Lecture 21:

Heterogeneous Parallelism and Hardware Specialization

Parallel Computer Architecture and Programming
CMU 15-418/15-618, Spring 2015

Tunes

Kanye West

Amazing

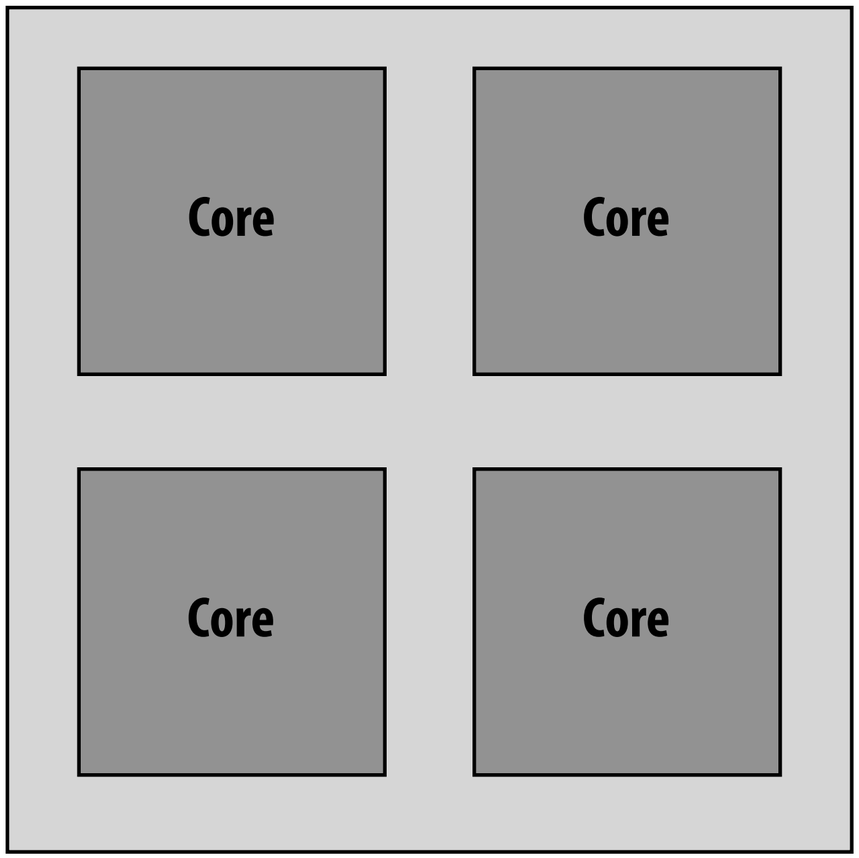
"Have you seen how fast Kim's iPhone battery dies if she tries to play back video in a software decoder? Amazing."

- Kanye

MAKE YOUR TEST DATASETS, TRACES, ETC. FOR YOUR FINAL PROJECT IN THE NEXT TWO WEEKS (trust me)

You want to be in a position where each time you tweak your code or try a more advanced algorithm you can run your system on all your examples, traces, etc. to get new results and graphs. (I'd even automate graph creation if possible.)

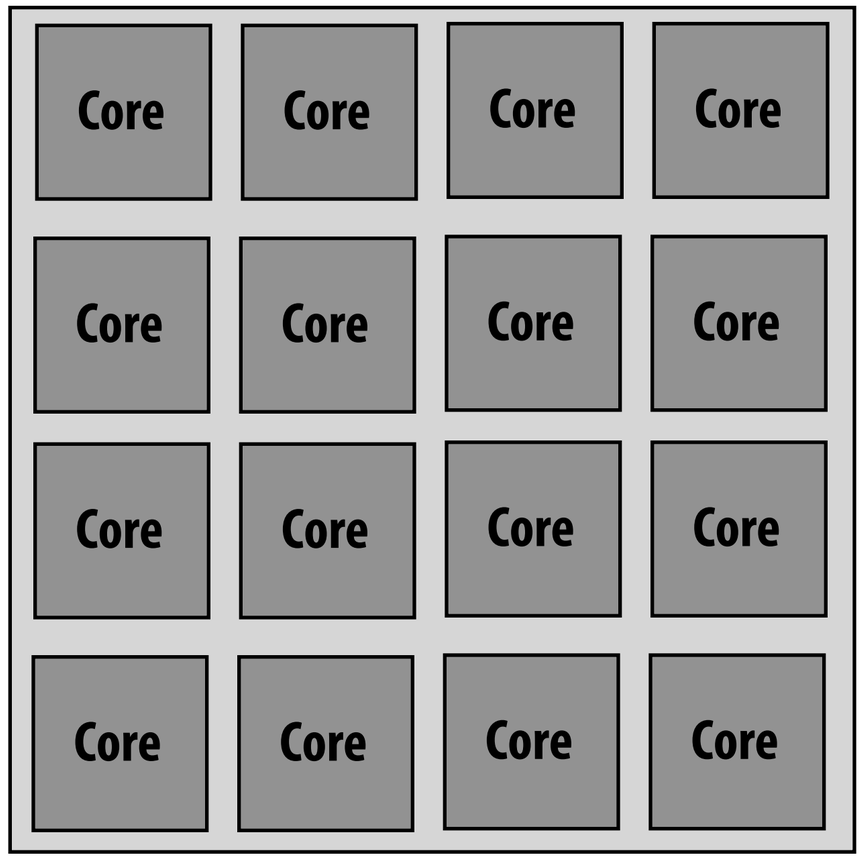
You need to buy a computer system



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

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

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

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

(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,

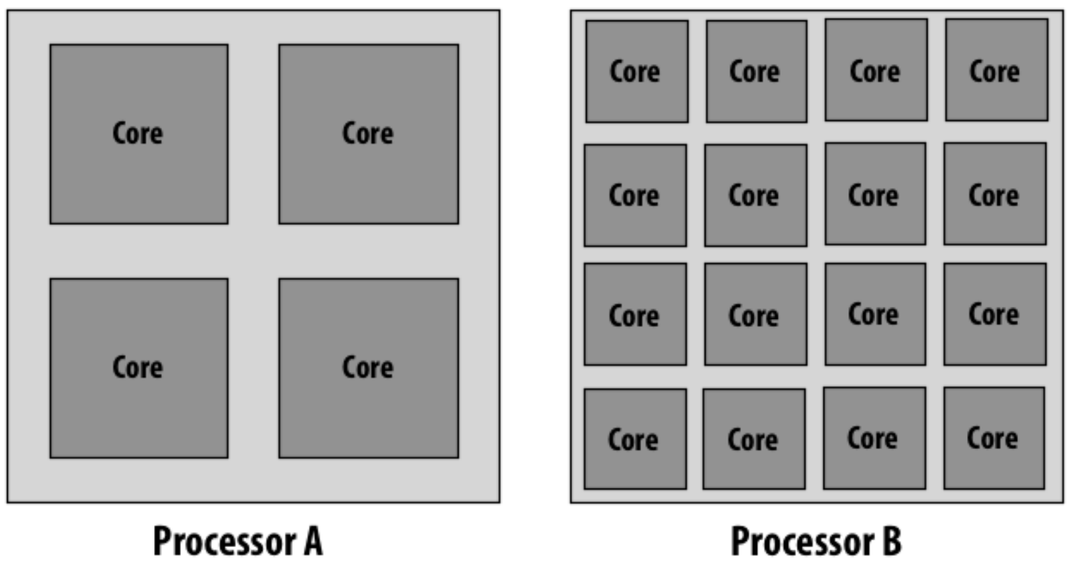
(each of the n/r cores has sequential performance $perf(r)$

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

Two examples where n=16

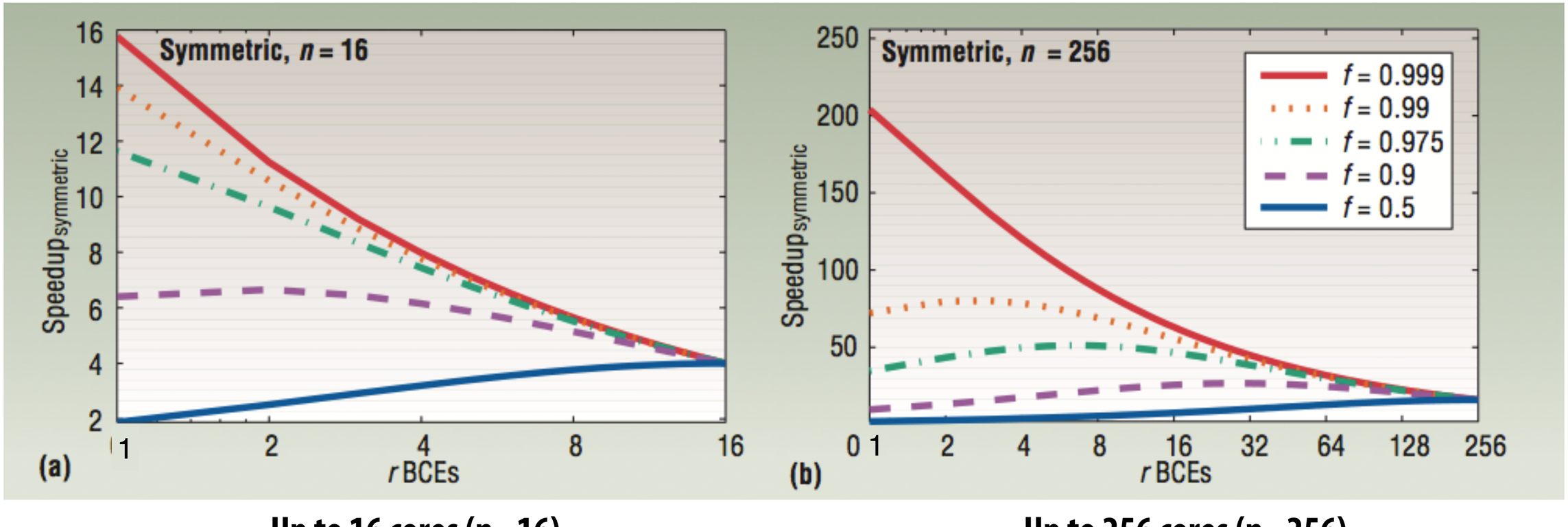
$r_A = 4$

$r_B = 1$



[Hill and Marty 08]

Speedup (relative to n=1)



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

Each line corresponds to a different chip configuration

All lines on the same graph correspond to configuration with same number of resources (constant n per chip)

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

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

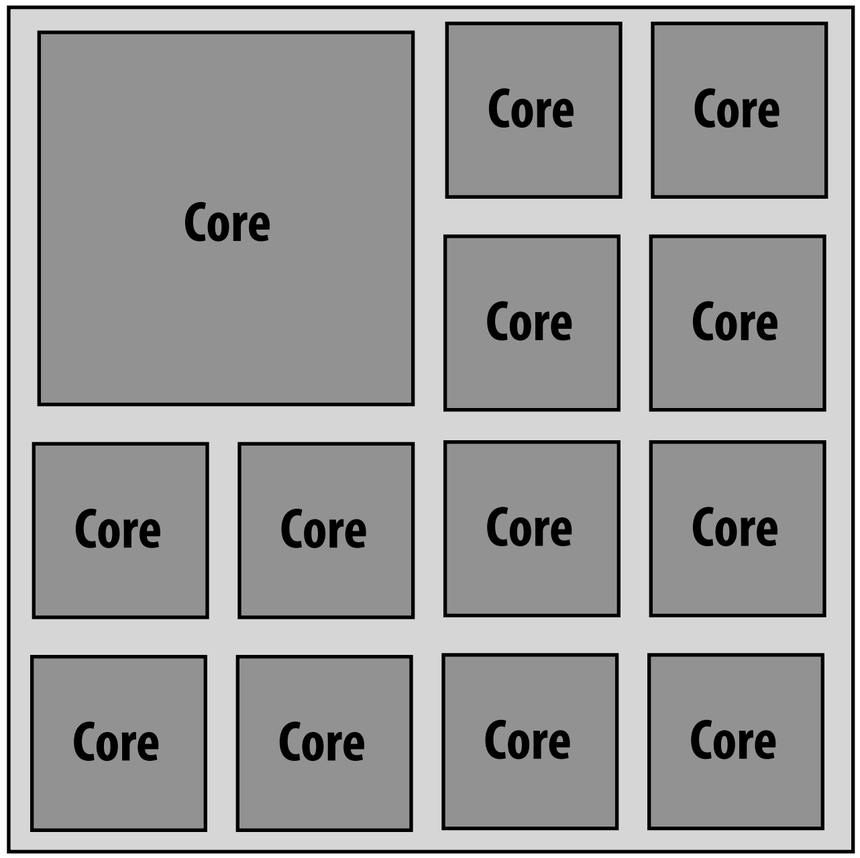
[Figure credit: Hill and Marty 08]

