

Performance (1):

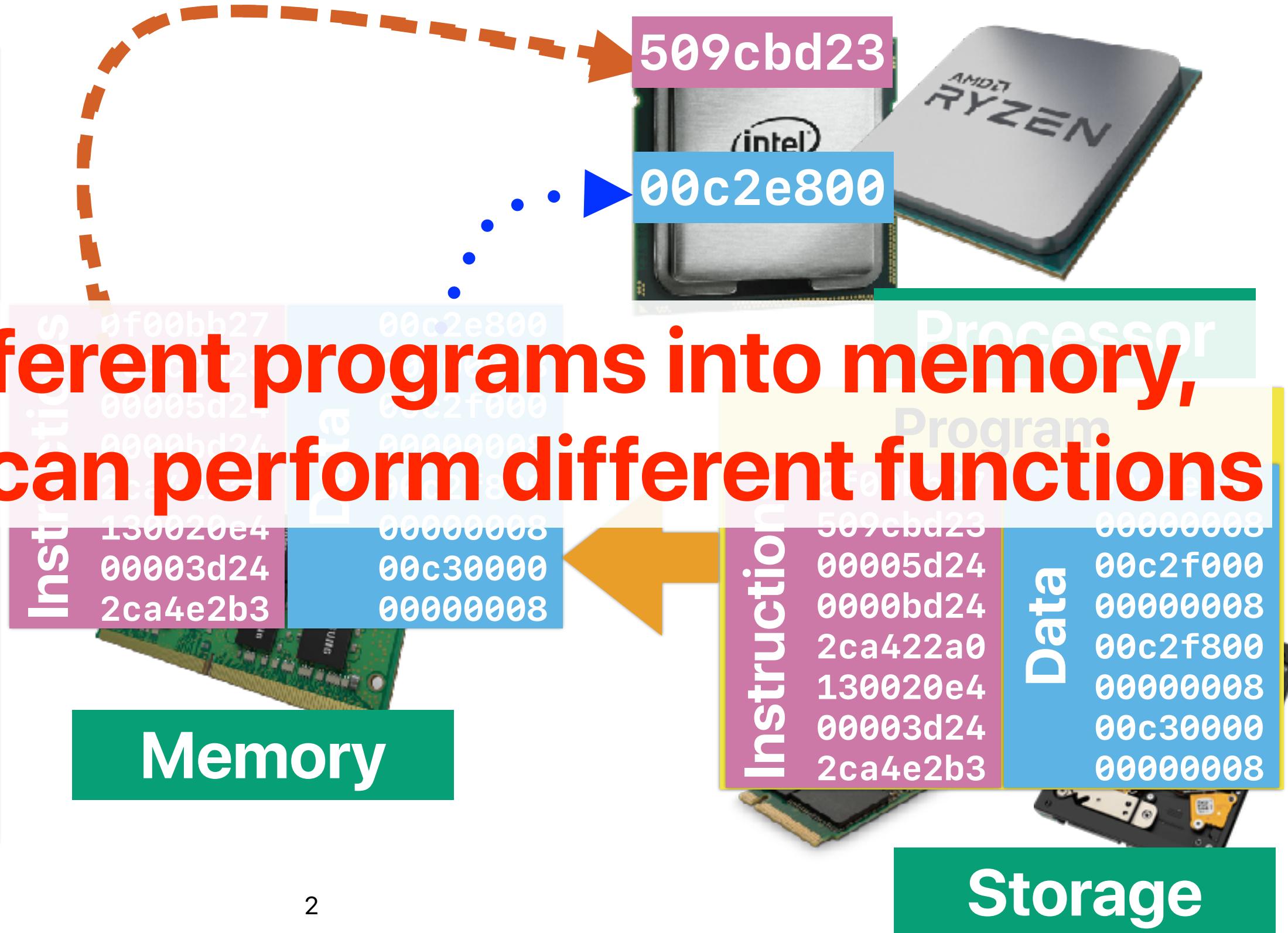
What does “perfect” mean?

