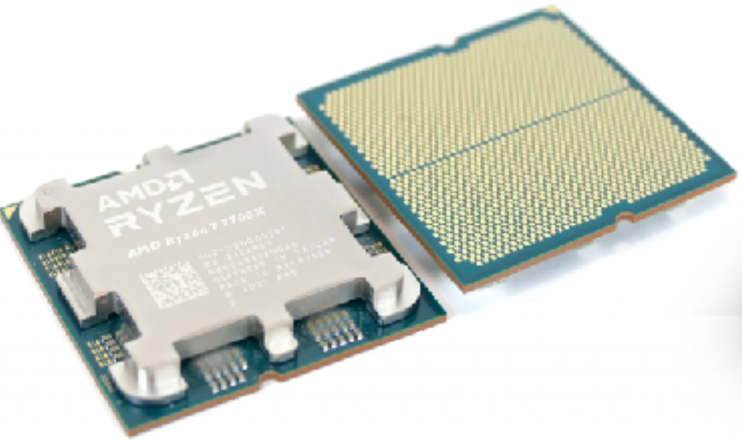
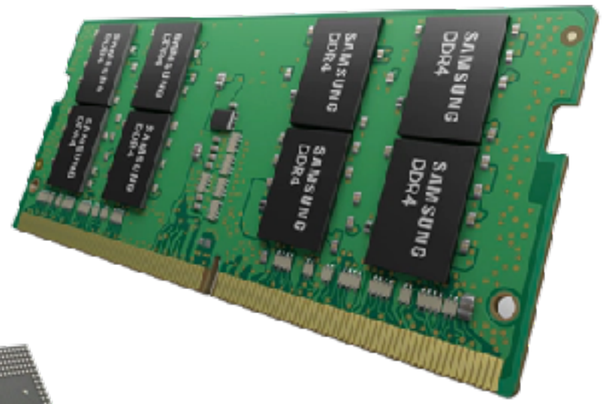
Performance (1): What does "perfect" mean?

Hung-Wei Tseng

Recap: Processors and memory modules are everywhere!

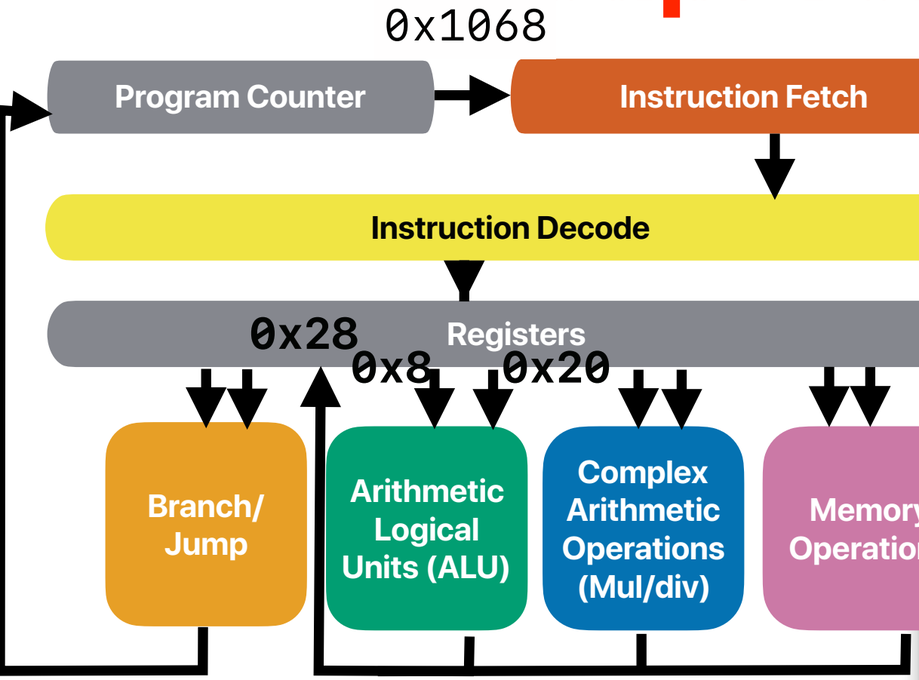
Processors



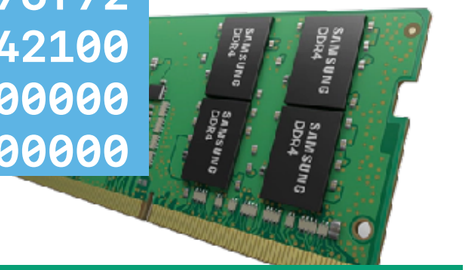


Memory

Recap: von Neumann architecture



Processor



Memory

Storage

```
int main(){
    pr
}
```

By loading different programs into memory, your computer can perform different functions

Recap: Demo

```
if(option)
    std::sort(data, data + arraySize);

for (unsigned c = 0; c < arraySize*1000; ++c) {
    int t = std::rand();

    (data[c%arraySize] >= t)
            sum ++;
    }
}
```

$O(nlog_2n)$

$O(n)$

if option is set to 1: $O(nlog_2n)$ **— but faster!!!**

otherwise, O(n): $O(n)$

Recap: Demo (2)

A

```
for(i = 0; i < ARRAY_SIZE; i++)
{
  for(j = 0; j < ARRAY_SIZE; j++)
  {
    c[i][j] = a[i][j]+b[i][j];
  }
}
```

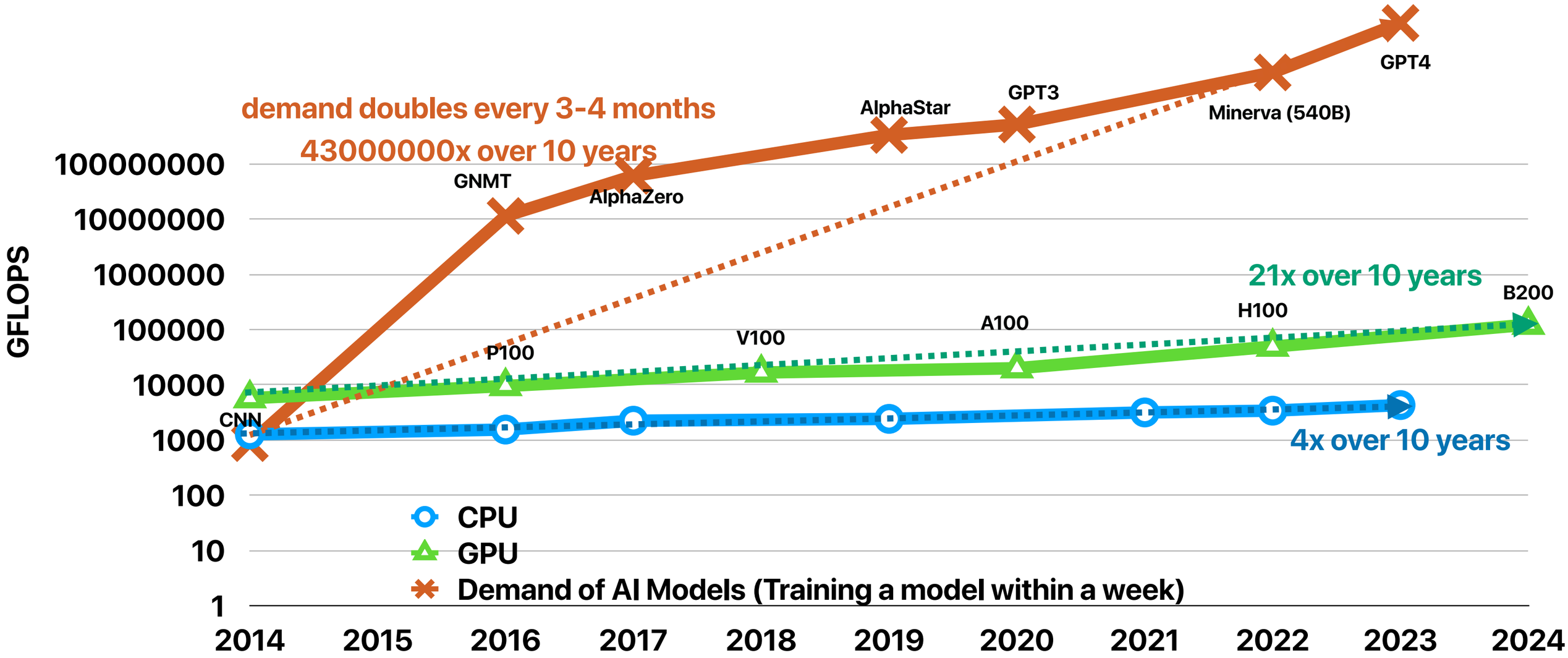
B

```
for(j = 0; j < ARRAY_SIZE; j++)
{
  for(i = 0; i < ARRAY_SIZE; i++)
  {
    c[i][j] = a[i][j]+b[i][j];
  }
}
```

