答題說明：本試卷包含兩個部分，共有 9 頁。 第一部分(1-8 題)必須回答在電腦閱卷答案卡；第二部分（9-12 題）必須回答在手寫答案卷。

Instructions: this exam consists of two parts and 12 questions in total. Part I (1 ~ 8) must be answered on the machine-graded answer sheet; Part II (9 ~ 12) must be answered on the hand-written answer sheet. **Only the answers following the above instructions will be graded.**

PART I: COMPUTER ARCHITECTURE (50 points)

Multiple Choices with at Least One Correct Choice: Each choice will be scored individually; only the correct choice receives the score.

G1. (15 points) Fundamental Computer Organization Knowledge

1. [5 points] Power management is important to the design of computer systems. Which of the following statements is/are true?

(A) To achieve the maximum performance, the clock frequency should be increased, but the supply voltage should be reduced at the same time, so that the processor would not be overheated.

(B) Suppose the active power consumption is 10W when the supply voltage is 1.3V. When the supply voltage is reduced to 1.1V, the active power consumption is changed to 7.2W.

(C) When the processor is executing a memory-bound program, reducing the clock frequency should not impact the performance significantly, assuming the memory bus frequency remains the same.

(D) As technology continues to shrink, leakage power will become a dominant factor. Dynamic voltage and frequency scaling (DVFS) is often used to reduce the leakage power.

(E) The processor caches can be turned off immediately at any time to save the power consumption as much as possible because the data can always be found in the memory.

2. [5 points] In this question, we examine how the pipeline affects the clock cycle time of the processor. Assume that individual stages of the datapath have the following latencies:

IF	ID	EX	MEM	WB
500 ps	700 ps	300 ps	600 ps	400 ps

Also, assume that instructions executed by the processor are broken down as the following:

ALU/Logic	40%
Jump/Branch	15%
Load	25%
re	20%

Which of the following statements is/are true?

(A) On a non-pipeline datapath, the store instruction needs 2500 ps to complete.

(B) The total latency for a load instruction is 1.4X higher on a pipelined datapath.

(C) Assuming there are no stalls or hazards, the utilization of the data memory is below 50%.

(D) Assuming there are no stalls or hazards, for the given instruction mix, it is possible for a pipelined datapath to provide 3X throughput over a non-pipeline datapath.

(E) Pipelining does not only improve throughput, but it also helps save the power consumption for embedded processors.

3. [5 points] At the age of AI and big data, memory hierarchy has become more and more important. Which of the following

見背面

statements is/are true?

(A) Programmers can ignore memory hierarchies in writing code because modern compilers can optimize the data accesses to reduce cache misses.
(B) The three Cs model is often used for understanding the behavior of memory hierarchies. In this model, all cache misses are classified into one of three categories: compulsory misses, capacity misses, and coherence misses.
(C) In a write-through cache, the modified block is written to the lower level of the hierarchy only when it is replaced.
(D) To support virtual memory, a cache can be virtually indexed and virtually tagged so that the cache can search for the data before the TLB translates the virtual address into a physical address.
(E) In a cache-coherent multiprocessor, the protocols to maintain coherence for multiple processors are called cache coherence protocols. With a write invalidate protocol, it invalidates copies in other caches on a write, so it is possible that a shared-memory multithreaded program generates more cache misses when it executes on a multiprocessor.

G2. **(15 points) Large Language Models and Computer Architecture**

A large language models (LLM) is a machine learning algorithm that can perform a variety of natural language processing tasks and is known for its ability to achieve general-purpose language understanding and generation. Because of its importance, many manufacturers are optimizing computer architectures for LLM. Thus, benchmarks are developed to evaluate the performance of computers on LLMs. For example, as the following news article published on *IEEE Spectrum* shows, a benchmark suite called *MLPerf* has added a LLM benchmark and chips from two major chip manufacturing companies, Intel and Nvidia, are compared. Please read this article and answer the following questions.