Asymmetric set of processing cores

Example: n=16

One core: $r = 4$

Other 12 cores: $r = 1$



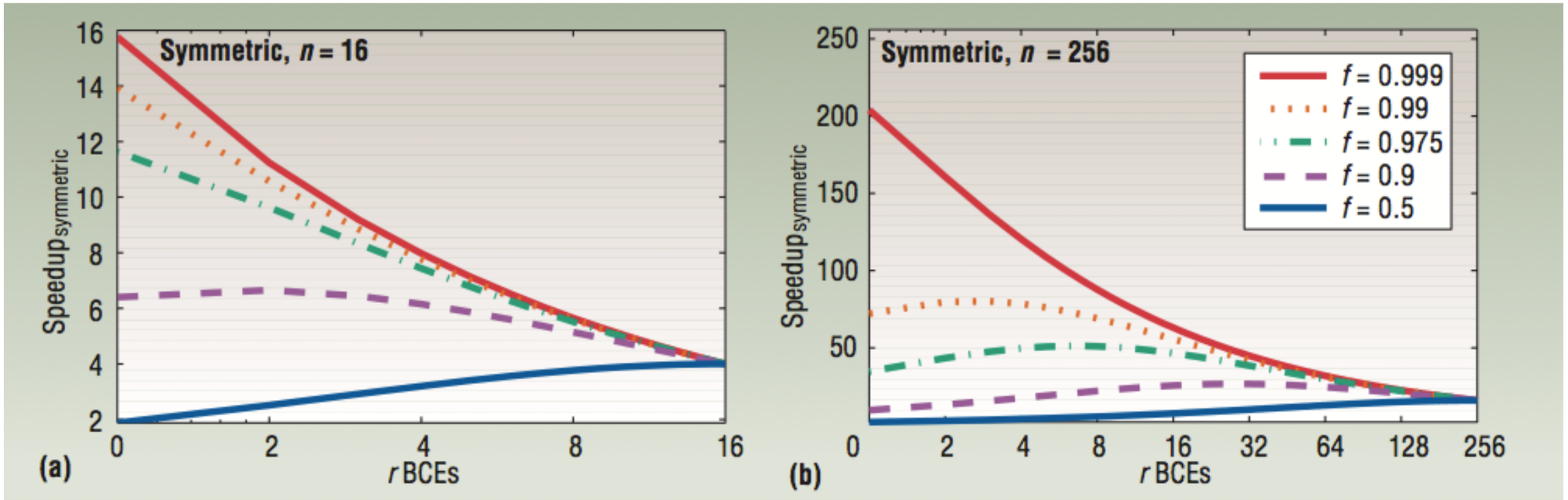
$$speedup(f,n,r) = \frac{1}{\frac{(1-f)}{perf(r)} + \frac{f}{perf(r)+(n-r)}}$$

(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

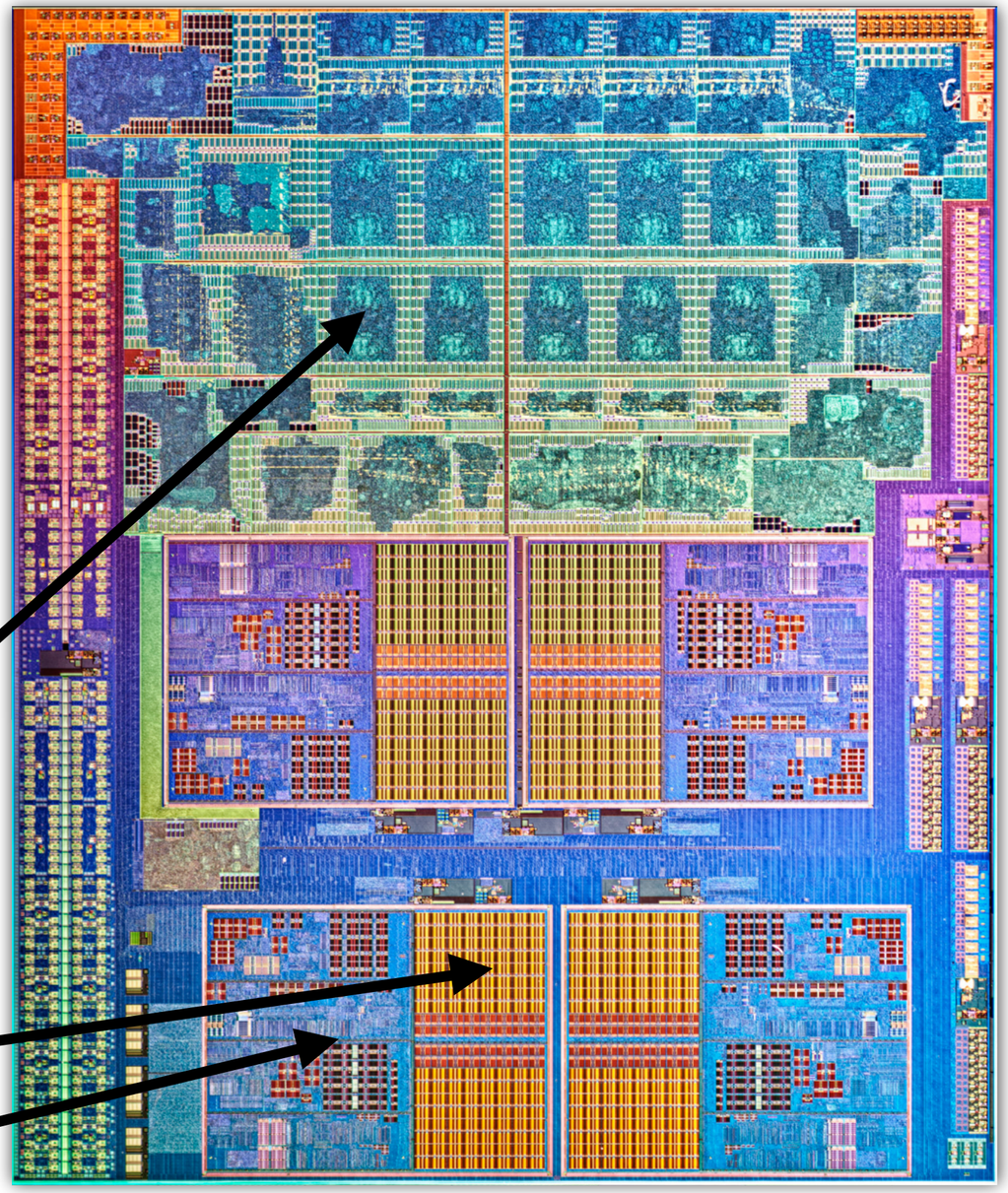
And components with unpredictable access, but those accesses might cache well.

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

Example: AMD Fusion

- "APU": accelerated processing unit
- Integrate CPU cores and GPU-style cores on same chip
- Share memory system
  - Regions of physical memory reserved for graphics (not x86 coherent)
  - Rest of memory is x86 coherent
  - CPU and graphics memory spaces are not coherent, but at least there is no need to copy data in physical memory (or over PCIe bus) to communicate between CPU and graphics

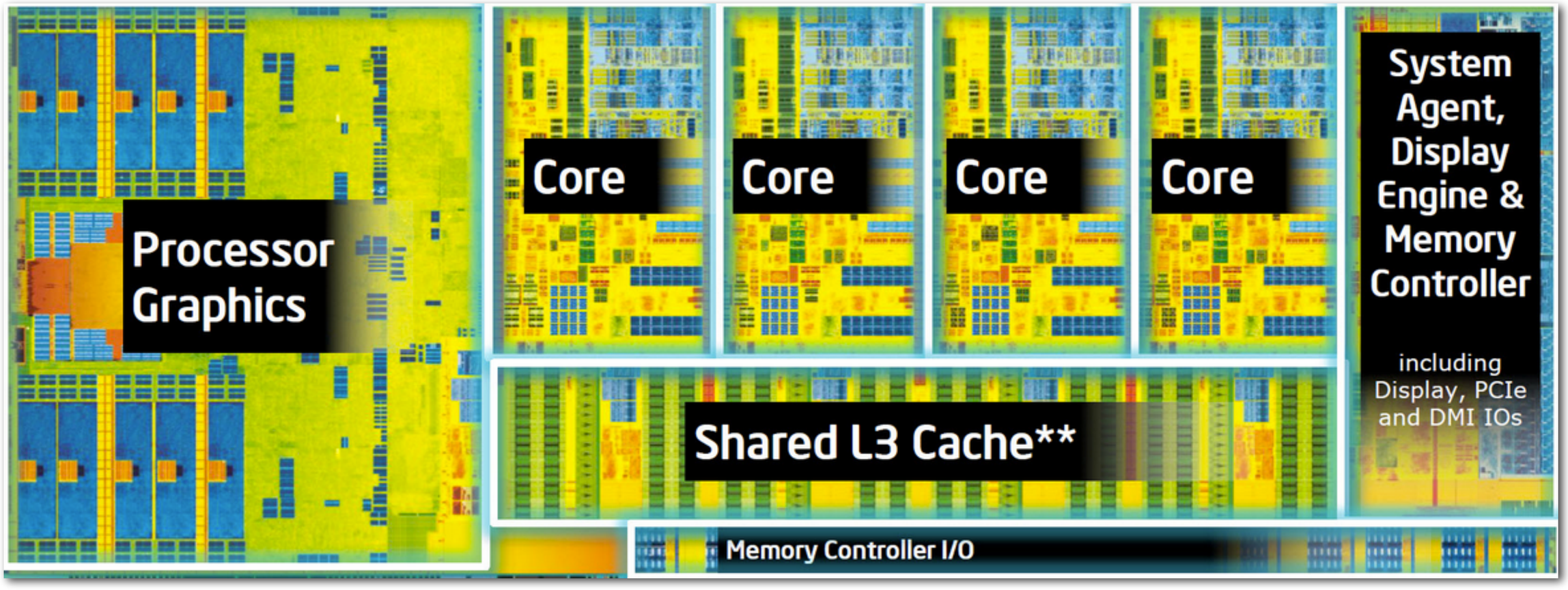


AMD Llano
(4 CPU cores + integrated GPU cores)

Example: Intel “Haswell" (2013)