$O(n^2)$ **Complexity** $O(n^2)$

A Lot Better! **Performance?** **Worse**

Recap: Mis-matching AI/ML demand and general-purpose processing

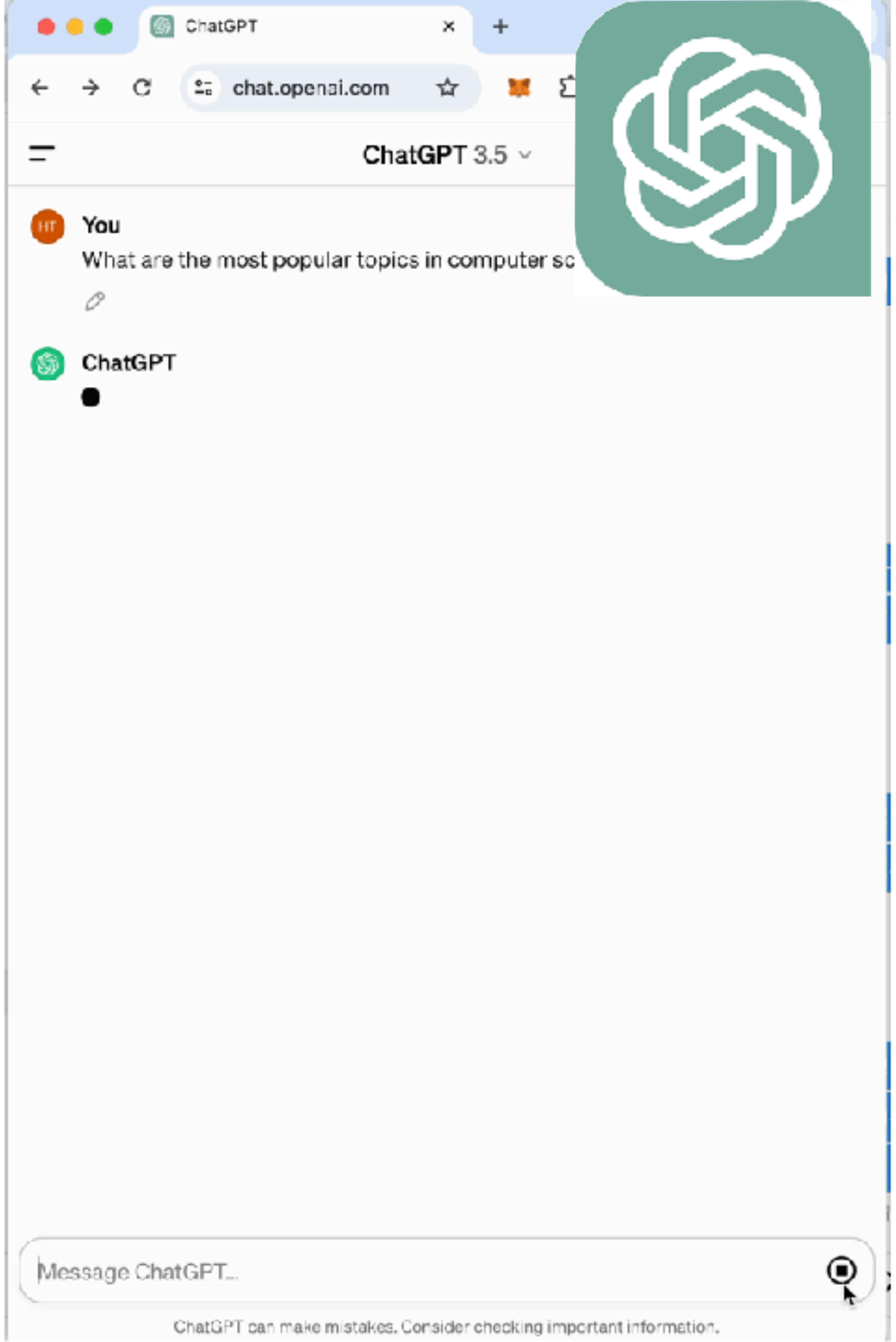


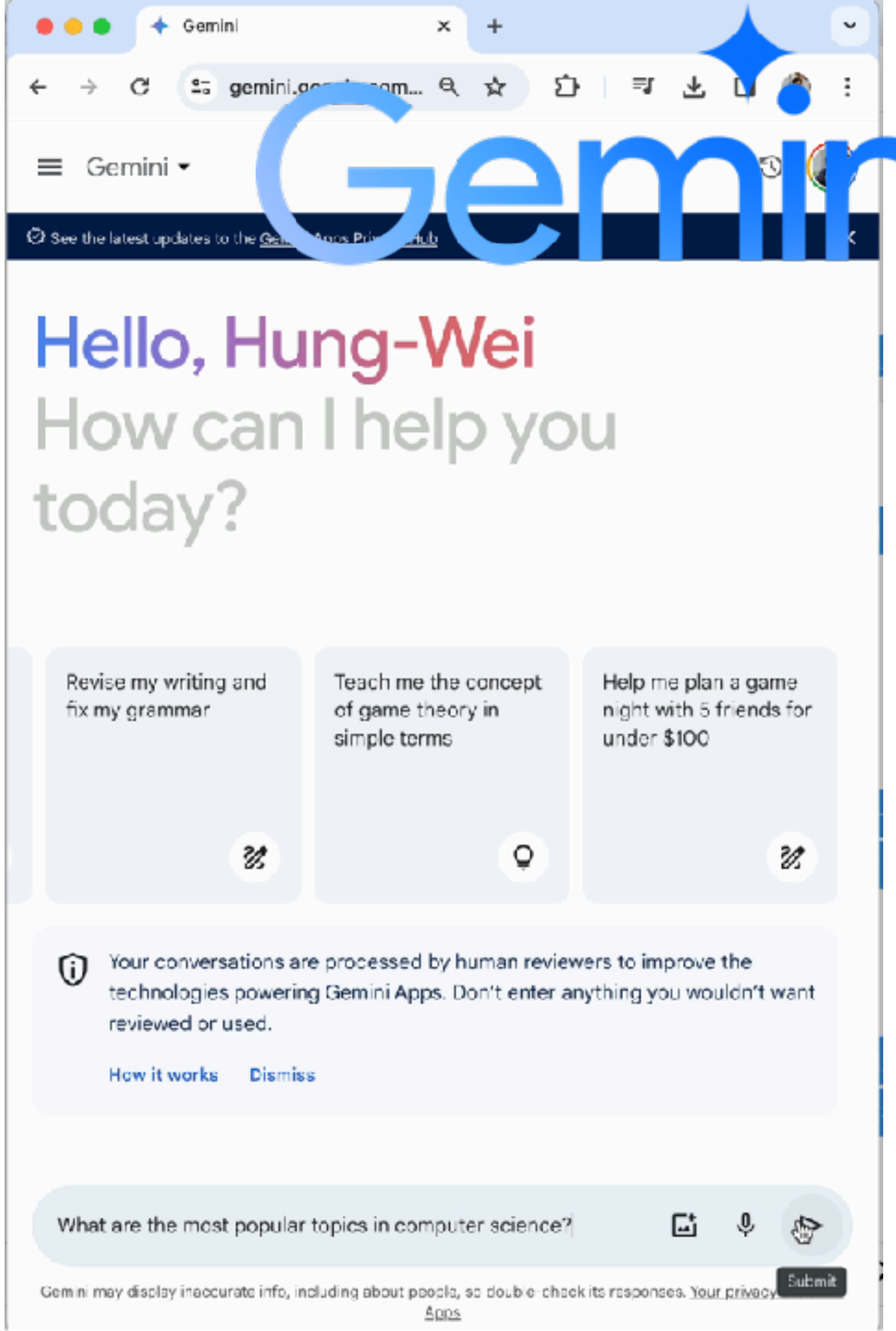
https://ourworldindata.org/grapher/artificial-intelligence-training-computation

Outline

- Definition of “Performance”
- The classical CPU performance equation
- Other important metrics

What does it really mean by performance?


ChatGPT
chat.openai.com
ChatGPT 3.5
You
What are the most popular topics in computer sc
ChatGPT
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.