NEWS ARTIFICIAL INTELLIGENCE

Intel and Nvidia Square Off in GPT-3 Time Trials

>MLPerf provides LLM testbed for Nvidia's H100 and top Intel chipsets

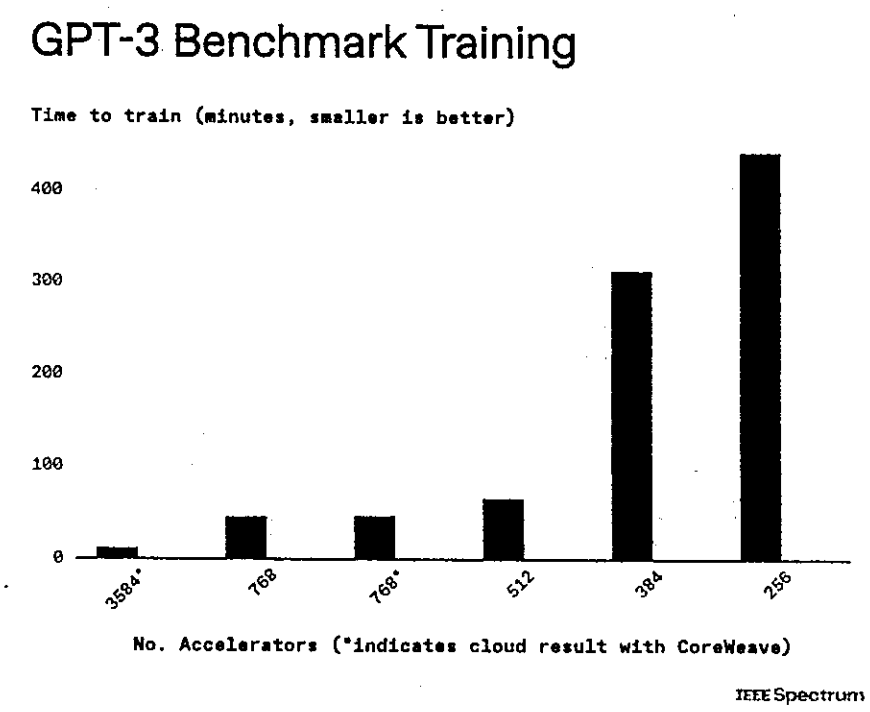
BY SAMUEL K. MOORE | 28 JUN 2023 | 4 MIN READ |

For the first time, a large language model—a key driver of recent AI hype and hope—has been added to MLPerf, a set of neural-network training benchmarks that have previously been called the Olympics of machine learning. Computers built around Nvidia's H100 GPU and Intel's Habana Gaudi2 chips were the first to be tested on how quickly they could perform a modified train of GPT-3, the large language model behind ChatGPT.

A 3,584-GPU computer run as a collaboration between Nvidia and cloud provider CoreWeave performed this task in just under 11 minutes. The smallest entrant, a 256-Gaudi2 system, did it in a little over 7 hours. On a per-chip basis, H100 systems were 3.6-times as fast at the task as Gaudi2. However, the Gaudi2 computers were operating "with one hand tied behind their back," says Jordan Plawner, senior director of AI products at Intel, because a capability called mixed precision has not yet been enabled on the chips.

Computer scientists have found that for GPT-3's type of neural network, called a transformer network, training can be greatly accelerated by doing parts of the process using less-precise arithmetic. Versions of 8-bit floating point numbers (FP8) can be used in certain layers of the network, while more precise 16-bit or 32-bit numbers are needed in others. Figuring out which layers are which is the key. Both H100 and Gaudi2 were built with mixed-precision hardware, but it's taken time for each company's engineers to discover the right layers and enable them. Nvidia's system in the H100 is called the transformer engine, and it was fully engaged for the GPT-3 results.

接次頁



Habana engineers will have Gaudi2's FP8 capabilities ready for GPT-3 training in September, says Plawner. At that point, he says, Gaudi2 will be "competitive" with H100, and he expects Gaudi2 to beat H100 on the combination of price and performance. Gaudi2, for what it's worth, is made using the same process technology—7 nanometers—as the H100's predecessor, the A100.

