Lecture 18:

Heterogeneous Parallelism and Hardware Specialization

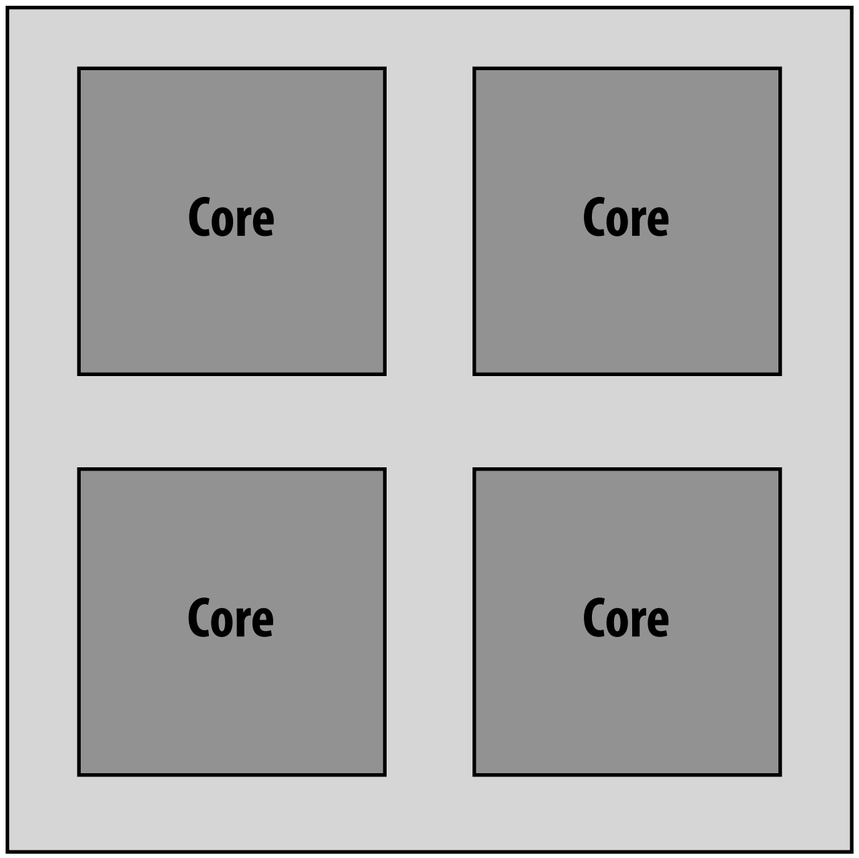
Parallel Computer Architecture and Programming
CMU 15-418/15-618, Fall 2024

Learning Objectives

- Describe the heterogeneous characteristics of parallel programs
- Explain how hardware design exploits parallel program characteristics
- Identify the benefits from heterogeneous execution
- Analyze shortcomings in Amdahl's Law based calculations

You need to buy a new computer…

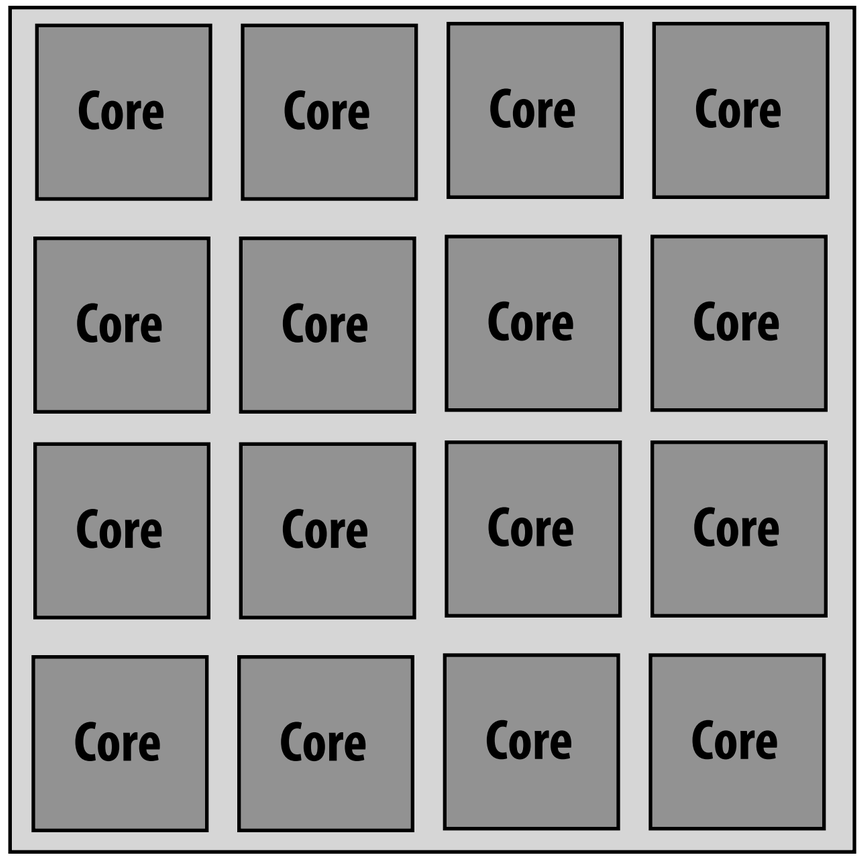
You need to buy a computer system



Processor A

4 cores

Each core has sequential performance P



Processor B

16 cores

Each core has sequential performance P/2

All other components of the system are equal.

Which do you pick?

Amdahl's law revisited

$$\text{speedup}(f, n) = \frac{1}{(1 - f) + \frac{f}{n}}$$

f = **fraction of program that is parallelizable**

n = **parallel processors**

Assumptions:

Parallelizable work distributes perfectly onto n processors of equal capability

Rewrite Amdahl's law in terms of resource limits

$$\text{speedup}(f, n, r) = \frac{1}{\frac{1-f}{\text{perf}(r)} + \frac{f}{\text{perf}(r) \cdot \frac{n}{r}}}$$

Relative to processor with 1 unit of resources, n=1.
Assume perf(1) = 1

f = **fraction of program that is parallelizable**

n = **total processing resources (e.g., transistors on a chip)**

r = **resources dedicated to each processing core,**

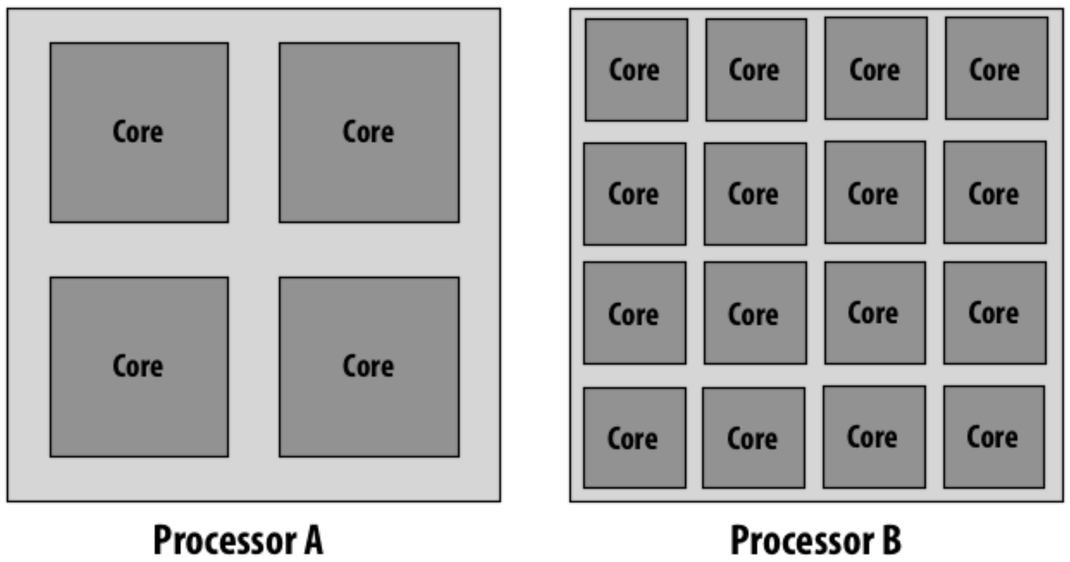
each of the n/r cores has sequential performance $perf(r)$

More general form of Amdahl's Law in terms of f, n, r

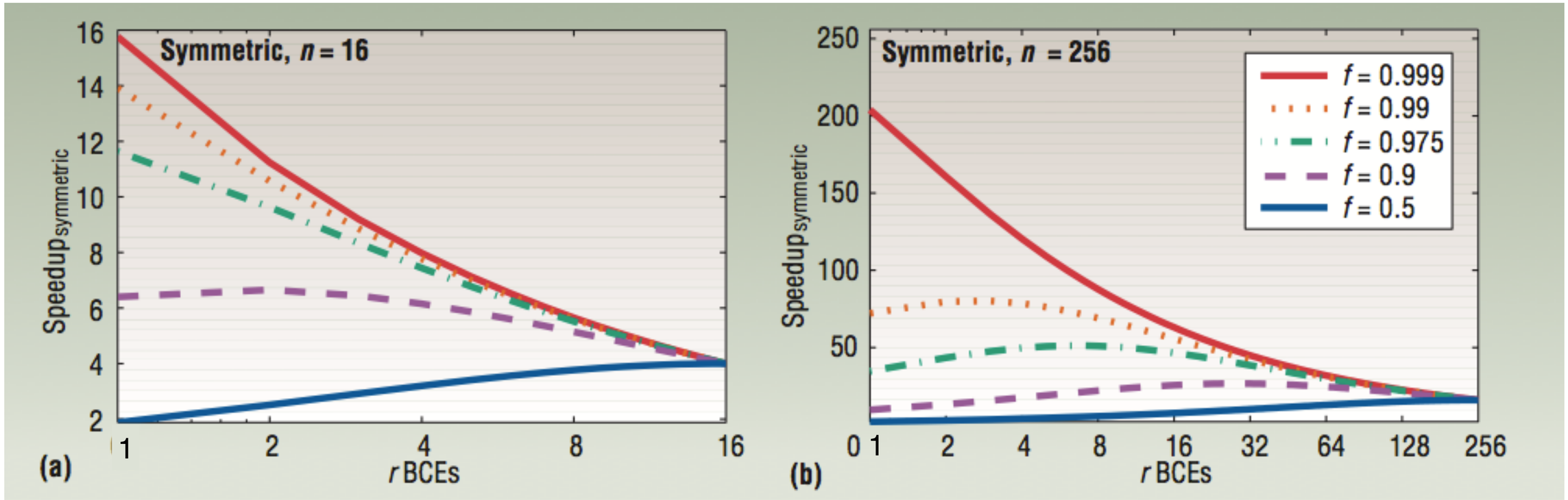
Two examples where n=16

$r_A = 4$

$r_B = 1$



Speedup (relative to n=1)



Up to 16 cores (n=16) Up to 256 cores (n=256)

X-axis = r (chip with many small cores to left, fewer "fatter" cores to right)

Each line corresponds to a different workload

Each graph plots performance as resource allocation changes, but total chip resources kept the same (constant n per graph)

$perf(r)$ modeled as $\sqrt{r}$

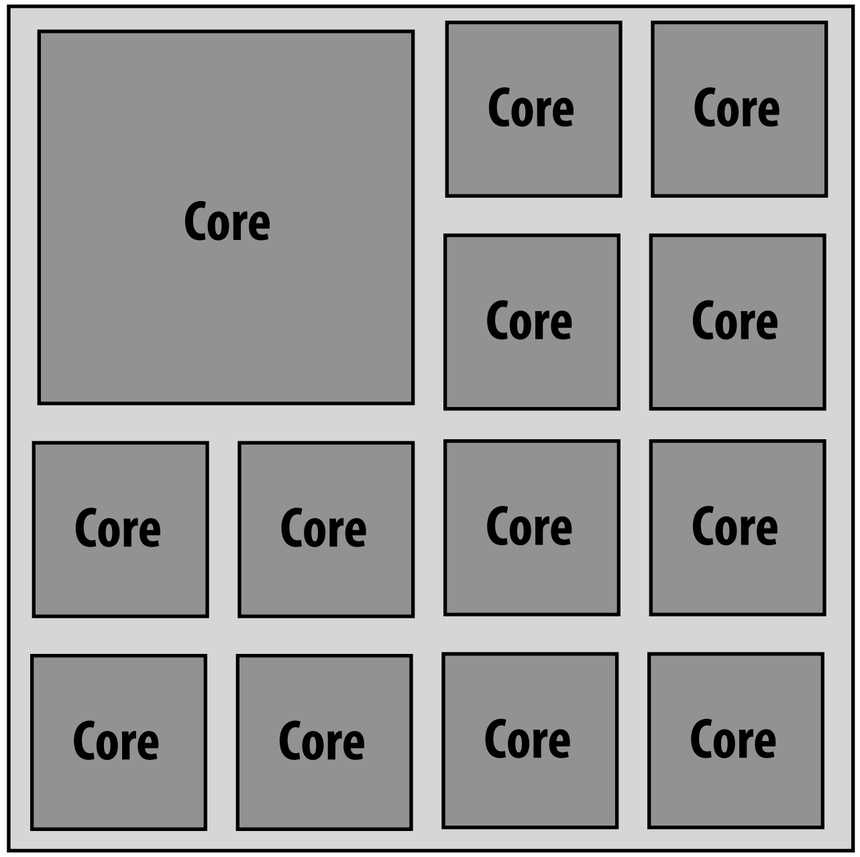
[Figure credit: Hill and Marty 08]

