Undergraduate Computer Architecture, Fall 2022

Final Exam, 2022-12-20

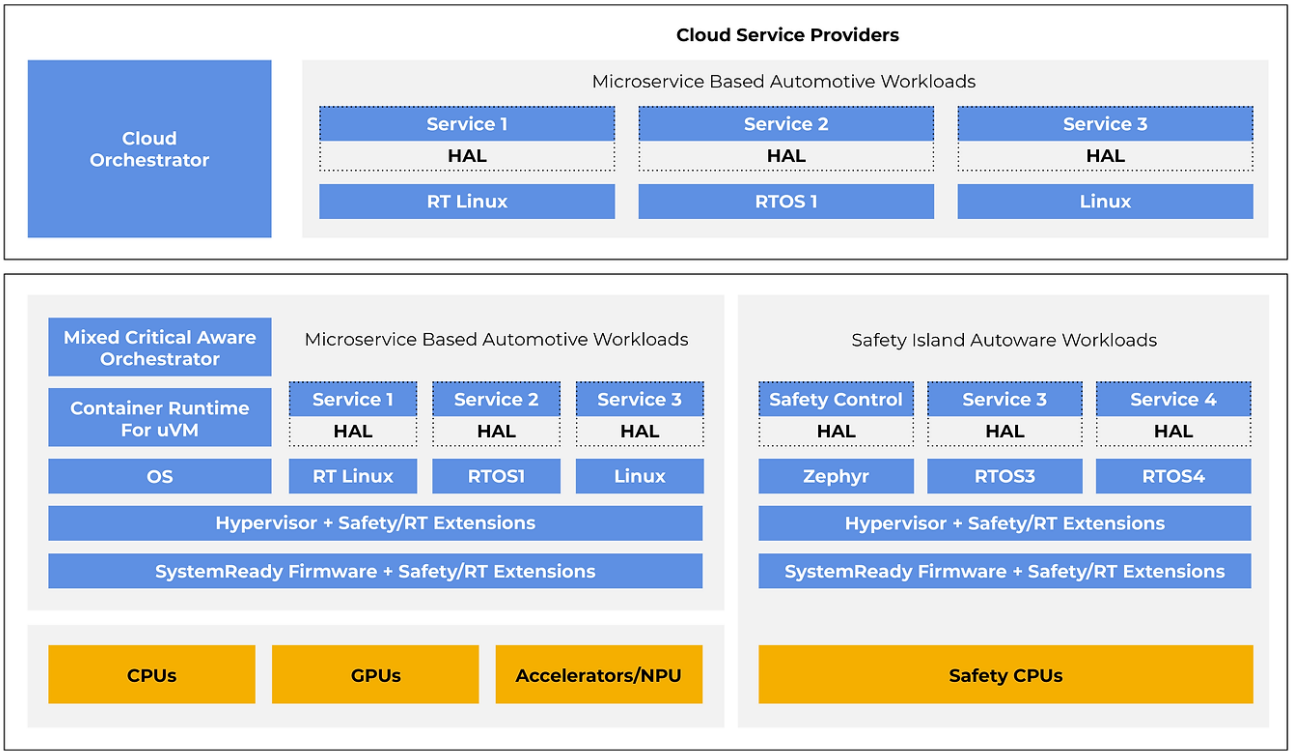
If you agree with the following sentence, please sign your name below it.

I have not cheated nor have I received any help from other students in the exam.

*Student ID & Name:*_____

1. (20 pts) Complex Hardware-Software Interactions

Autonomous Driving (AD) is an area which has been actively researched and developed recently. The following figure is from the Autoware open source project, which aims to provide a reference framework for software-defined vehicle (SDV) development of commercial AD solutions. As shown in the figure, the hardware and software architecture are somewhat different from traditional computing systems, but you should be able to comprehend the big picture of hardware-software interactions with the knowledge from this course.