Making GPT-3 work

Large language models "and generative AI have fundamentally changed how AI is used in the market," says Dave Salvatore, Nvidia's director of AI benchmarking and cloud computing. So finding a way to benchmark these behemoths was important.

But turning GPT-3 into a useful industry benchmark was no easy task. A complete training of the full 175-billion parameter network with an entire training dataset could take weeks and cost millions of dollars. "We wanted to keep the runtime reasonable," says David Kanter, executive director of MLPerf's parent organization, MLCommons. "But this is still far and away the most computationally demanding of our benchmarks." Most of the benchmark networks in MLPerf can be run on a single processor, but GPT-3 takes 64 at a minimum, he says.

Instead of training on an entire dataset, participants trained on a representative portion. And they did not train to completion, or convergence, in industry parlance. Instead, the systems trained to a point that indicated further training would lead to convergence.

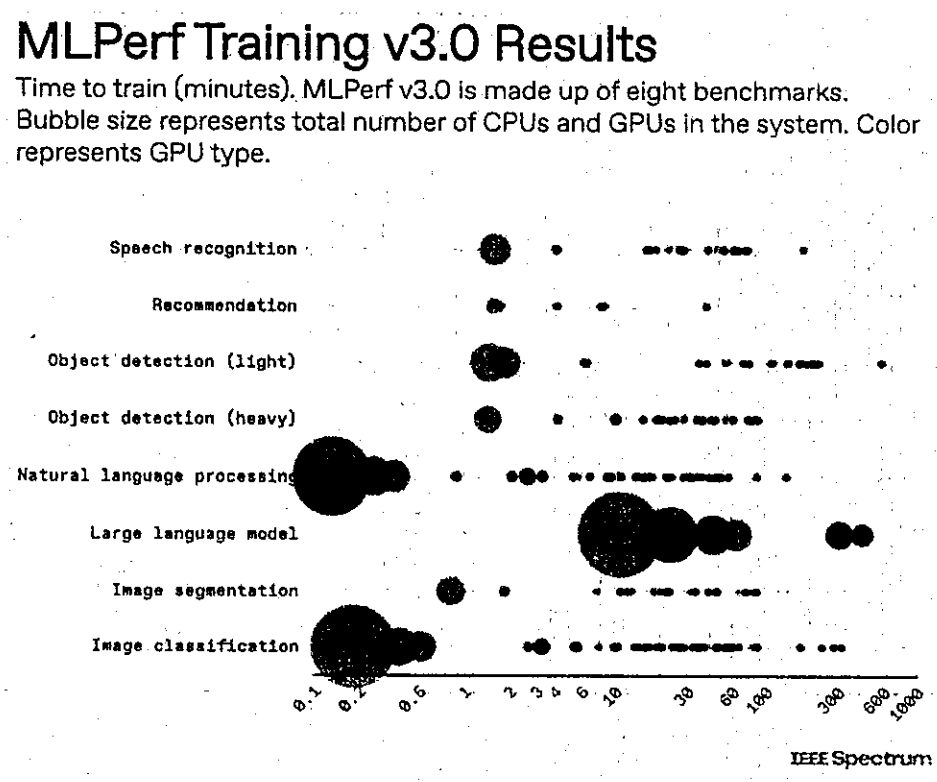
Figuring out that point, the right fraction of data, and other parameters so that the benchmark is representative of the full training task took "a lot of experiments," says Ritika Borkar, senior deep-learning architect at Nvidia and chair of the MLPerf training working group.

On Twitter, Abhi Venigalla, a research scientist at MosaicML, estimated that Nvidia and CoreWeave's 11-minute record would scale up to about two days of full-scale training.

H100 training records

見背面

This round of MLPerf wasn't just about GPT-3, of course; the contest consists of seven other benchmark tests: image recognition; medical-imaging segmentation; two versions of object detection; speech recognition; natural-language processing; and recommendation. Each computer system is evaluated on the time it takes to train the neural network on a given dataset to a particular accuracy. They are placed into three categories: cloud-computing systems, available on-premises systems, and preview systems, which are scheduled to become available within six months.