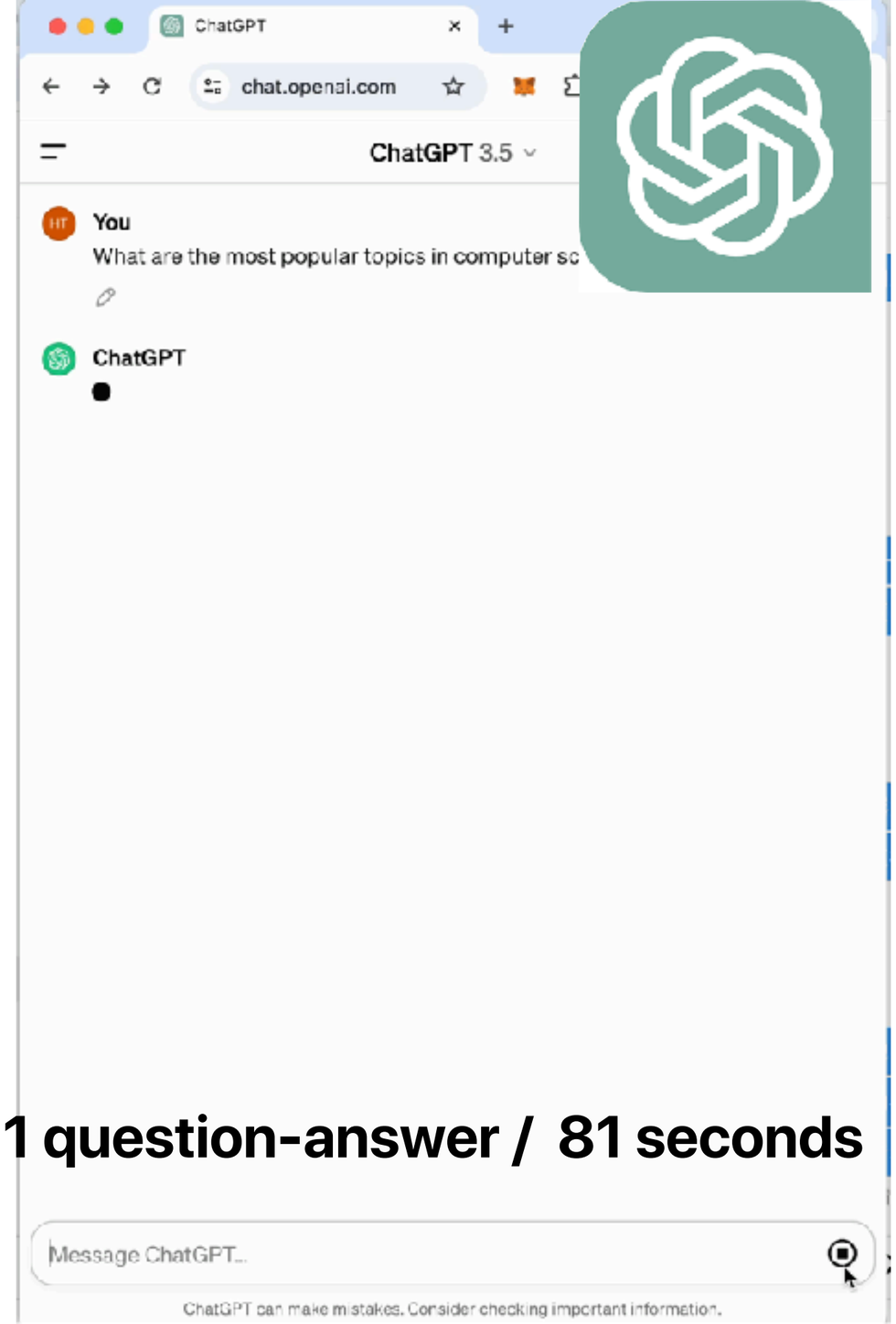

Gemini
Gemini
Hello, Hung-Wei
How can I help you today?
Revise my writing and fix my grammar
Teach me the concept of game theory in simple terms
Help me plan a game night with 5 friends for under $100
Your conversations are processed by human reviewers to improve the technologies powering Gemini Apps. Don't enter anything you wouldn't want reviewed or used.
How it works
Dismiss
What are the most popular topics in computer science?
Submit

Gemini v.s. ChatGPT

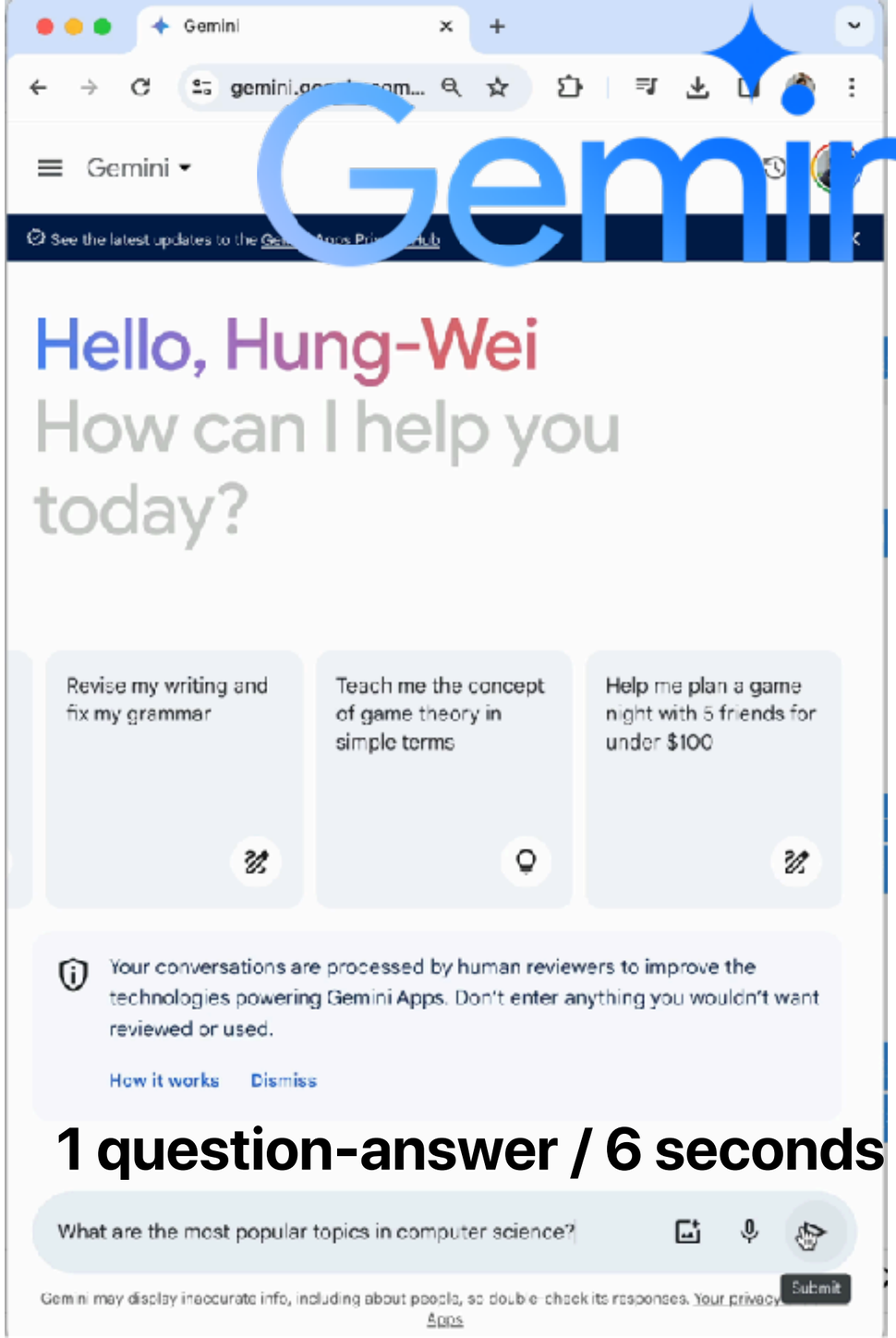
- Comparing the experiments we have done with Gemini and ChatGPT, how many of the following metrics does Gemini outperforms ChatGPT?
  - ① Response time
  - ② Throughput
  - ③ End-to-end latency (i.e., total execution time)
  - ④ Quality of results
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4

Gemini v.s. ChatGPT

- Comparing the experiments we have done with Gemini and ChatGPT, how many of the following metrics does Gemini outperforms ChatGPT?
  - ① Response time
  - ② Throughput
  - ③ End-to-end latency (i.e., total execution time)
  - ④ Quality of results
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4



1 question-answer / 81 seconds



1 question-answer / 6 seconds

Gemini v.s. ChatGPT

- Comparing the experiments we have done with Gemini and ChatGPT, how many of the following metrics does Gemini outperforms ChatGPT?
  ① Response time
  ② Throughput
  ③ End-to-end latency (i.e., total execution time)
  ④ Quality of results
  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Important performance metrics

- End-to-end latency — how much time the program/operation takes from the beginning to the end
- Response time — how much time the user starts to feel the program is running/finishing
- Throughput/bandwidth — the average amount of work/data can the program/system deliver within the execution time
- Energy consumption — the aggregated power during the execution time
- Cost of operation — the amount of money necessary for finishing an operation
- Quality of results — the human perception of the execution result
- Power consumption — the heat generation produced by the circuit

Important performance metrics

- End-to-end latency — how much **time** the program/operation takes from the beginning to the end
- Response time — how much **time** the user starts to feel the program is running/finishing
- Throughput/bandwidth — the average amount of work/data can the program/system deliver within the execution **time**
- Energy consumption — the aggregated power during the execution **time**
- Cost of operation — the amount of money necessary for finishing an operation (related to **time**)
- Quality of results — the human perception of the execution result
- Power consumption — the heat generation produced by the circuit

Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric

Let’s start with “end-to-end latency” as the default metric — how long it takes to execute a program?

https://www.pollev.com/hungweitseng close in 1:30

Performance equation

- Consider the following c code snippet and x86 instructions implement the code snippet

C	x86

```
for(i = 0; i < count; i++) {
    s += a[i];
}
```

```
.L3:
movslq  (%rdi), %rdx
addq    $4, %rdi
addq    %rdx, %rax
cmpq    %rcx, %rdi
jne     .L3
```

If (1) count is set to 1,000,000,000, (2) a memory instruction takes 5 cycles, (3) a branch/jump instruction takes 2 cycles, (4) other instructions takes 1 cycle on average, and (5) the processor runs at 4 GHz, how much time is it take to finish executing the code snippet?

