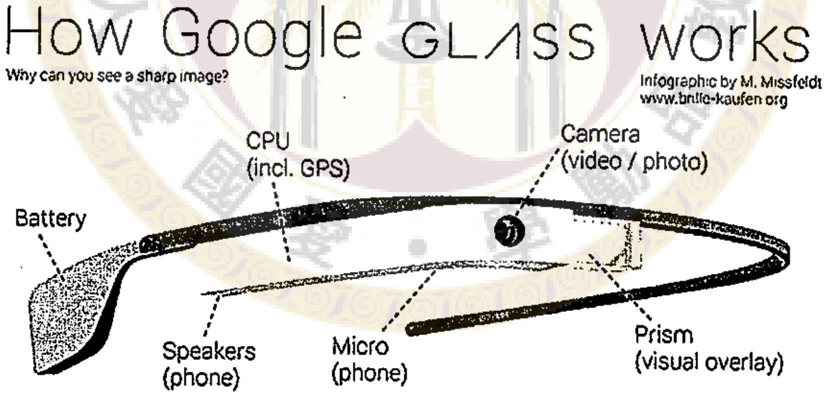
PART I: COMPUTER ARCHITECTURE

1. (10 pts) Wearable computing devices are very popular nowadays. One news articles says that an airline company is testing a new system where airline workers and gate agents use wearable devices such as Google Glass to streamline the check-in process for passengers. As illustrated in the picture below, the gate agent wears Google Glass. The Google Glass may recognize the person in front of her and tell her about this person, even without looking at his passport.



Assume that you are going to design the system. Please answer the following questions:

a. [5%] Draw the diagram for the computer architecture of Google Glass. The diagram should include the components in the picture and perhaps some other components not showing in the picture. The diagram should illustrate the connections between the components. Discuss the requirements of the Google Glass in this specific application and present your choice of the components, including CPU, memory, storage, etc.

b. [5%] Do you think that Google Glass can accomplish the job alone? Actually, Google Glass needs to work with a server to recognize the person and retrieve information.

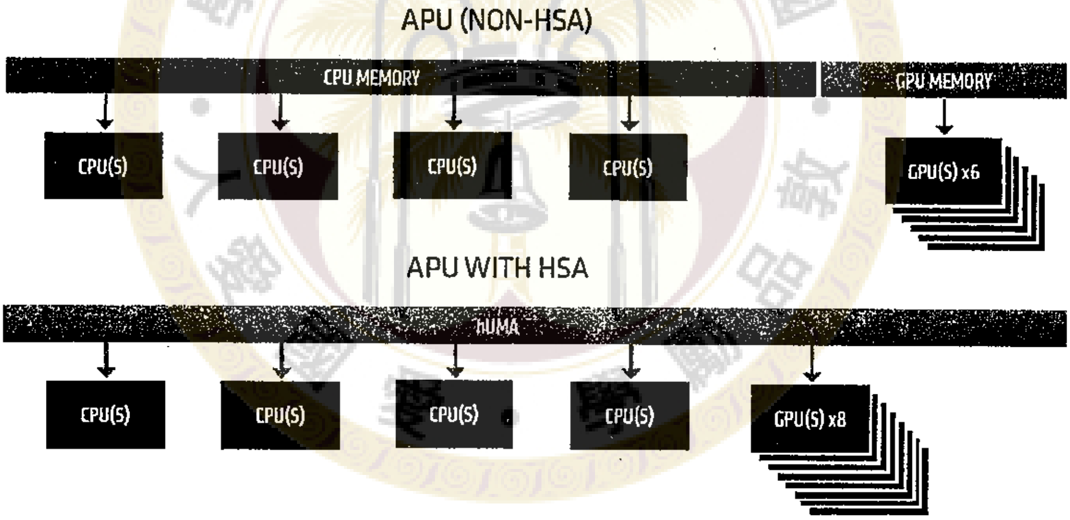
見背面

Please draw the diagram for the system infrastructure. The diagram should include the hardware and software components involved in the operation. Describe how the system works.