Hung-Wei Tseng

Recap: von Neumann Architecture



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



Recap: Start with this simple program in C

```
int A[] =  
{1,2,3,4,5,6,7,8,9,10,1,2,3,4  
,5,6,7,8,9,10};
```

Compiler

Contents of section .data:
0000 01000000 02000000 03000000 04000000
0010 05000000 06000000 07000000 08000000
0020 09000000 0a000000 0b000000 0c000000
0030 03000000 04000000 05000000 06000000
0040 07000000 08000000 09000000 0a000000

control flow
operations
logical
operations

```
int main()  
{  
    int i=0, sum=0;  
    for(i = 0; i < 20; i++)  
    {  
        sum += A[i];  
    }  
    return 0;  
}
```

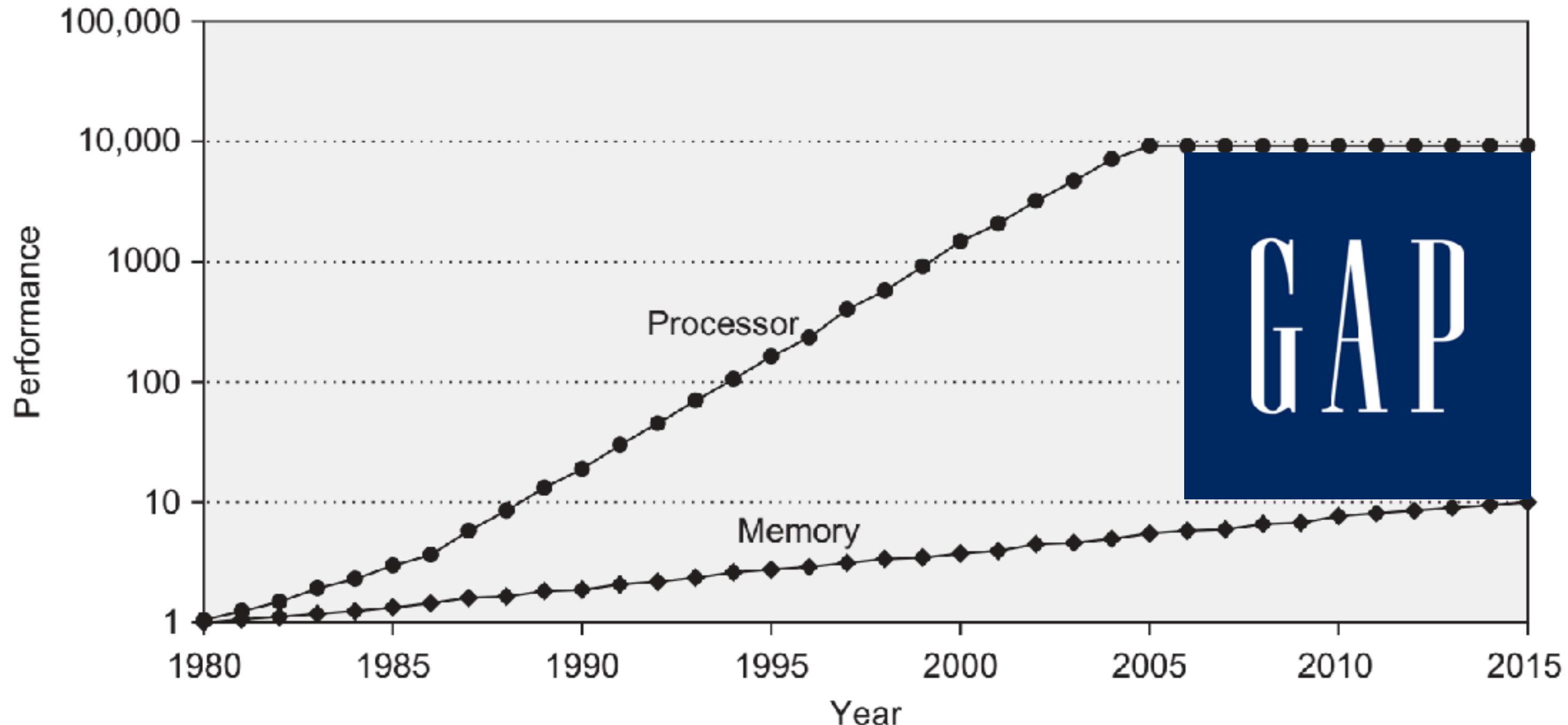
memory
access
arithmetic
operations

Compiler

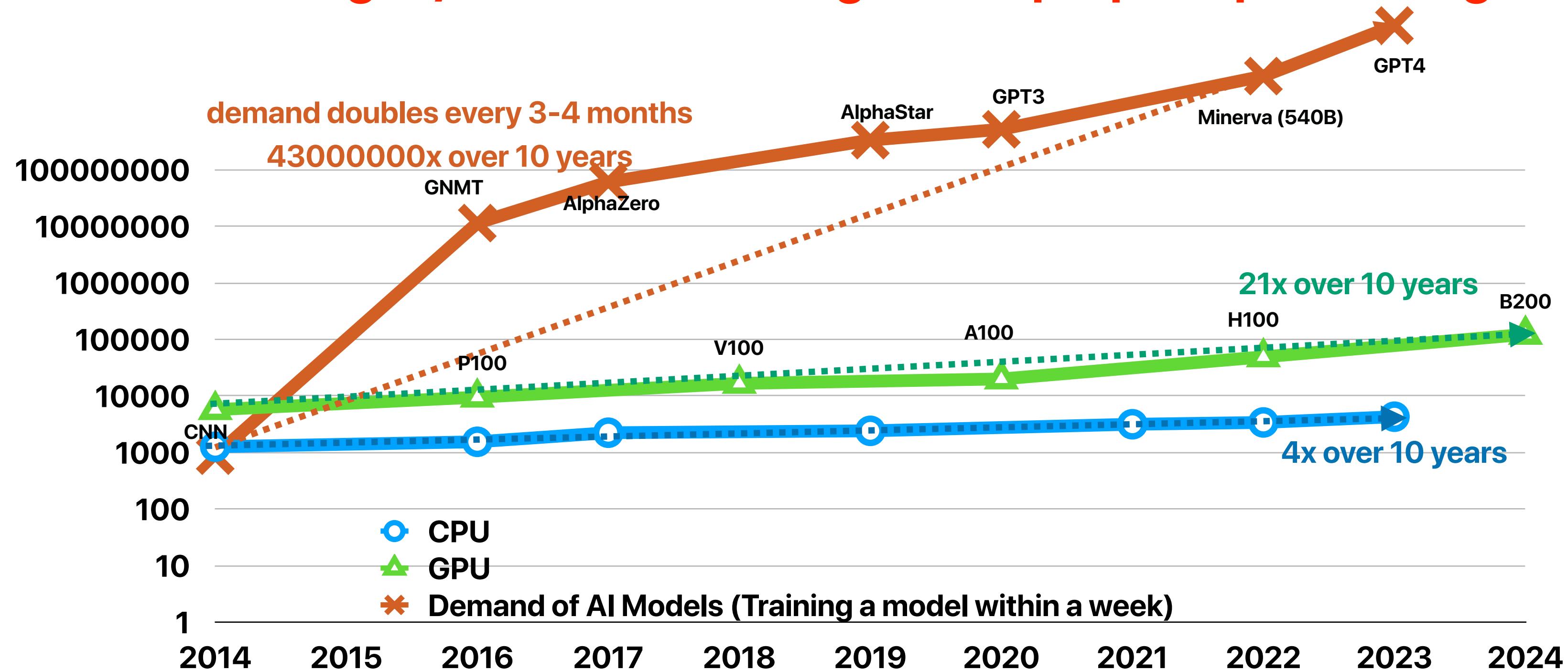
```
main:  
.LFB0:  
    endbr64  
    pushq %rbp  
    movq %rsp, %rbp  
    movl $0, -8(%rbp)  
    movl $0, -4(%rbp)  
    movl $0, -8(%rbp)  
    jmp .L2  
.L2:  
    cmpl $19, -8(%rbp)  
    jle .L3  
    movl $0, %eax  
    popq %rbp  
    ret  
.L3:  
    movl -8(%rbp), %eax  
    cltq  
    leaq 0(%rax,4), %rdx  
    leaq A(%rip), %rax
```

Contents of section .text:
0000 f30f1efa 554889e5 c745f800 000000c7
0010 45fc0000 0000c745 f8000000 00eb1e8b
0020 45f84898 488d1405 00000000 488d0500
0030 0000008b 04020145 fc8345f8 01837df8
0040 137edcb8 00000000 5dc3

Recap: Performance gap between Processor/Memory



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



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

Is UCSD the best university? Why
or why not?