A. 0.5 sec
B. 1 sec
C. 2.5 sec
D. 3.75 sec
E. 4 sec

Performance equation

- Consider the following c code snippet and x86 instructions implement the code snippet

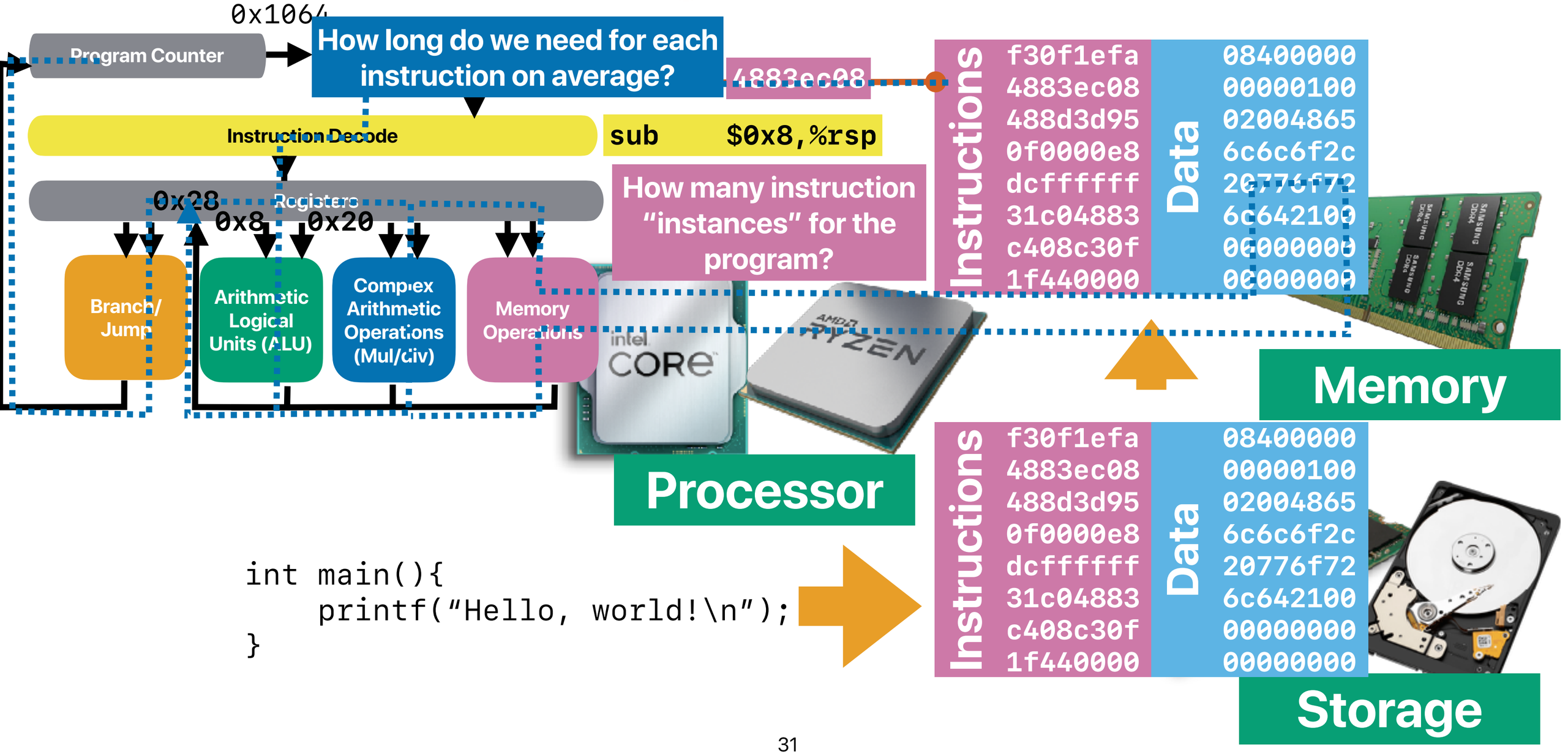
C	x86

`for(i = 0; i < count; i++) {`
`  s += a[i];`
`}`
	`.L3:`
	`movslq  (%rdi), %rdx`
	`addq  $4, %rdi`
	`addq  %rdx, %rax`
	`cmpq  %rcx, %rdi`
	`jne  .L3`

If (1) count is set to 1,000,000,000, (2) a memory instruction takes 5 cycles, (3) a branch/jump instruction takes 2 cycles, (4) other instructions takes 1 cycle on average, and (5) the processor runs at 4 GHz, how much time is it take to finish executing the code snippet?

A. 0.5 sec
B. 1 sec
C. 2.5 sec
D. 3.75 sec
E. 4 sec

Execution time of a program in the von Neumann model




```
int main(){
    printf("Hello, world!\n");
}
```

CPU Performance Equation

$$Performance = \frac{1}{Execution\ Time}$$

$$Execution\ Time = \frac{Instructions}{Program} \times \frac{Cycles}{Instruction} \times \frac{Seconds}{Cycle}$$

$$ET = IC \times CPI \times CT$$

$$1GHz = 10^9 Hz = \frac{1}{10^9} sec\ per\ cycle = 1\ ns\ per\ cycle$$

$$\frac{1}{Frequency(i.e., clock\ rate)}$$

Classic CPU Performance Equation (ET of a program)

How many instruction "instances" for the program? × **How long do we need for each instruction on average?**

$$Execution\ Time = \frac{Instructions}{Program} \times \frac{Cycles}{Instruction} \times \frac{Seconds}{Cycle}$$

C Code	x86 instructions

```
int init_data(int64_t *data, int
data_size) {
    register unsigned int i = 0;
    for(i = 0; i < data_size; i++)
{
       s+=data[i];
    }
    return s;
}

int main(int argc, char **argv) {
    int *data = malloc(8000000000);
    init_data(data, 1000000000);
    return 0;
}
```

```
init_data:
.LFB16:
    endbr64
    testl    %esi, %esi
    jle .L2
    leal     -1(%rsi), %ecx
    xorq     %rax, %rax
.L3:
→movslq   (%rdi), %rdx
    addq     $4, %rdi
    addq     %rdx, %rax
    cmpq     %rcx, %rdi
→jne       .L3
.L2:
    xorlq    %rax, %rax
    ret
```

memory inst.

branch inst.

If data memory access instructions takes 5 cycles, branch 2 cycles, others take only 1 cycle, CPU freq. = 4 GHz

$$CPI_{average} = 20\% \times 5 + 20\% \times 2 + 60\% \times 1) = 2$$

$$ET = (5 \times 10^9) \times 2 \times \frac{1}{4 \times 10^9}\ sec = 2.5\ sec$$

1000000000x

Performance equation

- Consider the following c code snippet and x86 instructions implement the code snippet

C	x86

`for(i = 0; i < count; i++) {`
`  s += a[i];`
`}`
	`.L3:`
	`movslq (%rdi), %rdx`
	`addq $4, %rdi`
	`addq %rdx, %tax`
	`cmpq %rcx, %rdi`
	`jne .L3`

If (1) count is set to 1,000,000,000, (2) a memory instruction takes 4 cycles, (3) a branch/jump instruction takes 3 cycles, (4) other instructions takes 1 cycle on average, and (5) the processor runs at 4 GHz, how much time is it take to finish executing the code snippet?

A. 0.5 sec
B. 1 sec
C. 2.5 sec
D. 3.75 sec
E. 4 sec

$$ET = IC \times CPI \times CT$$

$$ET = (5 \times 10^9) \times (20\% \times 5 + 20\% \times 2 + 60\% \times 1) \times \frac{1}{4 \times 10^9}\ sec = 2.5\ sec$$

total # of dynamic instructions

average CPI

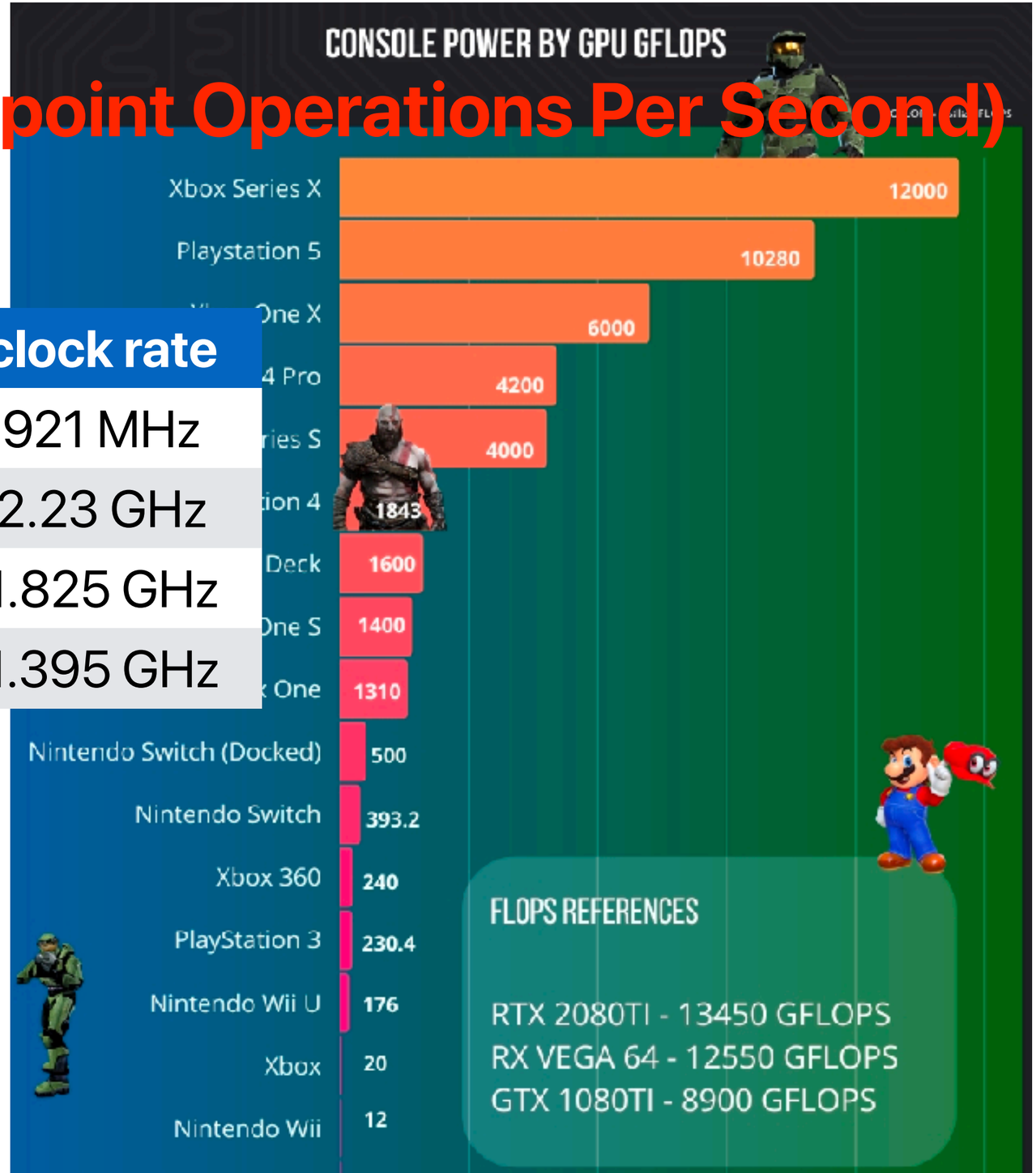
Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric
- Instruction count, cycles per instruction, cycle time define the latency of execution on CPUs

