CS4223 – Multi-core Architectures

Trevor E. Carlson
tcarlson@comp.nus.edu.sg
Assistant Professor – School of Computing

(Slides originally based on those from Tulika Mitra)

Who Am I?

- Assistant Professor here at NUS
  - – Looking to move computer architecture forward with new processors and accelerators
  - – Deliver more capabilities, accomplish more in a given area, energy or power budget

My Background

- IBM, USA
  - Global team leader, architecture validation
  - 4 patents: DRAM scrubbing, processor simulation
- IMEC, Belgium
  - Compiler/HW co-design
  - High-level hardware pathfinding
  - 3D-DRAM optimization
- Startups, Belgium & USA
  - University funding for computer architecture startup
  - IoT startup

My Background

- Ghent University, Belgium
  - PhD: Simulation and sampling methodologies
- Intel, MA, USA
  - PhD Internship to Speed up industry simulators
- Uppsala University, Sweden
  - Postdoc: Efficient architectures, simulation, modeling

Question – Who Are You?

- How many years have you been learning Computer Architecture?

Question – Who Are You?

- What do you find most interesting about Computer Architecture?
- Why did you sign up for this class?

Question – Who Are You?

- What experience do you have with Computer Architecture software / simulators?

Definition: Parallel Computer

- A parallel computer is a collection of processing elements that communicate and cooperate to solve a large problem fast

-- Almasi and Gottlieb 1989

Analyzing the definition

- Collection of processing elements
  - A processing element can be a functional unit, a thread context on a processor, a processor core, a processor chip, or an entire node (processors + memory + disk)
  - The de-factor processing elements are processors --- so parallel computers are also called multi-processors
  - When the processor cores are on a single chip, the system is referred to as multi-core

Analyzing the definition

- Communicate
  - Processing elements sending data to each other
  - Shared memory : parallel tasks running on processing elements communicate by reading and writing to common memory locations
  - Message passing: all data is local and parallel tasks must send explicit messages to each other to pass data
  - Communication medium and its structure determines communication latency, throughput, scalability, and fault tolerance

Analyzing the definition

- Cooperate
  - Synchronization of the progress of execution of a parallel task relative to other tasks
  - Synchronization allows sequencing of operations to ensure correctness
  - Synchronization granularity affects scalability and load balancing

Analyzing the definition

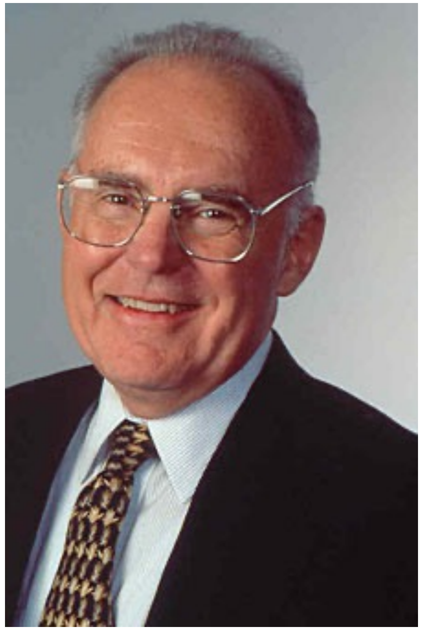
- Solve a large problem fast
  - Performance is a critical concern in parallel computing
  - Parallel computing attempts to solve challenging problems that cannot be handled by single processing elements

Why Multi-Core?

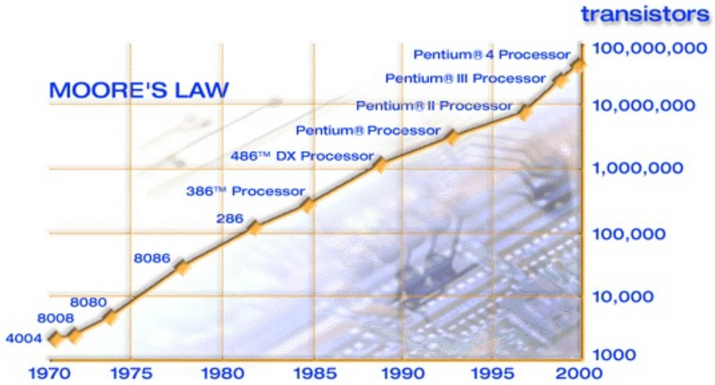
- Application demands
  - *Scientific computing*: CFD, Biology, Chemistry, Physics, ...
  - *General-purpose computing*: Video, Graphics, CAD, Databases, TP...
- Technology Trends
  - Number of transistors on chip growing rapidly
  - Clock rates expected to go up only slowly
- Architecture Trends
  - Instruction-level parallelism valuable but limited
  - Coarser-level parallelism, as in MPs, the most viable approach