Asymmetric set of processing cores

Example: n=16

One core: $r = 4$

Other 12 cores: $r = 1$



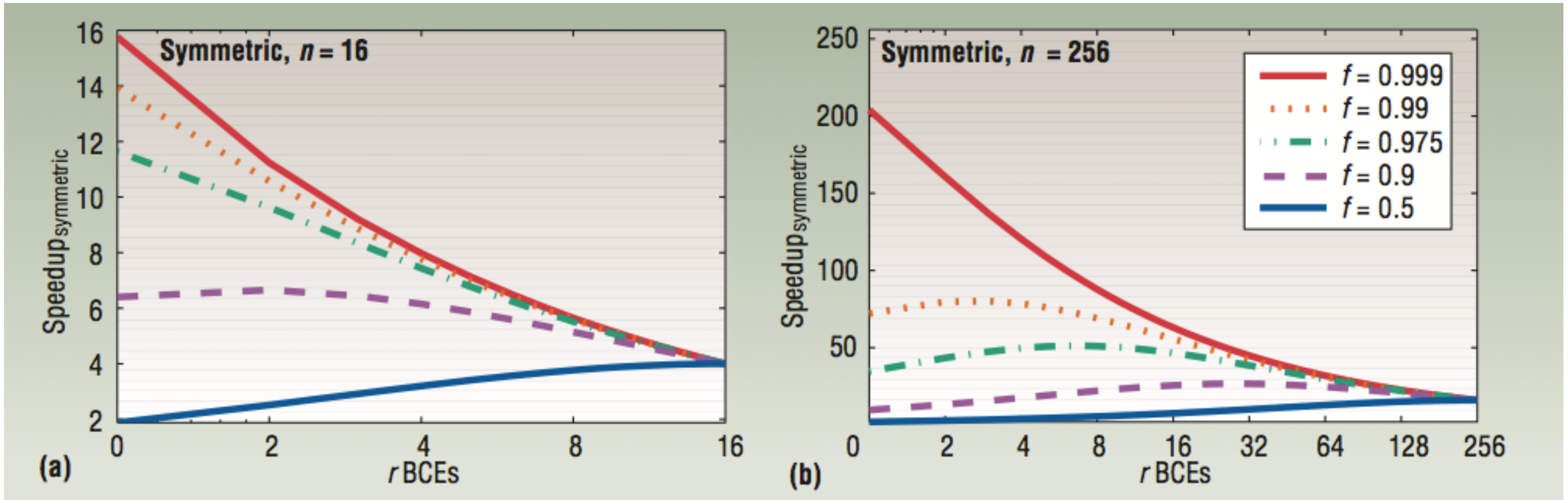
$$\text{speedup}(f, n, r) = \frac{1}{\frac{1-f}{\text{perf}(r)} + \frac{f}{\text{perf}(r)+(n-r)}}$$

(of heterogeneous processor with *n* recourses, relative to uniprocessor with one unit worth of resources, n=1)

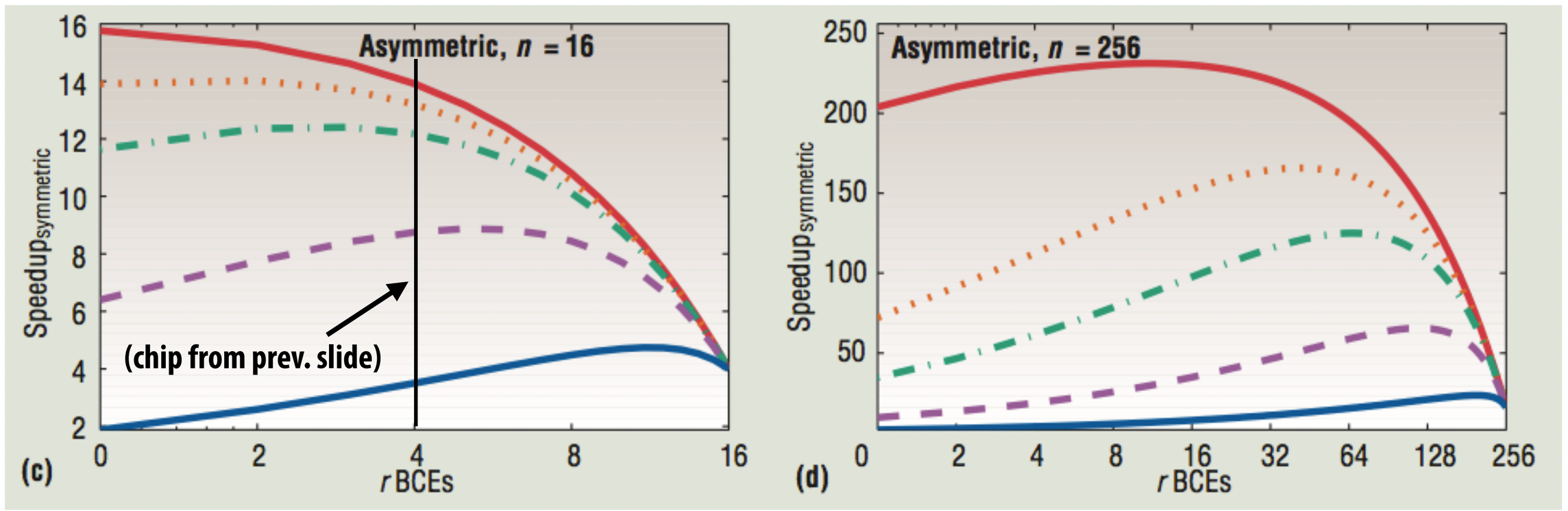
one perf(r) processor + (n-r) perf(1)=1 processors

Speedup (relative to n=1)

[Source: Hill and Marty 08]



X-axis for symmetric architectures gives r for all cores (many small cores to left, few "fat" cores to right)



X-axis for asymmetric architectures gives r for the single "fat" core (assume rest of cores are $r = 1$)

Heterogeneous processing

Observation: most "real world" applications have complex workload characteristics *

They have components that can be widely parallelized.

And components that are difficult to parallelize.

They have components that are amenable to wide SIMD execution.

And components that are not. (divergent control flow)

They have components with predictable data access

And components with unpredictable access, but those accesses might cache well.

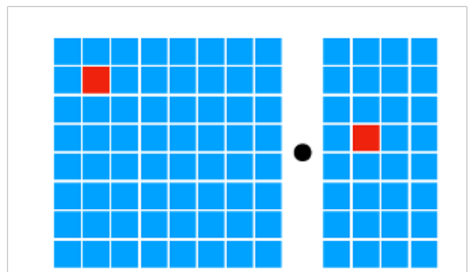
Idea: the most efficient processor is a heterogeneous mixture of resources ("use the most efficient tool for the job")

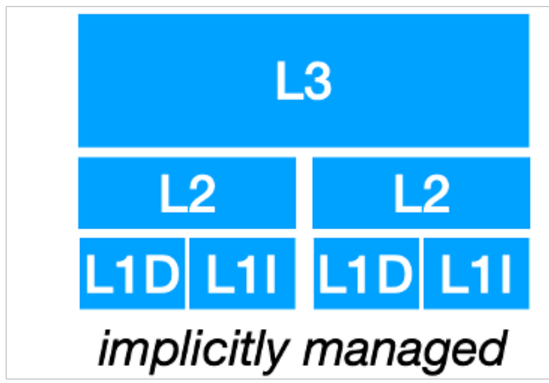
* You will likely make a similar observation during your projects

CPUs, GPUs, TPUs are all heterogeneous processors

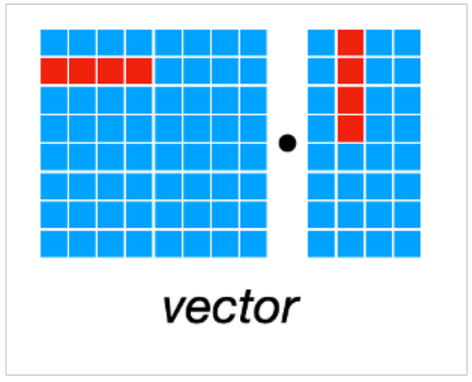
Compute Primitives | **Memory Hierarchy**

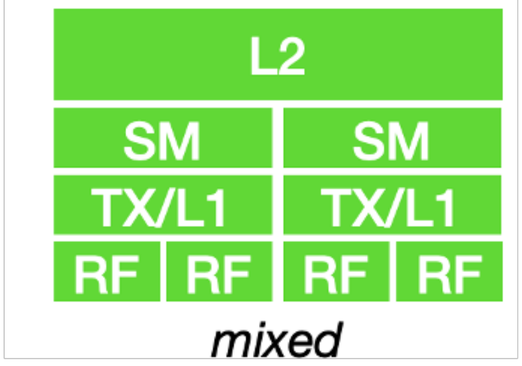
CPUs



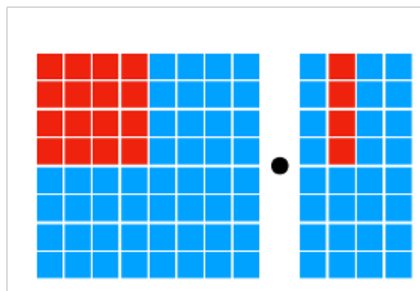


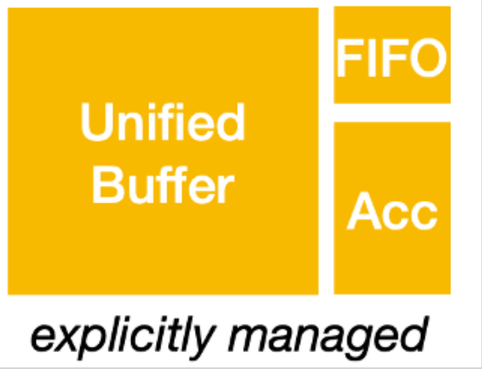
GPUs





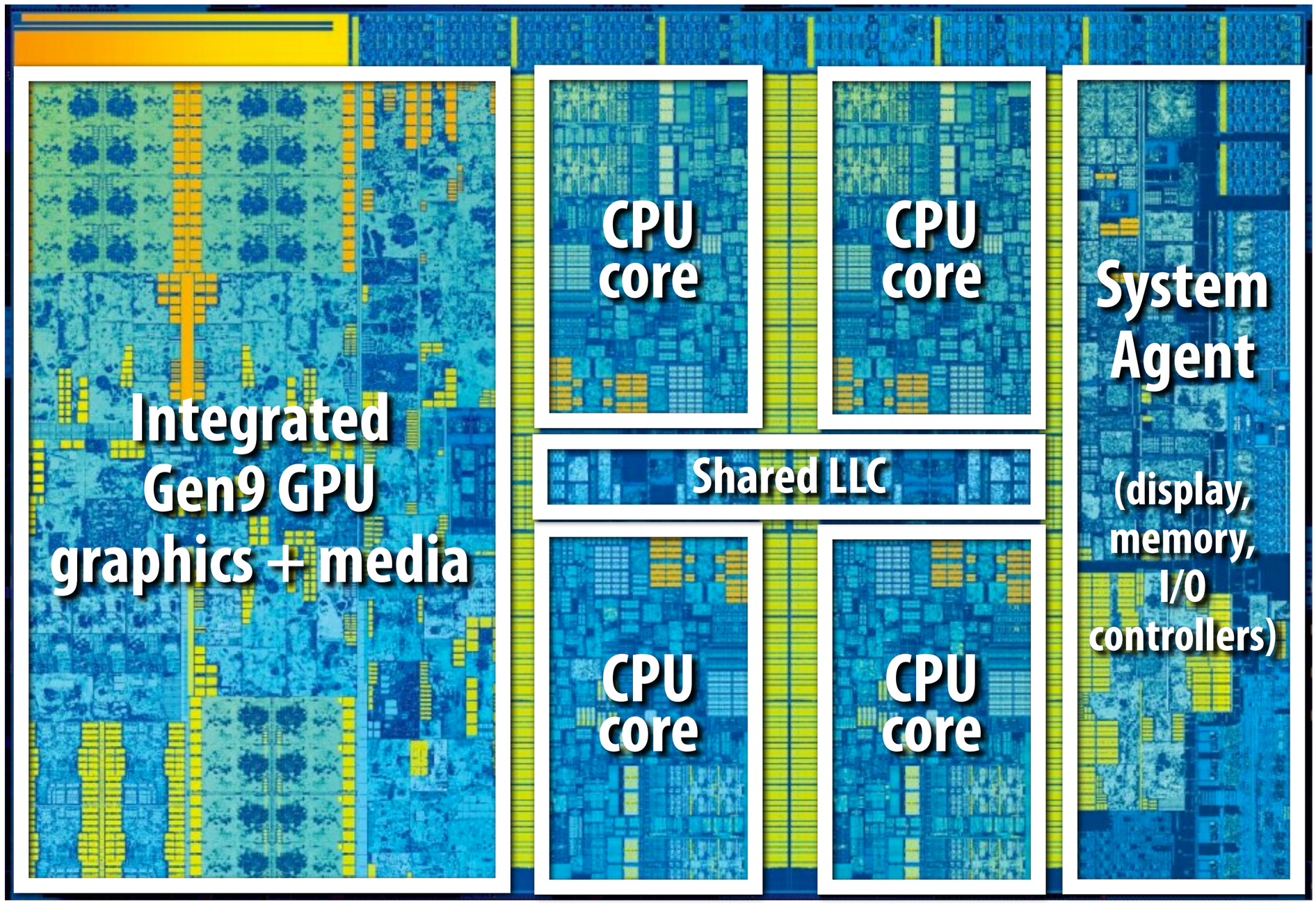
TPUs





Intel “Skylake" (2015)

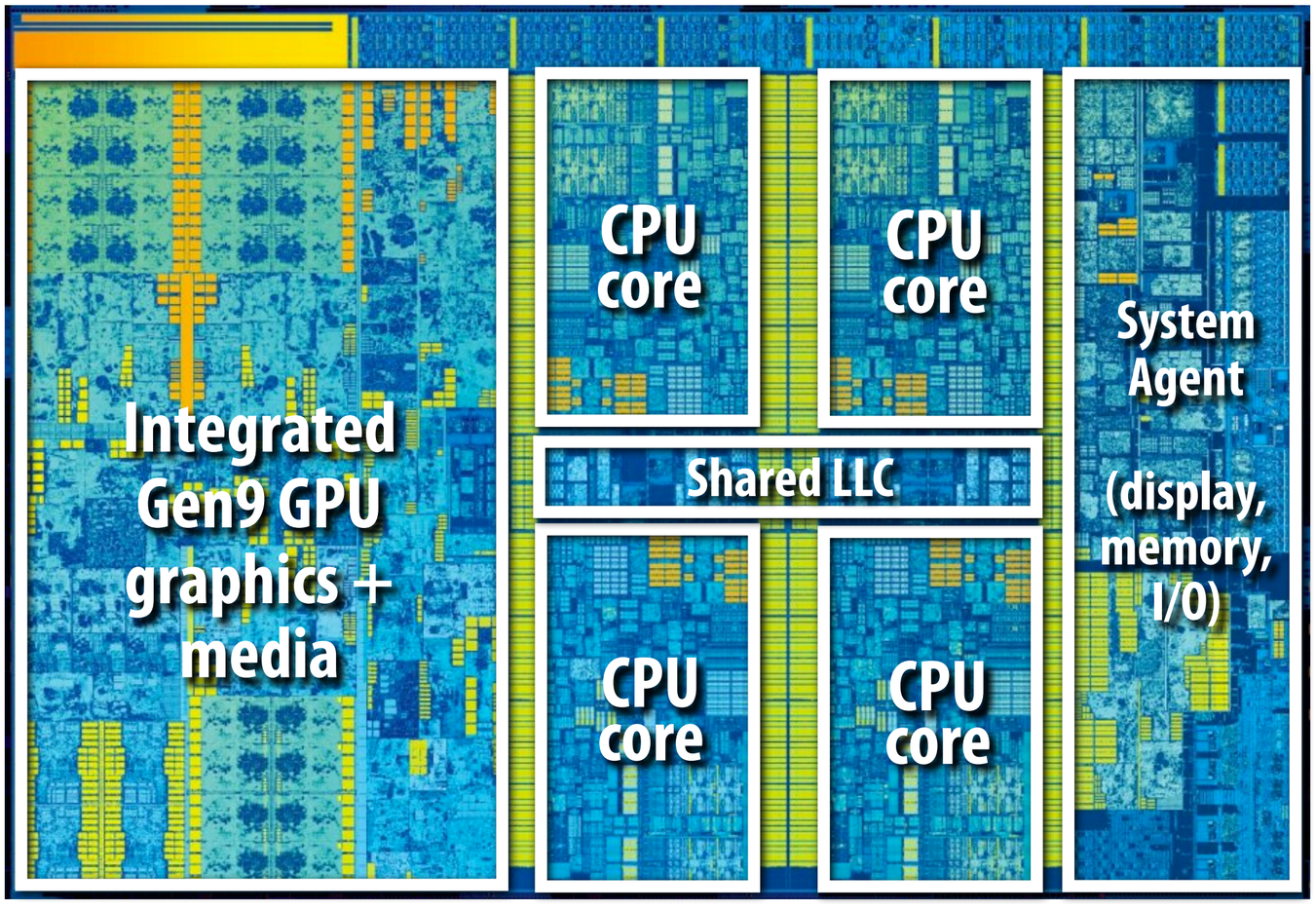
(6th Generation Core i7 architecture)