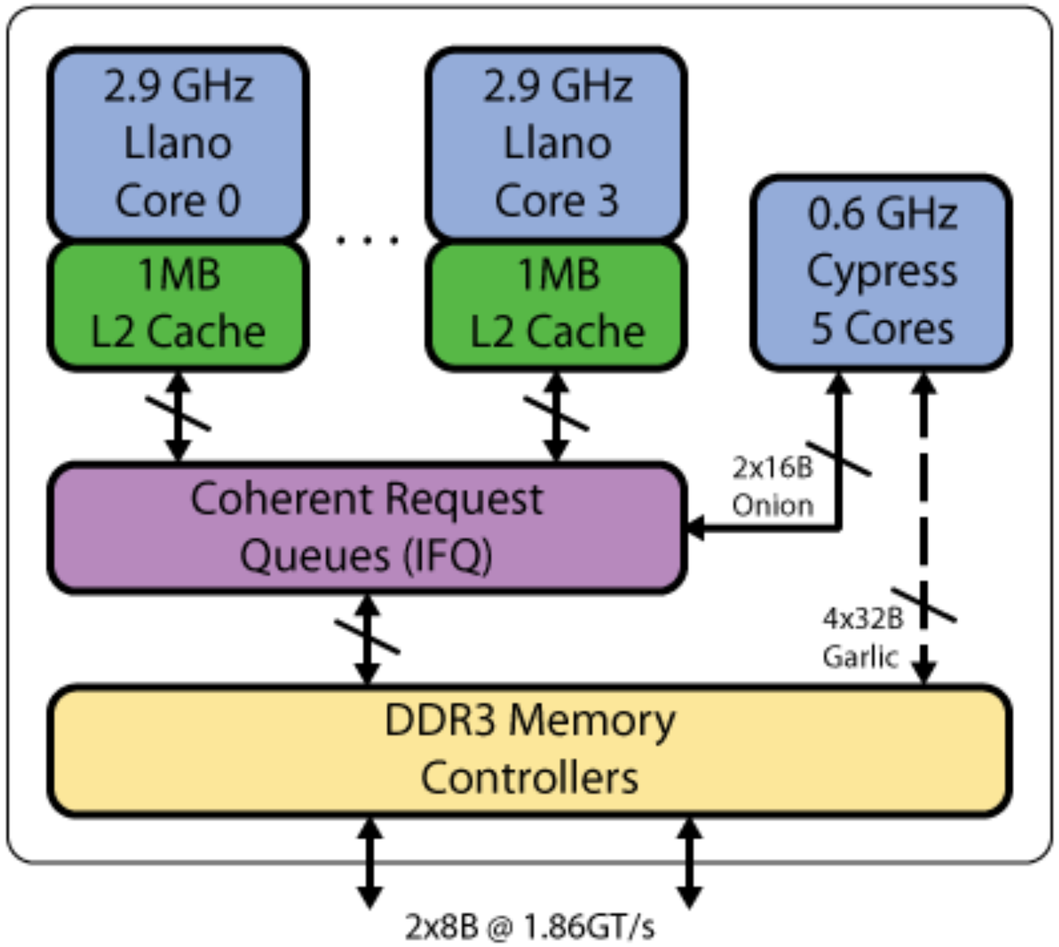
(4th Generation Core i7 architecture)

Four CPU cores + many GPU cores

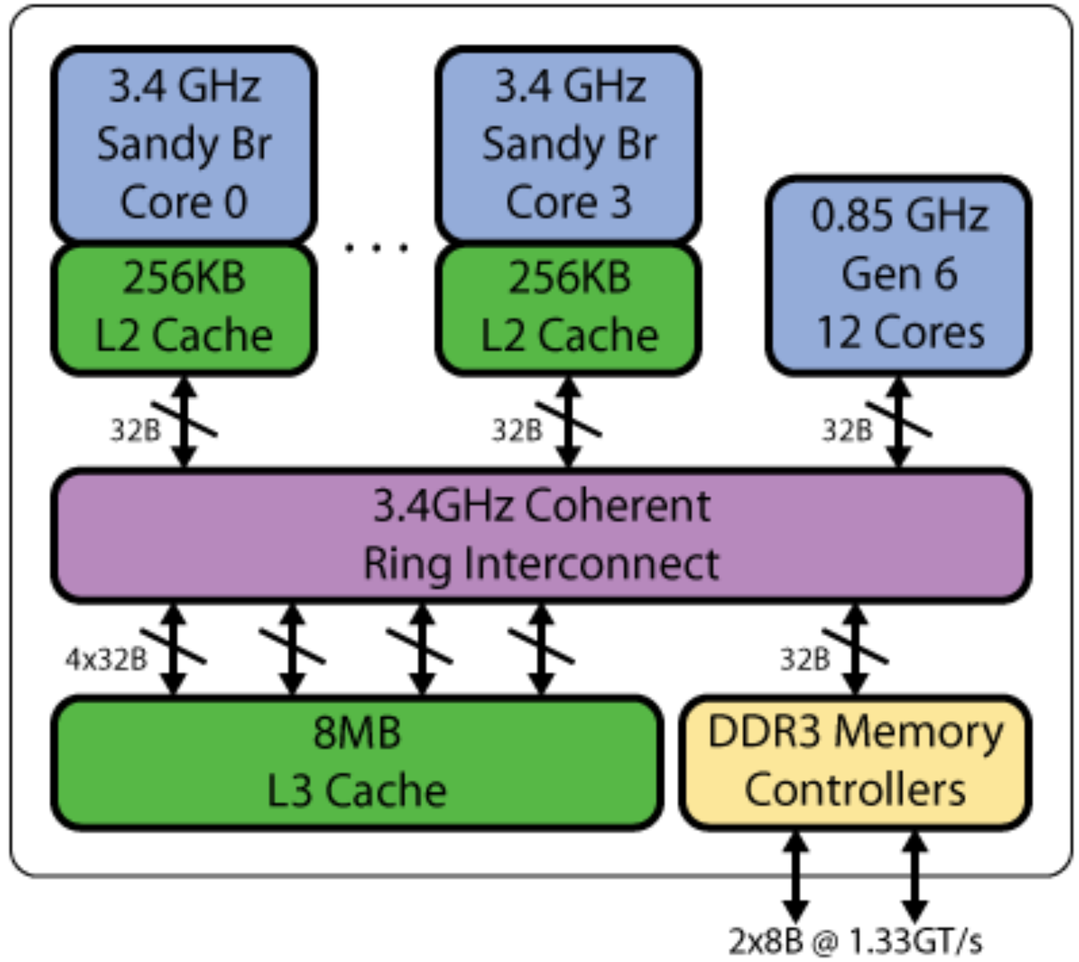


Multi-core CPU + integrated GPU

AMD Llano



Intel Sandy Bridge



More heterogeneity: add discrete GPU

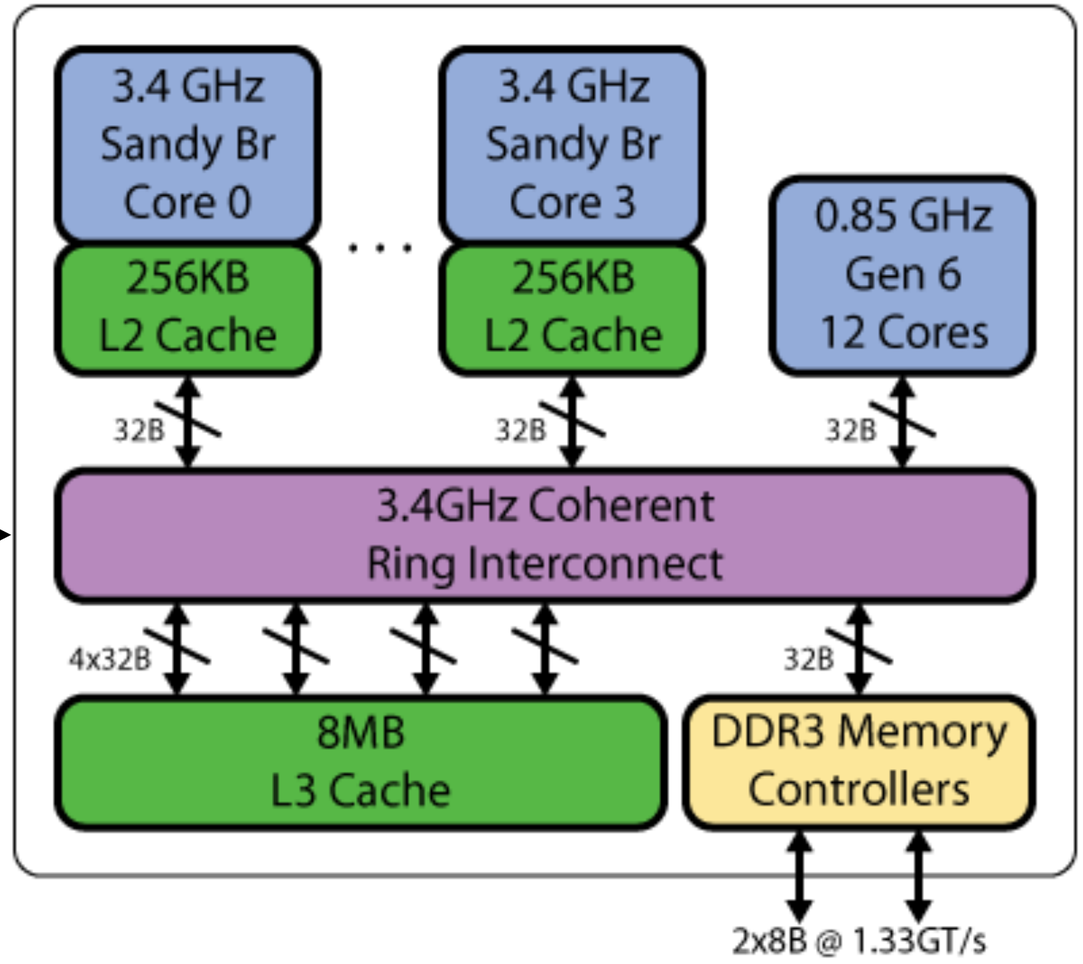
Keep discrete (power hungry) GPU unless needed for graphics-intensive applications
Use integrated, low power graphics for basic graphics/window manager/UI

Discrete high-end GPU (AMD or NVIDIA)