Moore’s Law

- Intel co-founder Gordon Moore predicted in 1965 that Transistor density will double every 18 months => Increase in clock frequency
- Moore’s Law 50th Anniversary

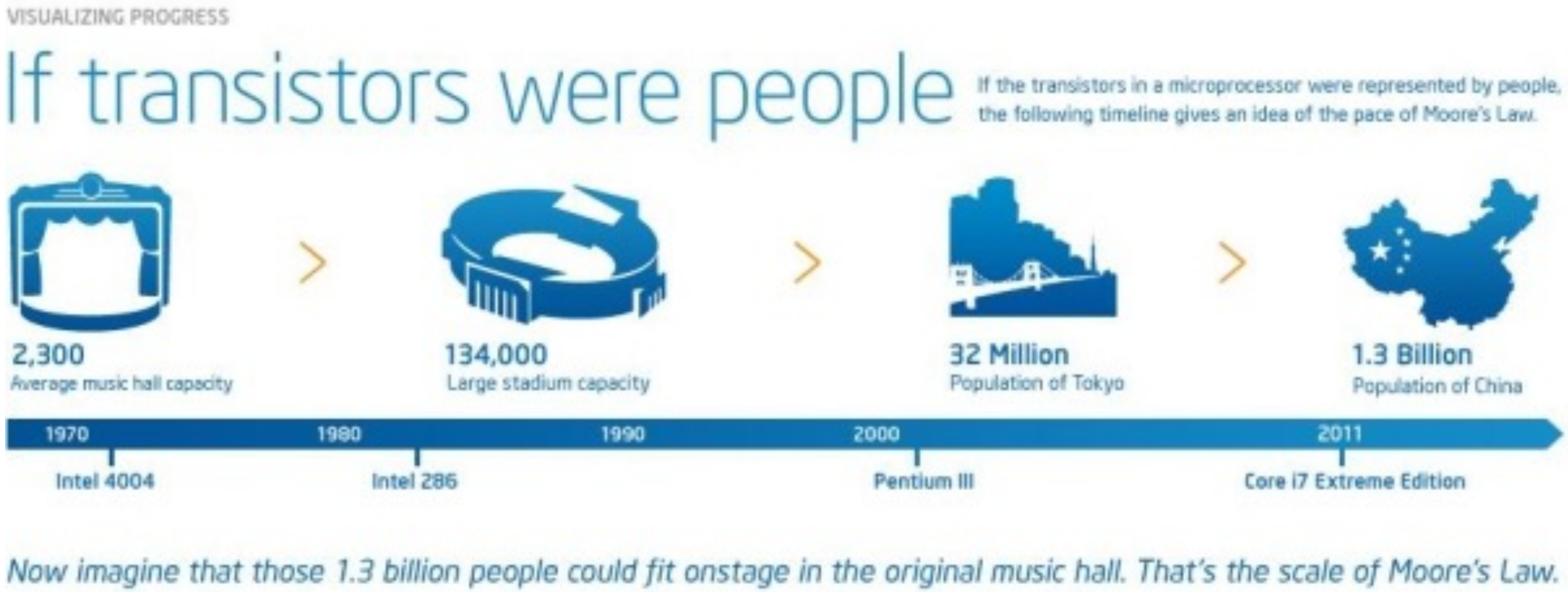


Gordon E. Moore, Co-founder, Intel Corporation.
Copyright © 2005 Intel