For these other benchmarks, Nvidia was largely involved in a proxy fight against itself. Most of the entrants were from system makers such as Dell, Gigabyte, and the like, but they nearly all used Nvidia GPUs. Eighty of 88 entries were powered by them, and about half of those used the H100, a chip made using Taiwan Semiconductors Manufacturing Co.'s 5-nanometer process that went to customers in the fourth quarter of 2022. Either Nvidia computers or those of CoreWeave set the records for each of the eight categories.

In addition to adding GPT-3, MLPerf significantly upgraded its recommender system test to a benchmark called DLRM DCN-V2. "Recommendation is really a critical thing for the modern era, but it's often an unsung hero," says Kanter. Because of the risk surrounding identifiable personal information in the dataset, "recommendation is in some ways the hardest thing to make a benchmark for," he says.

The new DLRM DCN-V2 is meant to better match what industry is using, he says. It requires five times the memory operations, and the network is similarly more computationally complex. The size of the dataset it's trained on is about four times as large as the 1 terabyte its predecessor used.

You can see all the results in https://mlcommons.org/en/training-normal-30/.

4. [5 points] According to the article, the use of low-precision calculation is critical to the transformer engine. The article mentions FP8, but there are two FP8 formats that are in use today. One is called *E4M3*, as it contains 4 exponent bits and 3 mantissa bits, and the other is called *E5M2*, which contains 5 exponent bits and 2 mantissa bits. Which of the following statements is/are true?

接次頁

(A) The maximum value that can be represented by E4M3 is larger than E5M2.
(B) The minimum number of E4M3 is smaller than E5M2, in terms of absolute value.
(C) The maximum value that can be represented with E4M3 is larger than an unsigned 8-bit integer.
(D) The maximum value that can be represented with E5M2 is larger than an unsigned 16-bit integer.
(E) The entire transformer network can be calculated with FP8 to save the memory space and accelerate the execution.

5. [5 points] Training a GPT-3 model was no easy task, as a complete training of the full 175-billion parameter network with an entire training dataset could take weeks and cost millions of dollars. According to the article, which of the following statements is/are true?

(A) All the parameters of GPT-3 can be placed on the memory of H100 GPU when most of the parameters are represented with FP8.
(B) One can use the LLM benchmark results to estimate the full training time and decide which solution would train GPT-3 faster.
(C) The LLM benchmark are designed to evaluate which computing platforms perform better in terms of accuracy.
(D) Based on the reported results, parallel computing is critical to speed up the training time.
(E) Based on the reported results, the efficiency of training is improved when the number of accelerators is increased from 256 to 512.

6. [5 points] The MLPerf benchmark suite contain 8 benchmarks related to machine learning, and the article presents a figure to show the benchmark results. Note that the X-axis in the figure represents the execution time. The GPT-3 LLM benchmark results show that, on a per-chip basis, H100 systems were much faster than Gaudi2, but the mixed precision capability has not yet been enabled on Gaudi2. Recently, Intel submitted new MLPerf results in November 2023 and Gaudi2 demonstrated a significant 2x performance leap, with the implementation of the FP8 data type on the GPT-3 training benchmark. According to article and the recently result, which of the following statements is/are true?

(A) GPU is always faster than CPU and the other accelerators for running the MLPerf benchmarks.
(B) Based on the recent result, with mixed precision capability, Gaudi2 is faster than H100 on per-chip basis.
(C) Most of the results are benchmarked on GPUs.
(D) Many of the results are running on chips manufactured by Taiwan Semiconductors Manufacturing Co.'s 5-nanometer process.
(E) Dataset is important to machine learning, and it can be hard to build a representative benchmark because of the risk in violating data privacy.

