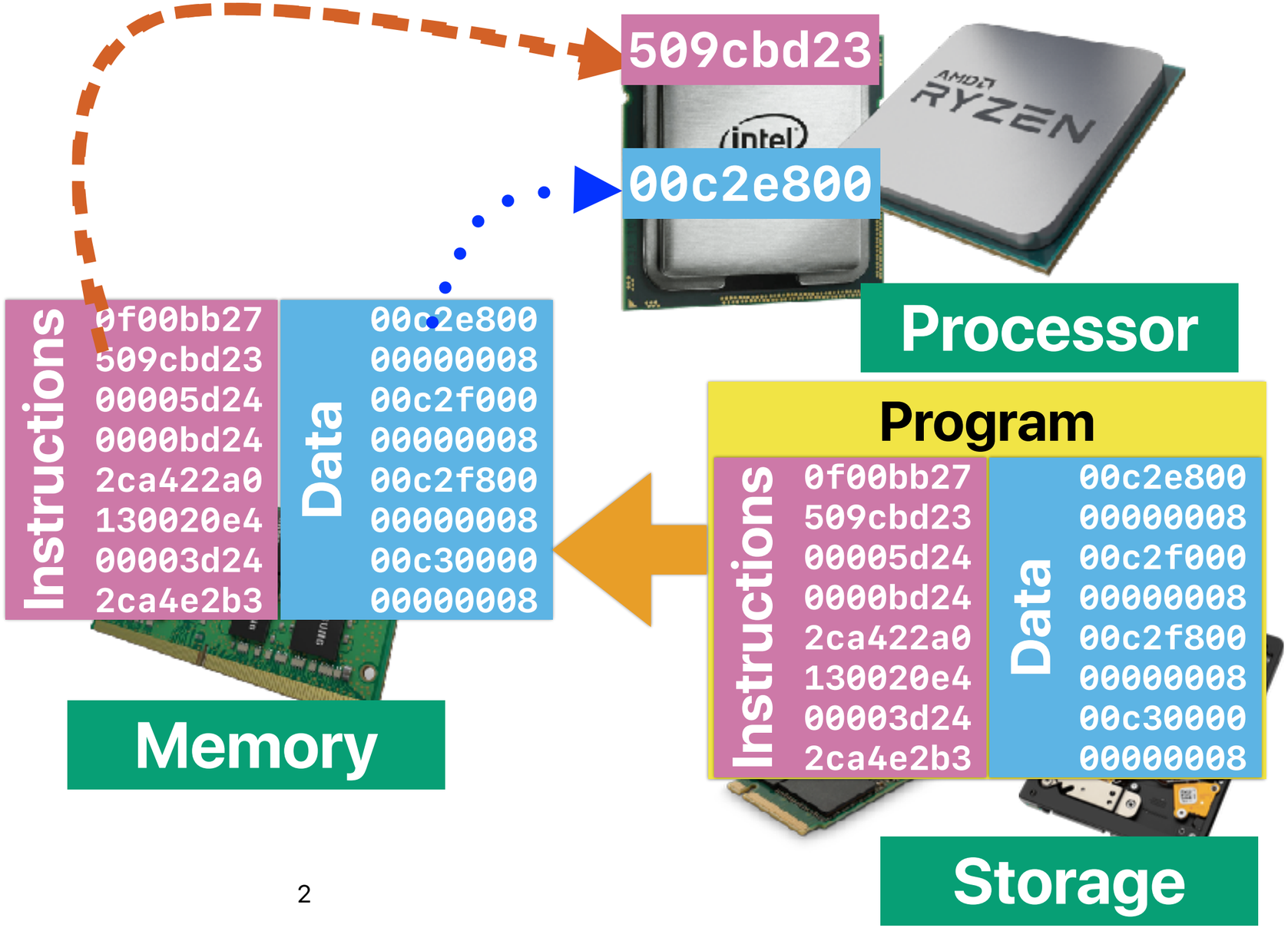
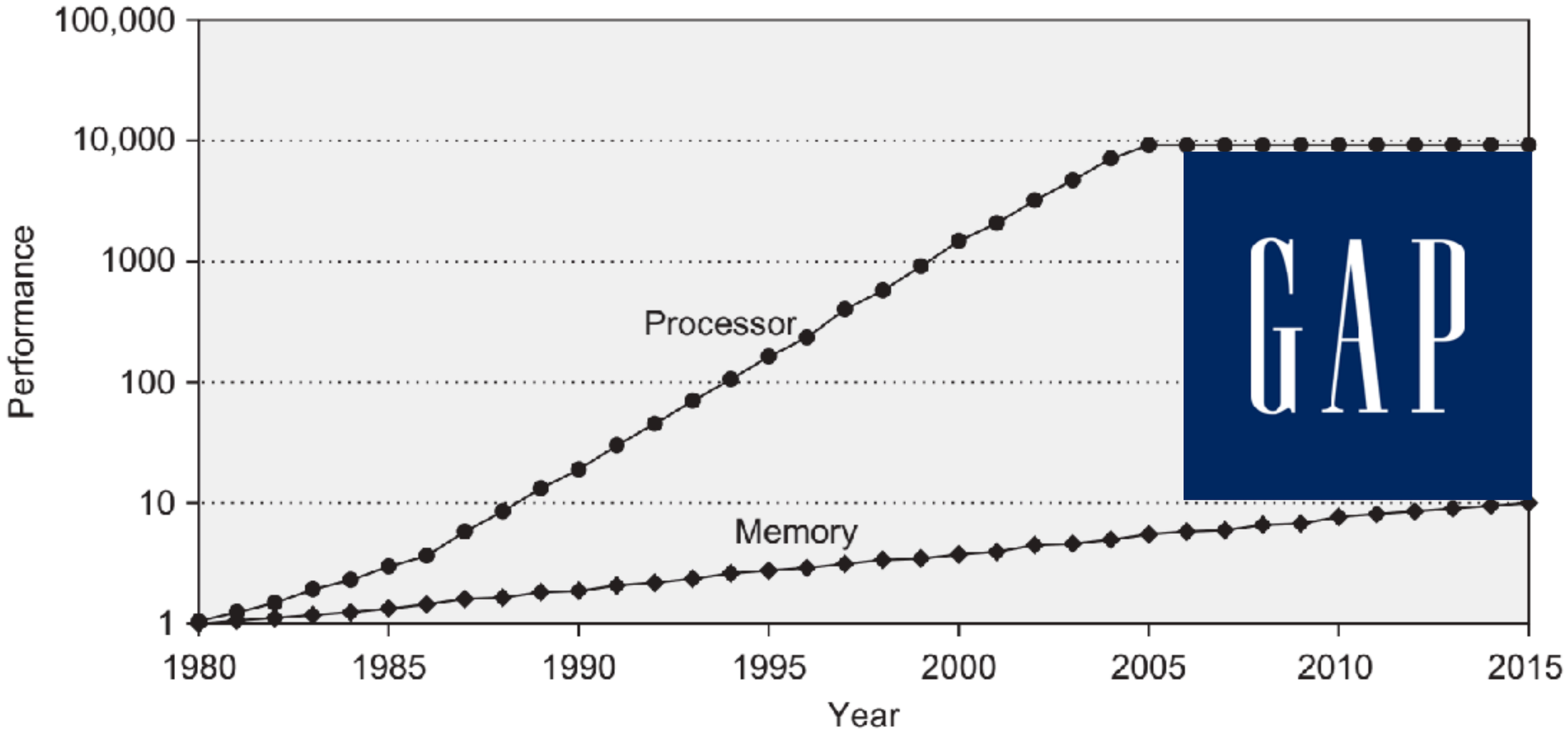
Performance (4) & Memory Hierarchy (1)