http://www.intel.com/research/silicon/mooreslaw.htm

Visualizing the Scale of Moore’s Law



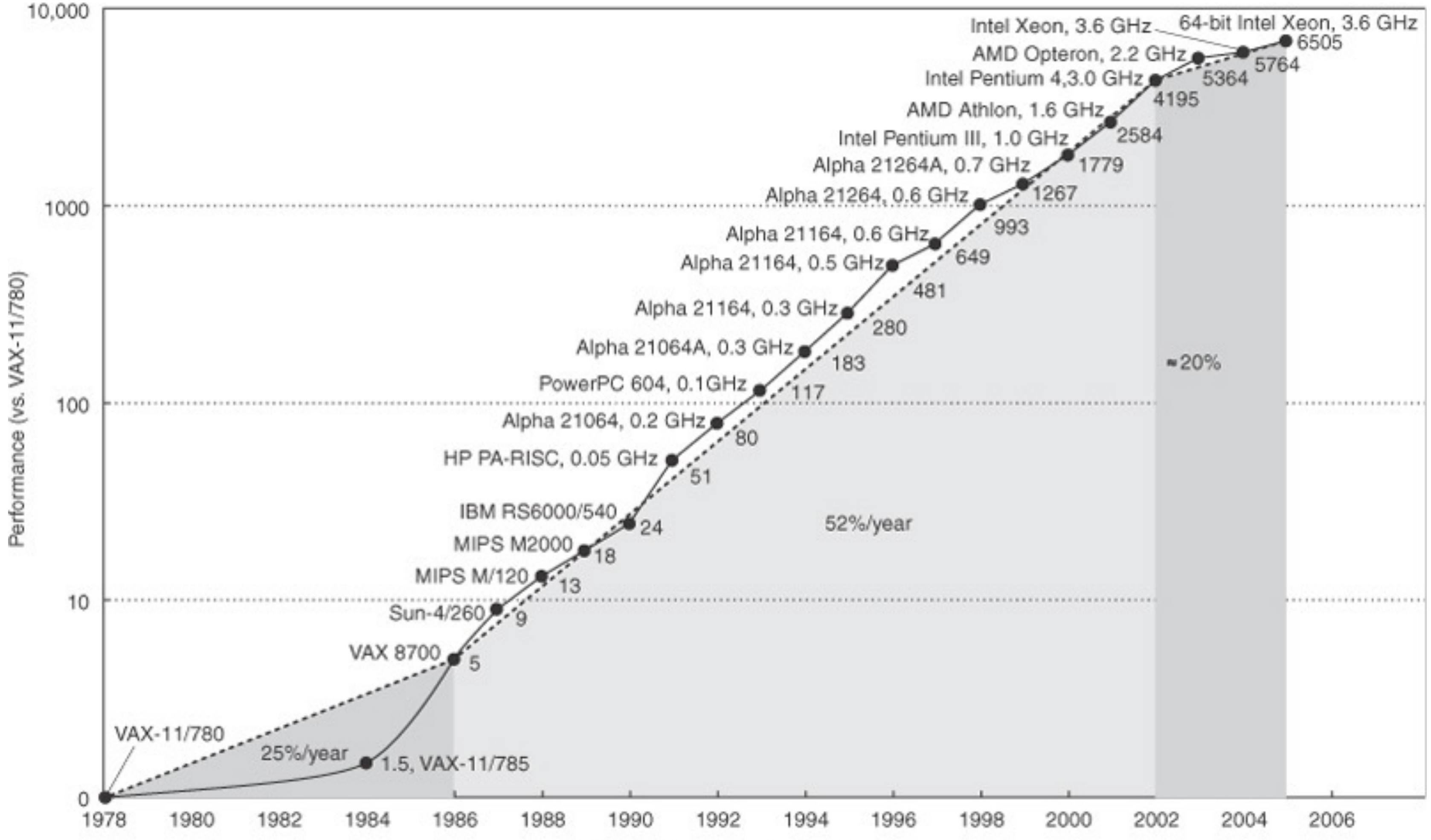
VISUALIZING PROGRESS

If transistors were people

If the transistors in a microprocessor were represented by people, the following timeline gives an idea of the pace of Moore’s Law.