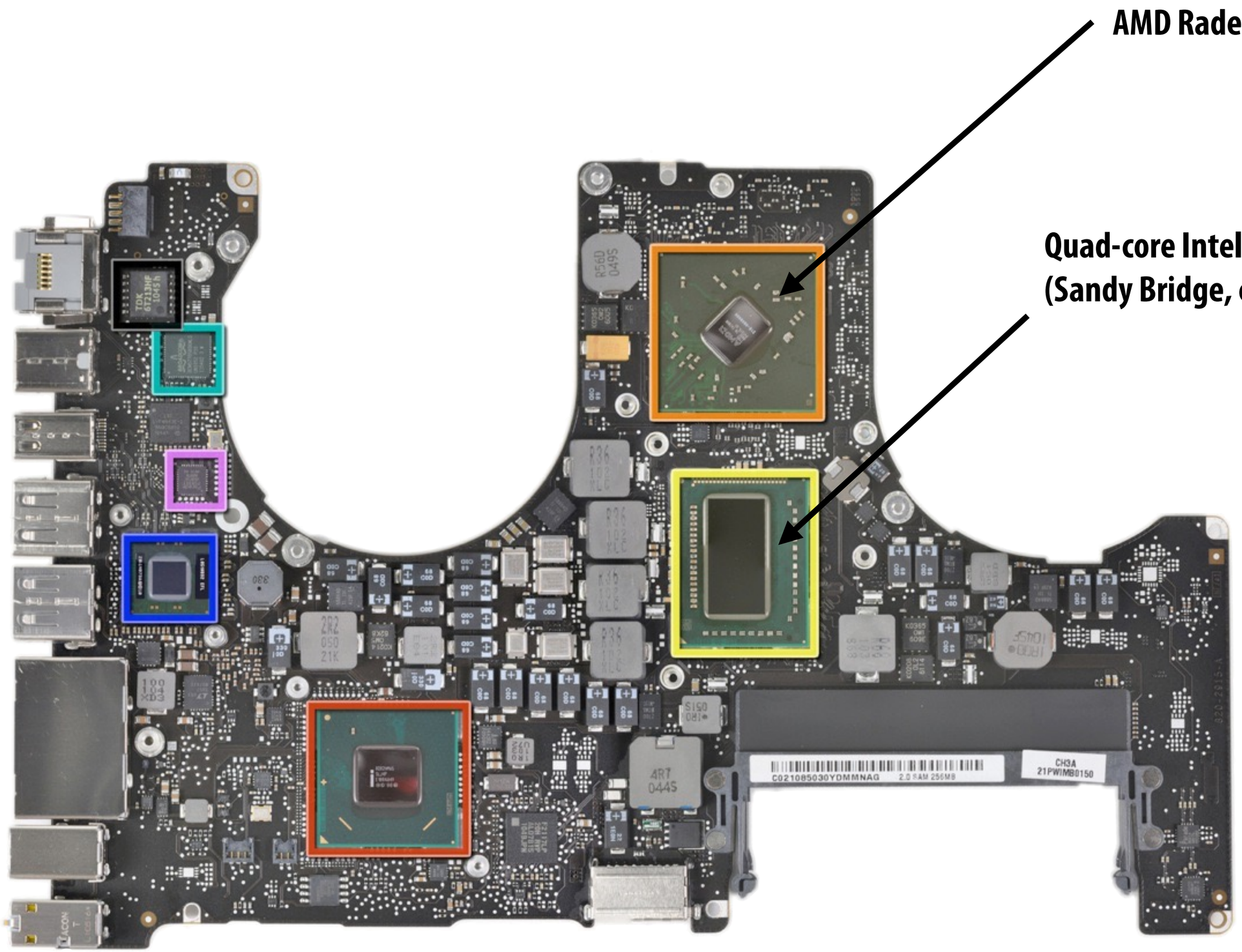
PCIe bus

DDR5 Memory

Intel Sandy Bridge

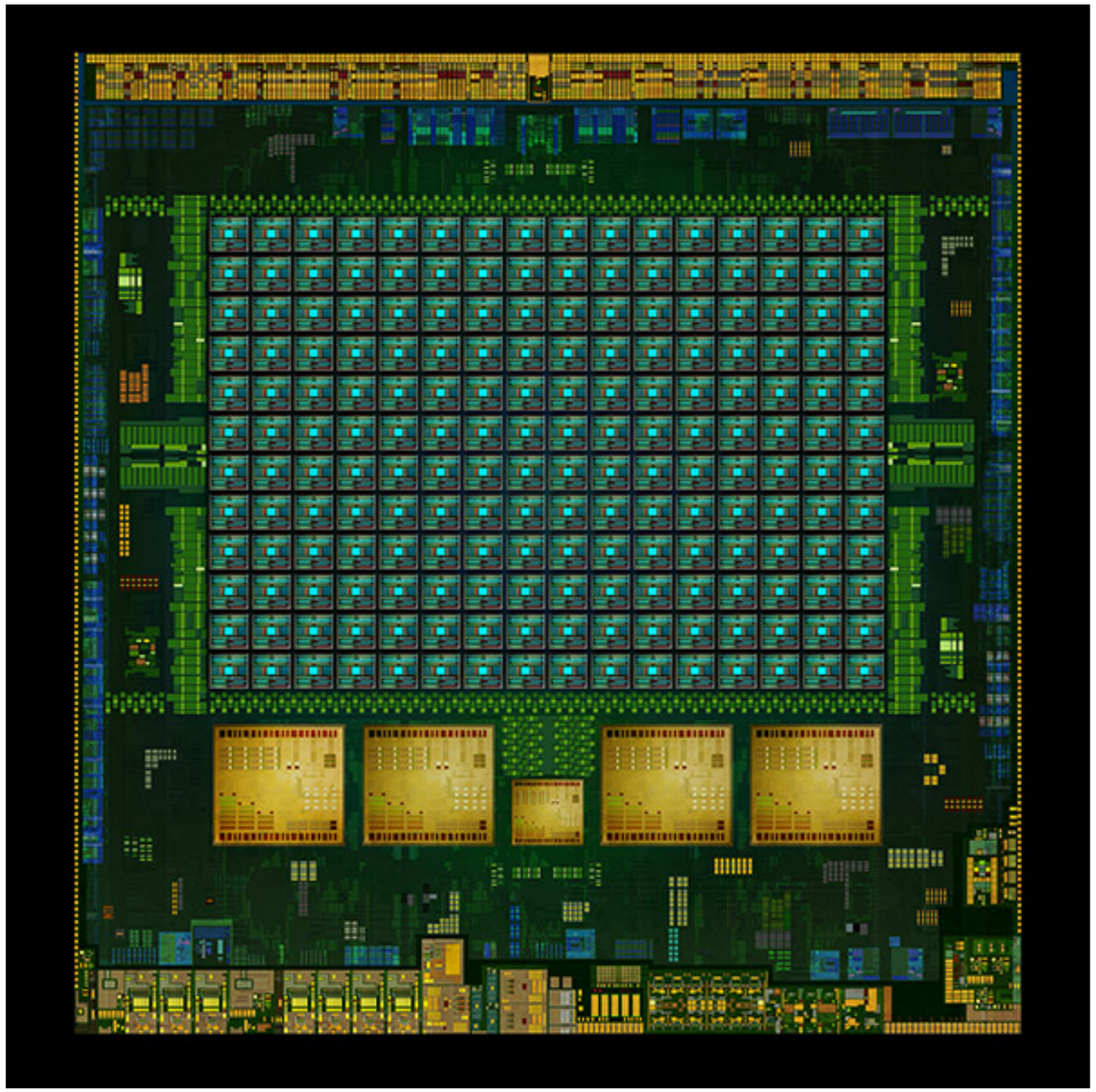


Macbook Pro 2011 (two GPUs)



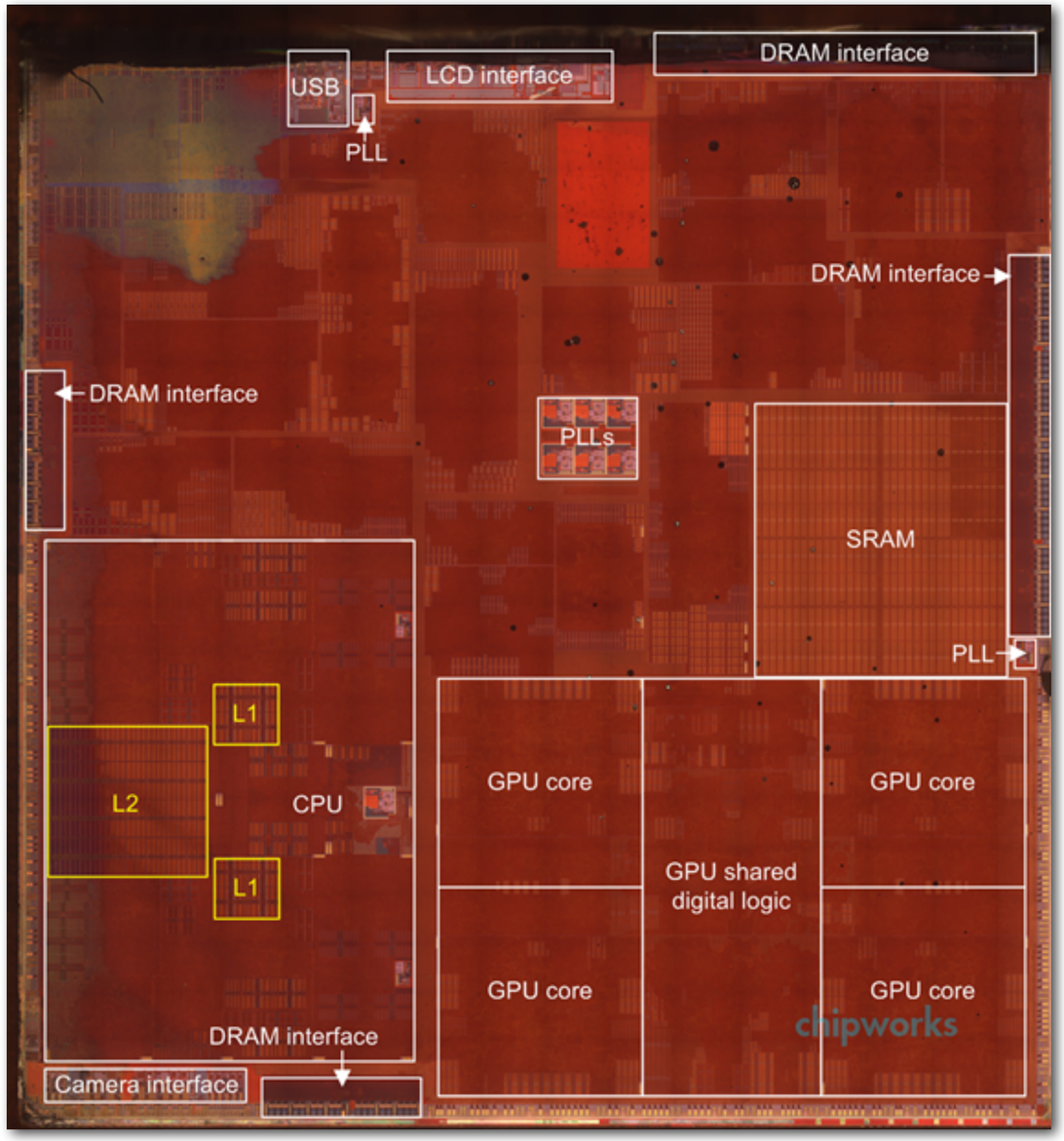
From ifixit.com teardown

Mobile heterogeneous processors



NVIDIA Tegra K1
Four 32-bit ARM Cortex A15 CPU cores for applications *
One low performance (low power) ARM CPU core
One Kepler SMX core (192 "CUDA" cores)

*** 64-bit dual core version also available**



Apple A7
Dual Core 64 bit CPU, 1.3GHz
GPU PowerVR 6430 @ 450 MHz

Image credit Chipworks, obtained from [http://thetechblock.com/history-apple-socs-predicting-a8-chip/]

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)**

GPU-accelerated supercomputing

- Oak Ridge Titan (world's #2)

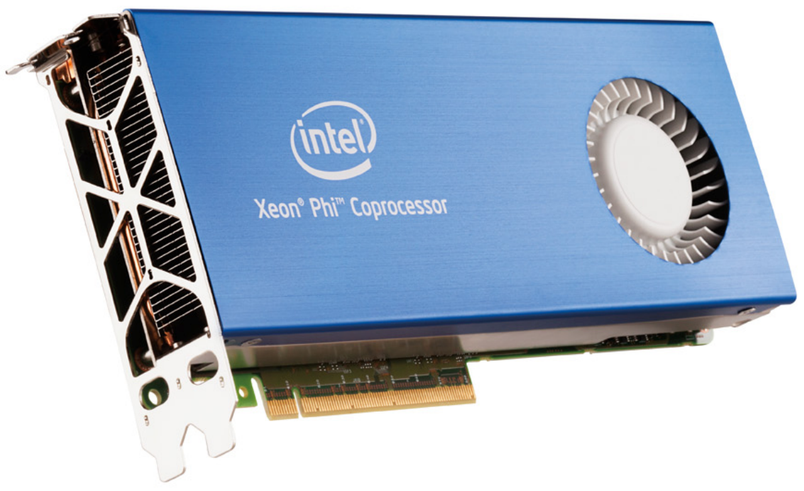
- 18,688 AMD Opteron 16-core CPUs
- 18,688 NVIDIA Tesla K20X GPUs
- 710 TB RAM

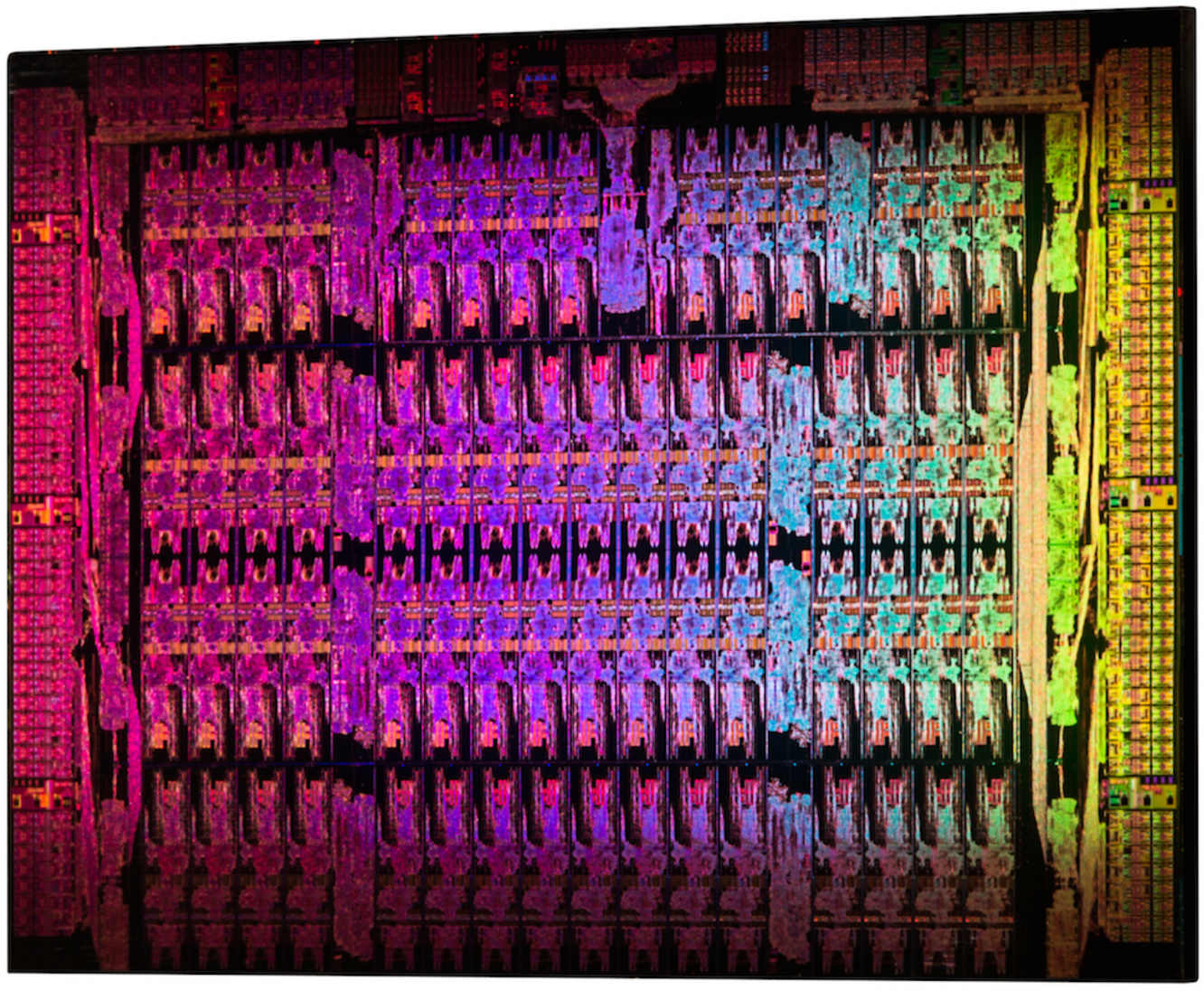
- Estimated machine cost $97M
- Estimated annual power/operating cost: ~ $9M *