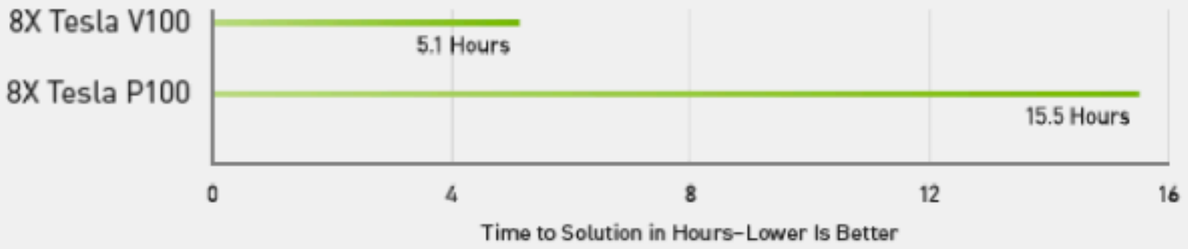
Metrics on throughputs

TFLOPS (Tera FLoating-point Operations Per Second)

	TFLOPS	clock rate
Switch	1	921 MHz
PS5	10.28	2.23 GHz
XBox Series X	12	1.825 GHz
GeForce RTX 3090	40	1.395 GHz



Deep Learning Training in Less Than a Workday



Server Config: Dual Xeon E5-2699 v4 2.6 GHz | 8X NVIDIA® Tesla® P100 or V100 | ResNet-50 Training on MXNet for 90 Epochs with 1.28M ImageNet Dataset.

AI TRAINING

From recognizing speech to training virtual personal assistants and teaching autonomous cars to drive, data scientists are taking on increasingly complex challenges with AI. Solving these kinds of problems requires training deep learning models that are exponentially growing in complexity, in a practical amount of time.

With 640 Tensor Cores, Tesla V100 is the world's first GPU to break the 100 teraFLOPS (TFLOPS) barrier of deep learning performance. The next generation of NVIDIA NVLink™ connects multiple V100 GPUs at up to 300 GB/s to create the world's most powerful computing servers. AI models that would consume weeks of computing resources on previous systems can now be trained in a few days. With this dramatic reduction in training time, a whole new world of problems will now be solvable with AI.

TFLOPS (Tera FLoating-point Operations Per Second)

$$TFLOPS = \frac{\#\ of\ floating\ point\ instructions \times 10^{-12}}{Exection\ Time}$$

Let's measure the TFLOPS of matrix multiplications

```
for(i = 0; i < ARRAY_SIZE; i++) {
  for(j = 0; j < ARRAY_SIZE; j++) {
    for(k = 0; k < ARRAY_SIZE; k++) {
      c[i][j] += a[i][k]*b[k][j];
    }
  }
}
```

Floating point operations per second (FLOP"S"):

Floating point operations (FLOP"s"):

$$i \times j \times k \times 2$$

Given $i = j = k = 2048$

$2^{3\times11} \times 2 = 2^{34}$ **FLOPs in total**

$$TFLOPS = \frac{2^{34}}{ET_{seconds} \times 2^{12}}$$

https://www.pollev.com/hungweitseng close in 1:30

How reflective is FLOPS?

- If you're given the FLOPS of an underlying GPU, how many situations below can the FLOPS be representative to the real performance?
  - ① The FLOPS remains the same on the same GPU even if we change the dataset
  - ② The FLOPS remains the same on the same GPU even if we change the data type to double
  - ③ The FLOPS remains the same on the same GPU if we change the algorithm implementation
  - ④ The ratio of FLOPS on two different GPUs reflects the ratio of execution time on these two GPUs when executing floating point applications

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Demo: matmul on GPU

Size	Latency	Relative Latency	Throughput (Output Numbers Per Second)	Relative Throughput
16x16x16	~ 0.09ms	1	0.09ms/256	1
32x32x32	~ 0.09ms	1	0.09ms/1024	4
64x64x64	~ 0.09ms	1	0.09ms/4096	16

Larger throughput doesn’t mean shorter latency!

How reflective is FLOPS?

- If you're given the FLOPS of an underlying GPU, how many situations below can the FLOPS be representative to the real performance?
  - ① The FLOPS remains the same on the same GPU even if we change the dataset
  - ② The FLOPS remains the same on the same GPU even if we change the data type to double
  - ③ The FLOPS remains the same on the same GPU if we change the algorithm implementation
  - ④ The ratio of FLOPS on two different GPUs reflects the ratio of execution time on these two GPUs when executing floating point applications

A. 0
B. 1
C. 2
D. 3
E. 4

TFLOPS (Tera FLoating-point Operations Per Second)

$$TFLOPS = \frac{\#\ of\ floating\ point\ instructions \times 10^{-12}}{Exection\ Time}$$

Is TFLOPS (Tera FLoating-point Operations Per Second) a good metric?

$$TFLOPS = \frac{\#\ of\ floating\ point\ instructions \times 10^{-12}}{Exection\ Time}$$

$$= \frac{IC \times \%\ of\ floating\ point\ instructions \times 10^{-12}}{IC \times CPI \times CT}$$

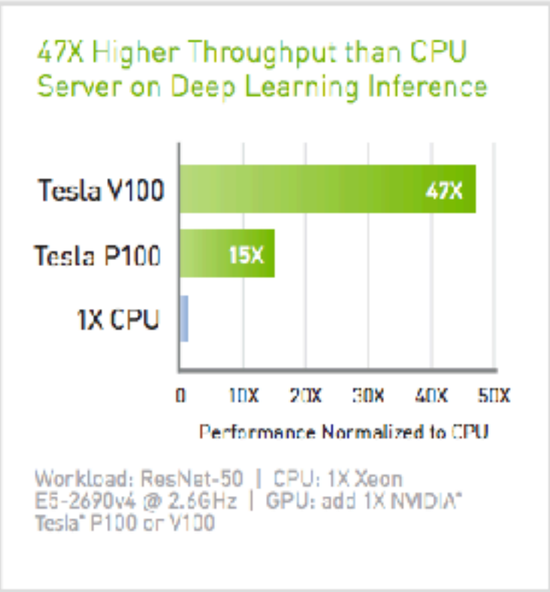
$$= \frac{\%\ of\ floating\ point\ instructions \times 10^{-12}}{CPI \times CT}$$

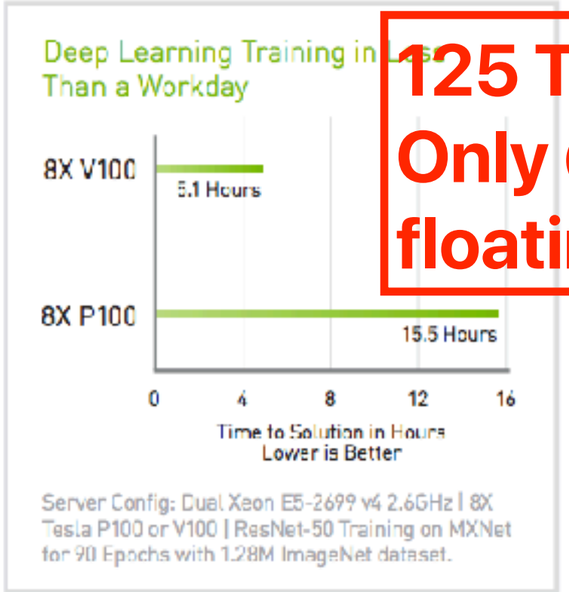
IC is gone!

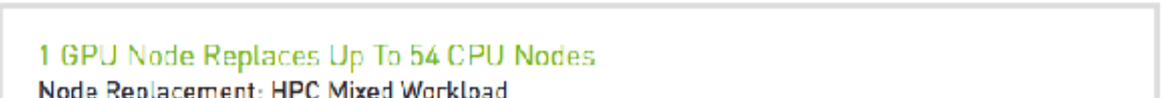
- If we have more iterations? Larger datasets? — potentially changes the IC
- What if the hardware trade (cheat) performance with accuracy?
- Cannot compare different ISA/compiler
  - What if the compiler can generate code with fewer instructions?
  - What if new architecture has more IC but also lower CPI?
- If floating point operations are not critical in the target application?

The Most Advanced Data Center GPU Ever Built.

NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta, the latest GPU architecture, Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once thought impossible.

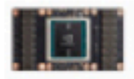






125 TFLOPS Only @ 16-bit floating point

SPECIFICATIONS



	Tesla V100 PCIe	Tesla V100 SXM2
GPU Architecture	NVIDIA Volta	
NVIDIA Tensor Cores	640	
NVIDIA CUDA® Cores	5,120	
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS
GPU Memory	32GB /16GB HBM2	
Memory Bandwidth	900GB/sec	
ECC	Yes	
Interconnect Bandwidth	32GB/sec	300GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink
Form Factor	PCIe Full Height/Length	SXM2
Max Power		

forbes.com/sites/arm/2023/12/04/the-compute-four

Click to go back, hold to see history

Forbes

AI training vs. inference

Much of the news coverage recently has been on LLMs, their development and their training – and the high cost and energy consumption required to do so. A recent study[1] estimated that Chat GPT3, comprised of 175 billion parameters, needed 1,287 MWh to train, and also emitted 552 tons of CO_2. This is roughly the equivalent to driving a car 1.3 million miles – but in one hour!