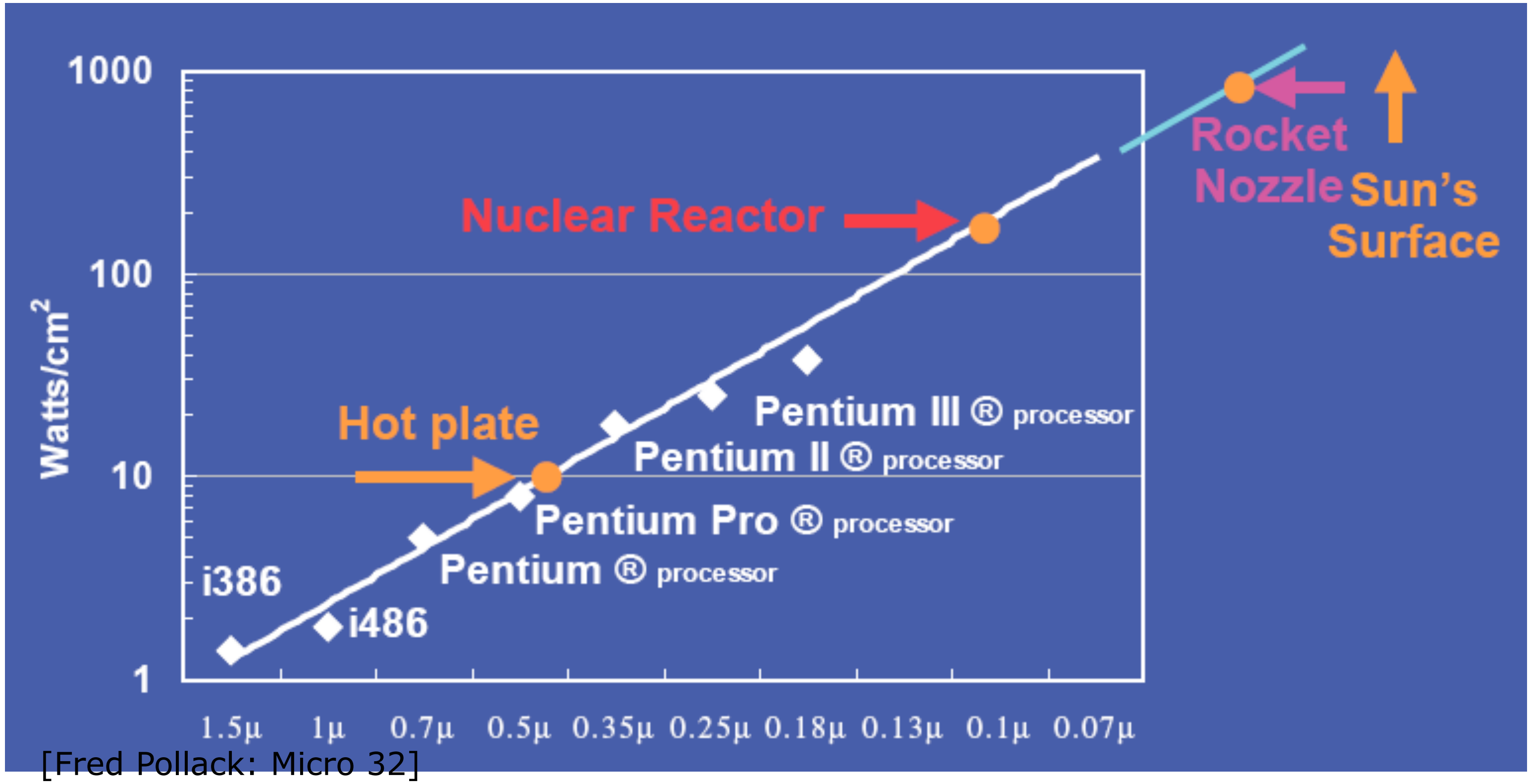
Now imagine that those 1.3 billion people could fit onstage in the original music hall. That’s the scale of Moore’s Law.