[Source: NPR]

Intel Xeon Phi compute "accelerator"

- 60 "simple" x86 cores (1.1 Ghz, derived from Pentium)
- 16-wide vector instructions (AVX-512), four threads per core
- Targeted as an accelerator for supercomputing applications





Heterogeneous architectures for supercomputing

Source: Top500.org Fall 2014 rankings

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
8	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301

Xeon Phi

GPU

54 PFLOPS, 17.8 MWatt

GPU

Xeon Phi

Green500: most energy efficient supercomputers

Source: Green500 Fall 2014 rankings

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	5,271.81	GSI Helmholtz Center	L-CSC - ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150 Level 1 measurement data available	57.15
2	4,945.63	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	37.83
3	4,447.58	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	35.39
4	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m Level 3 measurement data available	44.54
5	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
6	3,543.32	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60
7	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray CS300 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
8	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m	44.40
9	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
10	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41

Energy-constrained computing

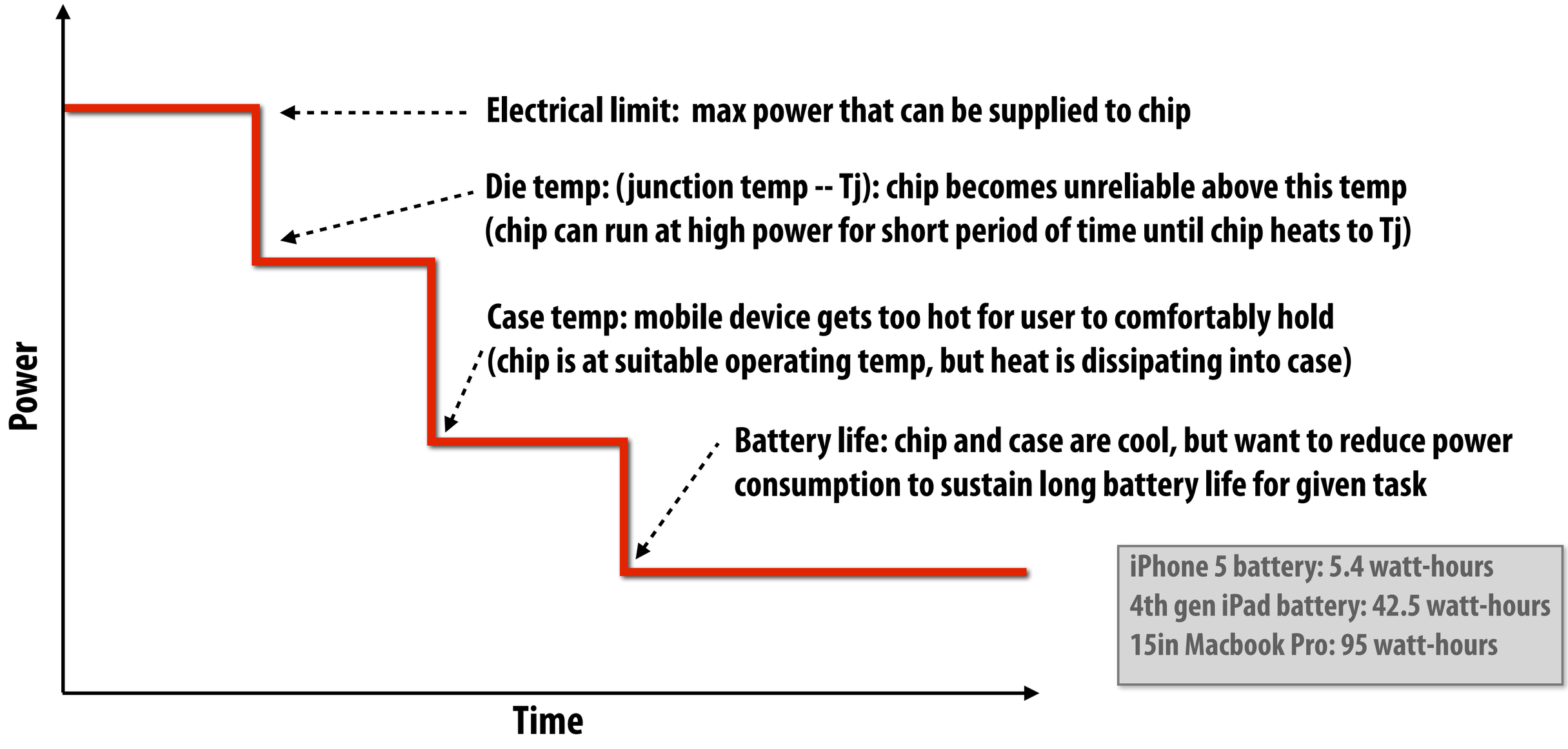
- Supercomputers are energy constrained
  - Due to shear scale
  - Overall cost to operate (power for machine and for cooling)

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

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

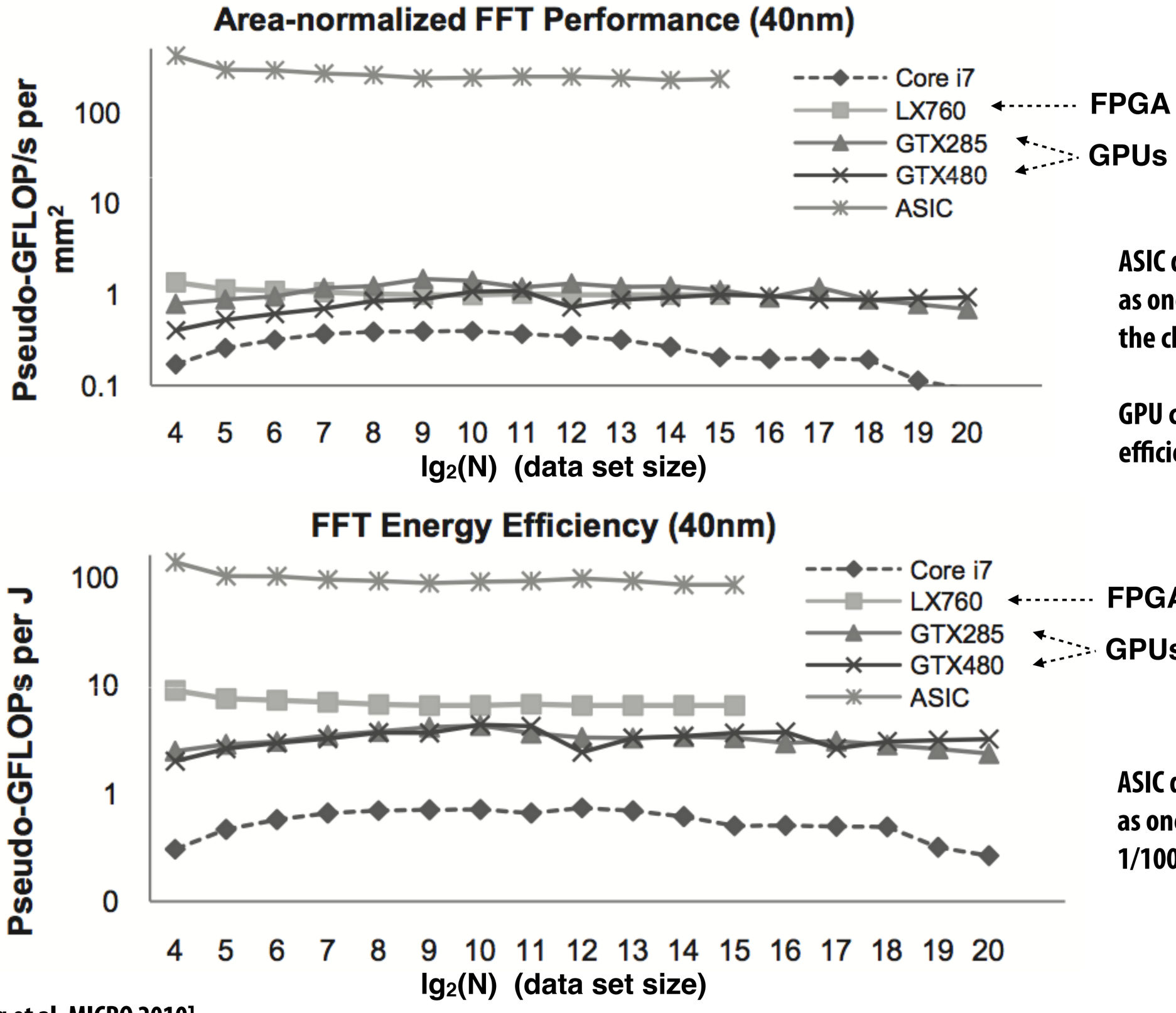
Limits on chip power consumption

- 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

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.