Outline

- Definition of “Performance”
- The performance equation
- Other metrics and are they good?

Best National

Schools in the National University are a full range of undergraduate research producing groundbreaking results.

To unlock full rankings, SAT/ACT scores

SUMMARY ▾



443

Scho

Sc

Loca

Cit

All

Rank

Nat

Best Computer Science Schools

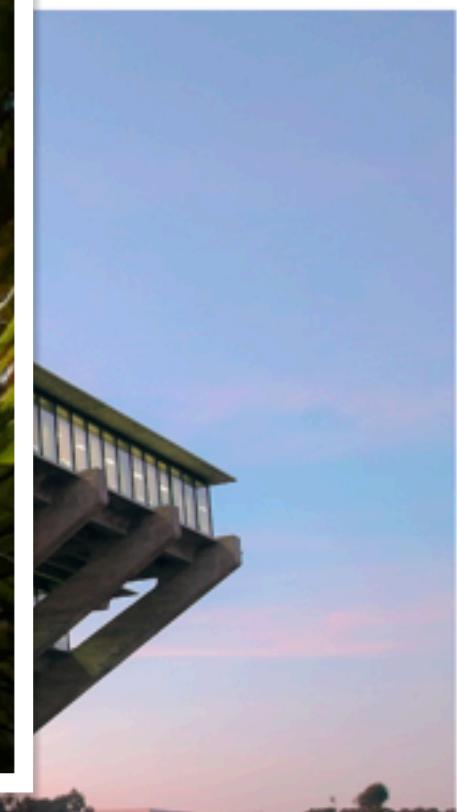
Ranked in 2022, part of Best Science Schools

Earning a graduate degree in computer technology companies and colleges are reflects its average rating on a scale from institutions. [Read the methodology](#) ▾



UC San Diego Ranked No. 1 Public University by Washington Monthly

Campus celebrated as a leader in social mobility, research and public



**What does it really mean by
“better” performance**

ChatGPT

chat.openai.com

ChatGPT 3.5

You
What are the most popular topics in computer science?

ChatGPT

Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

Gemini

gemini.google.com...

Gemini

See the latest updates to the Gemini Apps Preview Hub

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?

Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy Apps

Submit

Peer instruction

- Before the lecture — You need to complete the required **reading**
- During the lecture — I'll bring in activities to ENGAGE you in exploring your understanding of the material
 - Popup questions
 - Individual **thinking** — use polls in Zoom to express your opinion
 - Group **discussion**
 - Breakout rooms based on your residential colleges!
 - Use polls in Zoom to express your group's opinion
 - Whole-classroom **discussion** — we would like to hear from you

Read

Think

Discuss



Now, make sure you login to Poll Everywhere (through the App or the website) with your UCSD email (you're preregistered already)

**Now, you have at least 90 seconds
to answer the question!**



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 resultsA. 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?**