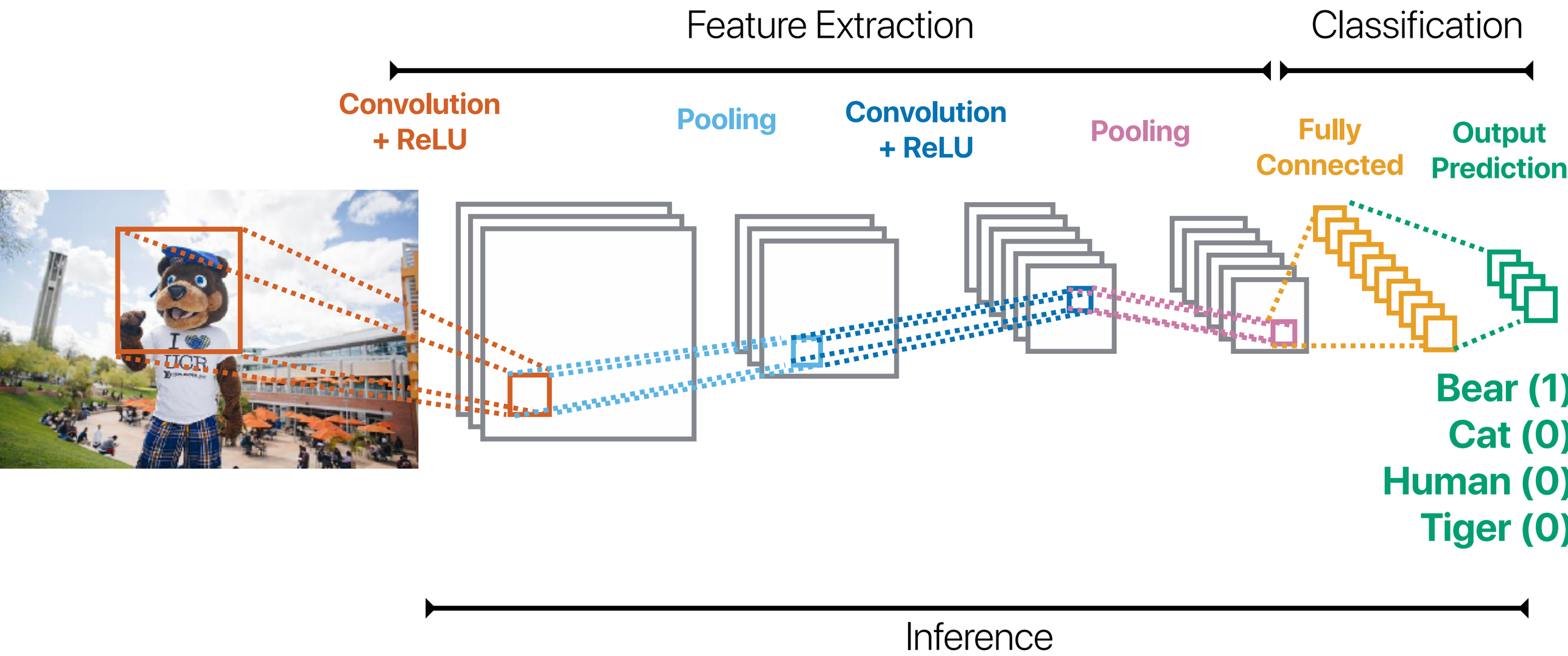
But those huge training workloads running in data centers represent just 15 percent of AI workloads today, according to Omdia's Data Center Compute Intelligence Service.[2]

The rest – 85 percent of all data center AI workloads – lies in AI inference, and that's not accounting for inference that's happening outside the data center on the edge. AI workloads like inference will remain a key workload in the cloud, but an increasing number will move to the edge as more efficient models continue to evolve. Inference, which is the process of using a trained model that has been deployed into a production environment to make predictions on new real-world

The rest – 85 percent of all data center AI workloads – lies in AI inference, and that's not accounting for inference that's happening outside the data center on the edge. AI workloads like inference will remain a key

https://www.forbes.com/sites/arm/2023/12/04/the-compute-foundation-powering-ais-present-and-future/

The Machine Learning Inference Pipeline



https://www.pollev.com/hungweitseng close in 1:30

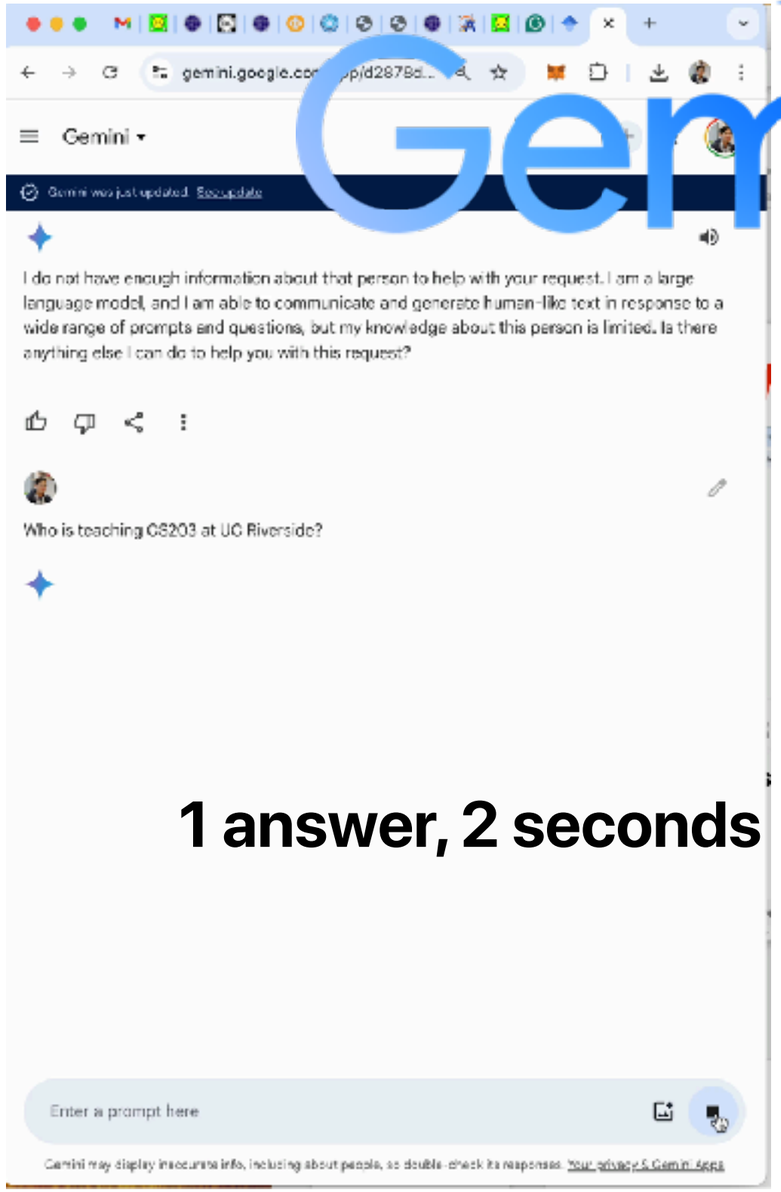
How reflective is “inferences per second”

- Regarding inferences per second (IPS), please identify how many of the following statements are correct
  - ① IPS can change if the application changes the ML model used
  - ② IPS can become worse if the application adds more features to improve the quality of the answer or the precision
  - ③ IPS can improve if the system applies a hardware that offers higher FLOPS
  - ④ IPS remains the same if the system/application answers questions 20x slower but can answer 20x more questions in parallel

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

What about inference per second?

1 answer, 2 seconds



1 answer, 2 seconds

Throughput v.s. goodput

- Throughput — the amount of work can be done within a given period of time (typically “something” per “timeframe” or the other way around)
  - Bandwidth (MB/Sec, GB/Sec, Mbps, Gbps)
  - IOPs (I/O operations per second)
  - FLOPs (Floating-point operations per second)
- Goodput — the amount of useful work can be done within a given period of time (e.g., error correction doesn’t count)
  - IPS (Inferences per second)
  - FPS (Frames per second)

Inferences per second

$$\frac{Inferences}{Second} = \frac{Inferences}{Operation} \times \frac{Operations}{Second}$$

$$= \frac{Inferences}{Operation} \times [\frac{operations}{cycle} \times \frac{cycles}{second} \times \#\_of\_PEs \times Utilization\_of\_PEs]$$

	Hardware	Model	Input Data
Operations per inference		v	
Operations per cycle	v		
Cycles per second	v		
Number of PEs	v		
Utilization of PEs	v	v	
Effectual operations out of (total) operations		v	v
Effectual operations plus unexploited ineffectual operations per cycle	v		