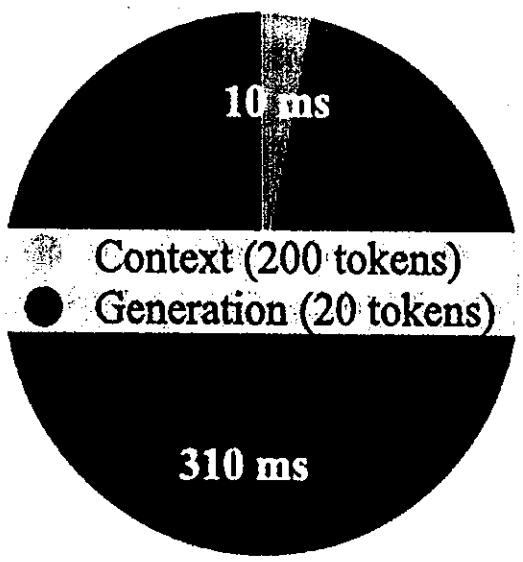
G3. (20 points) Designing Chips and Systems for Large Language Models

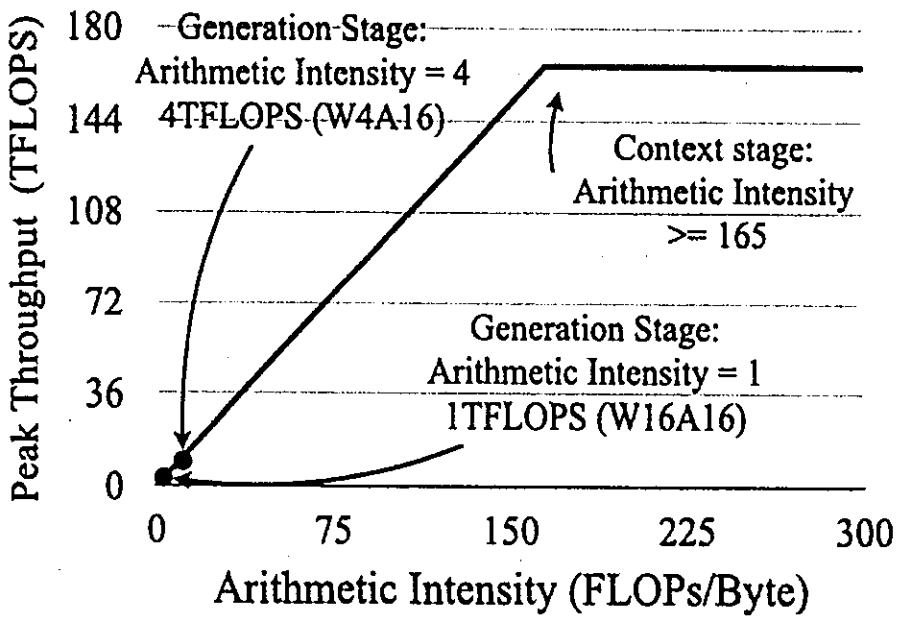
Now that we finished the previous problem set and have some basic understanding about LLM system benchmarking, assuming we are working for a chip design company which is trying to design accelerators for LLM, let us think about how to make a good design. It is important to obtain a representative workload of the application and characterize the workload in the design process and constantly evaluate the chip design with the workload to identify performance bottlenecks.

7. [10 points] A group of researchers at Massachusetts Institute of Technology developed a LLM called *TinyChat* to run on weak, power-constrained edge devices (https://hanlab.mit.edu/blog/tinychat). First, they profiled the execution time of TinyChat in the target application scenario to understand the workload. The LLM workload was broken down into two stages, Context stage and Generation Stage, and the profile revealed that the Generation Stage dominated the execution time, as Generation occupies 310ms of the total execution time (340ms). Furthermore, they used arithmetic intensity and roofline

見背面

model to characterize the workload, and the figure below reported the arithmetic intensity for the two stages.





Suppose we are designing a processor chip and fine-tune the LLM for a similar application, which of the following statements is/are true?