4 CPU cores + graphics cores + media accelerators

Intel “Skylake" (2015)

(6th Generation Core i7 architecture)

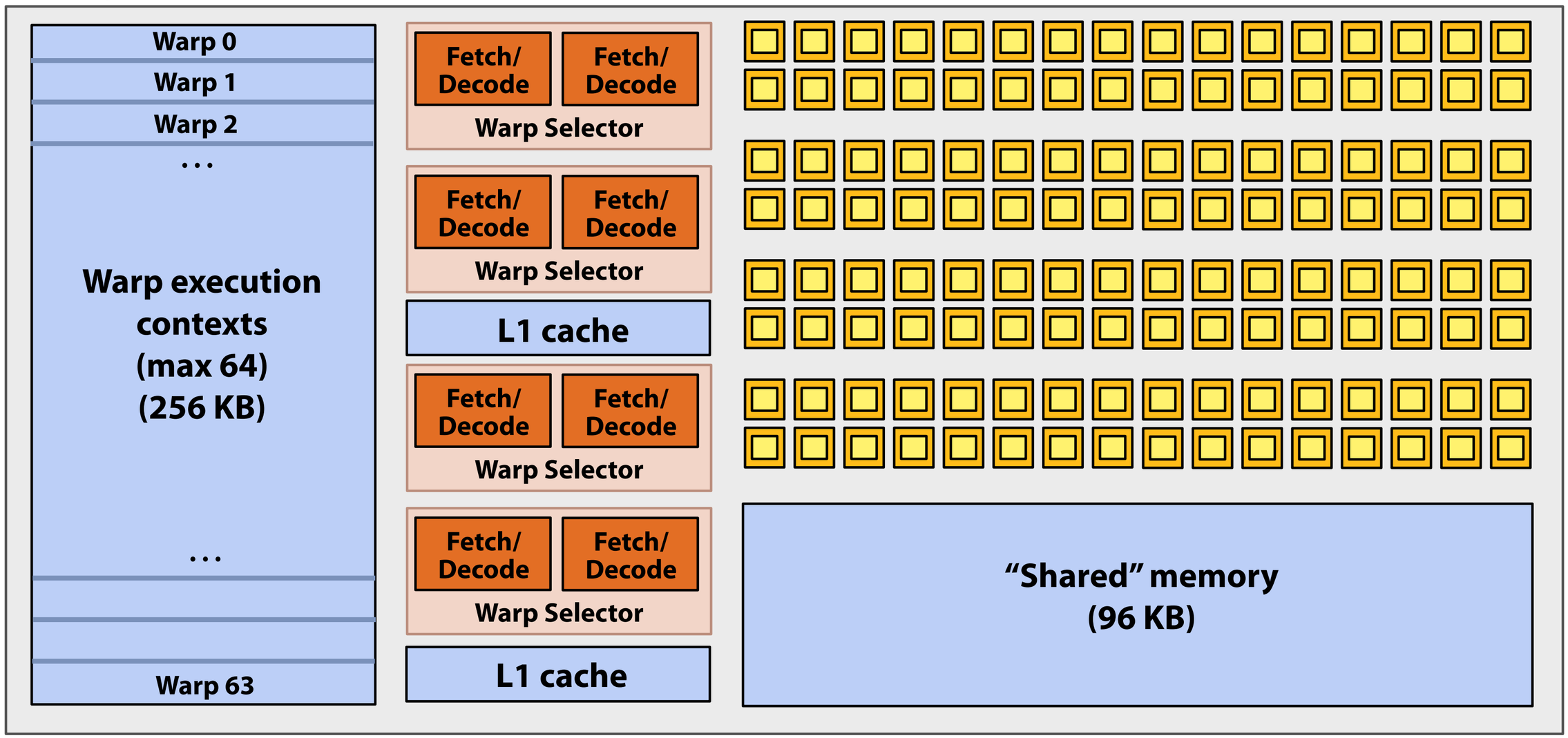


- CPU cores and graphics cores share same memory system
- Also share LLC (L3 cache)
  - Enables, low-latency, high-bandwidth communication between CPU and integrated GPU
- Graphics cores cache coherent with CPU

Revisit GPU Architecture: NVIDIA GTX 980

NVIDIA Maxwell GM204 architecture SMM unit (one "core")

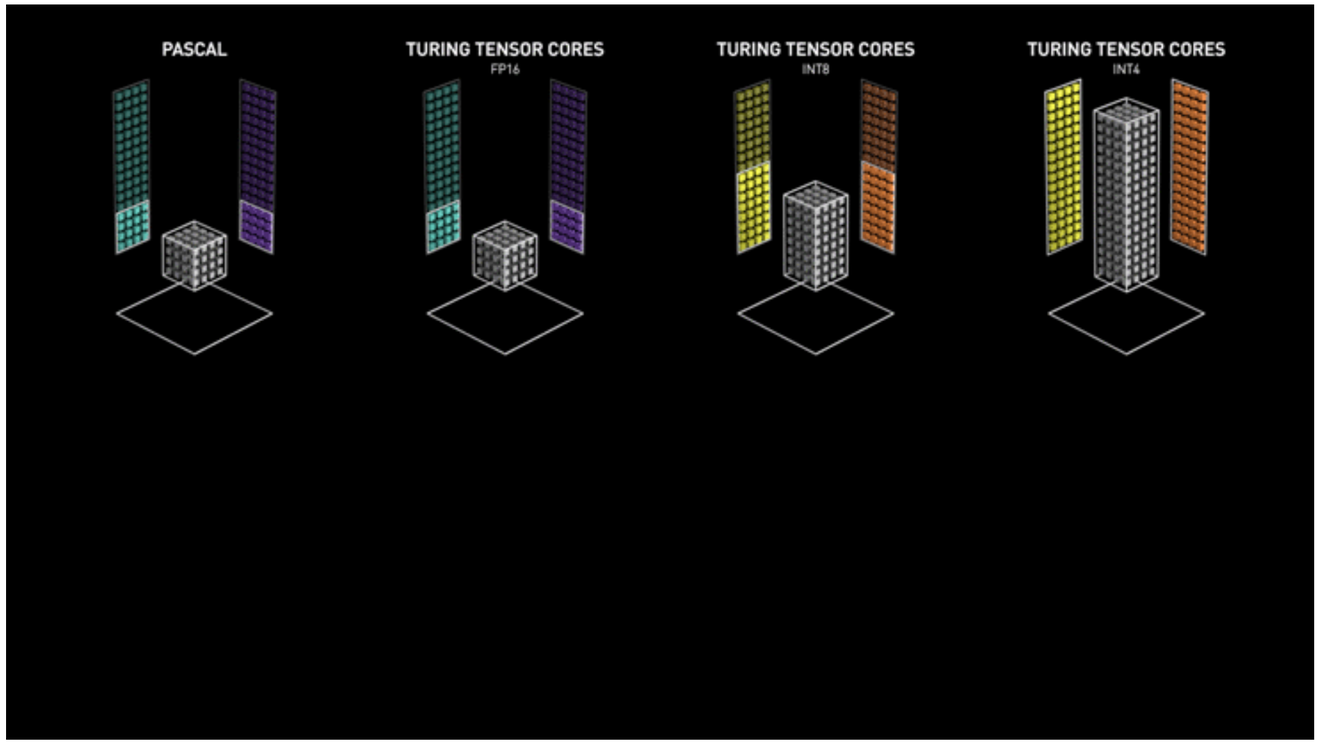
A warp is a set of 32 threads executing the same instruction



= SIMD functional unit, control shared across 32 units (1 MUL-ADD per clock)

Example: NVIDIA A100 Architecture

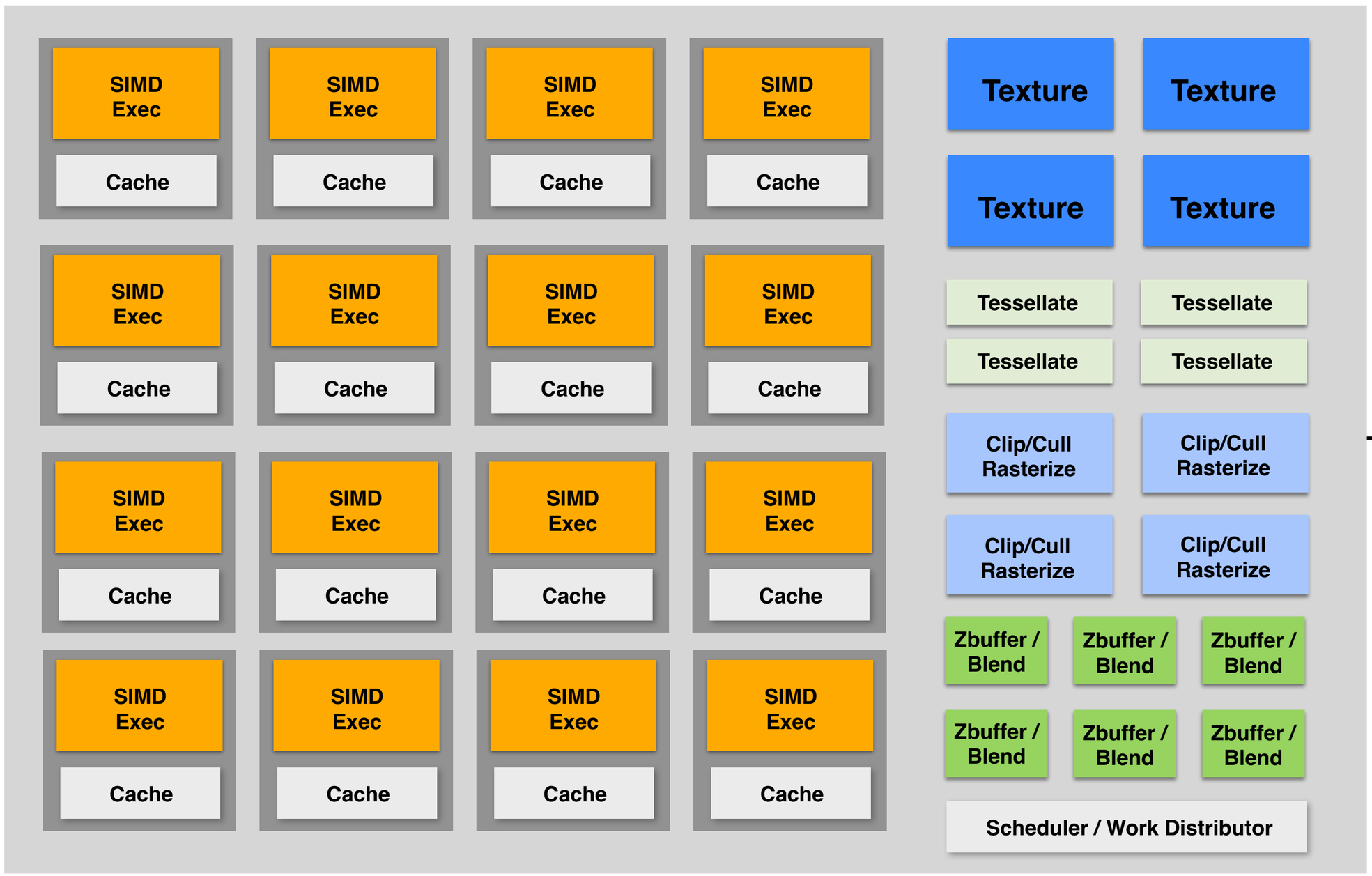
- Tensor Cores for Matrix Multiplication



GPU's are heterogeneous multi-core processors

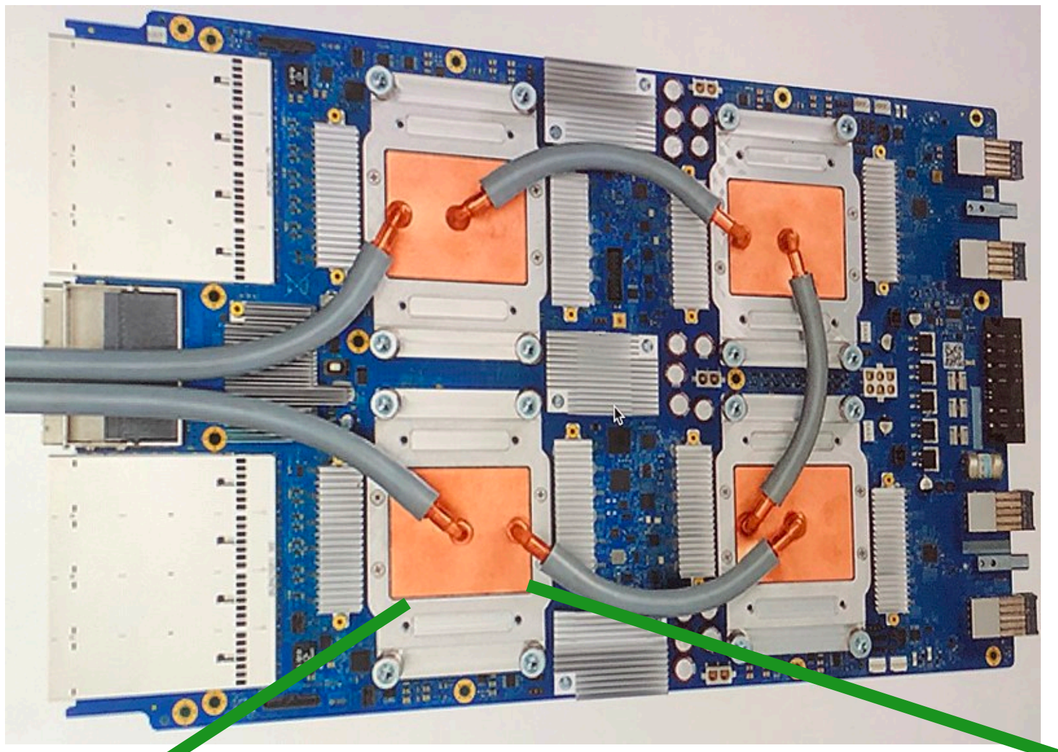
Compute resources your CUDA programs used in assignment 2