2. (10 pts)以下摘自 Engaget 中文版（2014-02-11）：

在這一代發表的 Kaveri APU，AMD 又再更進了一步，提出了「異質系統架構（Heterogeneous System Architecture，HSA）」，讓 CPU 和 GPU 可以共同分享處理器上的記憶體，並且模糊了 CPU 和 GPU 之間的分界。一直以來 AMD 就強調使用 GPU 來協助 CPU 做運算，而非只是單純的進行顯示工作而已，而在 Kaveri 上，GPU 的地位已經提升和 CPU 同等的地位，所以 AMD 把它也稱做處理器的「運算核心」。舉例來說，最高階的 Kaveri 處理器 A10-7850K 就有 4 個 CPU、8 個 GPU，成為成為「12 運算核心」的處理器。不過要能利用到 HSA 的最佳效能，還有待軟體方面跟上就是了。

下圖來自 AMD 的網站，點出 HSA 與傳統架構的不同處：



請回答以下問題：

a. [2%] 在傳統的 NON-HSA 系統架構中，使用者如何讓 GPU 幫忙 CPU 進行運算工作？（HINT: 處理機之間如何交換資料，管道為何?）

b. [2%] HSA 所提出的異質處理機架構設計，有什麼好處？

c. [3%] HSA 讓 CPU 和 GPU 共同分享處理器上的記憶體，在處理機架構上必須有什麼改變？改變之後的優缺點為何？請就效能、成本、技術、複雜度等面相討論之。

d. [3%]文中提到『要能利用到 HSA 的最佳效能，還有待軟體方面跟上』。其中的一項挑戰，而且讓為 HSA 編寫的程式具備高度的可攜性（portability），讓一個程式可在不同 ISA 的 GPU 上執行。因此 HSA 所提出一套稱為 HSAIL（HSA Intermediate Language）的組合語言，定義出新的 ISA，建議開發者先將程式編譯成 HSAIL 後提供給使用者。

接次頁

使用者在執行程式時，再利用 Just-in-Time (JIT) Optimization 的技術，對於可以解決此類問題。請解釋 JIT Optimization 技術，並且討論 HSA 這項技術的優缺點。

3. (5 pts) *MediaTek*, a Taiwanese company which develops SoC (System-on-Chip) solutions for smartphones, is also interested in HSA (See Question 2). Since many users play video games on their smartphones, GPU is very important to the design of smartphone chips. Assume that you are helping MediaTek in designing processor chips for the smartphones of Year 2020.
   a. [3%] You are told that the bandwidth of wide-area wireless network will be increased by 100 times, from today's 10mbps to 1Gbps. How would that affect your design? Draw the architecture for your chip design, including the specifications of major components, and explain your design.
   b. [2%] Would you adopt HSA for the design? Why?

4. (4 pts) Use an example to support your YES or NO answer. Note that you need to first write down "YES" or NO" and then give an example to support YES or NO. Both answering-YES-or-NO and an example are required. If you just give examples without writing down "YES" or "NO", you receive no point. If you just write down YES or NO, you also receive no point.
   a. [2 pts] Looping (that is, a loop code) is one reason to cause *spatial locality.* Answer **YES** or **NO** and then give an code example to illustrate your answer.
   b. [2 pts] Two instructions with data dependencies will cause data hazard in pipelining execution. Answer **YES** or **NO** and then illustrate your answer with an example.

5. (8 pts) Explain **why** true or **why** false. Note that you need to first write down "TRUE" or 'FALSE" and then and explain **why.** Both answering-true-or-false and explanation are required. If you just explain without writing down "TRUE" or "FALSE", you receive no point. If you just answer true or false, you also receive no point.
   a. [2 pts] The recent iPhone 5S uses the 64-bit Apple A7 processor which has higher benchmark numbers (in terms of GeekBench and Dhrystone) than other recent 32-bit processors such as Qualcomm's Snapdragon 400. A reason is because most programs run much faster with 64-bit instructions (vs. only 32-bit instructions). If the previous sentence is true, write down **TRUE** and then explain **why** true. Otherwise, write down **FALSE** and then explain **why** false.
   b. [2 pts] If two instructions within a basic block do not have data dependencies on each other, we want the hardware scheduler to be able to reorder the two if reordering results in higher performance. But today's x86 hardware will *not* reorder the instructions if it detects that any of the two instructions might cause an exception. If the previous sentence is true, write down **TRUE** and then explain **why** true. Otherwise, write down **FALSE** and then explain **why** false.

見背面

c. [2 pts] Multicores such as 4-core MTK6589 and 8-core MTK6592 from Mediatek have become popular in mobile hardware. It is because most existing applications on your phone will benefit from the additional processor cores. If the previous sentence is true, write down **TRUE** and then explain **why** true. Otherwise, write down **FALSE** and then explain **why** false.

d. [2 pts] Single-issue machines will *not* benefit from software-pipelining or global instruction scheduling. Answer **TRUE** or **FALSE** and explain your answer.