Hung-Wei Tseng

von Neumman Architecture



Recap: Performance gap between Processor/Memory



Recap: Summary of CPU Performance Equation

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

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

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

$$Speedup = \frac{Execution\ Time_X}{Execution\ Time_Y}$$

- IC (Instruction Count)
  - ISA, Compiler, algorithm, programming language, **programmer**
- CPI (Cycles Per Instruction)
  - Machine Implementation, microarchitecture, compiler, application, algorithm, programming language, **programmer**
- Cycle Time (Seconds Per Cycle)
  - Process Technology, microarchitecture, **programmer**

Amdahl's Law

$$Speedup_{enhanced}(f, s) = \frac{1}{(1 - f) + \frac{f}{s}}$$

- Corollary #1: Maximum speedup

$$Speedup_{max}(f, \infty) = \frac{1}{(1 - f)}$$

- Corollary #2: Make the common case fast
  - Common case changes all the time

$$Speedup_{max}(f_1, \infty) = \frac{1}{(1 - f_1)}$$

$$Speedup_{max}(f_2, \infty) = \frac{1}{(1 - f_2)}$$

$$Speedup_{max}(f_3, \infty) = \frac{1}{(1 - f_3)}$$

$$Speedup_{max}(f_4, \infty) = \frac{1}{(1 - f_4)}$$

- Corollary #3: Single-core performance still matters

$$Speedup_{parallel}(f_{parallelizable}, \infty) = \frac{1}{(1 - f_{parallelizable})}$$

- Corollary #4: Exploiting more parallelism from a program is the key to performance gain in modern architectures

$$Speedup_{parallel}(f_{parallelizable}, \infty) = \frac{1}{(1 - f_{parallelizable})}$$

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!

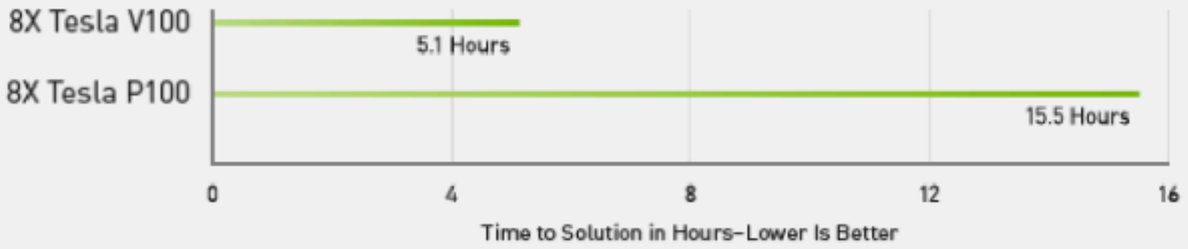
A good performance metric must cover IC, CPI, CT!

- 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

Outline

- How to make “fair comparisons” or … how to fool others with performance metrics
- The Basic Idea behind Memory Hierarchy
- How cache works

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.

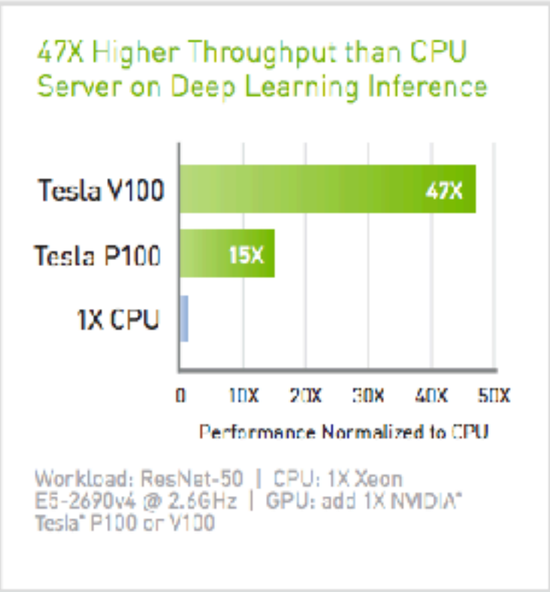
AI TRAINING

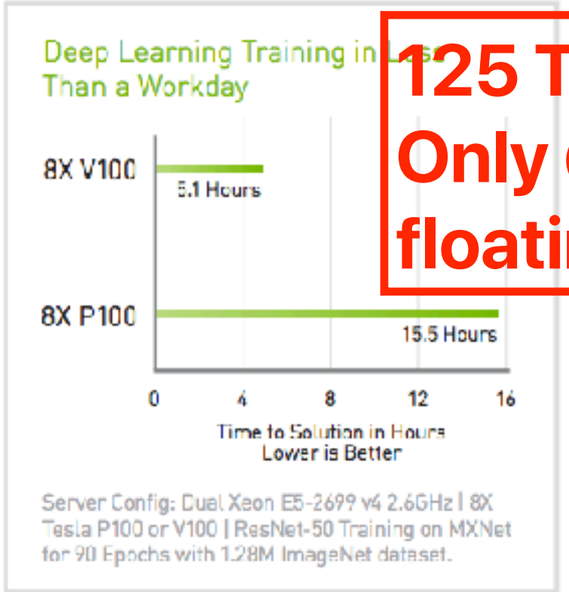
From recognizing speech to training virtual personal assistants and teaching autonomous cars to drive, data scientists are taking on increasingly complex challenges with AI. Solving these kinds of problems requires training deep learning models that are exponentially growing in complexity, in a practical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