Notice that the framework consists of two parts: the cloud service provider (on the top) and the vehicle (on the bottom). Various types of microservices can be downloaded from the cloud service provider to the vehicle to execute under specific environments. The vehicle executes microservices on regular CPUs, GPUs, and Accelerators/NPU, while some Safety Island workloads must be executed on safety CPUs. Functional safety is a critical element for any system deployed within a vehicle, robot, factory, and beyond. It's the system's ability to detect, diagnose, and safely mitigate the occurrence of any fault, preventing harm to people and the

environment. There are international functional safety guidelines such as ISO 26262 and IEC 61508. Also, the framework include real-time (RT) extensions since many of the tasks demand real-time processing.

Please answer the questions below.

(a) (5%) It would not be safe if the safety CPU fails. Please describe how to realize safety CPU with multiprocessors in this case.

(b) (5%) It is hard to guarantee real-time processing for an application in a regular operating system especially when there are many concurrent tasks within an OS. Can you utilize multiprocessors to improve real-time processing? How?

(c) (5%) Please explain why and how hypervisor may help improve the interactions between the vehicle and the cloud service providers.

(d) (5%) Please explain why and how hypervisor may help improve the safety and real-time processing in this case.

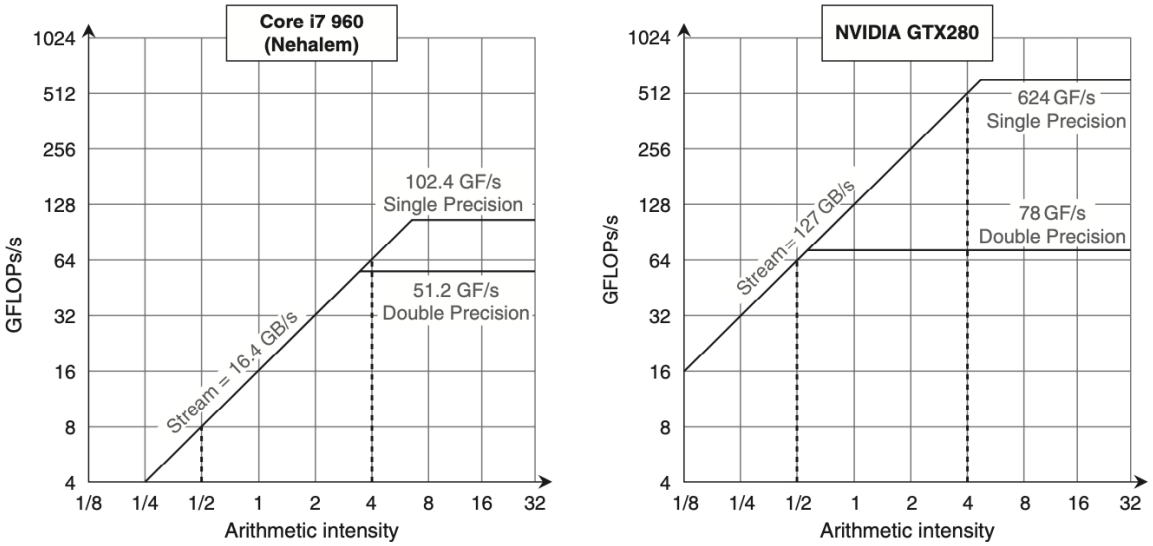
2. (50 pts) Heterogeneous Computing

In future systems, we expect to see heterogeneous computing platforms constructed out of heterogeneous CPUs. As in the autonomous driving case, we have begun to see some appear in the embedded processing market in systems that contain both CPU and DSP in a multichip module package.

Assume that you have three classes of CPU in a multichip module:
CPU A—A fast single-core CPU that can execute two single-precision floating-point (FP) instruction per cycle.
CPU B—A slow vector CPU that can execute multiple copies of the same single-precision FP instruction per cycle.

(e) (5%) Assume that our processors are implemented as the following: CPU A runs at 1 GHz and can execute two single-precision FP instructions per cycle, and CPU B run at 250MHz and can execute 16 single-precision FP instructions (through the same instruction) per cycle. Please calculate the peak performance for both CPUs, in terms of GFLOPs/s.