6. (2 pts) Assuming we are going to fetch some data in the memory hierarchy to CPU for computation, sort the following elements of the memory hierarchy in order of increasing access time (in other words, start from the fastest.) (a) Solid State Disk (b) L1 Cache (c) L2 Cache (d) Network (e) Main Memory (f) CPU Registers (g) Hard Disk

7. (4 pts) Given an application that runs for 100 seconds, and 20 out of those 100 seconds must run sequentially:

a. [1 pt.] If we run the application on the recent 4-core MTK6589, what's the maximum speedup we can expect?

b. [1 pt.] If we run the application on the recent 8-core MTK6592, what's the maximum speedup we can expect?

Next, given an application that runs for 100 seconds, and 60 out of those 100 seconds must run sequentially:

c. [1 pt.] If we run the application on the recent 4-core MTK6589, what's the maximum speedup we can expect?

d. [1 pt.] If we run the application on the recent 8-core MTK6592, what's the maximum speedup we can expect?

8. (1 pt) Other than RAW data hazards, what are other data hazard types?

9. (6 pts) Instruction scheduling refers to re-ordering of instructions to avoid stalls due to control or data hazards without changing the program semantics. The goal is to improve performance without affecting program correctness. Following is our machine model:

Each machine cycle can execute 2 operations:

A. one ALU operation or branch operation

Op dst,src1,src2　executes in 1 clock

B. one load (LD) or store (ST) operation:

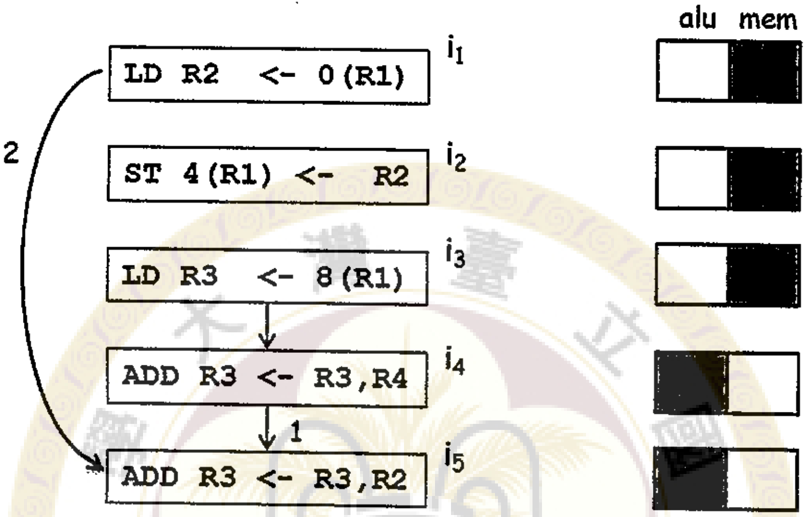
LD dst, addr　　result is available in 2 clocks

Note that load is pipelined: can issue LD next clock

ST src, addr　　executes in 1 clock cycle

接次頁

Below are 5 instructions: i_1 to i_5. The figure includes all but one important dependency edges. That is, one edge is intentionally left out. Please find the missing edge before answering the following questions.



a. [0.5 pt.] Schedule the 5 instructions above in order to finish the 5 instructions in the shortest possible time. Then, answer: What's the minimal number of cycles required to execute these 5 instructions?

b. [0.5 pt.] In your schedule above, if you answer "n" cycles to the question above, number your cycles as cycle_1, cycle_2, ..., cycle_n. In your schedule, which cycle is "i_1" being executed?

c. [0.5 pt.] In your schedule, which cycle is "i_2" being executed?

d. [0.5 pt.] In your schedule, which cycle is "i_3" being executed?

e. [0.5 pt.] In your schedule, which cycle is "i_4" being executed?

f. [0.5 pt.] In your schedule, which cycle is "i_5" being executed?

g. [1 pt] List all the data hazard types between i_1 and i_2.

h. [1 pt] List all the data hazard types between i_3 and i_4.

i. [1 pt] List all the data hazard types between i_4 and i_5.