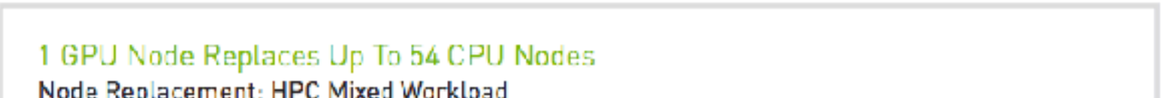
With 640 Tensor Cores, Tesla V100 is the world's first GPU to break the 100 teraFLOPS (TFLOPS) barrier of deep learning performance. The next generation of NVIDIA NVLink™ connects multiple V100 GPUs at up to 300 GB/s to create the world's most powerful computing servers. AI models that would consume 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.

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.

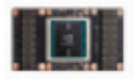






125 TFLOPS Only @ 16-bit floating point

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		

They try to tell it's the better AI hardware

https://blogs.nvidia.com/blog/2017/04/10/ai-drives-rise-accelerated-computing-datacenter/

	K80 2012	TPU 2015	P40 2016
Inferences/Sec <10ms latency	$^1/_{13}$X	1X	2X
Training TOPS	6 FP32	NA	12 FP32
Inference TOPS	6 FP32	90 INT8	48 INT8
On-chip Memory	16 MB	24 MB	11 MB
Power	300W	75W	250W
Bandwidth	320 GB/S	34 GB/S	350 GB/S

Instances per batch

Inference per second

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

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

	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?
- 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].

ImageNet Performance

operation. Ideally, the camera-to-recognition latency per frame should not substantially exceed the inter-frame time of the input images (e.g., 25 milliseconds for a 40 FPS camera).

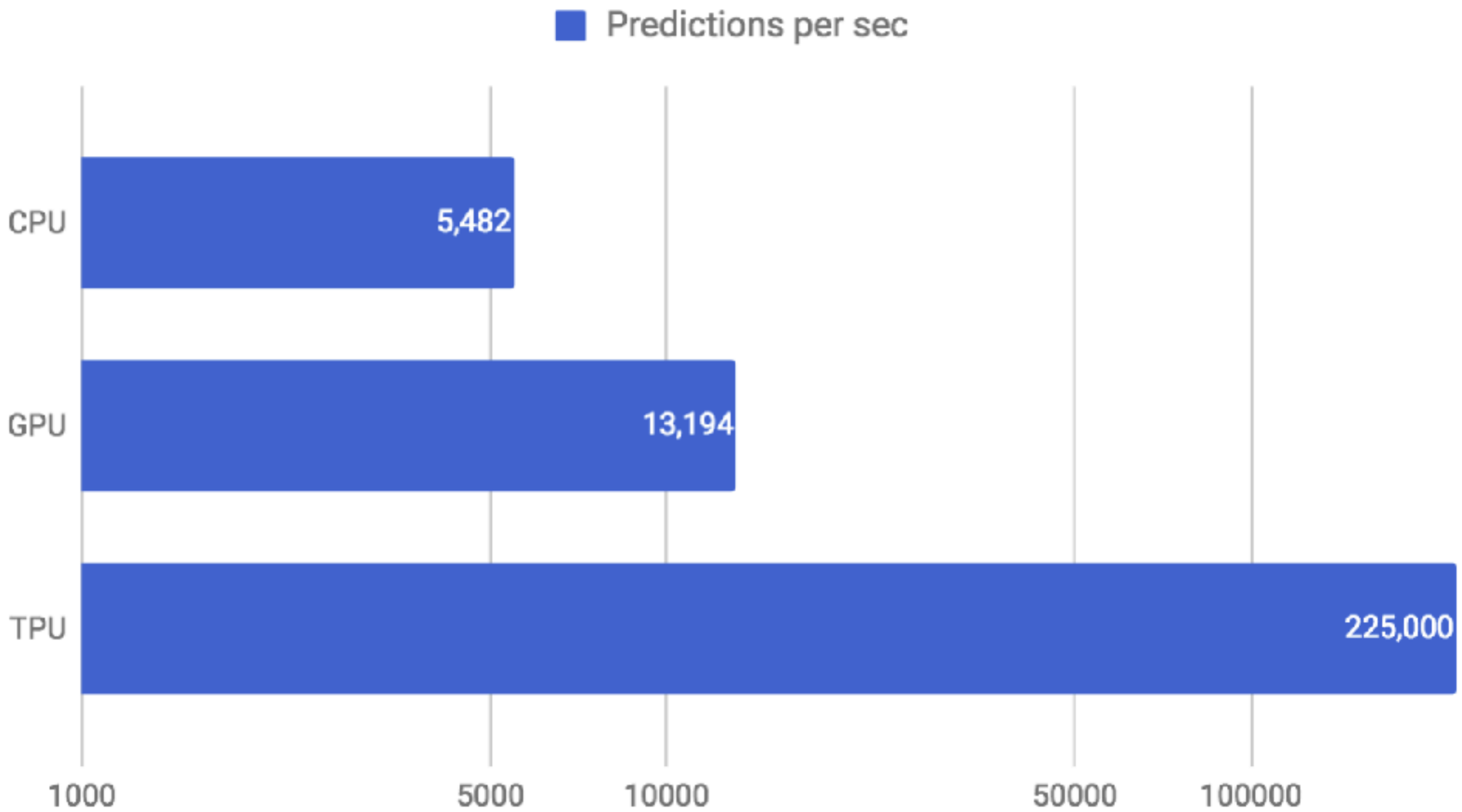
	Intel® Xeon® Platinum 9200 processors (CPU)	Google Cloud TPU v3 (TPU)	NVIDIA/Supermicro 4029GP-TRT-OTO-28 8xT4 (GPU)
Inferences per second	5,965.62	32,716.00	44,977.80
Cores	112 processors * 2-way SMT	2 MXU	320*8 MXU
Number of Maximum Parallel Inferencing Instances	224	2x256	320*8 = 2560