Graphics-specific, fixed-function compute resources

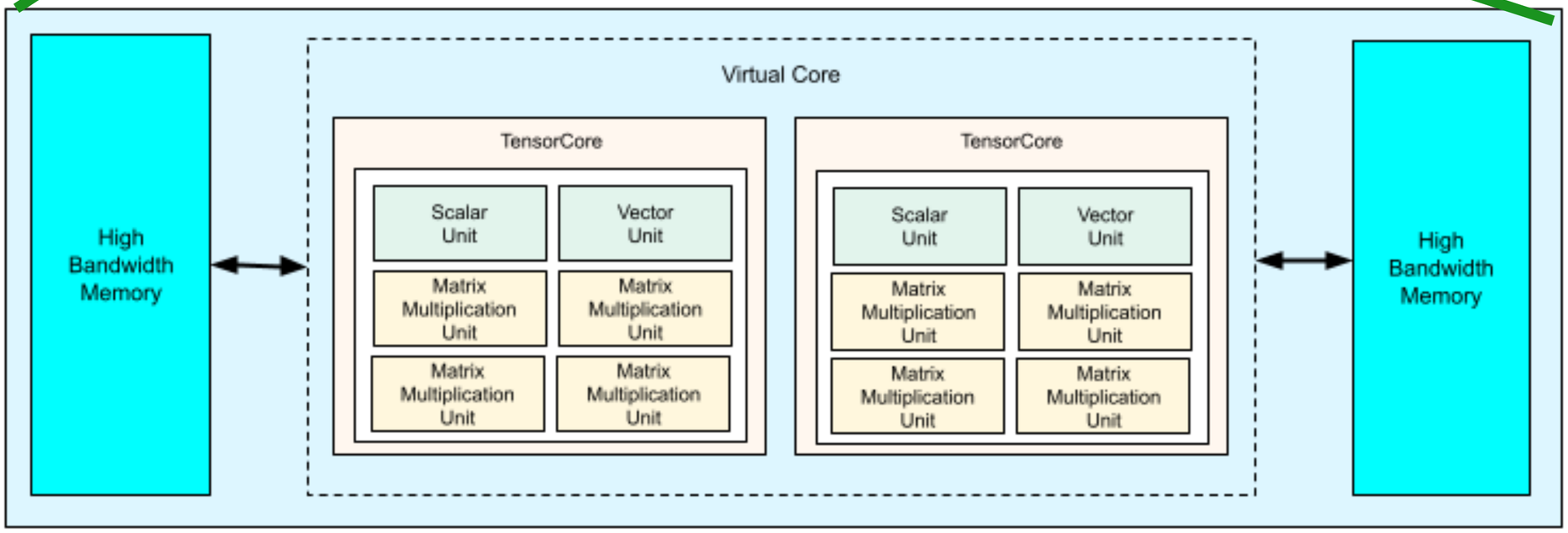


GPU

TPU's Heterogeneous Architecture

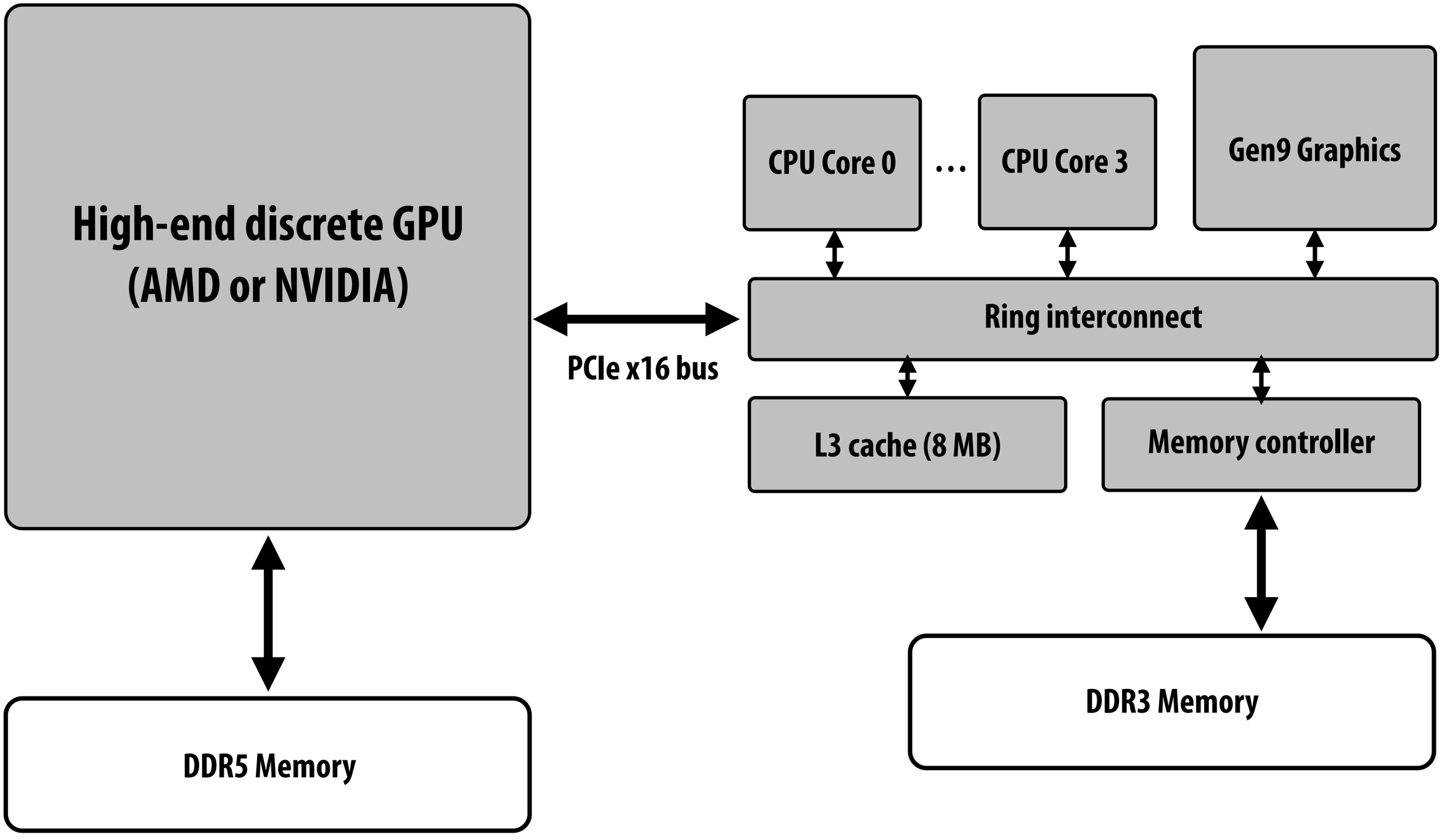


- Google's TPU v4 Architecture with 4 chips
- Each v4 TPU chip contains two TensorCores.
- Each TensorCore has four MXUs, a vector unit, and a scalar unit.

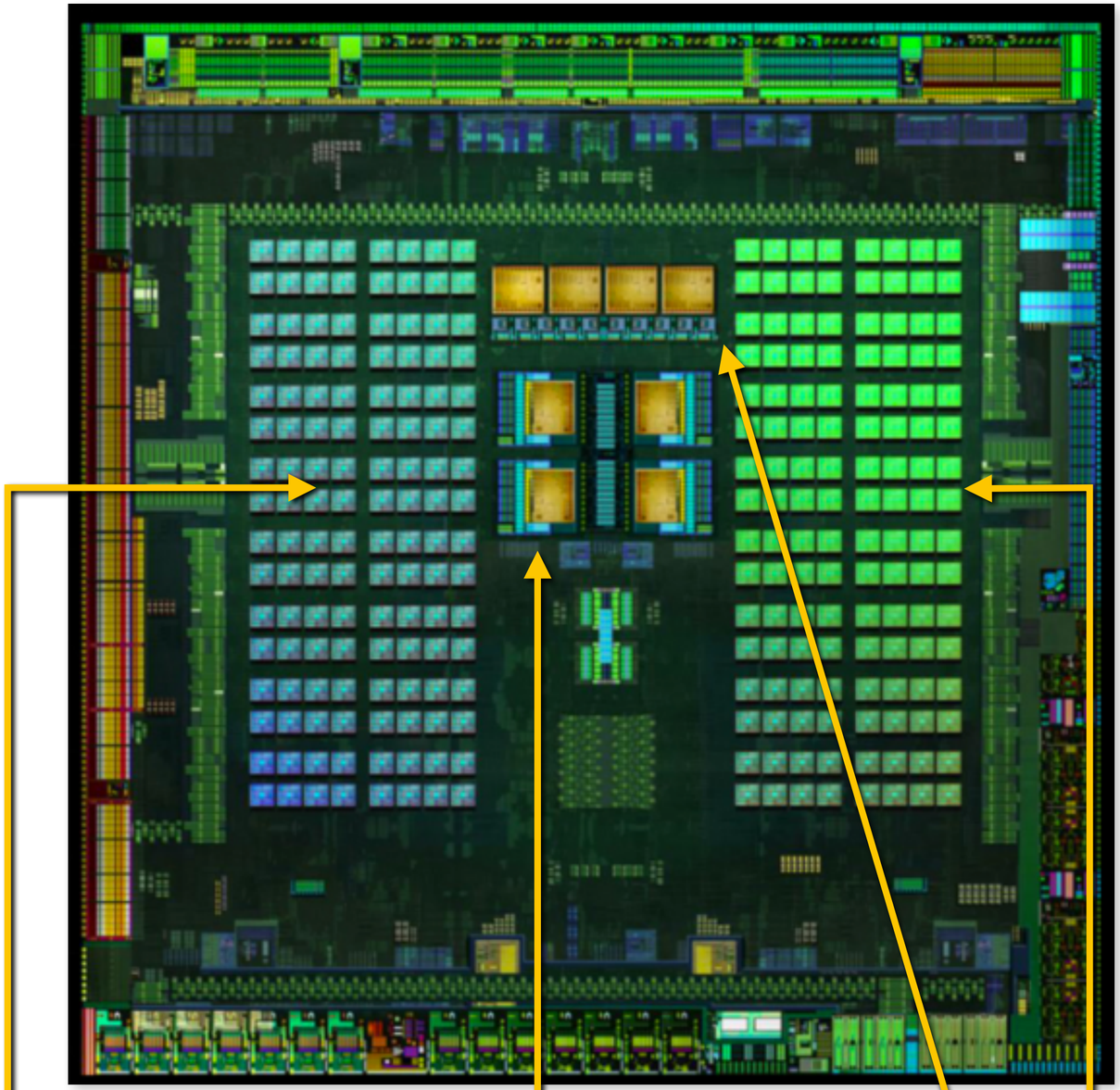


Modern Heterogeneous Platforms

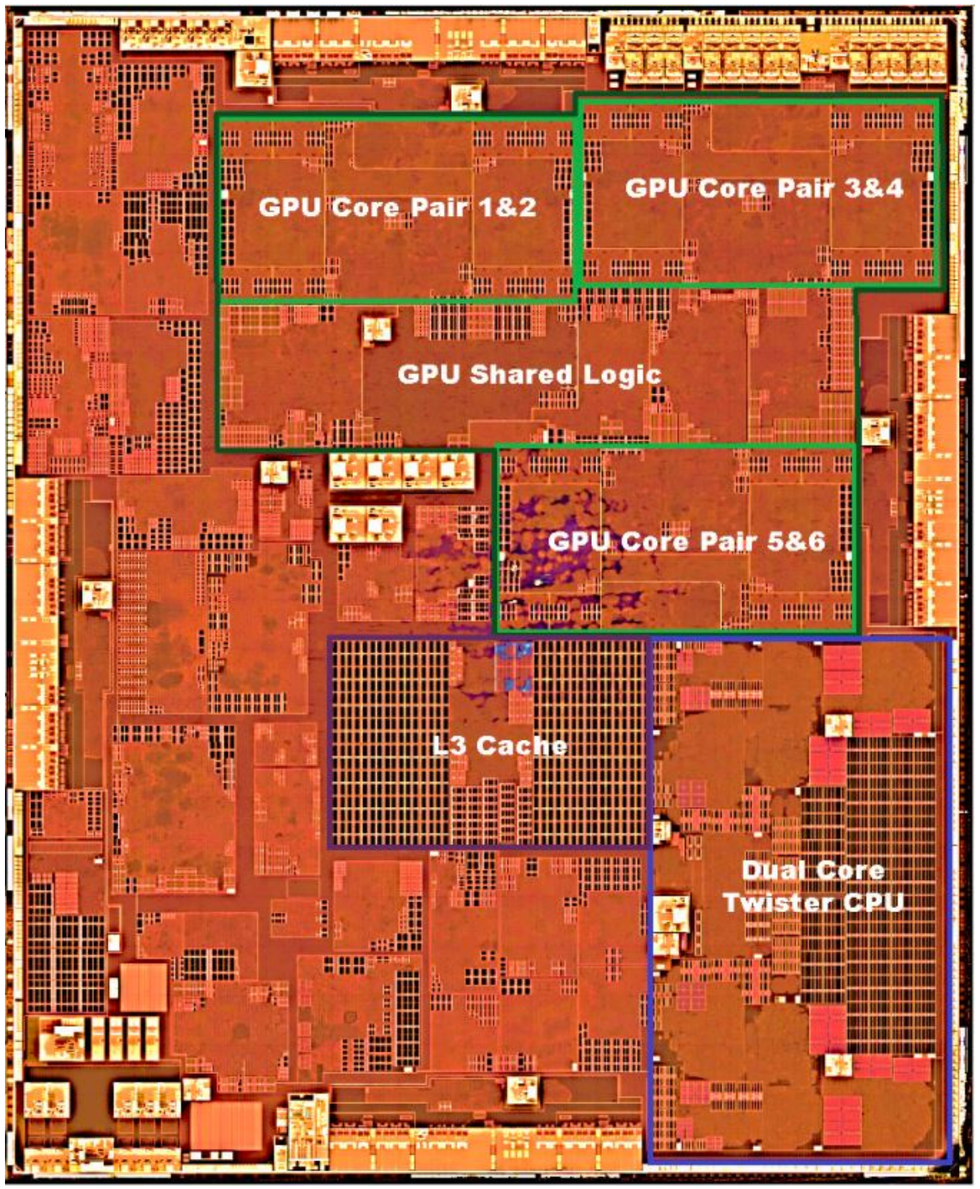
Keep discrete (power hungry) GPU unless needed for graphics-intensive applications
Use integrated, low power graphics for basic graphics/window manager/UI



Mobile heterogeneous processors



NVIDIA Tegra X1
Four ARM Cortex A57 CPU cores for applications
Four low performance (low power) ARM A53 CPU cores
One Maxwell SMM (256 "CUDA" cores)



Apple A9
Dual Core 64 bit CPU
GPU PowerVR GT6700 (6 "core") GPU

A9 image credit Chipworks, obtained from
http://www.anandtech.com/show/9686/the-apple-iphone-6s-and-iphone-6s-plus-review/3

Supercomputers use heterogeneous processing

Los Alamos National Laboratory: Roadrunner

Fastest US supercomputer in 2008, first to break Petaflop barrier: 1.7 PFLOPS

Unique at the time due to use of two types of processing elements

(IBM's Cell processor served as "accelerator" to achieve desired compute density)

- **6,480 AMD Opteron dual-core CPUs (12,960 cores)**
- **12,970 IBM Cell Processors (1 CPU + 8 accelerator cores per Cell = 116,640 cores)**
- **2.4 MWatt (about 2,400 average US homes)**

Why no GPUs?