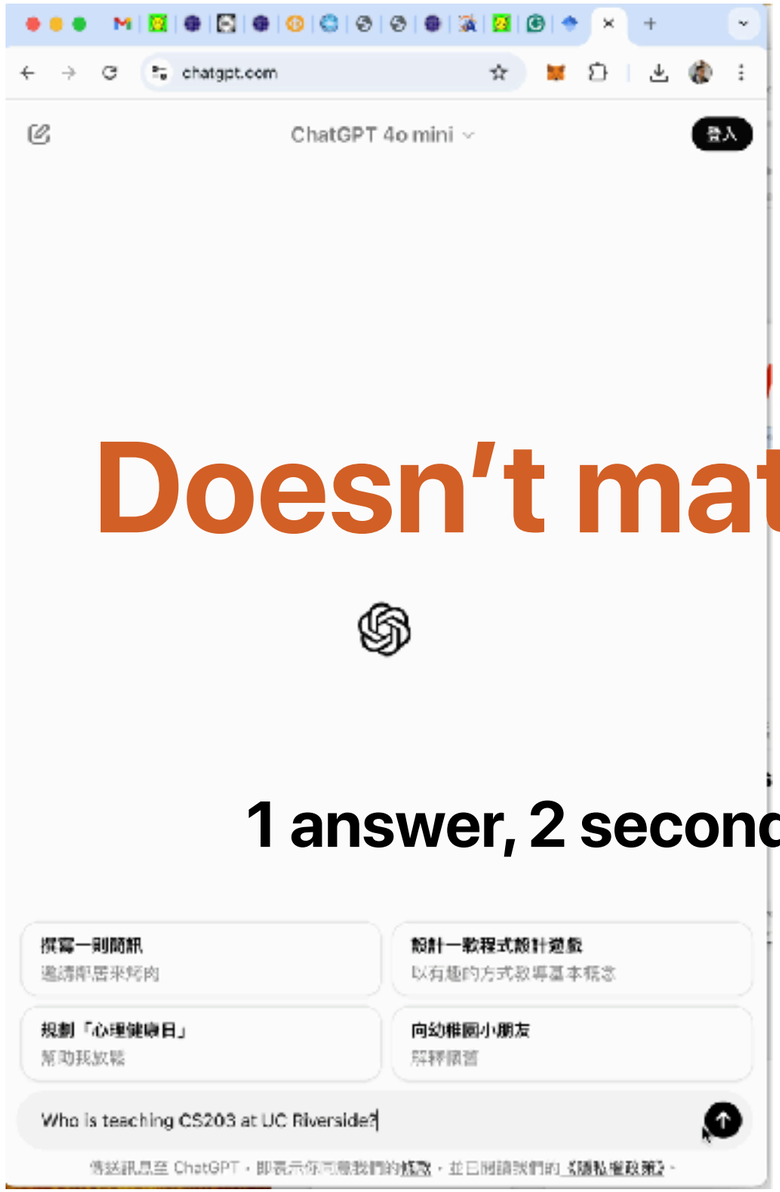
What's wrong with inferences per second?

- There is no standard on how they inference — but these affect!
  - What model?
  - What dataset?
  - Quality?
- That's why Facebook is trying to promote an AI benchmark — MLPerf

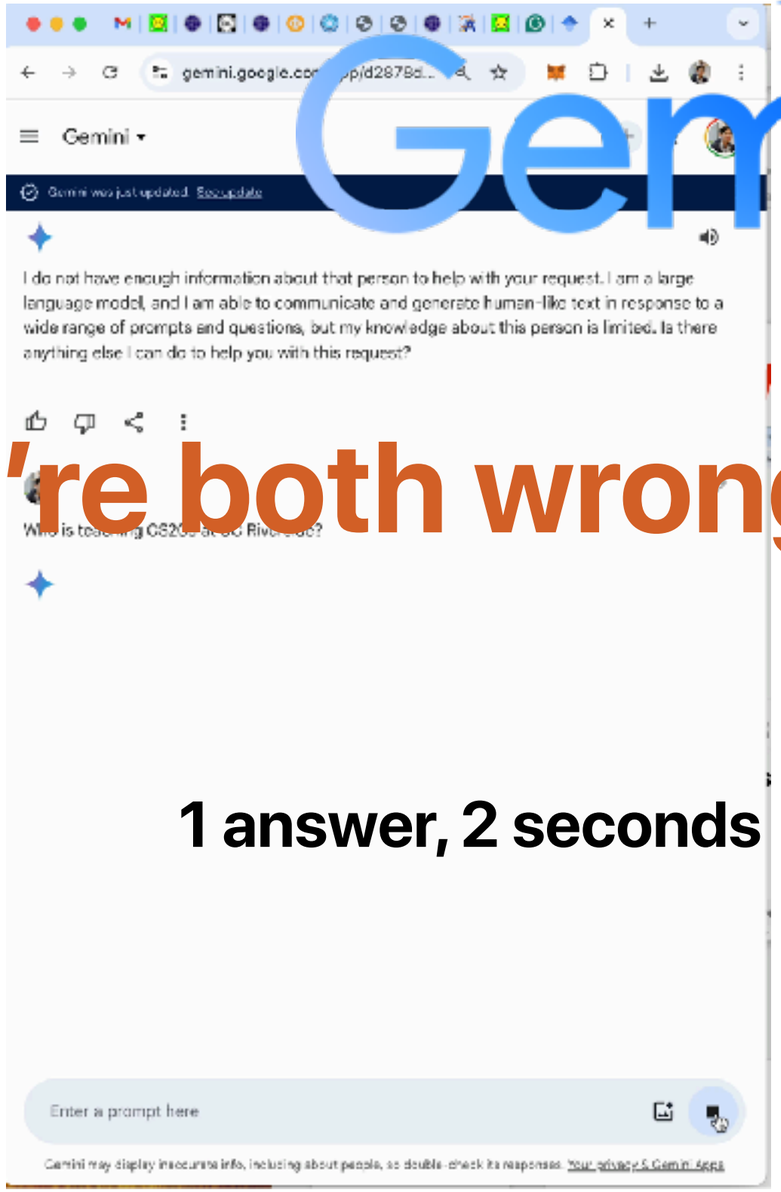
- *Pitfall: For NN hardware, Inferences Per Second (IPS) is an inaccurate summary performance metric.*

Our results show that IPS is a poor overall performance summary for NN hardware, as it's simply the inverse of the complexity of the typical inference in the application (e.g., the number, size, and type of NN layers). For example, the TPU runs the 4-layer MLP1 at 360,000 IPS but the 89-layer CNN1 at only 4,700 IPS, so TPU IPS vary by 75X! Thus, using IPS as the single-speed summary is *even more misleading* for NN accelerators than MIPS or FLOPS are for regular processors [23], so IPS should be even more disparaged. To compare NN machines better, we need a benchmark suite written at a high-level to port it to the wide variety of NN architectures. Fathom is a promising new attempt at such a benchmark suite [3].

What about inference per second?



1 answer, 2 seconds



1 answer, 2 seconds

Doesn't matter — they're both wrong :)

How reflective is “inference per second”

- Regarding inferences per second (IPS), please identify how many of the following statements are correct
  - ① IPS can change if the application changes the ML model used
  - ② IPS can become worse if the application adds more features to improve the quality of the answer or the precision
  - ③ IPS can improve if the system applies a hardware that offers higher FLOPS
  - ④ IPS remains the same if the system/application answers questions 20x slower but can answer 20x more questions in parallel

What if the model uses integers? At lower precisions?

A. 0
B. 1
C. 2
D. 3
E. 4

gamespot.com/articles/nvidia-accused-of-cheating-in-3dmark-03/1100-6028694/#:~:text="We%20have%20now%20established%20that,dr

Nvidia accused of cheating in 3DMar

Futuremark, the maker of 3DMark 03, releases a patch for the g to correct the way Nvidia drivers manipulate results in the popu

By David Becker on May 23, 2003 at 4:27PM PDT

Saratoga, California-based Futuremark on Friday said in a statement that Nvidia tweaked the software needed to run its new GeForce FX 5900 processor to distort performance in Futuremark's 3DMark 03 testing application. The company said drivers for the new Nvidia chip were altered to detect activity characteristic of a benchmark and adjust performance accordingly.

Recently, there have been questions and some confusion regarding 3DMark 03 results obtained with certain Nvidia products, Futuremark said in the statement. "We have now established that Nvidia's Detonator FX drivers contain certain detection mechanisms that cause an artificially high score when using 3DMark 03," the statement read.

Futuremark has released the version 330 patch for 3DMark 03, which prevents the Nvidia drivers from identifying the benchmark. By Futuremark's measure, the performance of the Nvidia GeForce FX 5900 Ultra drops 24 percent with the patch, compared with a drop of less than 2 percent with ATI's Radeon 9800 Pro with the latest ATI drivers.

A representative at Nvidia questioned the validity of Futuremark's conclusions. "Since Nvidia is not part of the Futuremark beta program (a program which costs of hundreds of thousands of dollars to participate in), we do not get a chance to work with Futuremark on writing the shaders like we would with a real applications developer," the representative said. "We don't know what they did, but it looks like they have intentionally tried to create a scenario that makes our products look bad."

https://www.gamespot.com/articles/nvidia-accused-of-cheating-in-3dmark-03/1100-6028894/#:~:text=%22We%20have%20now%20established%20that,drivers%20from%20identifying%20the%20benchmark.