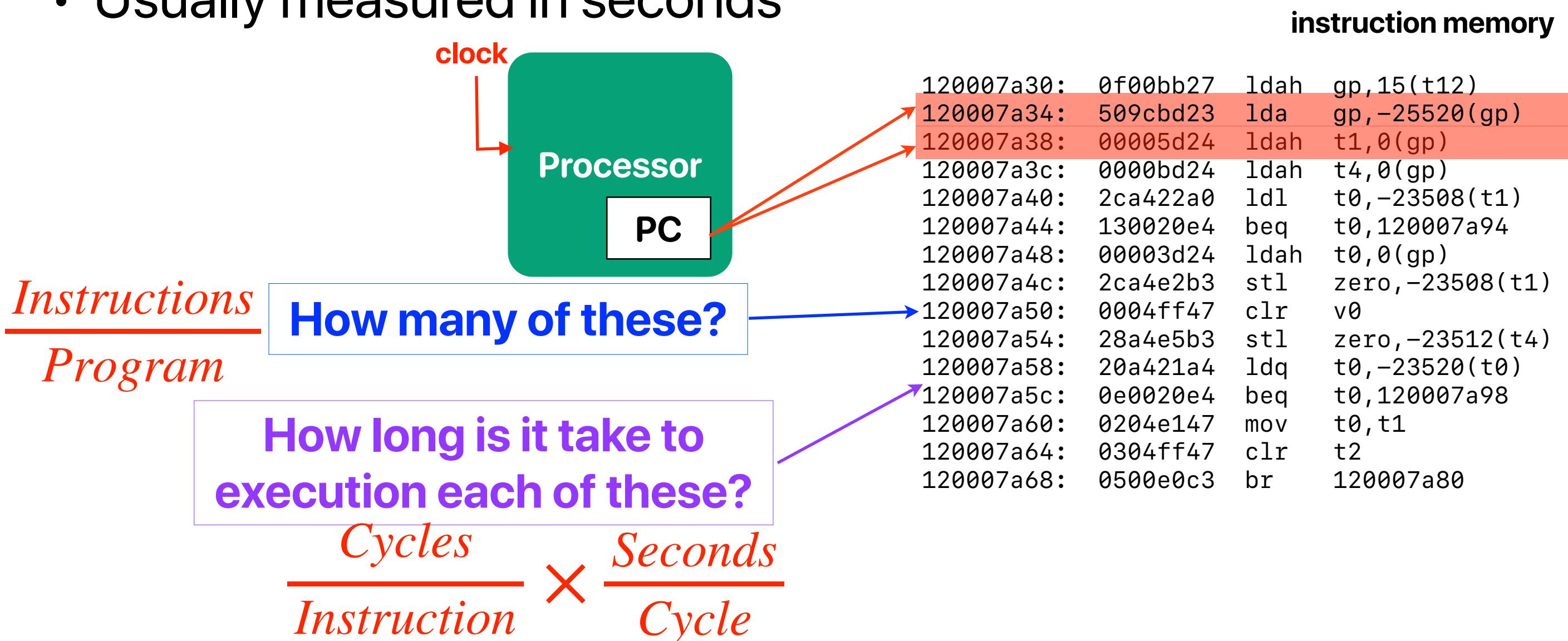


CPU Performance Equation (X)

- Assume that we have an application composed with a total of **5,000,000,000** instructions, in which **20%** of them are “Type-A” instructions with an average **CPI of 4** cycles, **20%** of them are “Type-B” instructions with an average **CPI of 3** cycles and **the rest** instructions are “Type-C” instructions with average **CPI of 1** cycle. If the processor runs at **4 GHz**, how long is the execution time?
 - 1.25 sec
 - 2.5 sec
 - 3.75 sec
 - 7.5 sec
 - 40 sec

Execution Time

- The simplest kind of performance
- Shorter execution time means better performance
- Usually measured in seconds



CPU Performance Equation

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

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

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

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

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

Performance Equation (X)

- Assume that we have an application composed with a total of **5,000,000,000** instructions, in which **20%** of them are “Type-A” instructions with an average **CPI of 4** cycles, **20%** of them are “Type-B” instructions with an average **CPI of 3** cycles and **the rest** instructions are “Type-C” instructions with average **CPI of 1** cycle. If the processor runs at **4 GHz**, how long is the execution time?