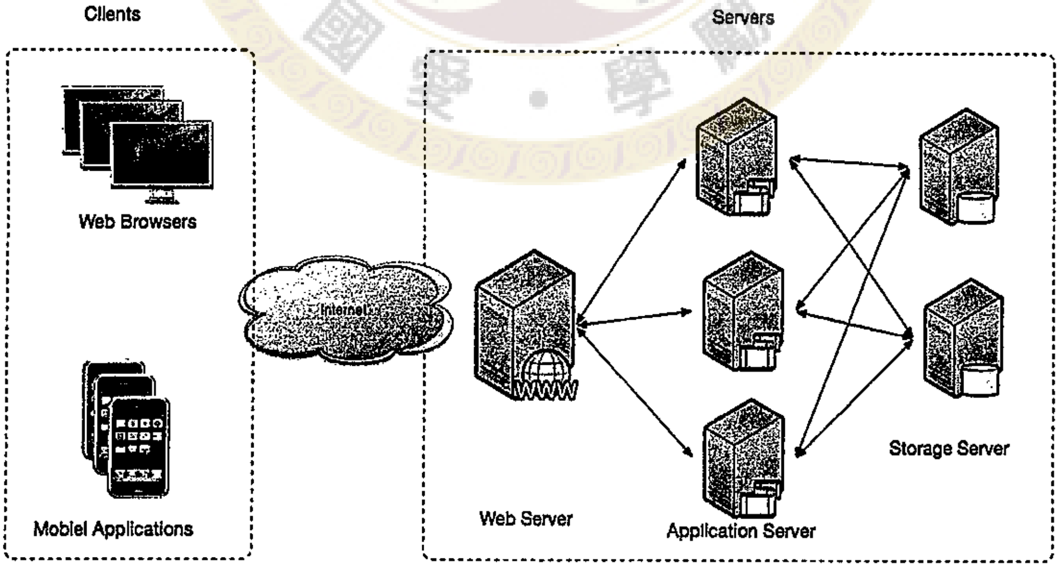
見背面

PART II – OPERATING SYSTEMS

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

In 2014, Taiwan Highway Commission officially discontinue the highway toll ticket system in favor of an electronic toll collection system, called ETC for short. The service company develops several services to allow the users to query historical toll charge, the balance of their account if there is any, and the estimated charge via mobile applications and web services. We will go through the exercise of software design to construct a highly available and responsive similar system via proper system services and operating system configuration.

The client server model shown below is the general application model to build such services. Mobile applications or web browsers are the user interfaces of the services. To serve the queries, a three-tier server model is adopted. In the model, there are a web server, application servers, and storage servers. The web server accepts the requests from mobile applications or web browsers, formats the query results in proper ways, and returns the formatted results. The applications conduct the computation logic, query the database server for required data, and compute the results. The storage servers store the user data, vehicle and travel information, and charge to the vehicle.



接次頁

10. (15 pts) On the client side, the (mobile or web) application issues the requests, receives the requests from the server and presents the results to the user. Please answer the following questions.
   a. (5 pts) The application may use either blocking or nonblocking calls to send the requests. Please describe these two types of function calls including process behavior, data transmission, and storage access of both caller and callee processes.
   b. (5 pts) Suppose blocking functions are used to issue requests. When the servers are too busy to process the requests, the application will be stalled. Signals are usually used to stop the request. Please describe how to stop the request by signals.
   c. (5 pts) Please describe the steps of processing a signal to a process. The description should include process behavior in user space and kernel space.

11. (10 pts) All these servers are supposed to accept simultaneous requests, handled by either multiple processes or multiple threads in servers. Please answer the following questions.
   a. (5 pts) When the services are implemented as multiple process applications. Please describe the difference between fork() and vfork() defined in POSIX standard (or IEEE Std 1003) to create child processes and handle incoming requests. You should discuss their memory layout, scheduling policy, and data sharing between parent and child processes.
   b. (5 pts) There are several thread models supported by modern operating systems. Please define the one-to-one and many-to-many thread model.

12. (15 pts) There are over 300 toll gates (收費感應門架) on the highways to identify passing vehicles by detect an etag on each vehicle to calculate toll charge by mileage. The eTag is a passive RFID (no power) and relies on radio sensors on the gates to collect the information on it using larger power than previously used OBU (on-board unit, powered) module. There are also cameras on the gates to record video of passing vehicles to do automatic plate number identification for vehicles without eTag or manual double-checking. Each gate might cross multiple lanes and a vehicle might cross two lanes when passing gates. Therefore, a passing vehicle might be detected by more than one sensors and a sensor might need to track more than one vehicles at the same time.
   a. (5 pts) What locking or synchronization mechanism can help to avoid double charging to a passed vehicle?
   b. (5 pts) Using eTag, how to design sensor software to save power consumption in collecting information?
   c. (5 pts) Where in the whole system (might be different) are the most reasonable places to do the automatic or manual plate number identification? Why?

見背面

13. (10 pts) The collected eTag data and recorded video will be sent to storage servers for applications to access. For security, the information needs to be kept for at least 3 months. Suppose the system allows vehicles without eTag, so that plate number identification or other identification method is necessary.

   a. (5 pts) How do we store the collected data and video information from gates to servers so that users can access the correct information as soon as possible?
   Hint: Vehicle by vehicle or periodically? Stateless or stateful? Batched or Instantly?

   b. (5 pts) Please estimate the shortest time for any user to access how much toll is charged after passing a gate. Why?