic sense given that your income model is based on the number of advertisements served?

⑥ Repeat ⑤ for a large disk farm operated by an online backup company. Does upgrading to either RAID 0 or RAID1 make economic sense given that your income model is based on the availability of your server?

33) RAID 3, RAID4, RAID 5 all use parity system to protect blocks of data. Specifically, a parity block is associated with a collection of data blocks. Each row in the following table shows the values of the data and parity blocks, as described in the figure below. [15]



	New D0	D0	D1	D2	D3	P
a.	FEFE	00FF	A387	F345	FF00	4582
b.	AB9C	F457	0098	00FF	2FFF	A387

① Calculate the new parity P' for RAID 3.

② Calculate the new parity P' for RAID 4.

③ Is RAID 3 or RAID 4 more efficient? Are there reasons why RAID 3 would be preferable to RAID 4?

④ RAID 4 and RAID 5 use roughly the same mechanism to calculate and store parity for data blocks. How does RAID 5 differ from RAID 4 and for what applications would RAID 5 be more efficient?

⑤ RAID 4 and RAID 5 speed improvements grow with respect to RAID 3 as the size of the protected block grows. Why is this the case? Is there a situation where RAID 4 and RAID 5 would be no more efficient than RAID 3?

34) In future systems, we expect to see heterogeneous computing platforms constructed out of heterogeneous CPUs. We have begun to see some appear in the embedded processing market in systems that contain both floating-point DSPs and microcontroller CPUs in a multichip module package.

Assume that you have three classes of CPU:
CPU A – A moderate speed multicore CPU (with a floating-point unit) that can execute multiple instructions per cycle.
CPU B – A fast single-core integer CPU (i.e., no floating-point unit) that can execute a single instruction per cycle.
CPU C – A slow vector CPU (with floating-point capability) that can execute multiple copies of the same instruction per cycle.
Assume that our processors run at the following frequencies:

CPU A	CPU B	CPU C
1.5 GHz	3 GHz	500 MHz

CPU A can execute 2 instructions per cycle, CPU B can execute 1 instruction per cycle, and CPU C can execute 8 instructions (through the same instruction) per cycle. Assume all operations can complete execution in a single cycle of latency without any hazards.
All three CPUs have the ability to perform integer arithmetic, through CPU B cannot perform floating-point arithmetic directly. CPUs A and B have an instruction set similar to a MIPS processor. CPU C can only perform floating-point add and subtract operations, as well as memory loads and stores. Assume all CPUs have access to shared memory and that synchronization has zero cost.
The task at hand is to compare two matrices X and Y that each contain 1024x1024 floating-point elements. The output should be a count of the number indices where the value in X was larger than the value in Y. [10]

① Describe how you would partition the problem on the three different CPUs to obtain the best performance.

② What kind of instruction would you add to the vector CPU C to obtain better performance?

35) Answer following problems.[20]

① [Shared memory multiprocessors] With reference to following conditions, write comment each line about operation of sum reduction code in the box. [10]

Sum 100,000 numbers on 100 processor UMA
Each processor has ID: $0 \leq Pn \leq 99$
Partition 1000 numbers per processor
Initial summation on each processor

```
sum[Pn] = 0;
for (i = 1000*Pn; i < 1000*(Pn+1); i = i + 1)
   sum[Pn] = sum[Pn] + A[i];
```

Reduction procedures:

1. Now need to add these 100 partial sums
2. Half the processors add pairs, then quarter, …

```
/* Pn is the number identifying the processor Code for Pn; i and half are private variables */

sum[Pn] = 0;
for (i = 1000*Pn; i < 1000*(Pn+1); i = i + 1)
   sum[Pn] = sum[Pn] + A[i];
half = 100;
repeat
   synch();
   if (half%2 != 0 && Pn == 0)
      sum[0] = sum[0] + sum[half-1];
   half = half/2;
   if (Pn < half) sum[Pn] = sum[Pn] + sum[Pn+half];
until (half == 1);
```

② [A Message Passing Multiprocessor] With reference to following conditions, write comment each line about operation of sum reduction code in the box. [10]

Sum 100,000 on 100 processors
First distribute 100 numbers to each

The do partial sums
sum = 0;
for (i = 0; i<1000; i = i + 1)
  sum = sum + AN[i];

Reduction procedures:

1. Half the processors send, other half receive and add
2. The quarter send, quarter receive and add, …

```
limit = 100; half = 100;/* 100 processors */
repeat
  half = (half+1)/2;
  if (Pn >= half && Pn < limit)
    send(Pn - half, sum);
  if (Pn < (limit/2))
    sum = sum + receive();
  limit = half;
until (half == 1);
```

36) Following table is about the memory hierarchy structure of GPU which supports CUDA. Fill blank about memory name.[10]

Memory Name	Cache?	Cycle	read-only?
	L1/L2	200~400(cache miss)	R/W
	No	1~3	R/W
	Yes	1~3	READ ONLY
	Yes	~100	READ-ONLY
	L1/L2	200~400(cache miss)	R/W