A. 1.25 sec

B. 2.5 sec

C. 3.75 sec

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

D. 7.5 sec

$$ET = (5 \times 10^9) \times (20\% \times 4 + 20\% \times 3 + 60\% \times 1) \times \frac{1}{4 \times 10^9} \text{ sec} = 2.5 \text{ sec}$$

E. 40 sec

average CPI



Performance equation (round 2)

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

C	x86
<pre>for(i = 0; i < count; i++) { s += a[i]; }</pre>	<pre>.L3: movslq (%rdi), %rdx addq \$4, %rdi addq %rdx, %tax cmpq %rcx, %rdi jne .L3</pre>

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

Performance equation (round 2)

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

C	x86
<pre>for(i = 0; i < count; i++) { s += a[i]; }</pre>	<pre>.L3: movslq (%rdi), %rdx addq \$4, %rdi addq %rdx, %tax cmpq %rcx, %rdi jne .L3</pre>

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

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

$$ET = (5 \times 10^9) \times (20\% \times 4 + 20\% \times 3 + 60\% \times 1) \times \frac{1}{4 \times 10^9} \text{ sec} = 2.5 \text{ sec}$$

**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

**Choose the right metric — Latency
v.s. Throughput/Bandwidth**



Artificial Intelligence Computing Leadership from NVIDIA

CLOUD & DATA CENTER

PRODUCTS ▾

SOLUTIONS ▾

APPS ▾

FOR DEVELOPERS

TECHNOLOGIES ▾

Tesla V100

AI TRAINING

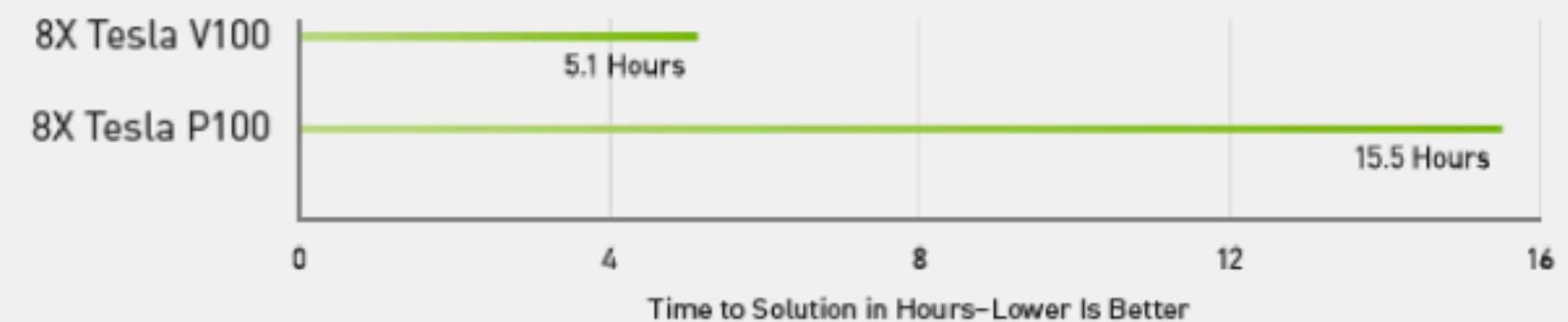
AI INFERENCE

HPC

DATA CENTER GPUs

SPECIFICATIONS

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.