Growth in Processor Performance



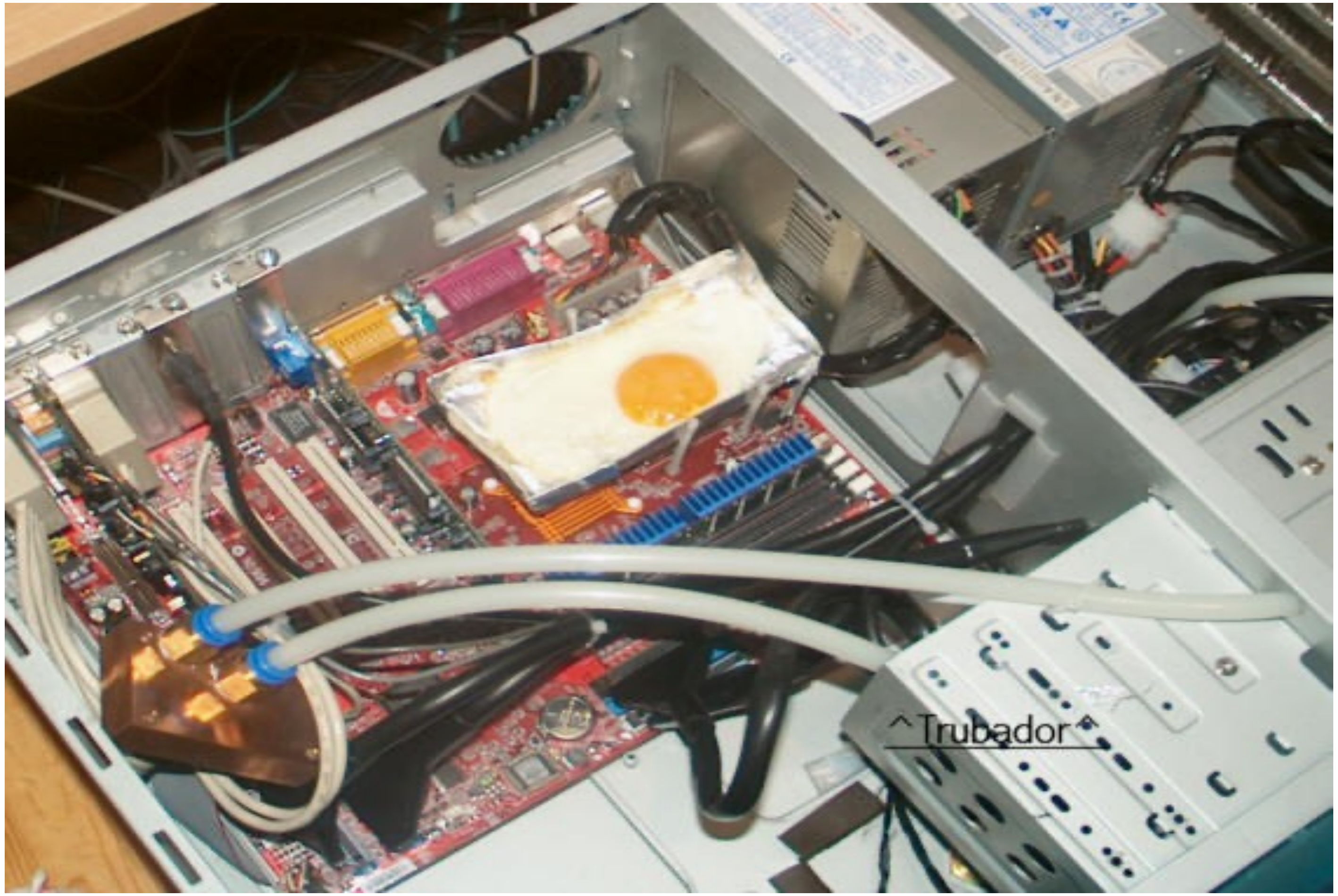
© 2007 Elsevier, Inc. All rights reserved.

Roadblock: Power density



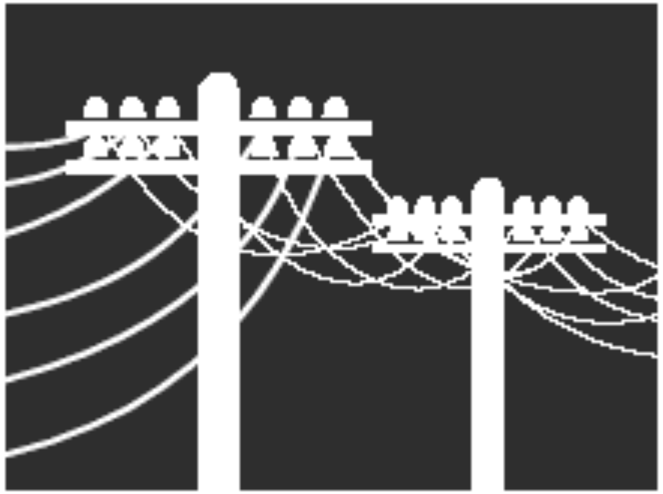
[Fred Pollack: Micro 32]

Got Heat?