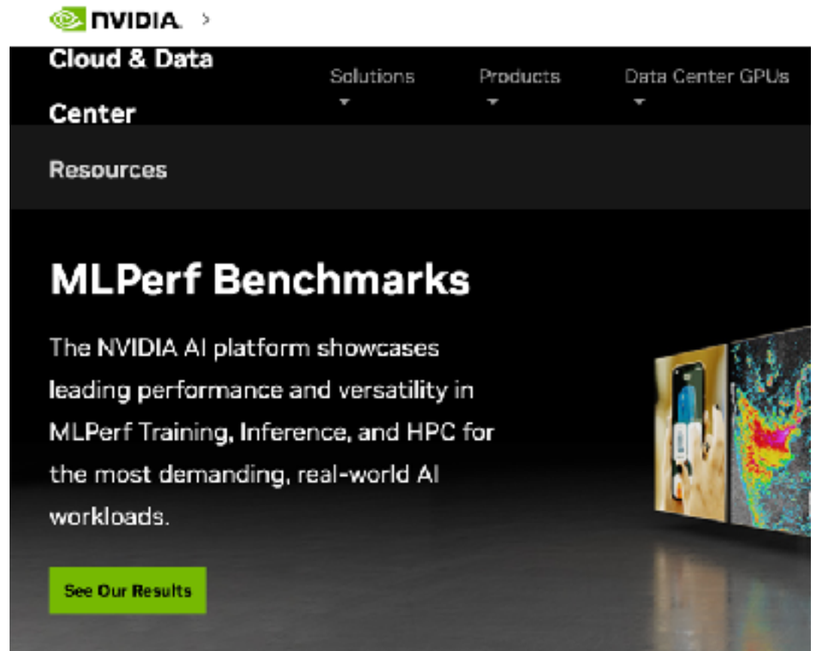
Nvidia is amid a hard-fought battle with rival ATI Technologies to claim the performance lead in PC graphics processors. After years of Nvidia dominating both in market share and performance, ATI

Fair comparison in computer architectures

- Metrics: you must consider the fact that performance is composed of IC, CPI, and CT. — any metric that misses one of them is misleading
- Only one variation in each comparison
  - Only change the processor, but not ISA (related to IC) and others
  - Only change the algorithm, but not others
  - The same dataset, must be the same outcome

The reason of “Benchmark Suites”

- Allowing people evaluate systems with exactly the same program and the same inputs and validate results from different machines
- Popular benchmark suites
  - SPEC — CPU benchmark
  - MLPerf — ML systems



Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric
- Classic CPU performance equation —Instruction count (IC), cycles per instruction (CPI), cycle time (CT) define the latency of execution on CPUs
- Performance metrics without considering all three factors in the classic performance equation can mislead — anything throughput typically miss one of them

What does the company really care about?

- Making profits!
  - Delivering features/services at an acceptable level
  - Reducing the cost as much as possible
    - TCO — total cost of ownership
    - OpEX — Operational expenses
    - CO2e — carbon footprint emissions

An increasingly important performance metric

un.org/en/climatechange/net-zero-coalition

ome to the United Nations العربية 中文 English França

Forbes

FORBES > INNOVATION

GenAI's Carbon Footprint: A New Challenge For Corporations

Nicola Sfondrini Forbes Councils Member

Forbes Technology Council COUNCIL POST | Membership (Fee-Based)

Mar 28, 2024, 08:15am EDT

Nicola Sfondrini - Partner Digital and Cloud Strategy at PWC.

zero-carbon/

-zero carbon

DOI:10.1145/3637867
Amin Vahdat, Xiaoyu Ma, and David Patterson

Opinion

New Computer Evaluation Metrics for a Changing World

To find a path to grow AI and cloud computing efficiently and responsibly, we ne[illegible] new metrics. Computing systems should be measured on their use of availa[illegible]

New Computer Evaluation Metrics for a Changing World

Amin Vahdat, Xiaoyu Ma, David Patterson
Google
Communications of the ACM, 2024

https://www.pollev.com/hungweitseng close in 1:30

Carbon dioxide equivalent emissions, CO_2e

- Regarding CO_2e, how many of the following statements are correct?
  - ① Lowering the system power consumption helps to reduce the operational expenses
  - ② Using a GPU with 2x more power consumption but shorten the execution time by 3x can reduce the operational CO_2e
  - ③ Continuous upgrade of GPUs will increase the data center's embodied emissions
  - ④ A system with higher FLOPS may not deliver better goodputs for applications

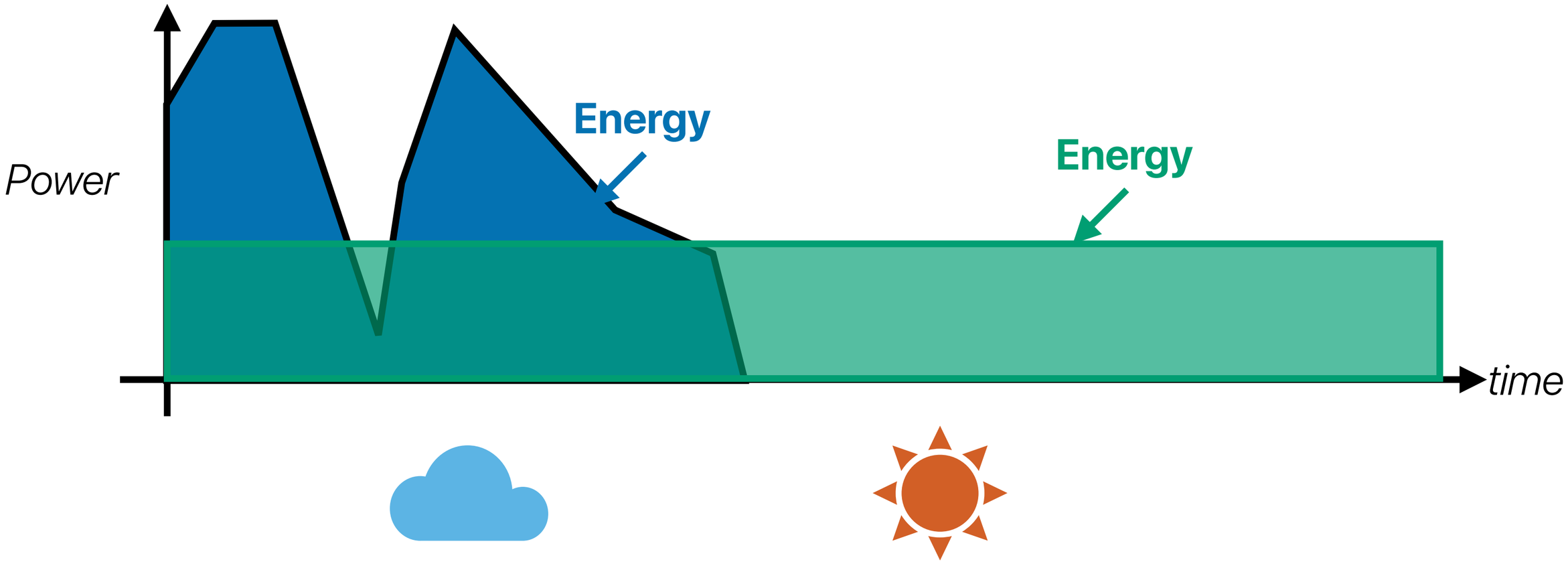
A. 0
B. 1
C. 2
D. 3
E. 4

Carbon dioxide equivalent emissions, CO_2e

- Regarding CO_2e, how many of the following statements are correct?
  - ① Lowering the system power consumption helps to reduce the operational expenses
  - ② Using a GPU with 2x more power consumption but shorten the execution time by 3x can reduce the operational CO_2e
  - ③ Continuous upgrade of GPUs will increase the data center's embodied emissions
  - ④ A system with higher FLOPS may not deliver better goodputs for applications
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4

Power/Energy/Carbon footprint

The Green can be more if power is not low enough



If we run the task when there is no green energy— more carbon footprint!

$$Energy = Power \times Execution\_Time$$

Demo — changing the max frequency and performance

- Change the maximum frequency of the intel processor — you learned how to do this when we discuss programmer's impact on performance
- LIKWID a profiling tool providing power/energy information
  - likwid-perfctr -g ENERGY [command_line]
  - Let's try blockmm and see what's happening!

Carbon dioxide equivalent emissions, CO_2e

- Regarding CO_2e, how many of the following statements are correct?
  - ① Lowering the system power consumption helps to reduce the operational expenses
  - ② Using a GPU with 2x more power consumption but shorten the execution time by 3x can reduce the operational CO_2e
  - ③ Continuous upgrade of GPUs will increase the data center's embodied emissions
  - ④ A system with higher FLOPS may not deliver better goodputs for applications

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

partially true — it reduces the cooling cost, but may increase operational energy!

Reduce energy by ⅓ ! — ET matters

embodied emissions — the CO_2e when manufacturing

throughput (e.g., FLOPS) may include overhead, goodput does not

Takeaways: What does "perfect" mean?

- Latency is the most fundamental performance metric
- Classic CPU performance equation —Instruction count (IC), cycles per instruction (CPI), cycle time (CT) define the latency of execution on CPUs
- Performance metrics without considering all three factors in the classic performance equation can mislead — anything throughput typically miss one of them
- CO_2e, energy efficiency and power become increasingly important — but latency/execution time still matters as

$$Energy = Power \times Execution\_Time$$

Announcement

- Programming assignment 1 **due this Thursday** & Assignment 1 **due next Thursday**
  - We cannot help you at the last minute — please start early
  - Watch before you start https://youtu.be/m7OoY8y_Isk
  - Please always make sure you follow the exact steps in the readme and the notebook
  - Submit to the right item on Gradescope
  - Please visit an office hour if you need more assistance
- Reading quiz due next Tuesday before the lecture — we will drop two of your least performing reading quizzes
- Check our website for slides, Gradescope for assignments, discord for discussions
- Check your grades at https://www.escalab.org/my_grades
  - If you don't have any grade, you need to make sure your gradescope account is associated with your UCRNetID@ucr.edu
  - You have to submit a course agreement to receive scores
- Youtube channel for lecture recordings: https://www.youtube.com/c/ProfUsagi/playlists

Computer
Science &
Engineering

203

つづく