$$\frac{5965.62}{224} = 26.63 \qquad \frac{32716}{2} = 63.8984375 \qquad \frac{44977.8}{2560} = 17.5694531$$

$$\frac{1}{26.63} = 37.55ms \qquad \frac{1}{0.0156498349} = 15.64ms \qquad \frac{1}{17.5694531} = 56.91ms$$

https://mlcommons.org/en/inference-datacenter-05/

IPS with strict 7ms limitation



https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

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

Latency v.s. Bandwidth/Throughput

- Latency — the amount of time to finish an operation
  - Access time
  - Response time
- Throughput — the amount of work can be done within a given period of time
  - Bandwidth (MB/Sec, GB/Sec, Mbps, Gbps)
  - IOPs (I/O operations per second)
  - FLOPs (Floating-point operations per second)
  - IPS (Inferences per second)

NETFLIX

N SERIES

SQUID GAME

Squid Game

2021 | TV-MA | 1 Season | TV Thrillers

Hundreds of cash-strapped players accept a strange invitation to compete in children's games. Inside, a tempting prize awaits — with deadly high stakes.

Starring: Lee Jung-jae, Park Hae-soo, Wi Ha-jun
Creators: Hwang Dong-hyuk

Latency/Delay v.s. Throughput

	Toyota Prius	100 Gb Network
	•100 miles (161 km) from UCSD •75 MPH on highway! •Max load: 374 kg = 2,770 hard drives (2TB per drive)	•100 miles (161 km) from UCSD •Lightspeed! — $3*10^8$m/sec •Max load:4 lanes operating at 25GHz
bandwidth	290GB/sec	100 Gb/s or 12.5GB/sec
total latency	3.5 hours	2 Peta-byte over 167772 seconds = 1.94 Days
latency in getting the first moivie	You see nothing in the first 3.5 hours	100GB/100Gb = 8 secs! You can start watching the first movie in 8 secs!

Extreme Multitasking Performance

- Dual 4K external monitors
- 1080p device display
- 7 applications

Qualcomm Snapdragon is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

What's missing in this video clip?

- The ISA of the "competitor"
- Clock rate, CPU architecture, cache size, how many cores
- How big the RAM?
- How fast the disk?

12 ways to Fool the Masses When Giving Performance Results on Parallel Computers

- Quote only 32-bit performance results, not 64-bit results.
- Present performance figures for an inner kernel, and then represent these figures as the performance of the entire application.
- Quietly employ assembly code and other low-level language constructs.
- Scale up the problem size with the number of processors, but omit any mention of this fact.
- Quote performance results projected to a full system.
- Compare your results against scalar, unoptimized code on Crays.
- When direct run time comparisons are required, compare with an old code on an obsolete system.
- If MFLOPS rates must be quoted, base the operation count on the parallel implementation, not on the best sequential implementation.
- Quote performance in terms of processor utilization, parallel speedups or MFLOPS per dollar.
- Mutilate the algorithm used in the parallel implementation to match the architecture.
- Measure parallel run times on a dedicated system, but measure conventional run times in a busy environment.
- If all else fails, show pretty pictures and animated videos, and don't talk about performance.

Performance of modern DRAM

			Best case access time (no precharge)			Precharge needed
Production year	Chip size	DRAM type	RAS time (ns)	CAS time (ns)	Total (ns)	Total (ns)
2000	256M bit	DDR1	21	21	42	63
2002	512M bit	DDR1	15	15	30	45
2004	1G bit	DDR2	15	15	30	45
2006	2G bit	DDR2	10	10	20	30
2010	4G bit	DDR3	13	13	26	39
2016	8G bit	DDR4	13	13	26	39

Figure 2.4 Capacity and access times for DDR SDRAMs by year of production. Access time is for a random memory word and assumes a new row must be opened. If the row is in a different bank, we assume the bank is precharged; if the row is not open, then a precharge is required, and the access time is longer. As the number of banks has increased, the ability to hide the precharge time has also increased. DDR4 SDRAMs were initially expected in 2014, but did not begin production until early 2016.

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

The impact of "slow" memory

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, consider we have DDR4 and the program is well-behaved that precharge is never necessary — the access latency is simply 26 ns. What's the average CPI (pick the most close one)?
  A. 9
  B. 17
  C. 27
  D. 35
  E. 69

The impact of "slow" memory

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, consider we have DDR4 and the program is well-behaved that precharge is never necessary — the access latency is simply 26 ns. What's the average CPI (pick the most close one)?