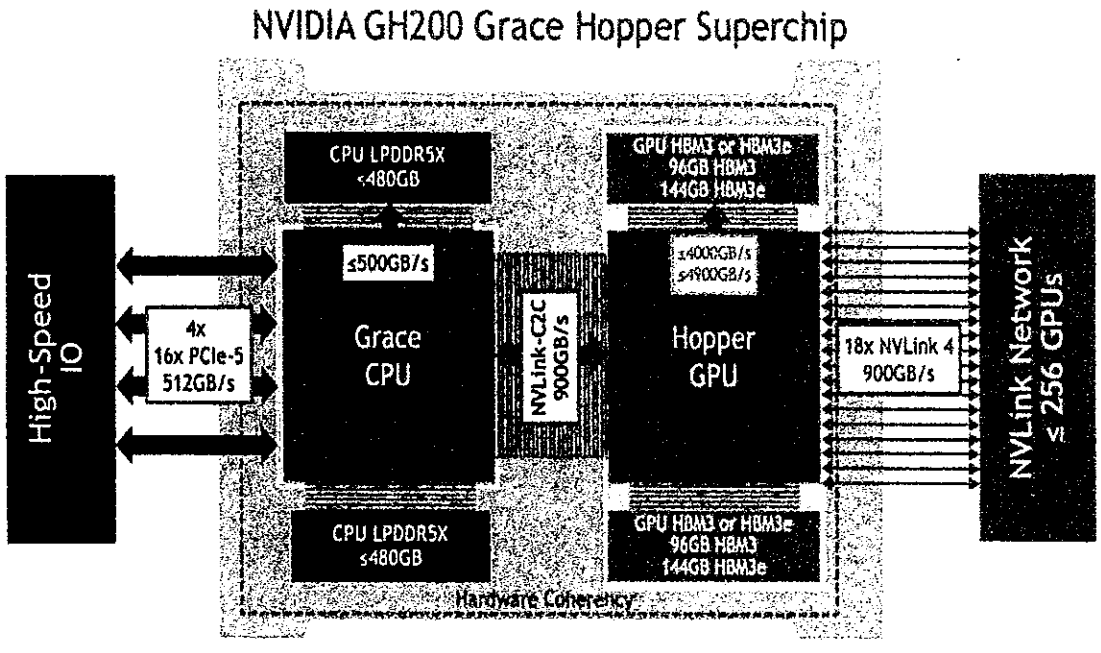
(A) We should include more arithmetic units in the chip design as it will increase the peak throughput and effectively accelerate the LLM.
(B) Because the performance of the LLM is mostly memory-bound, increasing the memory bandwidth should improve the performance.
(C) Increasing the clock rate for the processor should improve the power efficiency in terms of throughput per watt.
(D) Improving the branch predictor should greatly increase the performance because better branch prediction can reduce the control hazards.
(E) Suppose the original LLM uses FP16 to store the parameters, compressing the model to use FP8 parameters should accelerate the LLM by approximately 2 times.

8. [10 points] Suppose we are designing a computing system that aims to train LLMs with multi-trillion of parameters. For that, let us review the latest system architecture design. NVIDIA recently announced *DGX GH200 AI Supercomputer*. According to its specifications (shown below), it is capable of 128 petaFLOPS of FP8 AI performance with 32 GH200 superchips and 19.5 TB of shared memory.

接次頁

DGX GH200 Technical Specifications	
CPU and GPU	32x NVIDIA Grace Hopper Superchips
CPU Cores	2,304 Arm® Neoverse V2 Cores with SVE2 4X 128b
Shared Memory	19.5 TB
Performance	128 petaFLOPS of FP8 AI performance
Networking	32x OSFP single-port NVIDIA ConnectX-7 VPI with 400Gb/s InfiniBand 16x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA AI Enterprise (optimized AI software) NVIDIA Base Command (orchestration, scheduling, and cluster management) DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system)
Support	Three-year business-standard hardware and software support

However, to really exploit the potential of the DGX GH200, we need to understand the system architecture further by examining the *Grace Hopper Superchip*. As shown in the figure below, the GH200 chip features up to 480 GB of LPDDR5X CPU memory and supports up to 144 GB of HBM3e GPU memory, offering up to 624 GB of fast-access memory on a single GPU-CPU superchip. The CPU memory and GPU memory are connected via a cache-coherent interconnect (NVLink-C2C) to provide a unified, cache-coherent memory address space. The Hopper GPU can transfer data to/from Hopper GPUs on other superchips at 900GB/s via 18x NVLink 4 connections. (Note that the specifications for network and I/O in this figure are bi-directional bandwidth).