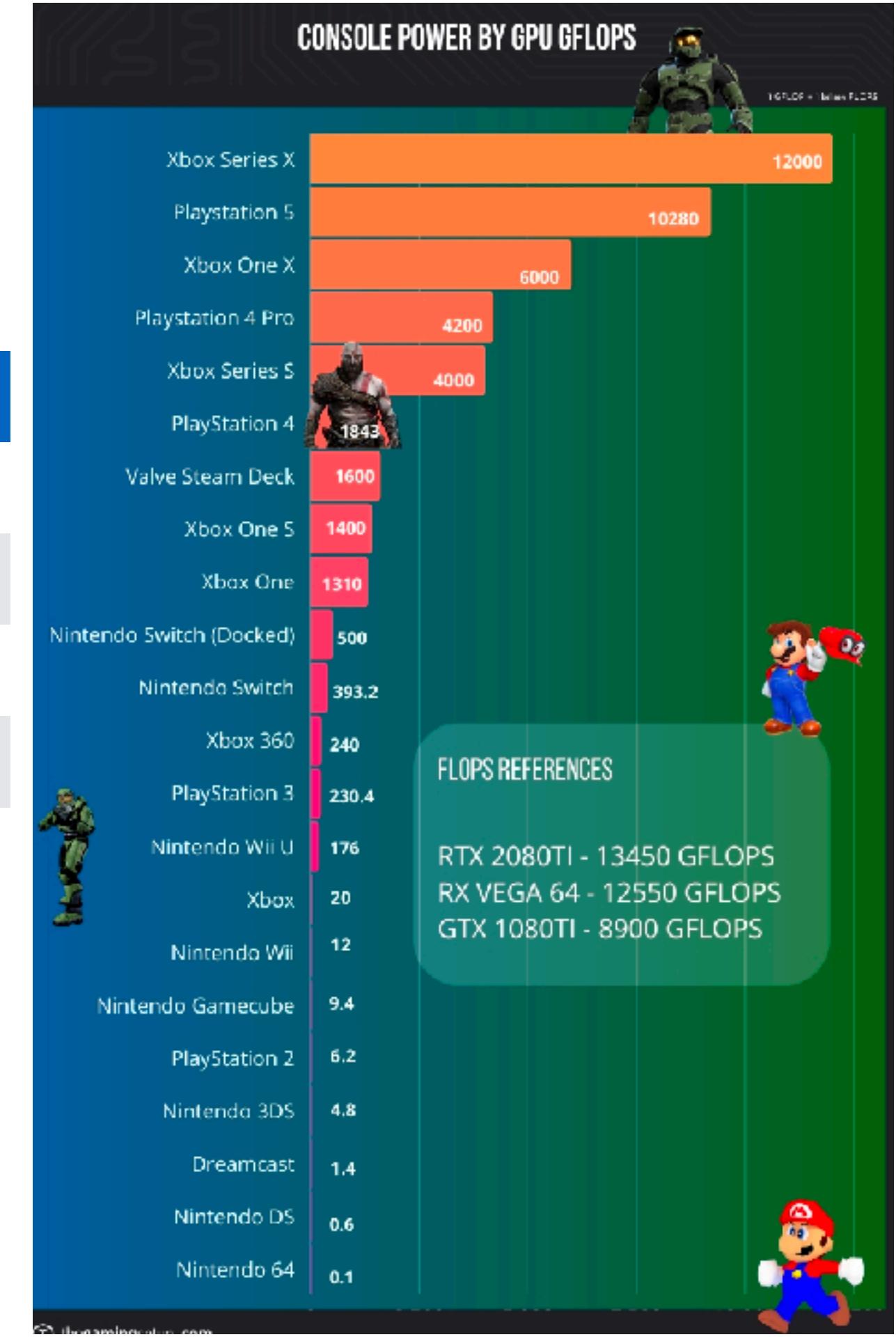
AI TRAINING

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	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



Let's measure the FLOPS 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 $i = j = k = 2048$

$$2^{3 \times 11} \times 2 = 2^{34} \text{ FLOPs in total}$$

$$FLOPS = \frac{2^{34}}{ET_{seconds}}$$



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 data size
 - 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 execution on these two GPUs when executing floating point applications

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



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 data size
 - 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 execution on these two GPUs when executing floating point applications
- A. 0
- B. 1
- C. 2
- D. 3
- E. 4