  A. 9
  B. 17
  C. 27
  D. 35
  E. 69

$$1 + 100\% \times (52) + 30\% \times 52 = 68.6\ cycles$$

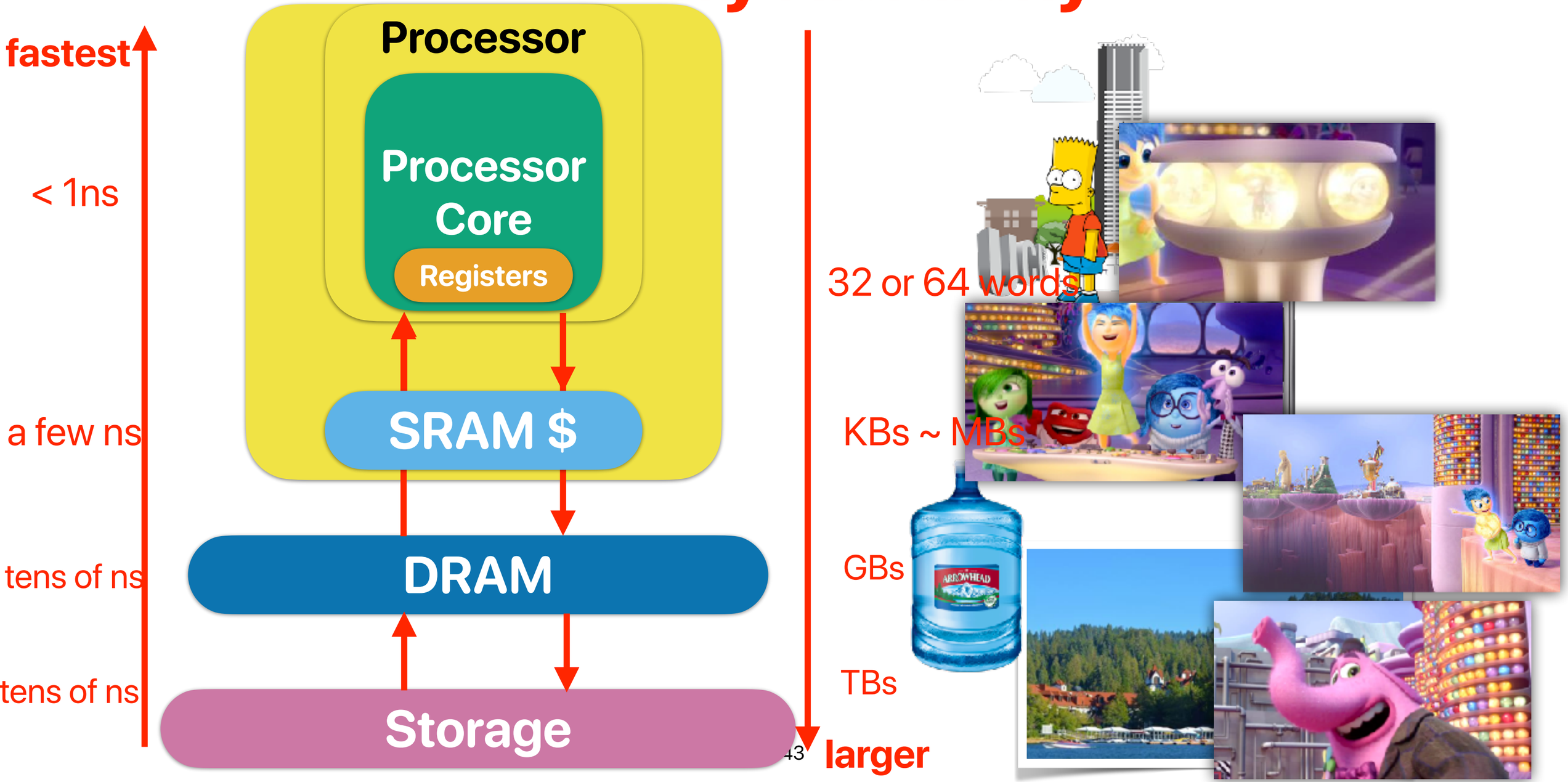
Alternatives?

Memory technology	Typical access time	$ per GiB in 2012
SRAM semiconductor memory	0.5–2.5 ns	$500–$1000
DRAM semiconductor memory	50–70 ns	$10–$20
Flash semiconductor memory	5,000–50,000 ns	$0.75–$1.00
Magnetic disk	5,000,000–20,000,000 ns	$0.05–$0.10

Fast, but expensive $$$


Disney+
Disney · PIXAR
INSIDE OUT
GET DISNEY+
TRAILER
PG 2015 • 1h 35m • Coming of age, Family, Animation
When 11-year-old Riley moves to a new city, her Emotions team up to help her through the transition. Joy, Fear, Anger, Disgust and Sadness work together, but when Joy and Sadness get lost, they must journey through unfamiliar places to get back home.