(f) (5%) Assume all CPUs have access to shared memory. The shared memory can provide data at 2 GB/s. Please draw the roofline models for CPU A and CPU B. For example, the figures below are the roofline models for Core i7 960 and GTX280.



(g) (5%) Suppose the task at hand is to compare two matrices X and Y that each contain 1024 × 1024 floating-point elements. The output should be a count of the number of indices where the value in X was larger or equal to the value in Y. Please write a C program to execute the task. Can you derive the arithmetic intensity for this program?

(h) (5%) Assuming X and Y are acquired by the sensor devices and stored in the main memory. Would the CPU cache be effective for the matrix comparison task? Why?

(i) (5%) Based on the roofline models, what would be the execution time on CPU A and CPU B?

(j) (5%) Suppose we are going to multiply the two matrices. Please write a C program with a naïve method (i.e. without loop blocking).

(k) (5%) Would the naïve matrix multiplication code uses the CPU cache effectively? Why?

(l) (5%) Please derive the arithmetic intensity for the naïve code and estimate the execution time for CPU A and CPU B based on the derived the arithmetic intensity.

(m) (5%) When you apply loop blocking, how would the arithmetic intensity be affected?

(n) (5%) Suppose our program does Z=Z+X*Y, where matrices X, Y, and Z each contain 1024 × 1024 floating-point elements. Can we partition the program to utilize both CPU A and B to obtain the best performance?

3. (30 pts) Stream Processing

In Question (e), comparing two matrices is an example of "streaming" workloads, which bring in large amounts of data but do not reuse much of it.

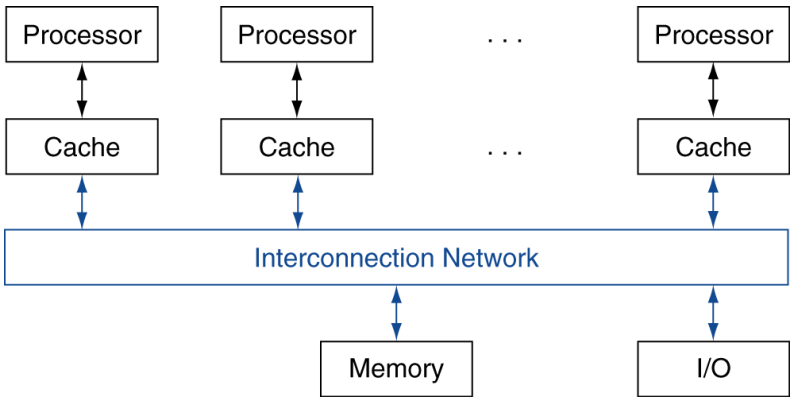
(o) (5%) Assume a 64 KiB direct-mapped cache with a 32-byte block. What is the miss rate for the address stream above?

(p) (5%) 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?

(q) (5%) Re-compute the miss rate when the cache block size is 16 bytes, 64 bytes, and 128 bytes. What kind of locality is this workload exploiting?

(r) (5%) "Prefetching" is a technique that leverages predictable address patterns to speculatively bring in additional cache blocks when a particular cache block is accessed. One example of prefetching is a stream buffer that prefetches sequentially adjacent cache blocks into a separate buffer when a particular cache block is brought in. If the data are found in the prefetch buffer, it is considered as a hit, moved into the cache, and the next cache block is prefetched. Assume a two-entry stream buffer; and, assume that the cache latency is such that a cache block can be loaded before the computation on the previous cache block is completed. What is the miss rate for the address stream above?

(s) (5%) Stream processing can often be easily parallelized to run on a symmetric multiprocessor (SMP) system such as the figure below. As opposed to heterogeneous computing, the processors in an SMP are the same (homogeneous). Can you parallelize this program (in pseudo code) to run on an SMP system?



(t) (5%) Stream processing programs are also called *data parallel* programs since the workload for each data piece can be processed by the same function kernel independently. Please explain why GPU is good at executing data parallel programs.