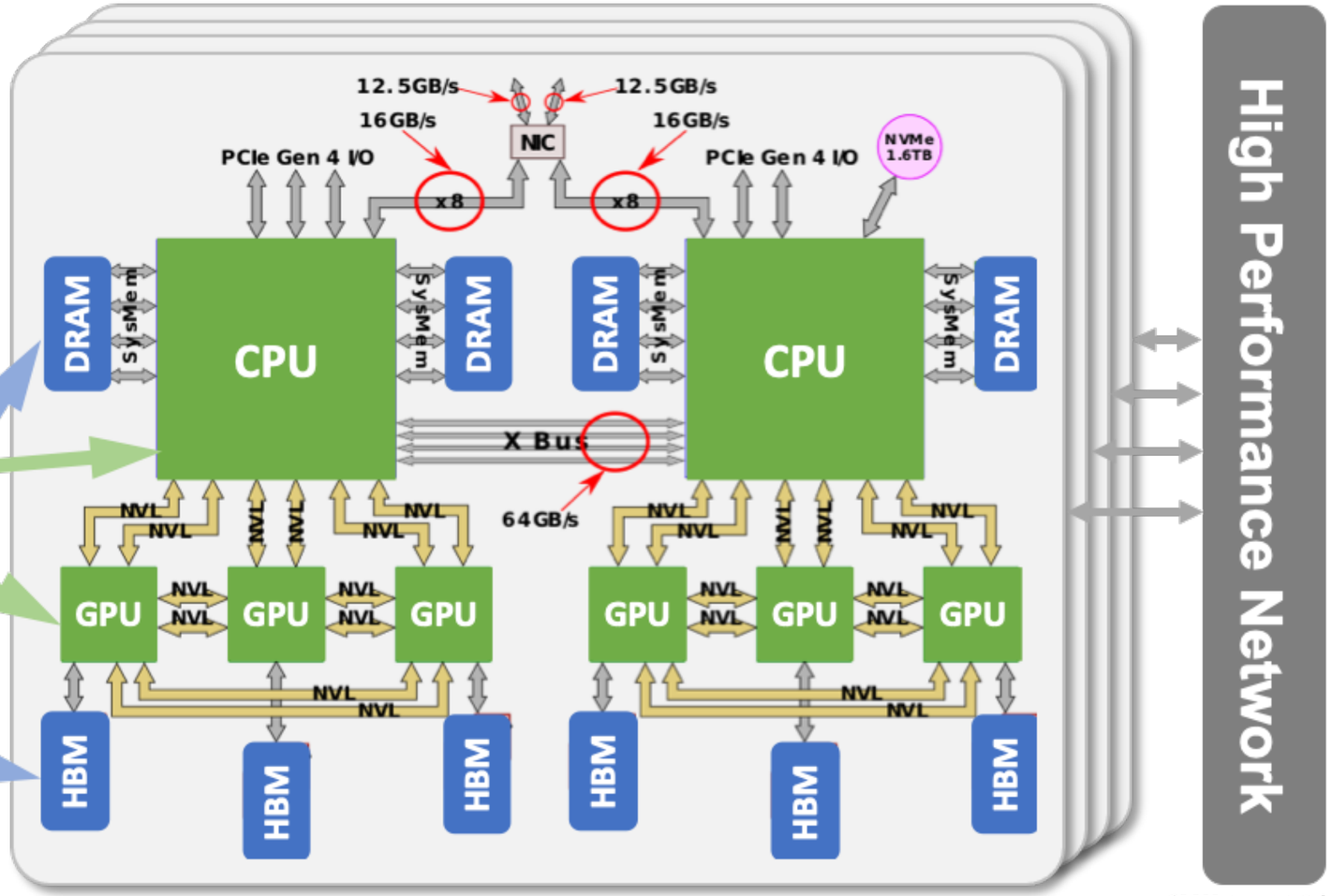
GPU-accelerated supercomputing

- Oak Ridge Summit (world's #4)
- Overall Throughput: 200 PFLOPS
- 9,216 POWER9 22-core CPUs
- 27,648 NVIDIA Tesla V100 GPUs
- 250 PB Storage
- Estimated Cost: 325 million USD

Summit's architecture

Heterogenous Processors

Heterogenous Memories



Heterogeneous architectures for supercomputing

Source: Top500.org Spring 2022 rankings

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942
4	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
5	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
6	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
7	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589

GPU

GPU

GPU

GPU

Green500: most energy efficient supercomputers

Efficiency metric: GFLOPS per Watt

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	29	**Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
2	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	21,100	52.227
3	3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	2,942	51.629
4	10	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	50.028
5	326	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, Preferred Networks Preferred Networks	1,664	2.18	53	40.901

Source: Green500 Spring 2022 rankings

Energy-constrained computing

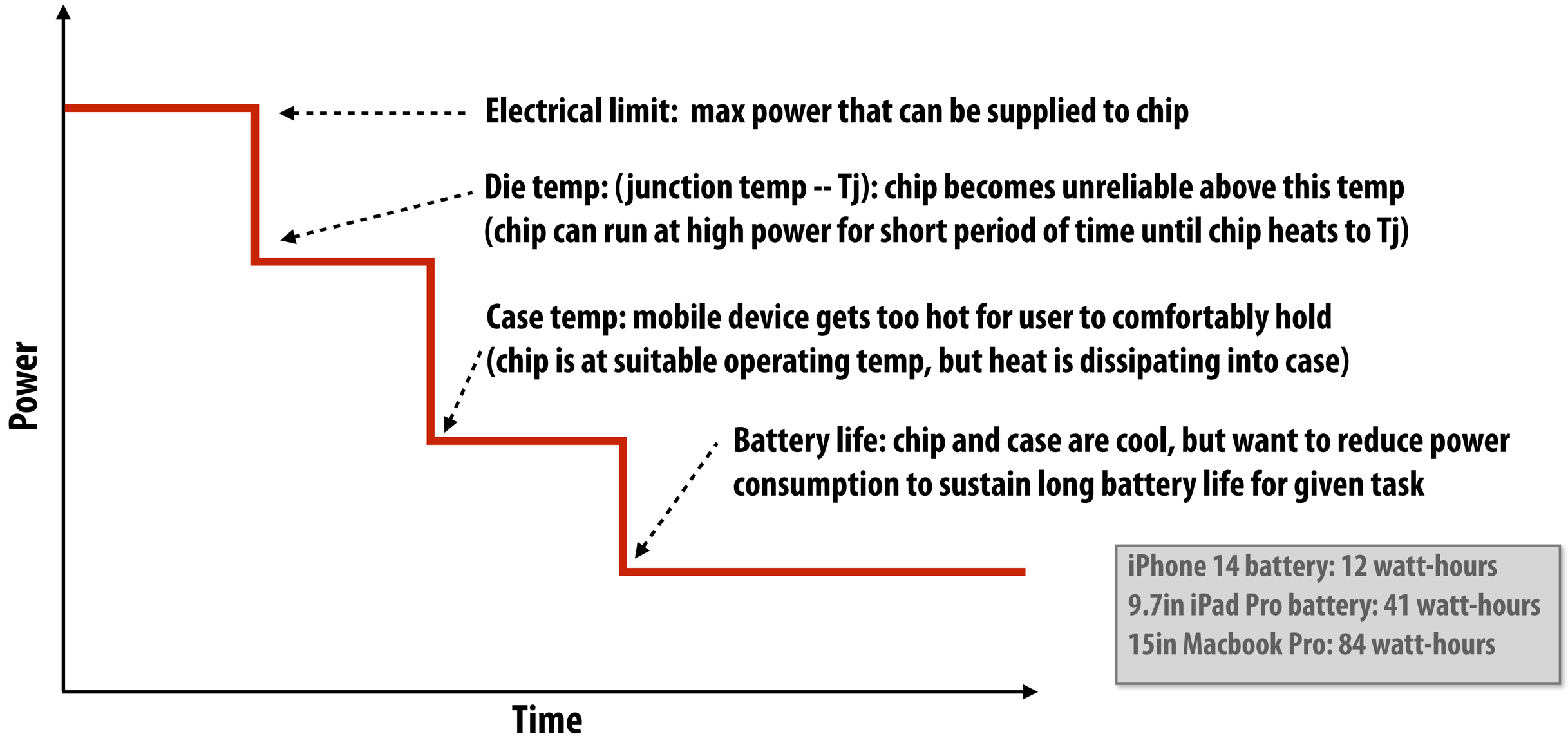
- Supercomputers are energy constrained
  - Due to shear scale
  - Overall cost to operate (power for machine and for cooling)

- Datacenters are energy constrained
  - Reduce cost of cooling
  - Reduce physical space requirements

- Mobile devices are energy constrained
  - Limited battery life
  - Heat dissipation

Limits on chip power consumption

- General in mobile processing rule: the longer a task runs the less power it can use
  - Processor's power consumption is limited by heat generated (efficiency is required for more than just maximizing battery life)



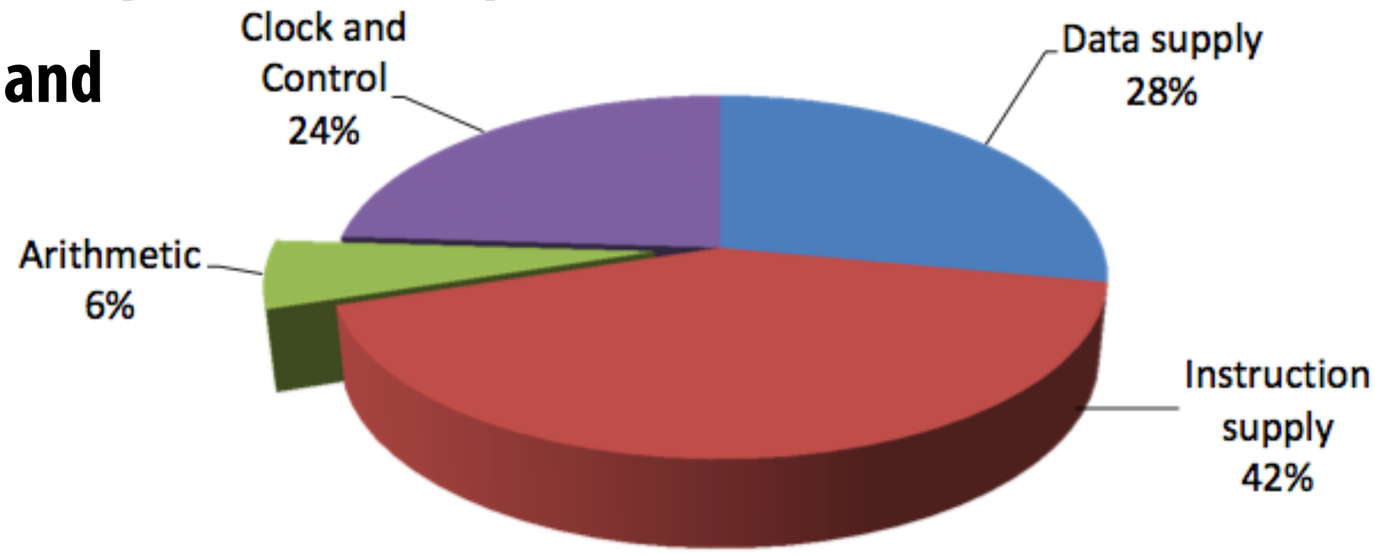
Slide credit: adopted from original slide from M. Shebanow: HPG 2013 keynote

Efficiency benefits of compute specialization

- Rules of thumb: compared to high-quality C code on CPU...