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

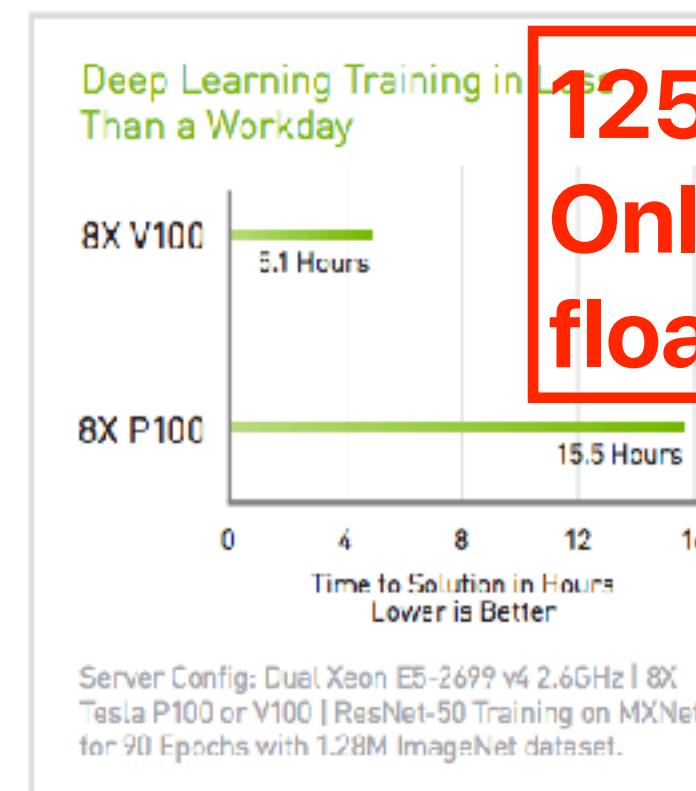
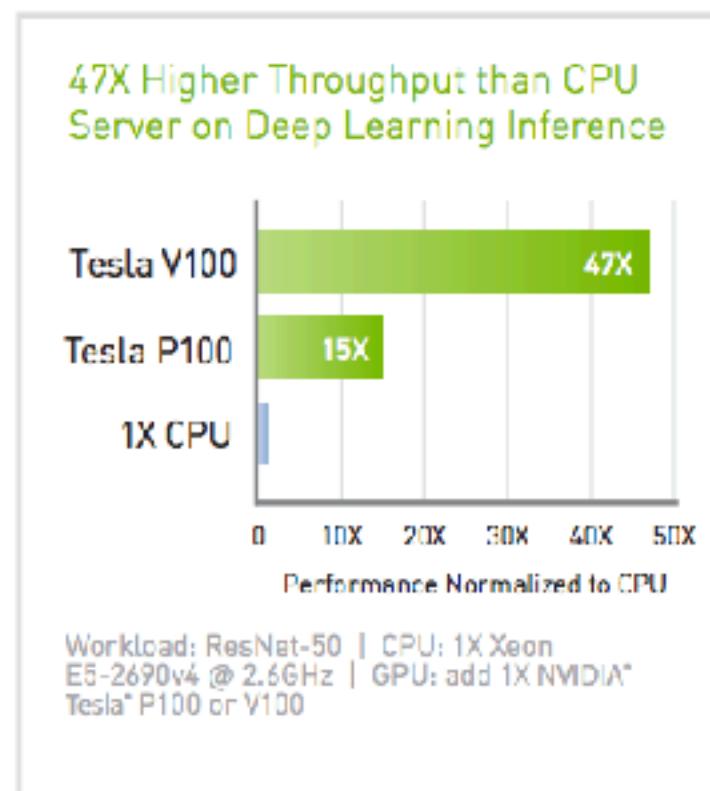
$$\begin{aligned}TFLOPS &= \frac{\# \text{ of floating point instructions} \times 10^{-12}}{\text{Execution Time}} \\&= \frac{IC \times \% \text{ of floating point instructions} \times 10^{-12}}{IC \times CPI \times CT} \\&= \frac{\% \text{ of floating point instructions} \times 10^{-12}}{CPI \times CT}\end{aligned}$$

IC is gone!

- 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?
- Does not make sense if the application is not floating point intensive

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	375W	300W

1 GPU Node Replaces Up To 54 CPU Nodes

Node Replacement: HPC Mixed Workload

Announcement

- Reading quiz due this Thursday before the lecture — we will drop two of your least performing reading quizzes
- Book your practice test with Triton Testing Center on either 8/7, 8/11, and 8/12 through <https://us.prairietest.com/pt/course/10381>
- Assignment #1 released
 - Unlimited submissions, autograded
 - Work/submit early to receive feedback and react early
- Check our website for slides, Gradescope for assignments, piazza for discussions
- Youtube channel for lecture recordings:
<https://www.youtube.com/c/ProfUsagi/playlists>
- Don't forget to check your access to escalab.org/datahub and piazza
- CSE142L lecture today 5p @ WLH 2205

Computer Science & Engineering

142

つづく