Memory Hierarchy

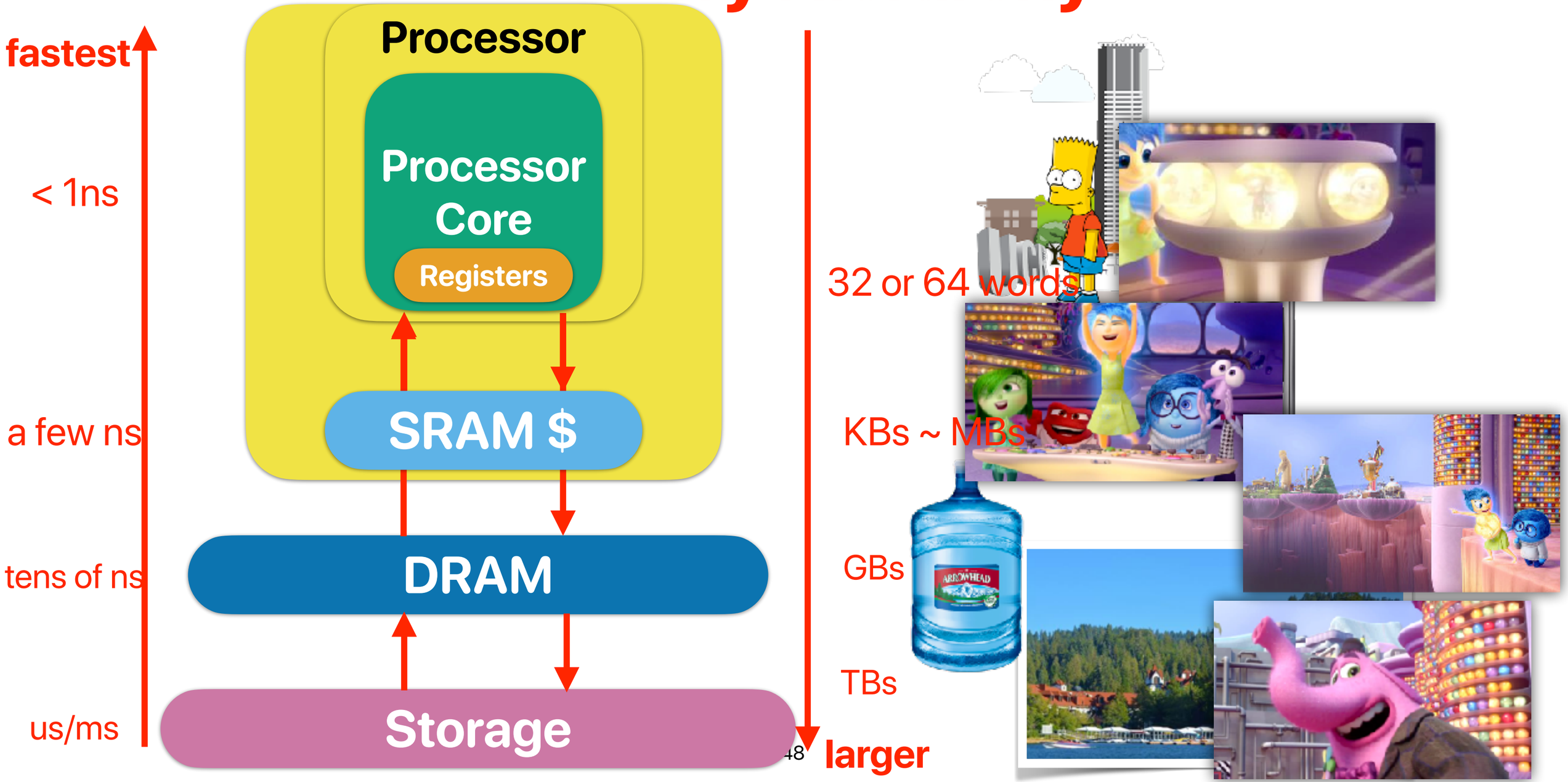


https://www.pollev.com/hungweitseng close in 1:30

How can memory hierarchy help in performance?

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, in addition to DDR4, whose latency 26 ns, we also got an SRAM cache with latency of just at 0.5ns and can capture 90% of the desired data/instructions. what's the average CPI (pick the most close one)?
  A. 2
  B. 4
  C. 8
  D. 16
  E. 32

Memory Hierarchy



How can memory hierarchy help in performance?

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, in addition to DDR4, whose latency 26 ns, we also got an SRAM cache with latency of just at 0.5ns and can capture 90% of the desired data/instructions. what's the average CPI (pick the most close one)?

