Part I: Architecture

1. (5 pts) For a 5-stage MIPS processor with separate L1 data/instruction caches (32KB/256KB), the latency for each stage is listed below:

Pipeline	IF	ID	EXE	MEM	WB
Latency	150ps	100ps	50ps	200ps	100ps

The CPI with a perfect cache is 1 and both I/D caches are blocking caches. The cache miss penalty is 200 CPU cycles. For the target benchmark suite, the I-Cache miss rate is 5%, D-cache miss rate is 10% and the frequency of the load/store instructions is 35%. The architecture team is considering to increase the data cache size to 512KB so the data miss rate could be reduced to 5% for the target benchmark. But it increases the MEM pipeline stage latency to 250ps. Please justify if increasing the data size is a good design decision or not.

2. (5 pts) A stride-based prefetching scheme exploits regular streams of memory accesses to hide memory latency. Which of the following code segments can get more benefit from stride-based prefetching? Please explain why.

Code A :

```
int Y[100];
int i,x;
for (i= 0; i++;i:<100)
{
    Y[i]=Y[i] + x;
    x++;
}
```

Code B:

```
struct node {
   int x;
   struct node *next;
};

while (node->next !=0) {
      node-> x = 1;
      node = node->next;
}
```

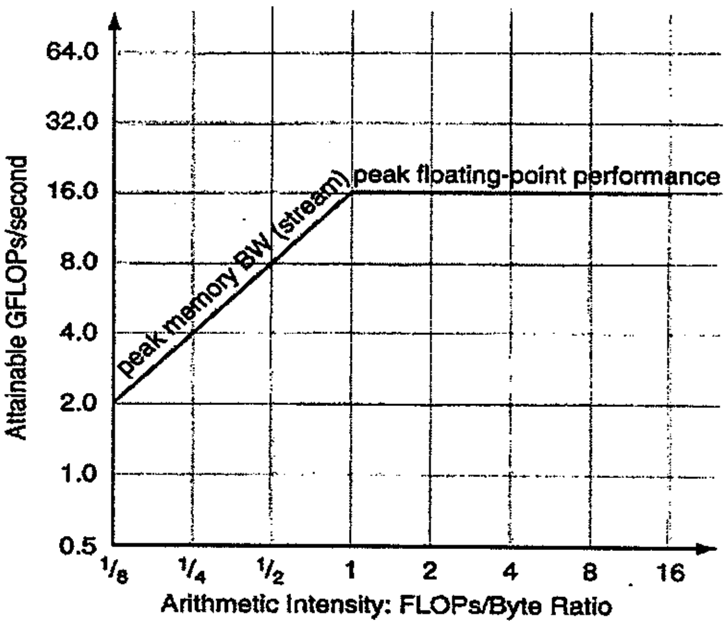
3. (5 pts) (Multiple Choices: no partial points) Figure 1 shows the roofline model of the target architecture (peak memory bandwidth:16GB/s and peak floating-point performance: 16GFLOPs/s). If the arithmetic intensity of an application falls between 1/2 FLOPs/byte and 2 FLOPs/byte in different program phases, which of the following techniques could be adopted to improve this application's performance:

(a) Software prefetching
(b) Loop unrolling
(c) Apply SIMD (Single instruction Multiple Data)

見背面

(d) Rewrite the codes for better data locality

Figure 1



4. (5 pts) (Multiple Choices: no partial points) For applications with high data-level parallelism, which of the following architectural features are effective for performance improvement:
   (a) SIMD (Single instruction Multiple Data)
   (b) SIMT (Single instruction Multiple Threads)
   (c) VLIW (Very Long Instruction Word)
   (d) SMT (Simultaneous Multi-Threading)
   (e) Vector Instruction Extension

5. (5 pts) For the MIPS instruction set (32-bit instruction/32 registers), if the register file size is increased to 128 registers, what is the total number of bits needed for a R-type format instruction (register instructions)? Could more registers decrease the size of a MIPS assembly program? Why?

Benchmarks	Machine 1		Machine 2		Observed CPI
	Seconds	Spec Ratio	Seconds	Spec Ratio	
400.perlbench	457	21.4	424	23.0	0.75
401.bzip2	539	17.9	622	15.5	0.85
403.gcc	338	23.9	293	27.5	1.72
429.mcf	280	32.6	204	44.7	10.00
445.gobmk	527	19.9	451	23.3	1.09
456.hmmer	270	34.6	440	21.2	0.80
458.sjeng	657	18.4	511	23.7	0.96
462.libquantum	96.5	215	190	109.0	1.61
464.h264ref	847	26.1	613	36.1	0.80
471.omnetpp	310	20.2	280	22.4	2.94
473.astar	380	18.5	447	15.7	1.79
483.xalancbmk	234	29.5	191	36.2	2.70

6. (8 pts) The above table gives measurements of running SPECint2006 on one Intel machine and one AMD machine. Execution time is reported in seconds. SPEC ratio is the reference time, supplied by SPEC, divided by the measured execution time.

接次頁

a) (5 pts) Which machine is faster? How much faster?
(You must show your calculation, without showing the calculation, your answer can only receive partial credits)
b) (3 pts) Which benchmarks you would target for some hardware/software optimizations? Suggest a few architecture and/or compiler optimizations that might speed up your selected programs.