- Throughput-maximized processor architectures: e.g., GPU cores
  - Approximately 10x improvement in perf / watt
  - Assuming code maps well to wide data-parallel execution and is compute bound

- Fixed-function ASIC ("application-specific integrated circuit")
  - Can approach 100-1000x or greater improvement in perf/watt
  - Assuming code is compute bound and and is not floating-point math

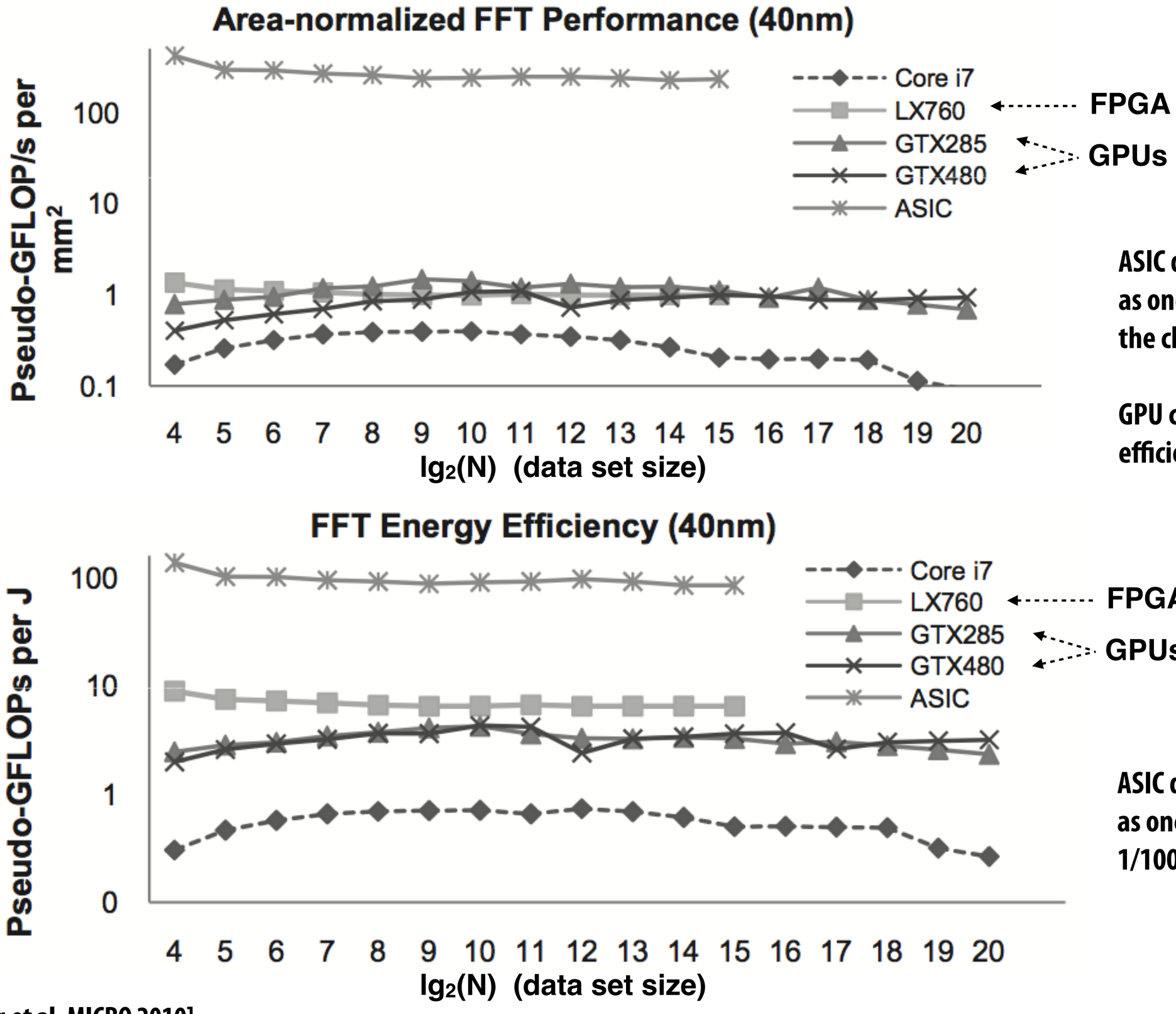


Efficient Embedded Computing [Dally et al. 08]

[Figure credit Eric Chung]

[Source: Chung et al. 2010 , Dally 08]

Hardware specialization increases efficiency



ASIC delivers same performance as one CPU core with ~ 1/1000th the chip area.

GPU cores: ~ 5-7 times more area efficient than CPU cores.

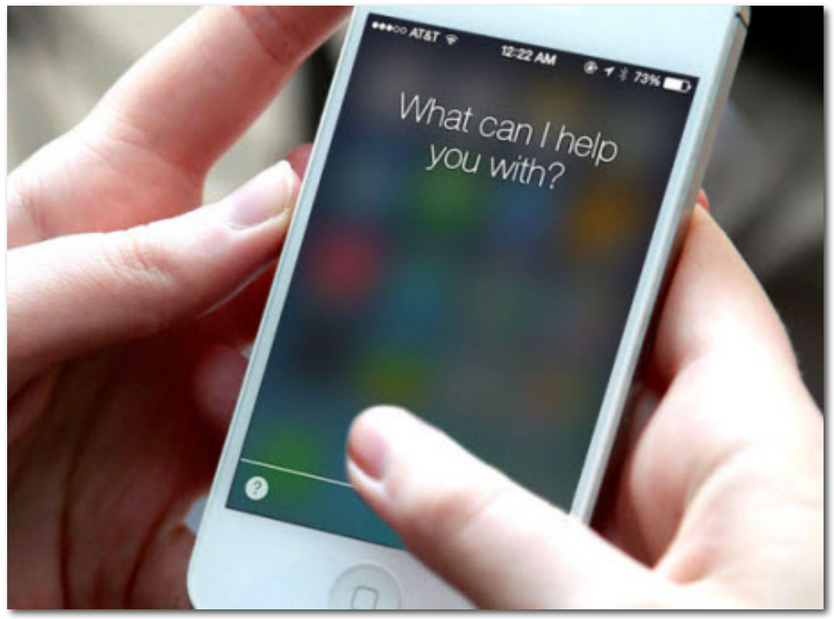
ASIC delivers same performance as one CPU core with only ~ 1/100th the power.

[Chung et al. MICRO 2010]

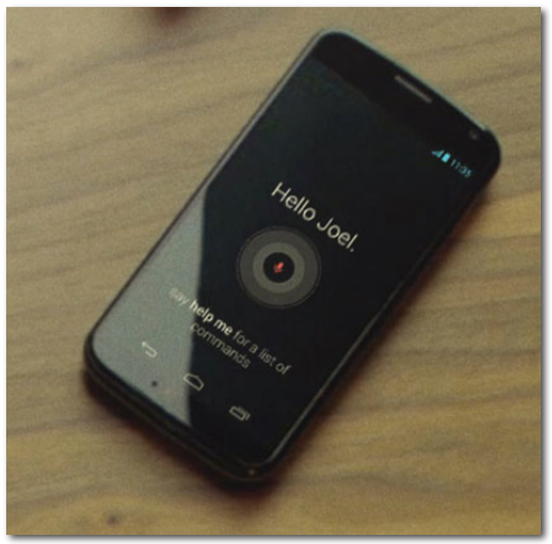
Benefits of increasing efficiency

- **Run faster for a fixed period of time**
  - Run at higher clock, use more cores (reduce latency of critical task)
  - Do more at once

- **Run at a fixed level of performance for longer**
  - e.g., video playback
  - Achieve "always-on" functionality that was previously impossible



iPhone:
Siri activated by button press or holding phone up to ear



Moto X:
Always listening for "ok, google now"
Device contains ASIC for detecting this audio pattern.

Google Glass: ~40 min recording per charge (nowhere near "always on")

Example: iPad Air (2013)

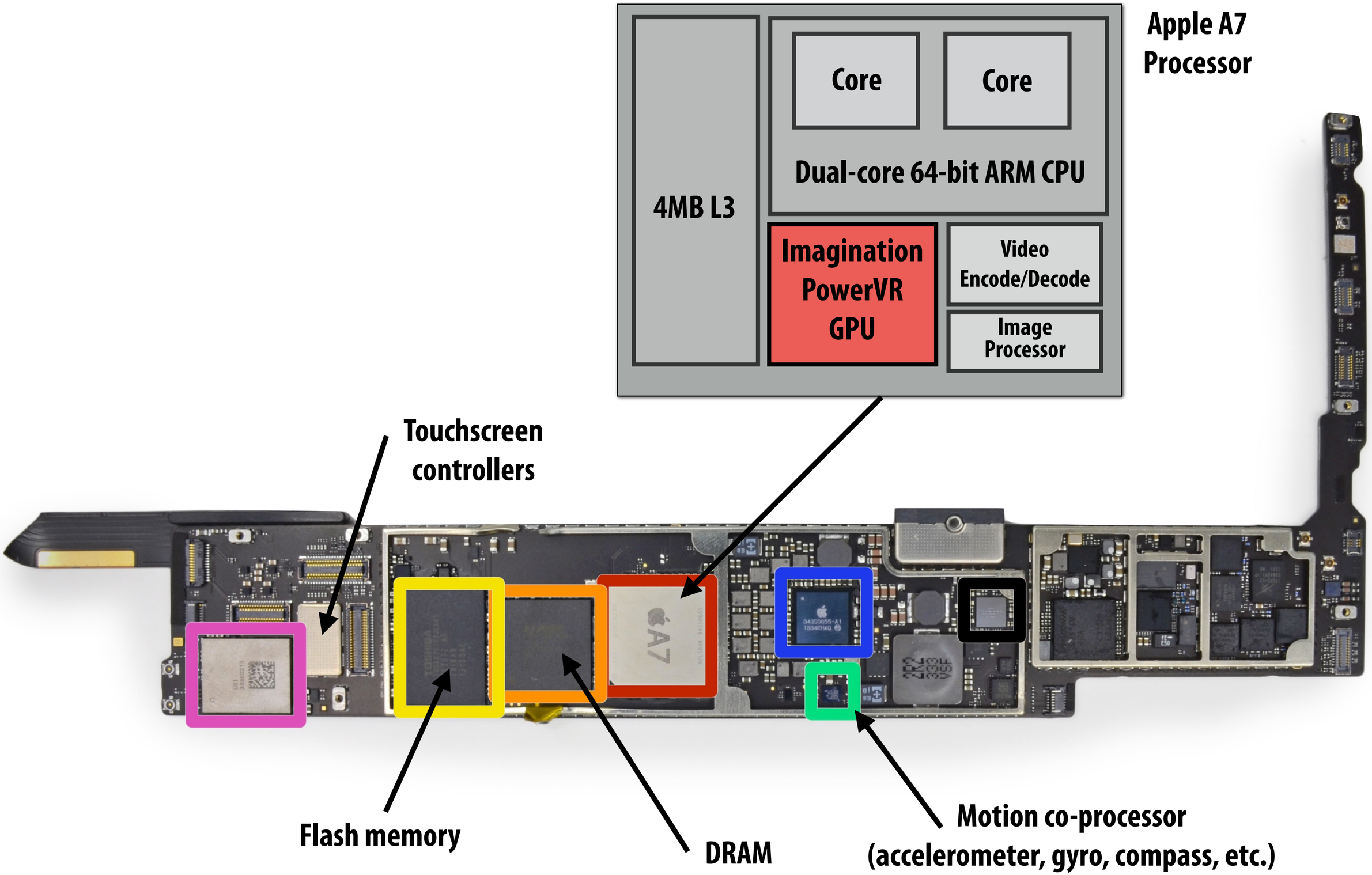


Image Credit: ifixit.com

Original iPhone touchscreen controller

Separate digital signal processor to interpret raw signal from capacitive touch sensor (do not burden main CPU)