  A. 2

  B. 4

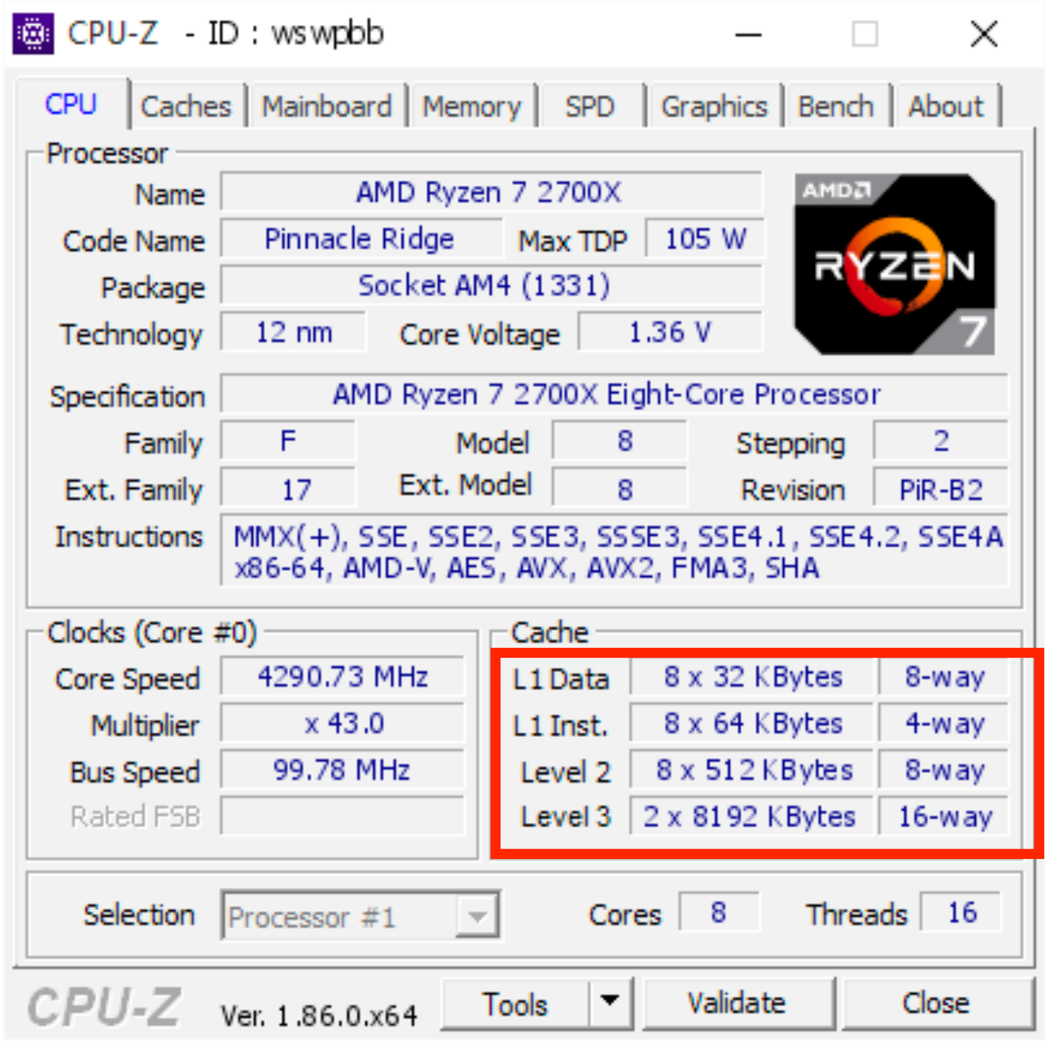
  C. 8

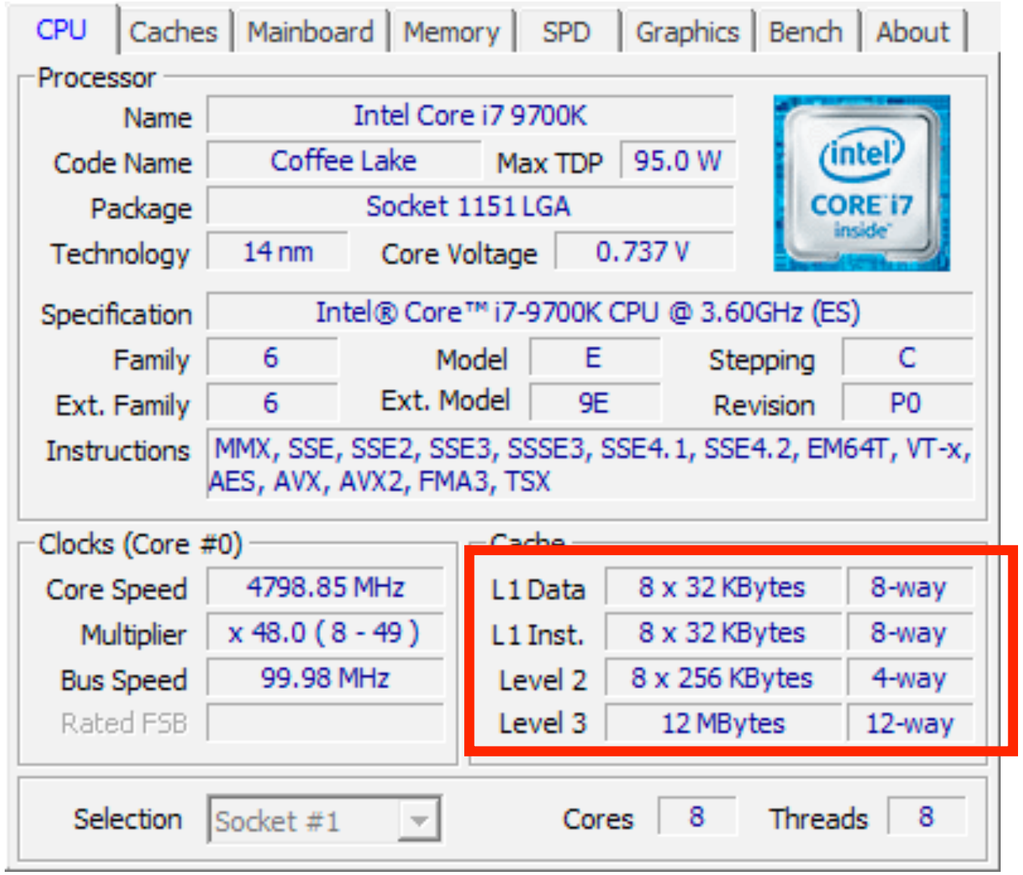
  D. 16

  E. 32

$$1 + (1 - 90\%) \times [100\% \times (52) + 30\% \times 52] = 7.76\ cycles$$

L1? L2? L3?


CPU-Z - ID : wswpbb
CPU
Caches
Mainboard
Memory
SPD
Graphics
Bench
About
Processor
Name
AMD Ryzen 7 2700X
Code Name
Pinnacle Ridge
Max TDP
105 W
Package
Socket AM4 (1331)
Technology
12 nm
Core Voltage
1.36 V
Specification
AMD Ryzen 7 2700X Eight-Core Processor
Family
F
Model
8
Stepping
2
Ext. Family
17
Ext. Model
8
Revision
PiR-B2
Instructions
MMX(+), SSE, SSE2, SSE3, SSSE3, SSE4.1, SSE4.2, SSE4A
x86-64, AMD-V, AES, AVX, AVX2, FMA3, SHA
Clocks (Core #0)
Core Speed
4290.73 MHz
Multiplier
x 43.0
Bus Speed
99.78 MHz
Rated FSB
Cache
L1 Data
8 x 32 KBytes
8-way
L1 Inst.
8 x 64 KBytes
4-way
Level 2
8 x 512 KBytes
8-way
Level 3
2 x 8192 KBytes
16-way
Selection
Processor #1
Cores
8
Threads
16
CPU-Z
Ver. 1.86.0.x64
Tools
Validate
Close


CPU
Caches
Mainboard
Memory
SPD
Graphics
Bench
About
Processor
Name
Intel Core i7 9700K
Code Name
Coffee Lake
Max TDP
95.0 W
Package
Socket 1151 LGA
Technology
14 nm
Core Voltage
0.737 V
Specification
Intel® Core™ i7-9700K CPU @ 3.60GHz (ES)
Family
6
Model
E
Stepping
C
Ext. Family
6
Ext. Model
9E
Revision
P0
Instructions
MMX, SSE, SSE2, SSE3, SSSE3, SSE4.1, SSE4.2, EM64T, VT-x,
AES, AVX, AVX2, FMA3, TSX
Clocks (Core #0)
Core Speed
4798.85 MHz
Multiplier
x 48.0 (8 - 49)
Bus Speed
99.98 MHz
Rated FSB
Cache
L1 Data
8 x 32 KBytes
8-way
L1 Inst.
8 x 32 KBytes
8-way
Level 2
8 x 256 KBytes
4-way
Level 3
12 MBytes
12-way
Selection
Socket #1
Cores
8
Threads
8

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

How can deeper memory hierarchy help in performance?

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, in addition to DDR4, whose latency 26 ns, we also got a 2-level SRAM caches with
  - it's 1st-level one at latency of 0.5ns and can capture 90% of the desired data/ instructions.
  - the 2nd-level at latency of 5ns and can capture 60% of the desired data/instructions

What's the average CPI (pick the most close one)?

A. 2
B. 4
C. 8
D. 16
E. 32

How can deeper memory hierarchy help in performance?

- Assume that we have a processor running @ 2 GHz and a program with 30% of load/store instructions. If the computer has "perfect" memory, the CPI is just 1. Now, in addition to DDR4, whose latency 26 ns, we also got a 2-level SRAM caches with
  - it's 1st-level one at latency of 0.5ns and can capture 90% of the desired data/ instructions.
  - the 2nd-level at latency of 5ns and can capture 60% of the desired data/instructions

  What's the average CPI (pick the most close one)?