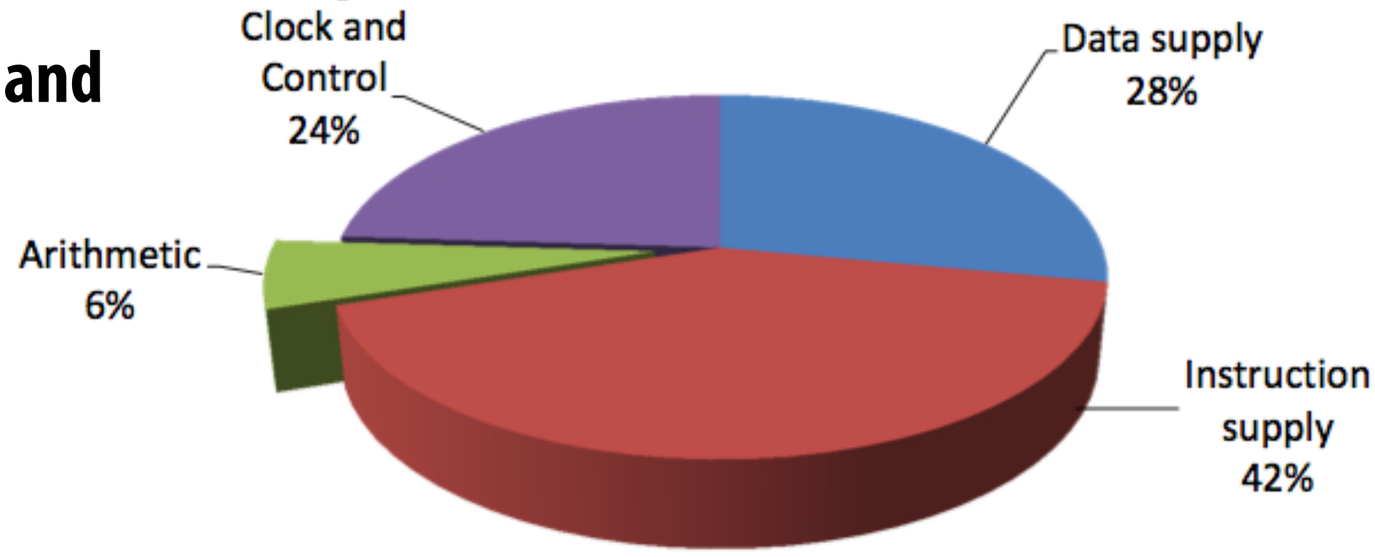
[Chung et al. MICRO 2010]

Efficiency benefits of compute specialization

- Rules of thumb: compared to good-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 100x 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]

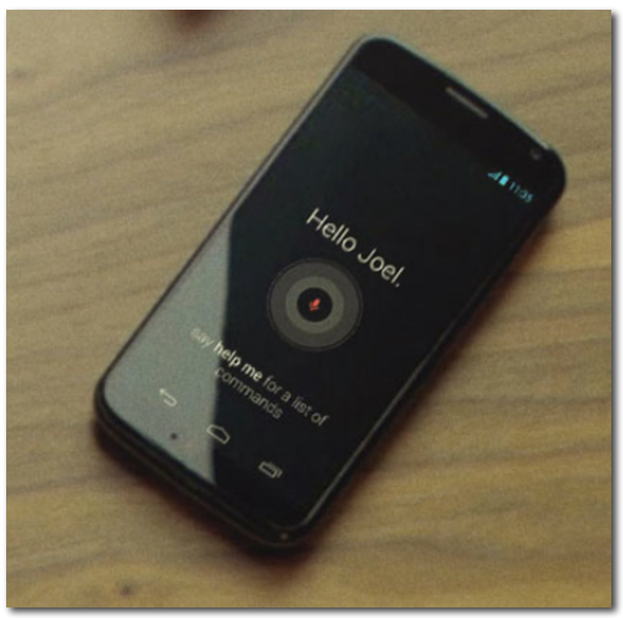
[Figure credit Eric Chung]

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

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.

Example: iPad Air (2013)

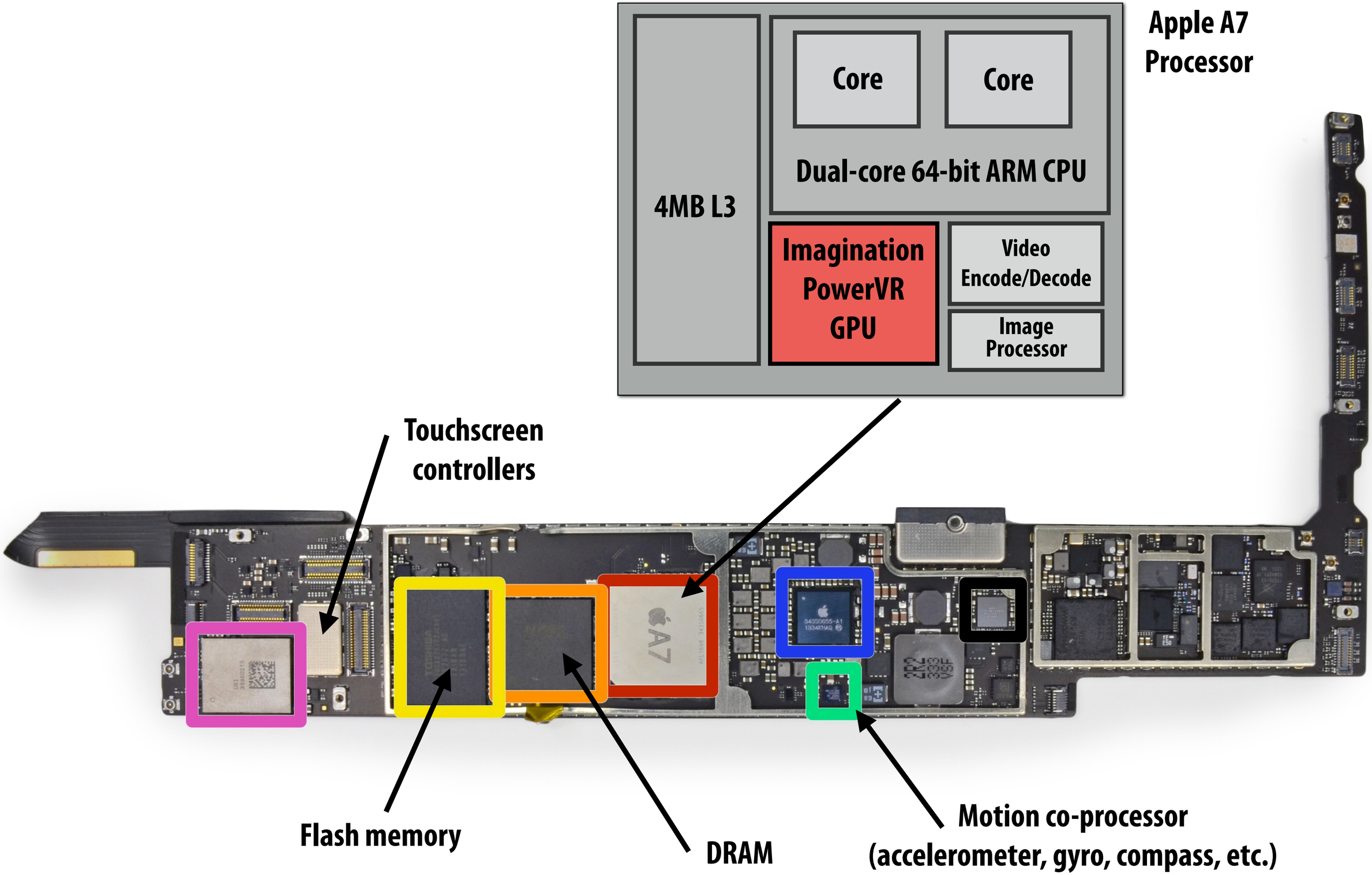


Image Credit: ifixit.com

Original iPhone touchscreen controller

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

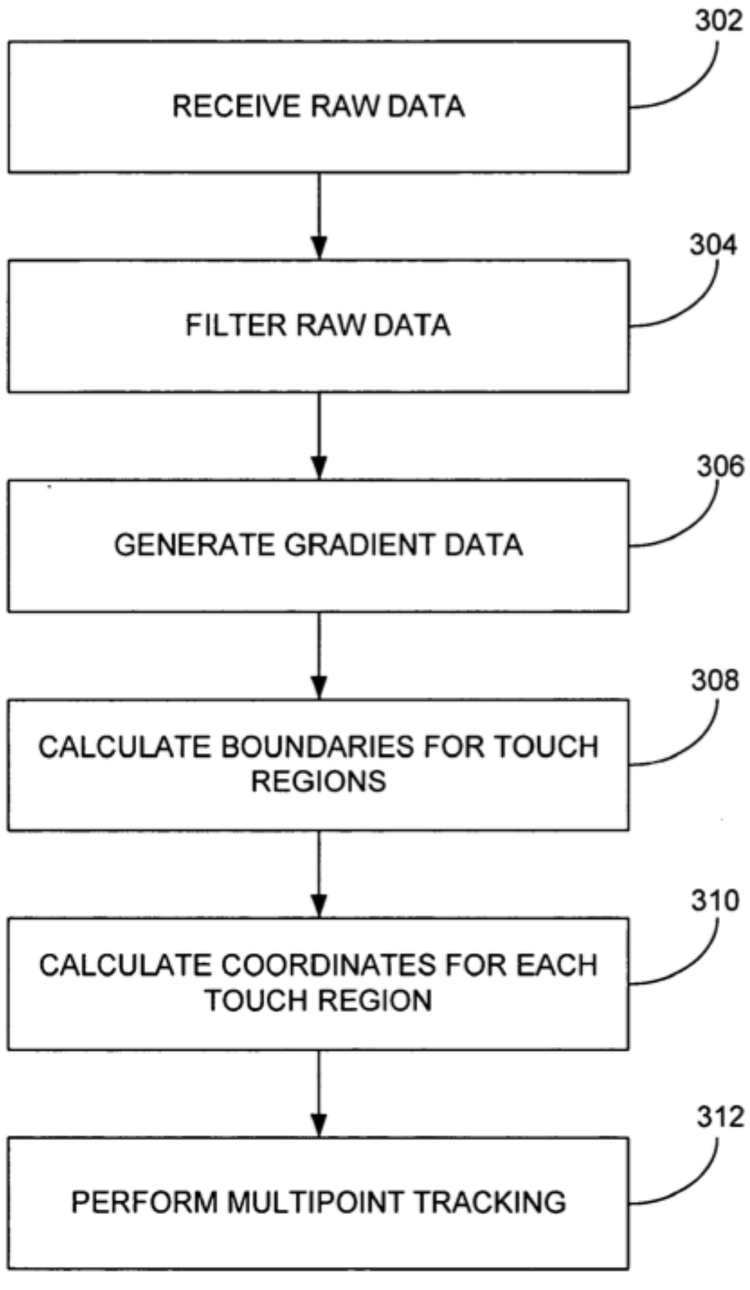


FIG. 16

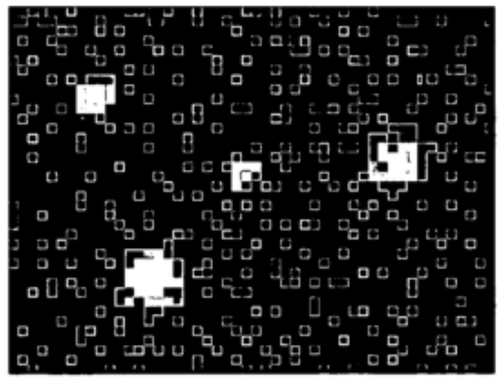


FIG. 17A

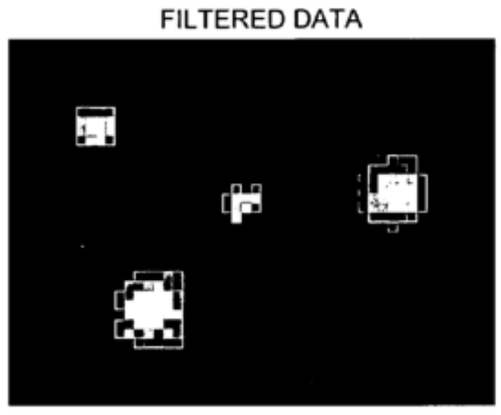


FIG. 17B

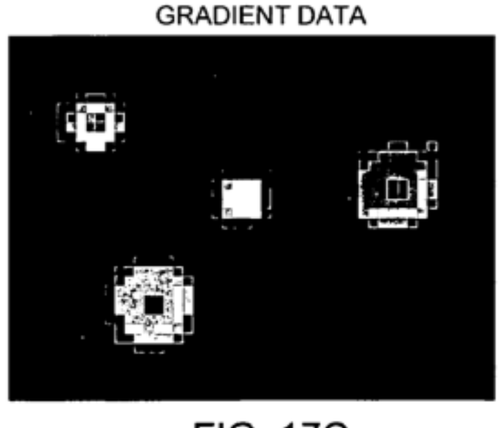


FIG. 17C

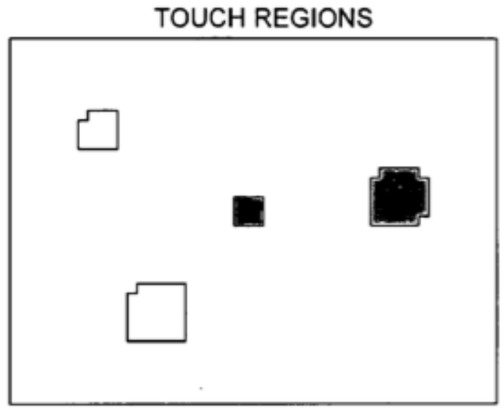


FIG. 17D

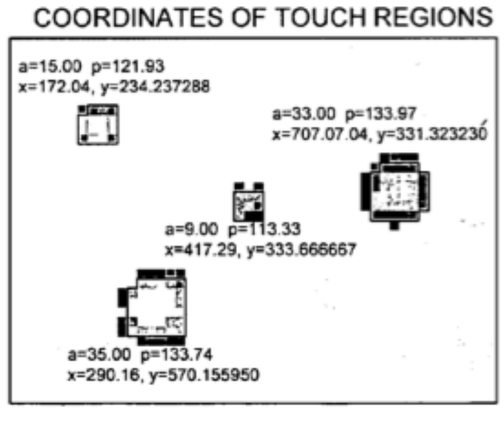


FIG. 17E

From US Patent Application 2006/0097991

Modern computing: efficiency often matters more than in the past, not less

Fourth, there's battery life.