(In comparison, today's H100 GPU is a standalone Hopper GPU chip which is connected to a CPU chip via a 16x PCIe-5 I/O bus at a bi-directional bandwidth of 128GB/s.)

見背面

For simplicity, assume that our target LLM uses an FP8 to represent each parameter, most of the calculations are matrix-vector multiplications performed in FP8, and the arithmetic intensity is 1 FLOPs/Byte for FP8. Suppose we use a DGX GH200 to train the LLM, which of the following statements is/are true?

(A) If we only use the GPU memory to store the input data and parameters, the total GPU memory capacity, 144GB x 32 = 4.6TB, should be enough to run a LLM with 1 trillion parameters, but the performance is limited to 32 x 4900 GB/s * 1 FLOPs = 5 petaFLOP/s, which is only 20% of the peak performance of the system.

(B) It is possible to accelerate the training process for a 1-trillion parameter LLM with 2 sets of DGX GH200, but the scalability should be poor, according to the IEEE Spectrum article shown in a previous question.

(C) Because the CPU memory is 3~4 times larger than the GPU memory, if we modify our software to utilize of the unified memory feature on each GH200 node, we can train a much larger LLM at the same speed.

(D) If we cannot afford to buy a DGX GH200, we can still try to train a 1 trillion parameter LLM with a high-performance storage on one GH200 chip as a poor man's solution. We can use a solid-state disk array to take advantage of the high-speed I/O bandwidth (256GB/s each direction is not far from the 500GB/s CPU memory bandwidth) and optimize the software to minimize the performance degradation.

(E) Speaking of software optimization, cache blocking is important to accelerate large-scale matrix-vector multiplications, and the same concept can be used here. The principle behind cache blocking is to increase the cache block size to reduce the number of cache misses.

接次頁

Part II: Operating System (50 points)

NOTE that in the questions, it is intended to provide redundant or miss certain assumptions to disguise you. Please describe your own assumptions if necessary to answer the questions. Answers not in the order of question number might NOT be scored. You can answer in English or Chinese.

9. (20 points) For each of the following statements, answer **Yes** if it is TRUE or **brief why it is wrong.**
   A. (2 points) The less threading, the less performance.
   B. (2 points) The less frames, the more page fault rates.
   C. (2 points) The smaller time quantum for scheduling, the longer average turnaround time.
   D. (2 points) Firmware can be executed faster in ROM than in RAM.
   E. (2 points) Working set model has to be supported by MMU.
   F. (2 points) A safe state will not go to a deadlocked state.
   G. (2 points) Context switch is usually supported in hardware instruction for performance.
   H. (2 points) Section Semantics: Once the file is closed, the changes are visible only to new sessions.
   I. (2 points) A real-time scheduler schedules tasks according to their real-time priorities.
   J. (2 points) The two-phase locking protocol guarantees serializability and prevents deadlock.

10. (5 points) When a mouse device is moved, a monitor reflects by redrawing the mouse icon in windows manager. What does the mouse driver (not ISR) do in this event? How is the driver called? How does the windows manager know to start redrawing?

11. (10 points) Intrusive linked lists are a variation of linked lists, where the links are embedded in the item structure that's being linked as follows. It is used in Linux kernel for better performance than traditional linked lists. Why? Hint: Consider memory related advantages.

```
typedef struct list {
  struct list *next;
} list;

typedef struct item {
  int val;    // data
  list links;
} item;
```

12. (15 points) How is a Round Robin scheduler implemented so that each task is preempted accurately every 1 millisecond and resume after indefinitely? (5 points) In the same RR scheduling, please write (pseudo) codes for a daemon (looping) task T with comments as succinct as possible so that T runs for 1 second, then blocked for another 1 second. (10 points) Hint: Asynchronous I/O might be needed.