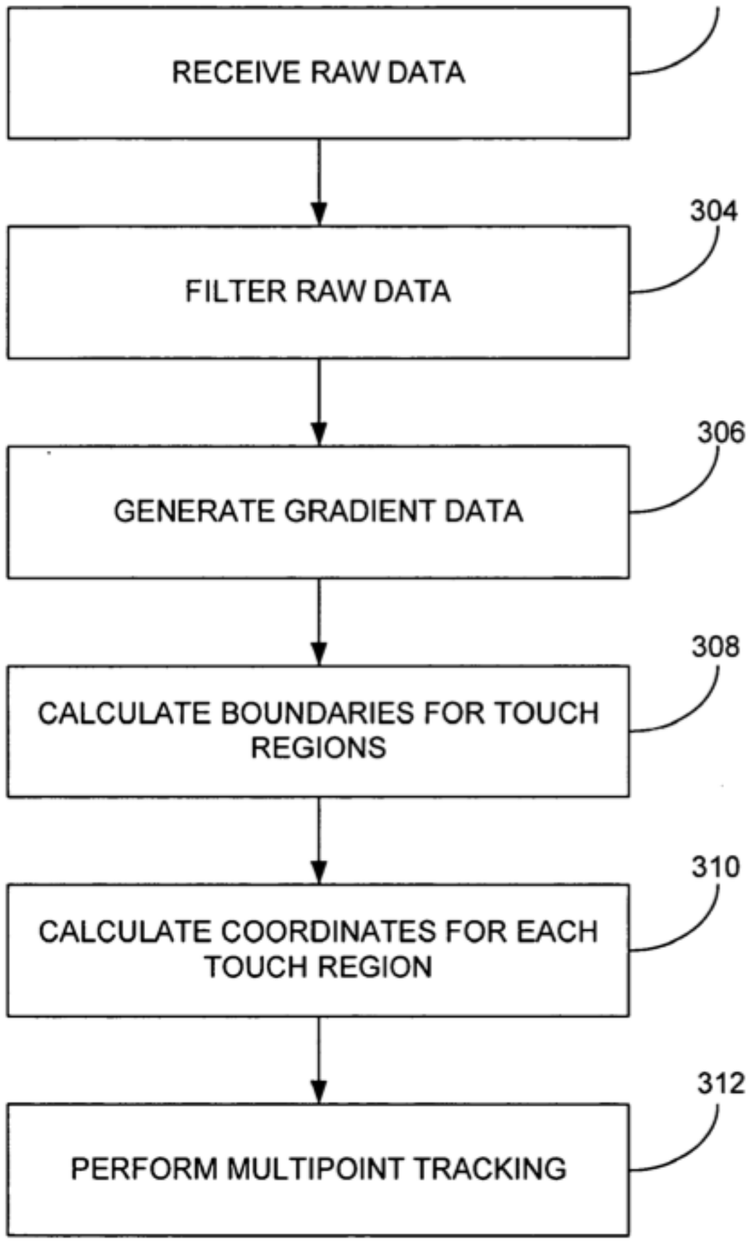


FIG. 16

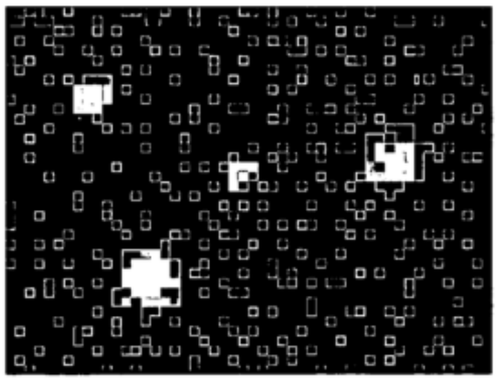


FIG. 17A

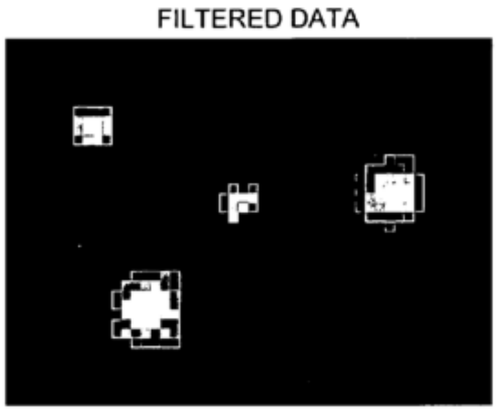


FIG. 17B

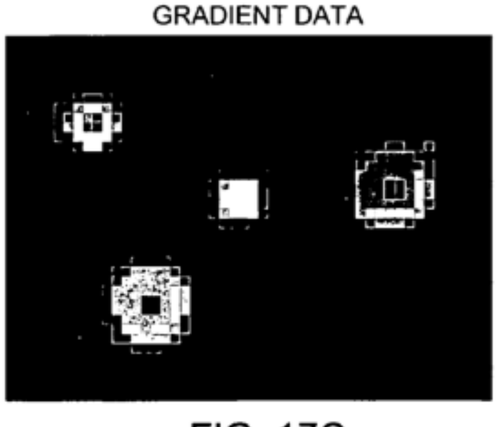


FIG. 17C

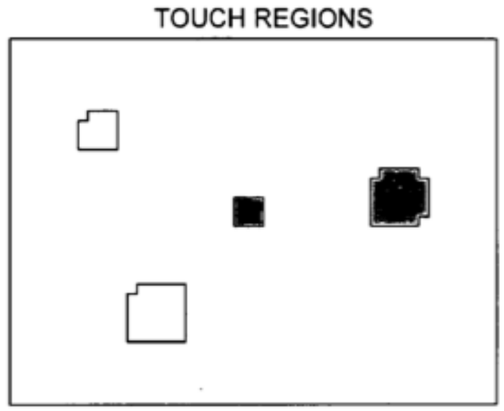


FIG. 17D

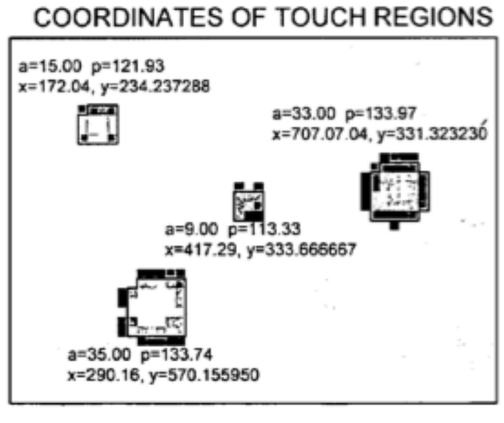


FIG. 17E

From US Patent Application 2006/0097991

Example: image processing on Nikon D7000

Process 16 MPixel RAW data from sensor to obtain JPG image:
On camera: ~ 1/6 sec per image
Adobe Lightroom on a quad-core Macbook Pro laptop: 1-2 sec per image

This is a older camera: much, much faster image processing performance on a modern smart phone (burst mode)

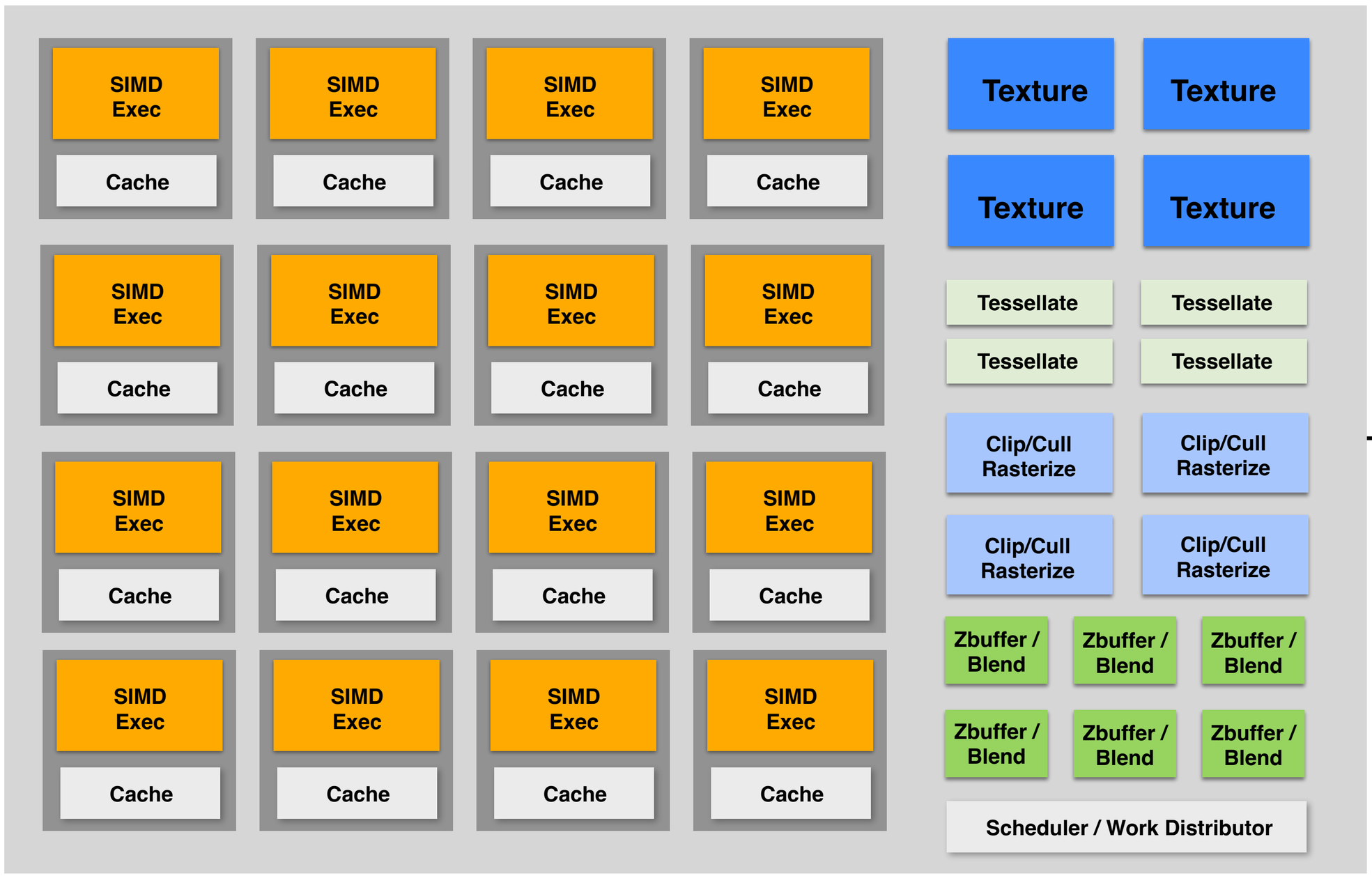
Trading Efficiency and Programmability

- Improved energy efficiency often comes at a cost
  - Programmability (e.g., consider debugging on a CPU v.s. GPU)
  - Applications (general-purpose v.s. domain-specific)

GPU's are heterogeneous multi-core processors

Compute resources your CUDA programs used in assignment 2

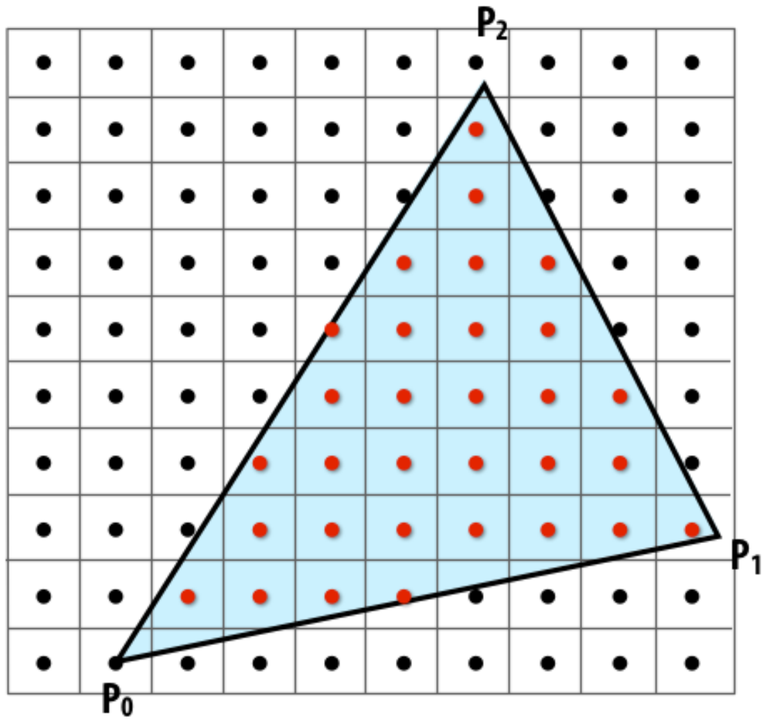
Graphics-specific, fixed-function compute resources



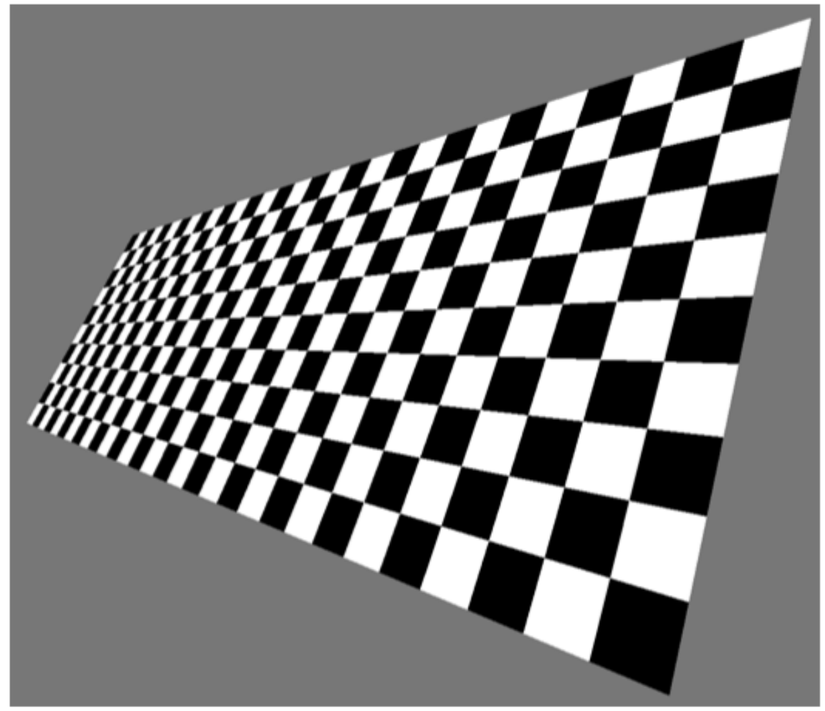
GPU

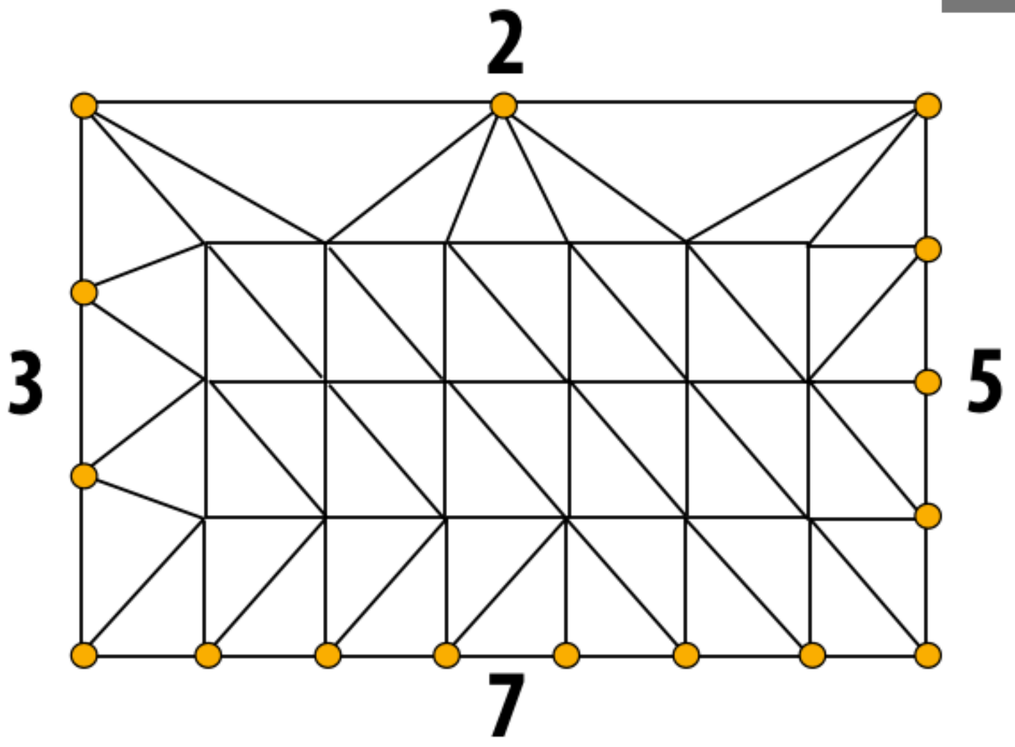
Example graphics tasks performed in fixed-function HW