To achieve long battery life when playing video, mobile devices must decode the video in hardware; decoding it in software uses too much power. Many of the chips used in modern mobile devices contain a decoder called H.264 – an industry standard that is used in every Blu-ray DVD player and has been adopted by Apple, Google (YouTube), Vimeo, Netflix and many other companies.

Although Flash has recently added support for H.264, the video on almost all Flash websites currently requires an older generation decoder that is not implemented in mobile chips and must be run in software. The difference is striking: on an iPhone, for example, H.264 videos play for up to 10 hours, while videos decoded in software play for less than 5 hours before the battery is fully drained.

When websites re-encode their videos using H.264, they can offer them without using Flash at all. They play perfectly in browsers like Apple's Safari and Google's Chrome without any plugins whatsoever, and look great on iPhones, iPods and iPads.

Steve Jobs' "Thoughts on Flash", 2010

http://www.apple.com/hotnews/thoughts-on-flash/

Example: image processing on my Nikon D7000

Process 16 MPixel RAW data from sensor to obtain JPG image:

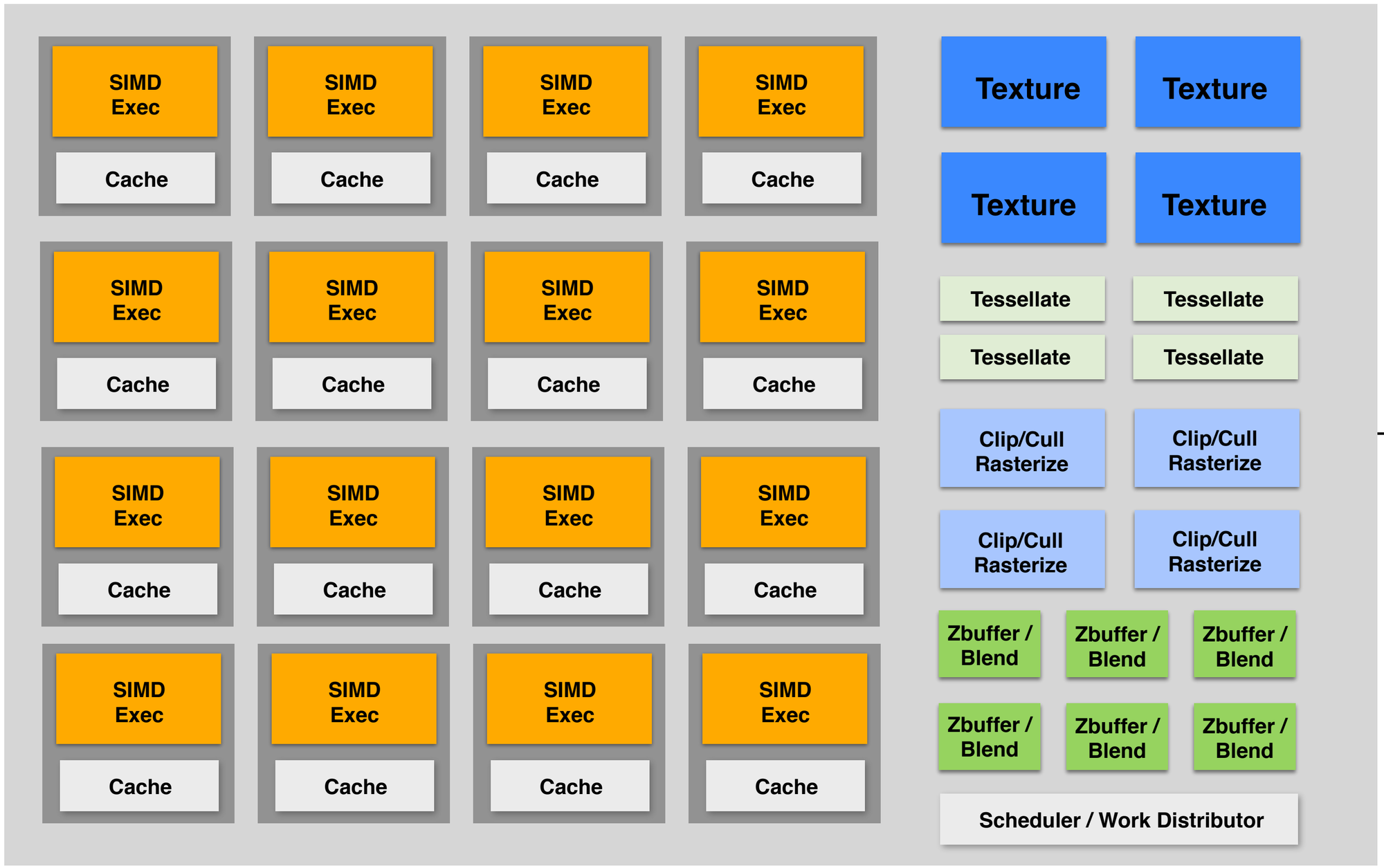
On camera: ~ 1/6 sec per image

Adobe Lightroom on my quad-core Macbook Pro laptop: 1-2 sec per image

GPU is itself a heterogeneous multi-core processor

Compute resources your CUDA programs used in assignment 2

Graphics-specific, fixed-function compute resources

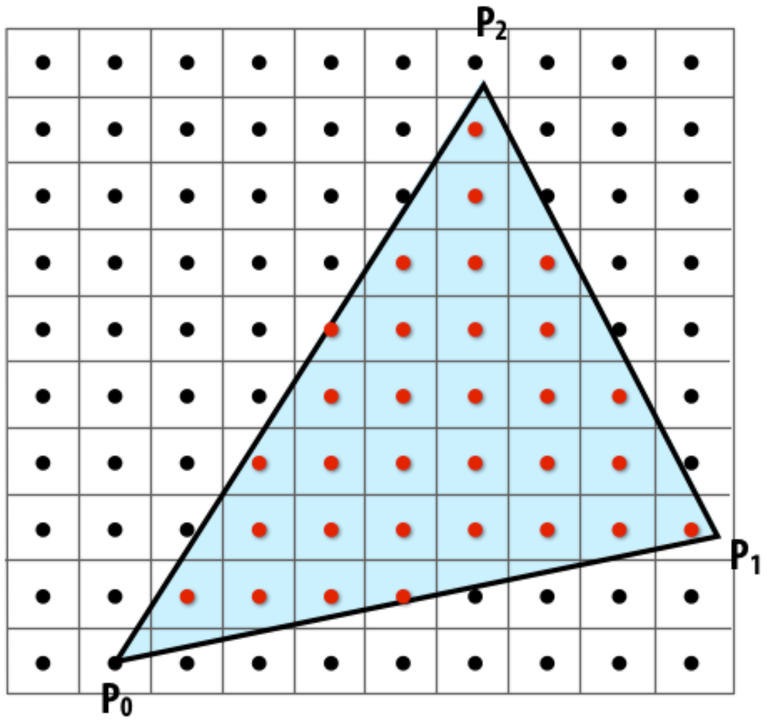


GPU

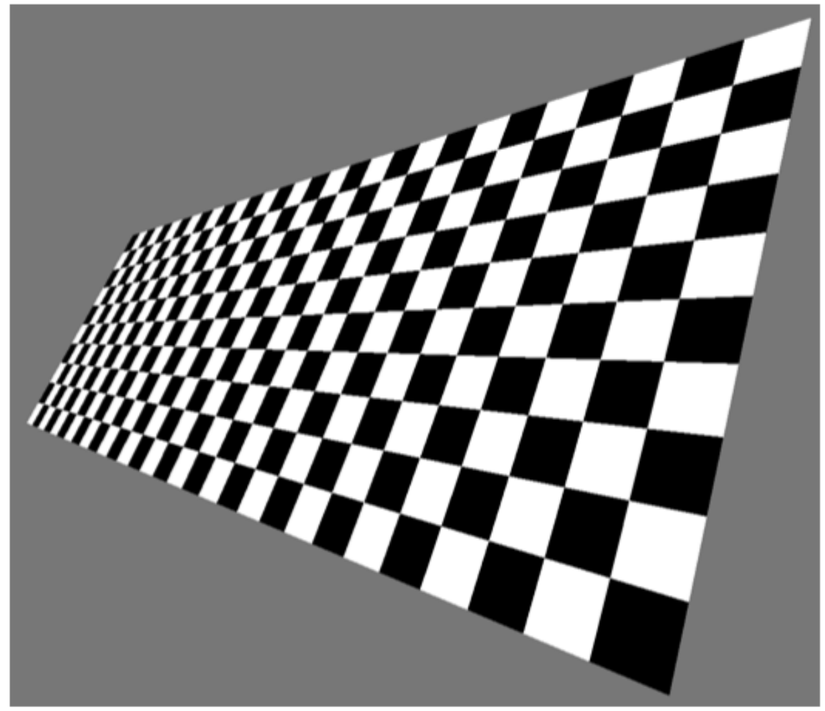
GPU Memory

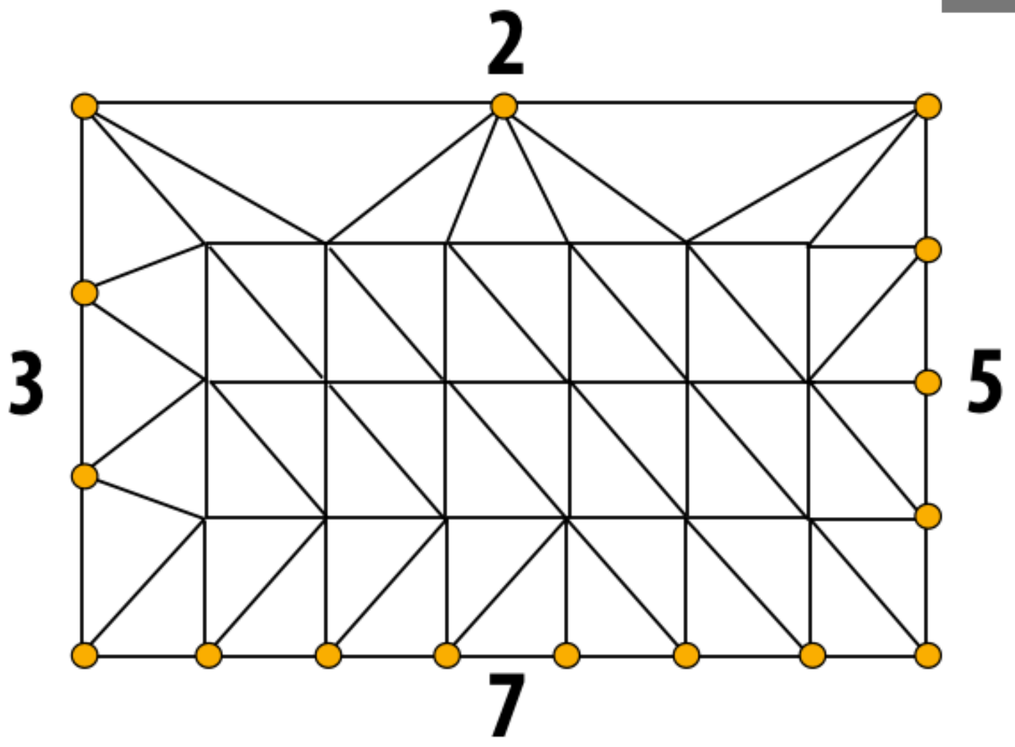
Example graphics tasks performed in fixed-function HW