7. (9 pts) New mobile phones are required to handle larger data sets, for example, recording 4K video and taking high-resolution pictures often drive the needs for 256GB of on-board memory. Buying larger memory configurations or extended storage with lightning flash drives are costly options. Recently, some companies offer CME (Cloud Memory Extension) solutions to make use of WiFi and Cloud storage to extend the memory of your mobile phones. The idea is to use the cloud storage as the secondary storage, and uses part of the on-board memory as a cache for the storage. Frequently used videos, photos, music can be cached and less frequently used files can be transferred to the cloud storage using WiFi. This approach is similar to the idea of disk caches which are often used to speed up disk I/O. Please answer the following questions of this CME cache design.

a) (1 pt) How is this CME cache different from on-chip caches in processor?
b) (2 pts) What should be the block size of a CME cache line? Why?
c) (2 pts) What should be the prefetch policy for such CME caches?
d) (1 pt) Should the cache be fully associative, set associative or direct-mapped? Why?
e) (2 pts) If set associative is the choice of (d), please suggest a replacement algorithm that works better than LRU. Explain why it could be better.
f) (1 pt) What should be used as the CME cache tag?

8. (8 pts) Branch instructions can be divided into conditional and unconditional branches, or direct and indirect branches. For example, bne $s1, $s2, 25 is a conditional and direct branch in the MIPS architecture. So a branch instruction can be classified as one of the following four types:
1) Conditional direct branch
2) Conditional indirect branch
3) Unconditional direct branch
4) Unconditional indirect branch
   a) (4 pts) For each branch instruction type, please gives one C construct which the compiler would generate that branch instruction
   b) (2 pts) Which one in the four types is easier to predict? Which one is most difficult to predict?
   c) (2 pts) What hardware structures are often used to predict type (2) and type (4) branches

見背面

Part II: Operating System

NOTE that in the question, it is intended to provide redundant or miss certain assumption to disguise you. Please make your own assumption if necessary to answer the questions.

Eva is designing an Internet-of-Things (IoT) service for Taipei city. The system is called Web of Things, which is designed to measure the vehicle flows on roads and pedestrian flows on sidewalk so as to minimize the waiting time for vehicle and pedestrian on the intersection subject to the safety constraint. In the meantime, Mayor Ke asked Eva to use minimum budget to design and deploy the systems to 2,500 intersections in Taipei City.

9. (15 pts) TaipeiBox is the device to be installed on the intersections. Eva has to choose the best configuration of hardware and software platforms between Arduino MEGA2560 and Intel Edison.
   (a) (5 pts) On Intel Edison, Eva can install Linux to manage physical and logical resources on the platforms, which has one dual-core Atom processor, 1GB memory, and 4GB storage. Suppose that one process requires at least 100MB memory. When there are 10 processes in the system, the system can still start another new process without showing 'out of memory'. Please illustrate the reason to allow the process to use more memory than the memory installed on the device, including name of the mechanism and how a memory address in a process is translated into a memory address on the device.

   (b) (5 pts) The memory space of a process, generated by C compiler, are divided into several segments. Please describes the segmentation mechanism, including their names and the data to be stored in each segment.

   (c) (5 pts) On Arduino MEGA2560 device, there are 256KB Flash and 8KB SRAM. Suppose that each function requires 50KB for binary code and variables. Can EVA deploy 10 functions to the device? If yes, please illustrate how this can be done. If not, please explain it.

10. (5 pts) The following code segment is the part of the program on TaipeiBox. Please answer the following questions.
    (a) What are the succeeding 3 program counters, in terms of line number in this program, after the program executes the instructions on Line 18?
    (b) When the system call is interrupted by a signal, will the system call be resumed after the signal is caught? If yes, please illustrate the procedure to resume. Otherwise, please describe how the operating system terminates the system call.

接次頁

```
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <unistd.h>

int
main(int argc, char **argv) {
 void catch(int);
 int a;

 printf("Hello Taipei!!\n");

 signal(SIGSEGV, catch);
 /*
     1st part: measure the traffic flow
  */

 a = *(int *) 0;

  /*
     2nd part: change the traffic light
  */
}

void
catch(int snum) {
 printf("Oooops, there is an illegal memory access(%d).\n", snum);
 raise(SIGSTOP);
 exit(0);
}

```

11. (5 pts) The 1st and 2nd part in the above code segment can be separated into multiple processes or multiple threads. Please describe how the measured flow in 1st part can be sent to 2nd part when multiple process and multiple thread are used.

12. (8 pts) To implement low-level locking mechanism for process synchronization utilities such as mutex, semaphore, or else, we need to consider using hardware solution or software solution, and in a single-processor environment or multiple-processor environment. Please describe the pros and cons.

13. (9 pts) When we specify or describe devices in Linux, there are several approaches such as devfs in /dev, sysfs in /sys, or using Device Tree. Please describe the differences of the three examples, especially, when to use it.

14. (8 pts) To collect information from part of or all intersections to make the best timing decision to change the traffic lights, Eva can have a centralized server solution, a decentralized distributed solution, or even a hybrid solution between them. Not in the order of priority and you might define your own, considering minimum budget, maximum safety, and minimum (total or maximum) waiting time, how you would help Eva to find a solution with as detail design features and justification as you can for the system, so that Mayor, drivers, and pedestrians are happy. Hint: scheduling, reaching consensus, memory, security, storage, communication ...

試題隨卷繳回