Rasterization:
Determining what pixels a triangle overlaps



Texture mapping:
Warping/filtering images to apply detail to surfaces

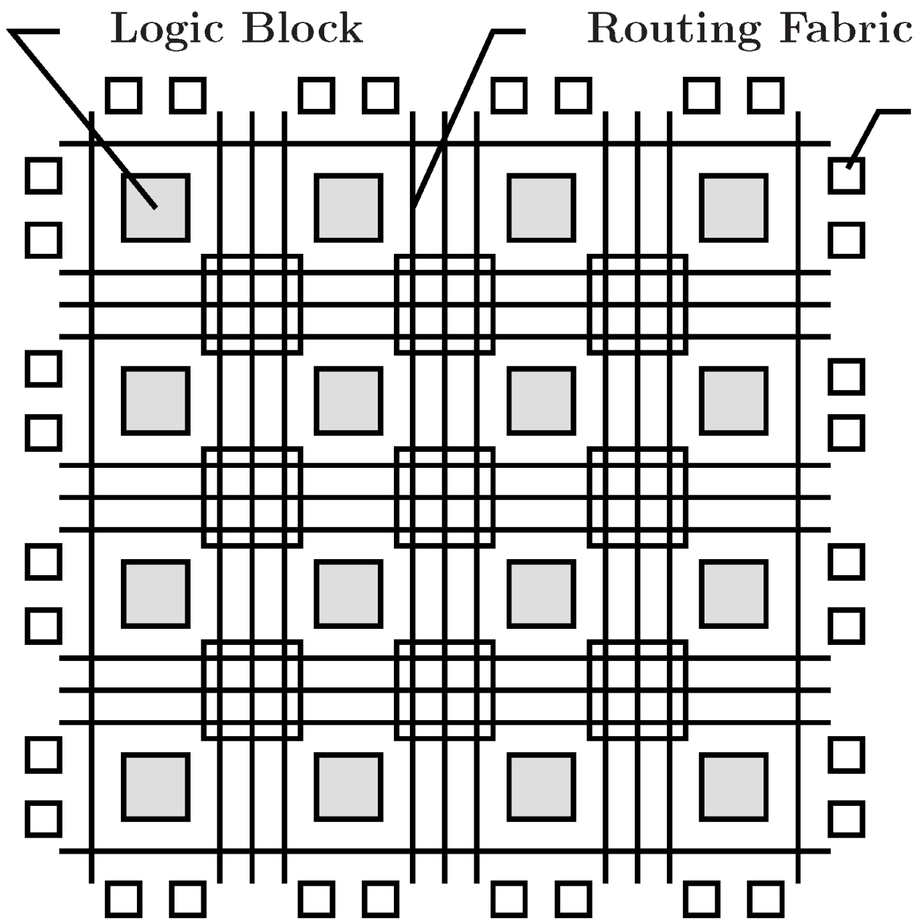




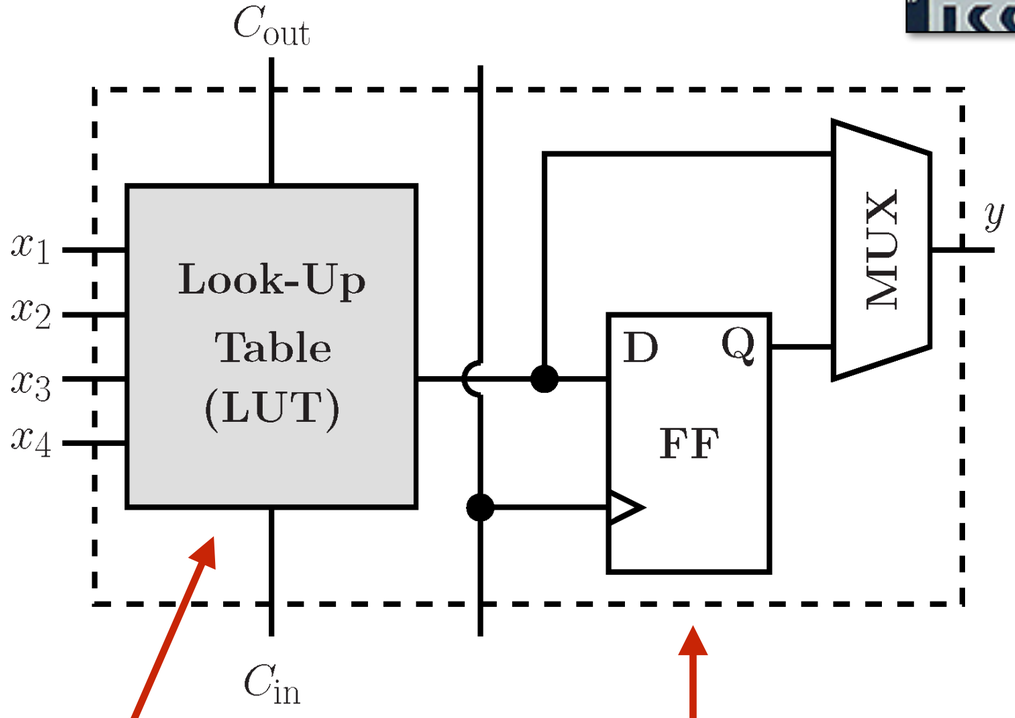
Geometric tessellation:
computing fine-scale geometry from coarse geometry

FPGAs (Field Programmable Gate Arrays)

- Middle ground between an ASIC and a processor
- FPGA chip provides array of logic blocks, connected by interconnect
- Programmer-defined logic implemented directly by FGPA



(a)



(b)

Programmable lookup table (LUT)

Flip flop (a register)

Image credit: Bai et al. 2014

Project Catapult [Putnam et al. ISCA 2014]

- Microsoft Research investigation of use of FPGAs to accelerate datacenter workloads
- Demonstrated offload of part of Bing Search's document ranking logic
- Now widely used to accelerate DNNs across Microsoft services

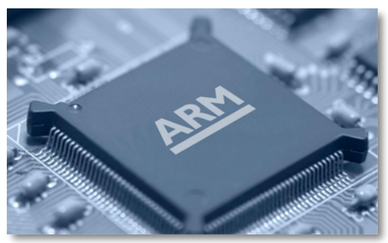
FPGA board

1U server (Dual socket CPU + FPGA connected via PCIe bus)

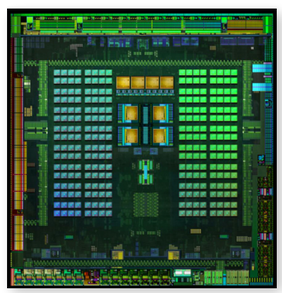
Two 8-core Xeon CPUs, 64 GB DRAM, 4 HDDs @ 2TB, 10Gb Ethernet

Summary: choosing the right tool for the job

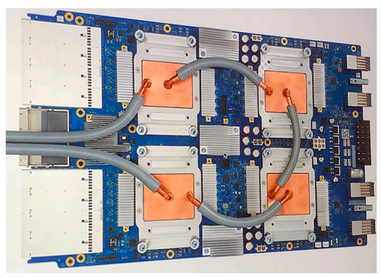
Energy-optimized CPU



Throughput-oriented processors (GPU)



(TPU)



FPGA/Future reconfigurable logic

ASIC

Video encode/decode, Audio playback, Camera RAW processing, neural nets (future?)

~10X more efficient

~100X??? (jury still out)

~100-1000X more efficient

Easiest to program

Difficult to program (making it easier is active area of research)

Not programmable + costs 10-100's millions of dollars to design / verify / create

Credit Pat Hanrahan for this taxonomy

Challenges of heterogeneous designs

Challenges of heterogeneity

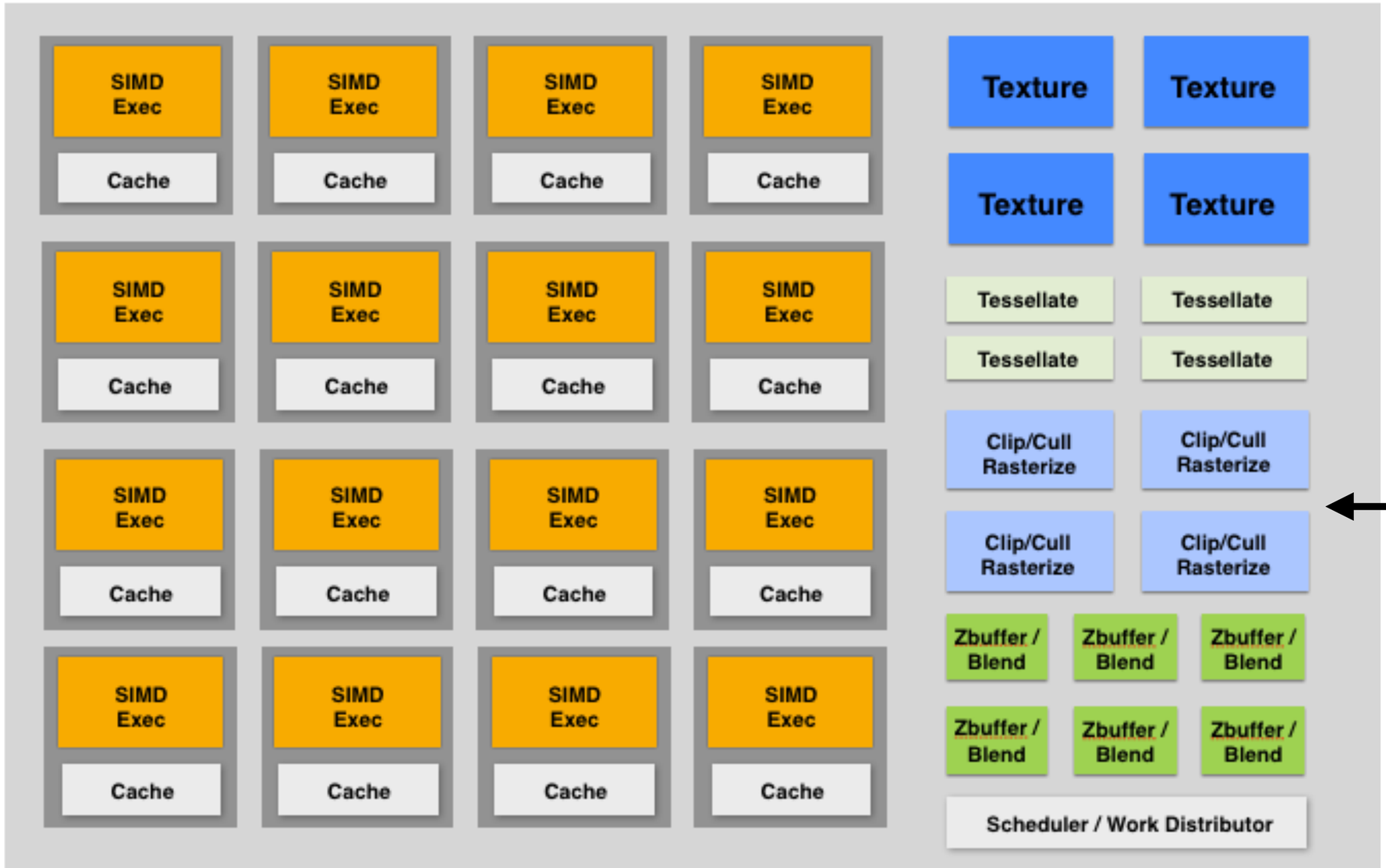
- So far in this course:
  - Homogeneous system: every processor can be used for every task
  - To get best speedup vs. sequential execution, "keep all processors busy all the time"

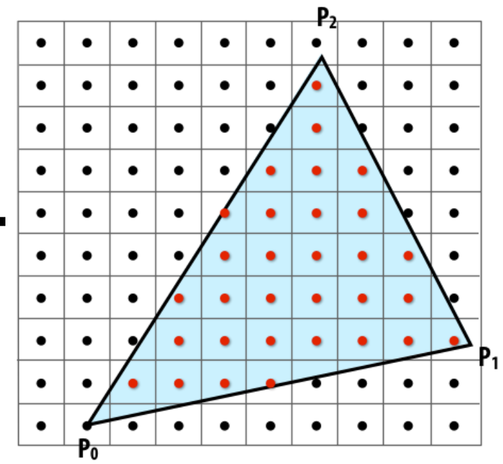
- Heterogeneous system: use preferred processor for each task
  - Challenge for system designer: what is the right mixture of resources to meet performance, cost, and energy goals?
    - Too few throughput-oriented resources (lower peak performance/efficiency for parallel workloads -- should have used resources for more throughput cores)
    - Too few sequential processing resources (get bitten by Amdahl's Law)
    - How much chip area should be dedicated to a specific function, like video? (these resources are taken away from general-purpose processing)

- Implication: increased pressure to understand workloads accurately at chip design time

Pitfalls of heterogeneous designs

[Molnar 2010]





Say 10% of the workload is rasterization
Let's say you under-provision the fixed-function rasterization unit on GPU:
Chose to dedicate 1% of chip area used for rasterizer, really needed 20% more throughput: 1.2% of chip area

Problem: rasterization is bottleneck, so the expensive programmable processors (99% of chip) are idle waiting on rasterization. So the other 99% of the chip runs at 80% efficiency!

Challenges of heterogeneity

- **Heterogeneous system: preferred processor for each task**
  - Challenge for hardware designer: what is the right mixture of resources?
    - Too few throughput oriented resources (lower peak throughput for parallel workloads)
    - Too few sequential processing resources (limited by sequential part of workload)
    - How much chip area should be dedicated to a specific function, like video? (these resources are taken away from general-purpose processing)
    - Work balance must be anticipated at chip design time
      - System cannot adapt to changes in usage over time, new algorithms, etc.

- Challenge to software developer: how to map programs onto a heterogeneous collection of resources?
  - Challenge: "Pick the right tool for the job": design algorithms that decompose well into components that each map well to different processing components of the machine
  - The scheduling problem is more complex on a heterogeneous system
  - Available mixture of resources can dictate choice of algorithm
  - Software portability nightmare

Reducing energy consumption idea 1: use specialized processing

Reducing energy consumption idea 2: move less data

Data movement has high energy cost

- Rule of thumb in mobile system design: always seek to reduce amount of data transferred from memory
  - Earlier in class we discussed minimizing communication to reduce stalls (poor performance). Now, we wish to reduce communication to reduce energy consumption

- "Ballpark" numbers [Sources: Bill Dally (NVIDIA), Tom Olson (ARM)]
  - Integer op: ~ 1 pJ *
  - Floating point op: ~20 pJ *
  - Reading 64 bits from small local SRAM (1mm away on chip): ~ 26 pJ
  - Reading 64 bits from low power mobile DRAM (LPDDR): ~1200 pJ ← Suggests that recomputing values, rather than storing and reloading them, is a better answer when optimizing code for energy efficiency!

- Implications
  - Reading 10 GB/sec from memory: ~1.6 watts
  - Entire power budget for mobile GPU: ~1 watt (remember phone is also running CPU, display, radios, etc.)
  - iPhone 14 battery: ~12 watt-hours
  - Exploiting locality matters!!!

* Cost to just perform the logical operation, not counting overhead of instruction decode, load data from registers, etc.

Three trends in energy-optimized computing

- **Compute less!**
  - Computing costs energy: parallel algorithms that do more work than sequential counterparts may not be desirable even if they run faster
- **Specialize compute units:**
  - Heterogeneous processors: CPU-like cores + throughput-optimized cores (GPU-like cores)
  - Fixed-function units: audio processing, "movement sensor processing" video decode/encode, image processing/computer vision?
  - Specialized instructions: expanding set of AVX vector instructions, new instructions for accelerating AES encryption (AES-NI)
  - Programmable soft logic: FPGAs
- **Reduce bandwidth requirements**
  - Exploit locality (restructure algorithms to reuse on-chip data as much as possible)
  - Aggressive use of compression: perform extra computation to compress application data before transferring to memory (likely to see fixed-function HW to reduce overhead of general data compression/decompression)

Summary

- Heterogeneous parallel processing: use a mixture of computing resources that each fit with mixture of needs of target applications
  - Latency-optimized sequential cores, throughput-optimized parallel cores, domain-specialized fixed-function processors
  - Examples exist throughout modern computing: mobile processors, servers, supercomputers

- Traditional rule of thumb in "good system design" is to design simple, general-purpose components
  - This is not the case with emerging processing systems (optimized for perf/watt)
  - Today: want collection of components that meet perf requirement AND minimize energy use

- Challenge of using these resources effectively is pushed up to the programmer
  - Current CS research challenge: how to write efficient, portable programs for emerging heterogeneous architectures?