Rasterization:
Determining what pixels a triangle overlaps



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

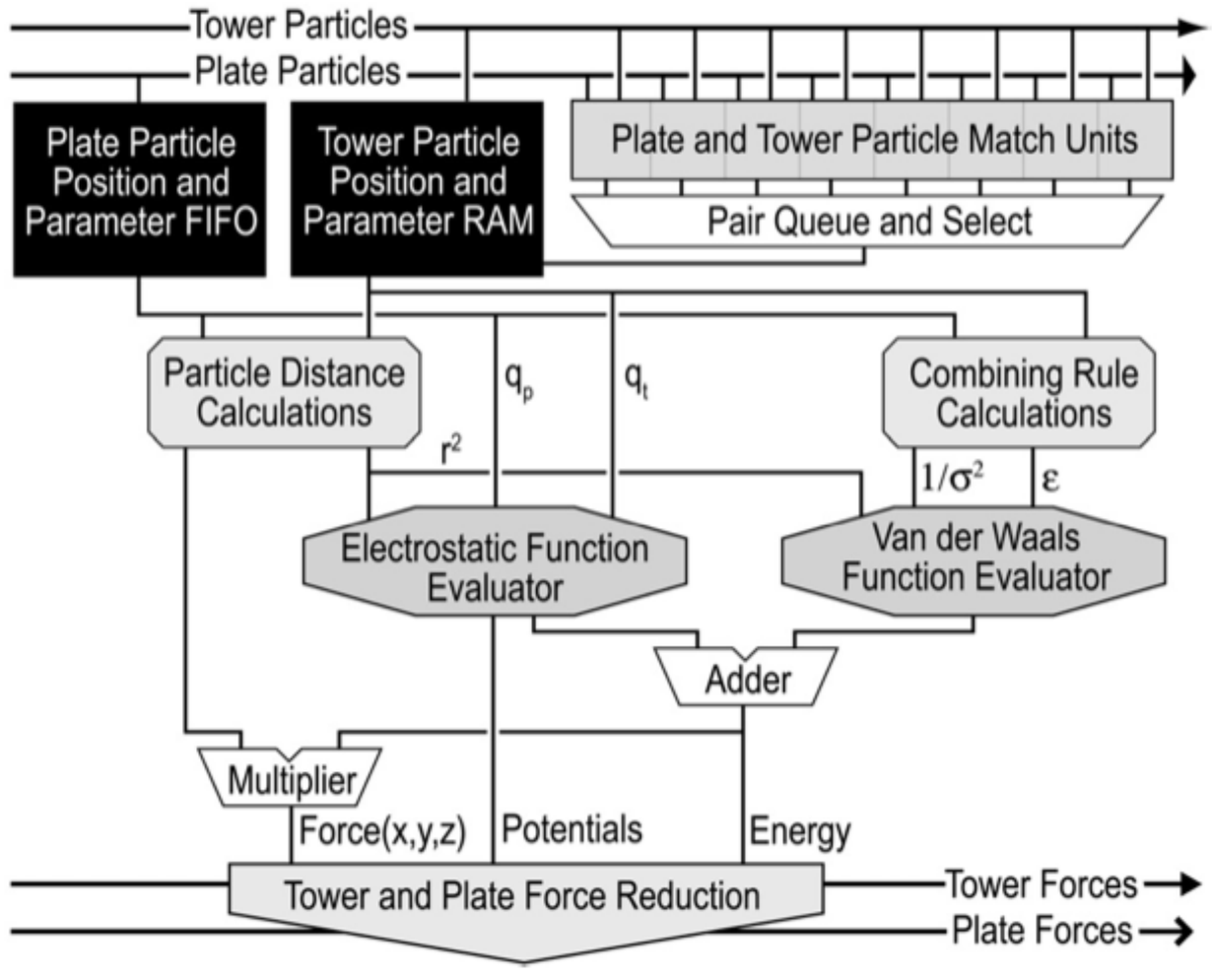




Geometric tessellation:
computing fine-scale geometry from coarse geometry

DESRES Anton supercomputer

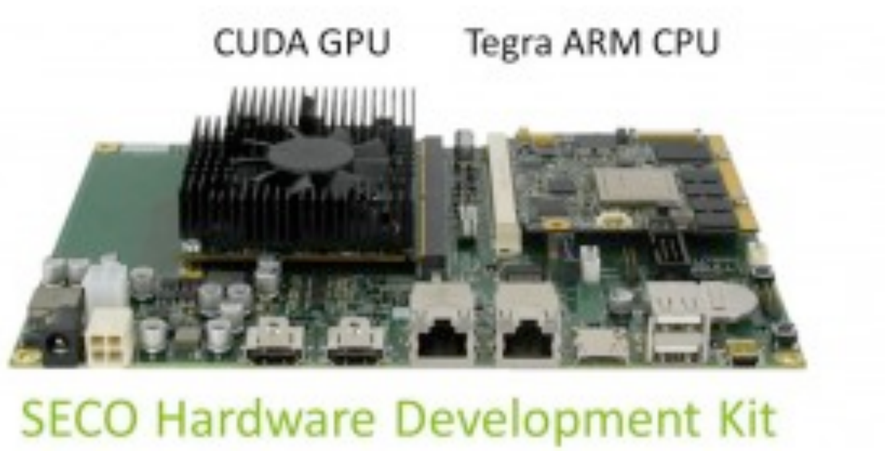
- Supercomputer highly specialized for molecular dynamics
  - Simulates time evolution of proteins
- ASIC for computing particle-particle interactions (512 of them in machine)
- Throughput-oriented subsystem for efficient fast-fourier transforms
- Custom, low-latency communication network designed for communication patterns of N-body simulations



Research: ARM + GPU Supercomputer

- Observation: the heavy lifting in supercomputing applications is the data-parallel part of workload
  - Less need for "beefy" sequential performance cores
- Idea: build supercomputer out of power-efficient building blocks
  - ARM CPUs (for control/scheduling) + GPU cores (primary compute engine)
- Goal: 7 GFLOPS/Watt efficiency
- Project underway at Barcelona Supercomputing Center

  http://www.montblanc-project.eu



Challenges of heterogeneity

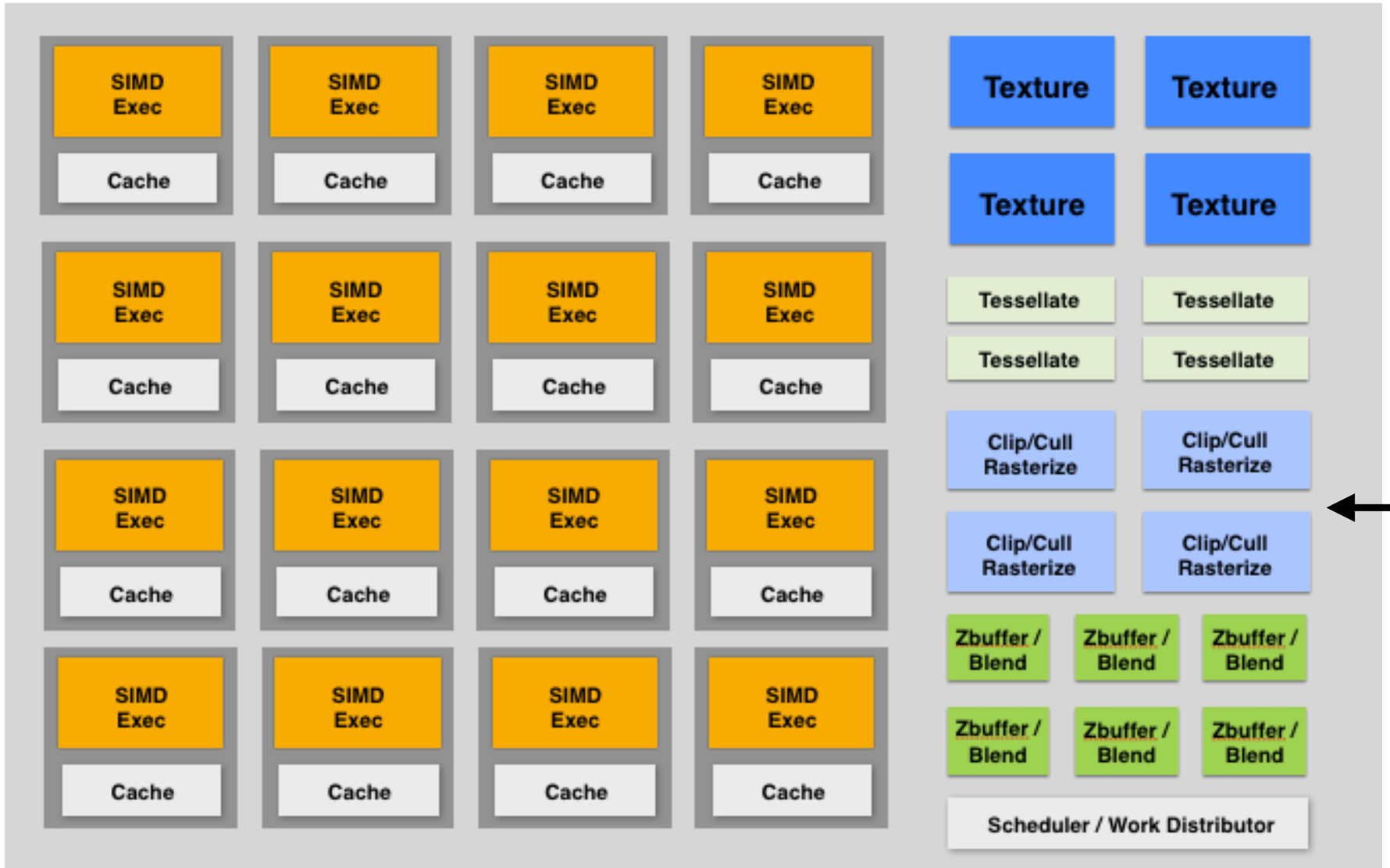
- So far in this course:
  - Homogeneous system: every processor can be used for every task
  - To get best speedup vs. sequential execution, "keep all processors busy all the time"

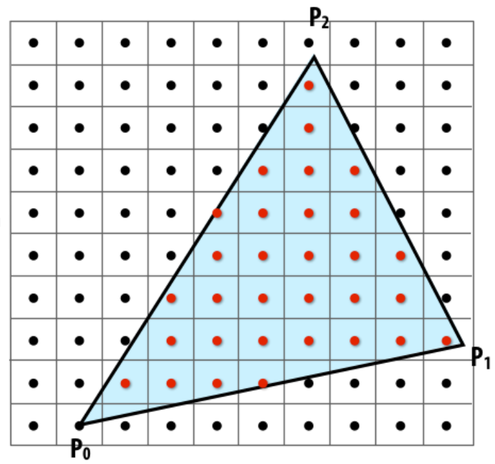
- Heterogeneous system: use preferred processor for each task
  - Challenge for system designer: what is the right mixture of resources to meet performance, cost, and energy goals?
    - Too few throughput-oriented resources (lower peak performance/efficiency for parallel workloads -- should have used 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)

- Increased pressure to understand workloads accurately at chip design time

Pitfalls of heterogeneous design

[Molnar 2010]





Say 10% of the computation is rasterization (most of graphics workload is computing color of pixels)
Let's say you under-provision the fixed-function rasterization unit on GPU:
(e.g., 1% of chip area used for rasterizer, really needed 20% mode: 1.2% of chip)

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!
Tendency is to be conservative, and over-provision fixed-function components (diminishing their advantage)

Challenges of heterogeneity

- **Heterogeneous system: preferred processor for each task**
  - Challenge for system 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 (we'll revisit in a future lecture on domain specific programming languages

Data movement has high energy cost

- Rule of thumb in mobile system design: 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
- 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 5 battery: ~5.5 watt-hours (note: my Macbook Pro laptop: 77 watt-hour battery)
  - 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-focused computing

- **Compute less!**
  - Computing more costs energy: parallel algorithms that do more work than sequential counterparts may not be desirable even if they run faster

- **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)

- **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

Fun reads about mobile power concerns

- **Power concerns in ARM Mali GPUs**
  - "How Low Can You Go? Building Low-Power, Low_bandwidth ARM Mali GPUs" - Tom Olson, ARM
  - http://community.arm.com/groups/arm-mali-graphics/blog/2012/08/17/how-low-can-you-go-building-low-power-low-bandwidth-arm-mali-gpus

- **Display Backlight measurements**
  - http://www.displaymate.com/iPad_ShootOut_1.htm#Backlight_Power

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?