  A. 2

  B. 4 $1 + (1 - 90\%) \times [10 + (1 - 60\%) \times 52 + 30\% \times (10 + (1 - 60\%) \times 52)] = 5\ cycles$

  C. 8

  D. 16

  E. 32

Memory Hierarchy

fastest

< 1ns

a few ns

tens of ns

us/ms

Processor

Processor Core

Registers

SRAM $

DRAM

Storage

fastest

L1 $

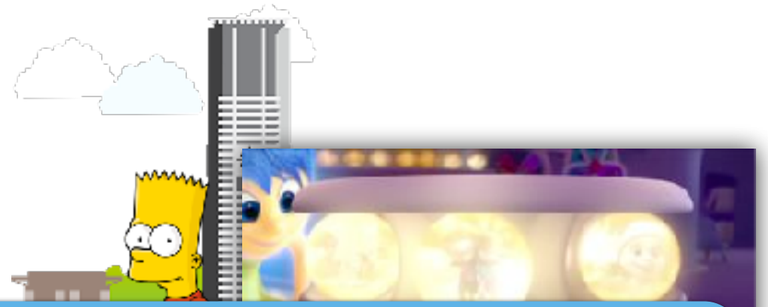
L2 $

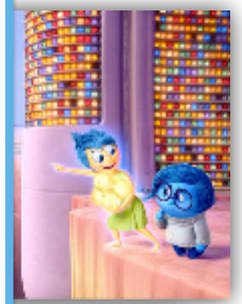
L3 $

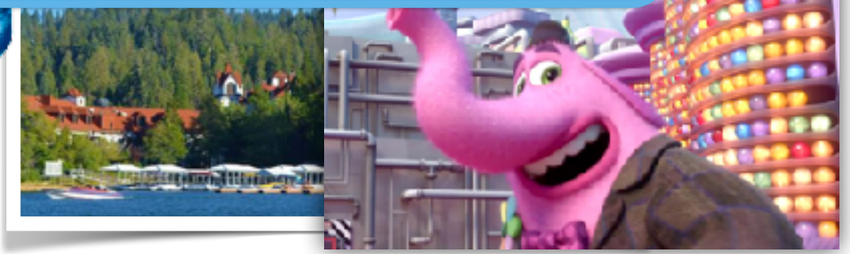
larger

TBs

larger







Why adding small SRAMs would work?

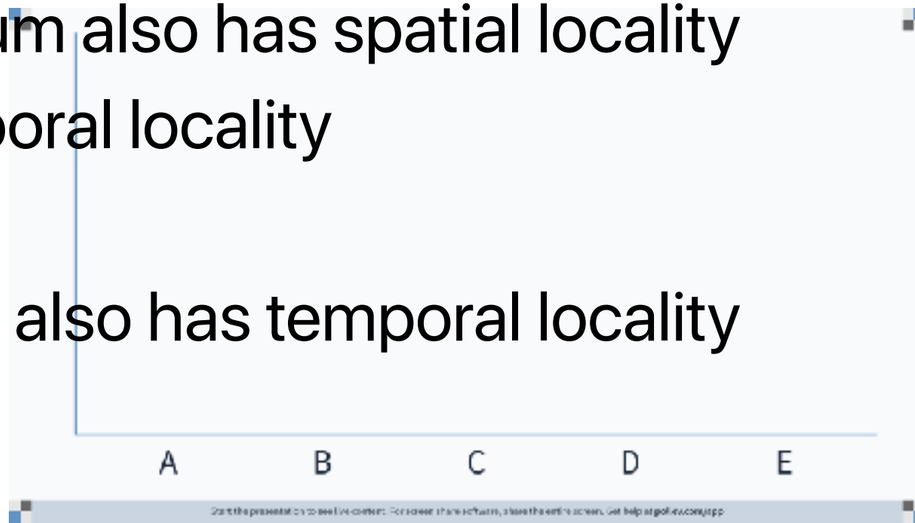
https://www.pollev.com/hungweitseng close in 1:30

Locality

- Which description about locality of arrays `sum` and `A` in the following code is the most accurate?

```
for(i = 0; i< 100000; i++)
{
    sum[i%10] += A[i];
}
```

A. Access of `A` has temporal locality, `sum` has spatial locality
B. Both `A` and `sum` have temporal locality, and sum also has spatial locality
C. Access of `A` has spatial locality, `sum` has temporal locality
D. Both `A` and `sum` have spatial locality
E. Both `A` and `sum` have spatial locality, and sum also has temporal locality



Locality

- Which description about locality of arrays `sum` and A in the following code is the most accurate?

```
for(i = 0; i< 100000; i++)
{
    sum[i%10] += A[i];
}
```

spatial locality:
A[0], A[1], A[2], A[3], ...
sum[0], sum[1], ... , sum[9]
temporal locality:
reuse of sum[0], sum[1], ... , sum[9]

A. Access of A has temporal locality, `sum` has spatial locality
B. Both A and `sum` have temporal locality, and sum also has spatial locality
C. Access of A has spatial locality, `sum` has temporal locality
D. Both A and `sum` have spatial locality
E. Both A and `sum` have spatial locality, and sum also has temporal locality

Locality

- Spatial locality — application tends to visit nearby stuffs in the memory
  - Code — the current instruction, and then PC + 4
  - Data — the current data, and then the next one
- Temporal locality — application revisit the same thing again and again
  - Code — loops, frequently invoked functions
  - Data — the same data can be read/write many times

Most of time, your program is just visiting a very small amount of data/instructions within a given window

Announcement

- Assignment #1 due this Wednesday
  - Assignments SHOULD BE done/submitted individually — if discussed with others, make sure their names on your submission
  - We will drop your least performing assignment as well
- Office Hours
  - Walk-in, no appointment is necessary
  - Hung-Wei/Prof. Usagi: MTu 2p-3p (WCH 406 or on Zoom)
  - Abenezer Wudenhe: WTh 3p-4p (Zoom only)

Computer
Science &
Engineering

203

つづく