Why worry about power?

Battery life



Environment

Thermal issues: cooling,
packaging,
reliability, timing

Power challenge for everybody

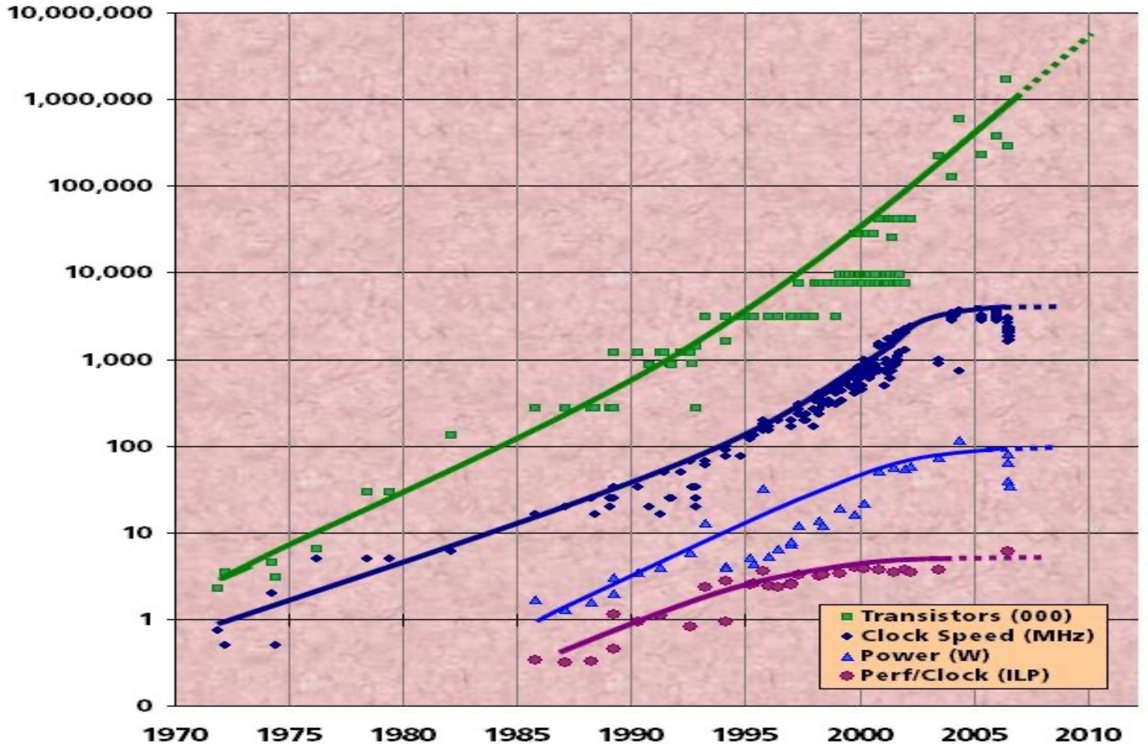
- Mobile/Portable (cell phone, laptop, PDA)
  - Battery life is critical
- Desktop
  - 400 million computers in the world
  - 0.16PW (PetaWatt = 10^{15} Watt) of power dissipation
  - Equivalent to 26 nuclear power plants
- Data centers
  - 1 single server rack is between 5 and 20 kW
  - 100s of those racks in a single room
  - 10 largest data centers in Singapore consume energy equivalent to 130,000 households
  - Contributes to 2% of world's total carbon emission – more than airline industry by 2020

Parallelism saves Power

- Power = C x V^2 x f
- Performance = cores x f
- Exploit explicit parallelism for reducing power using additional cores
  - Can increase cores (2x) and performance (2x)
  - Or increase cores (2x) but decrease frequency (f/2)
    - decrease in frequency by half decreases voltage by half
    - same performance at ¼ power

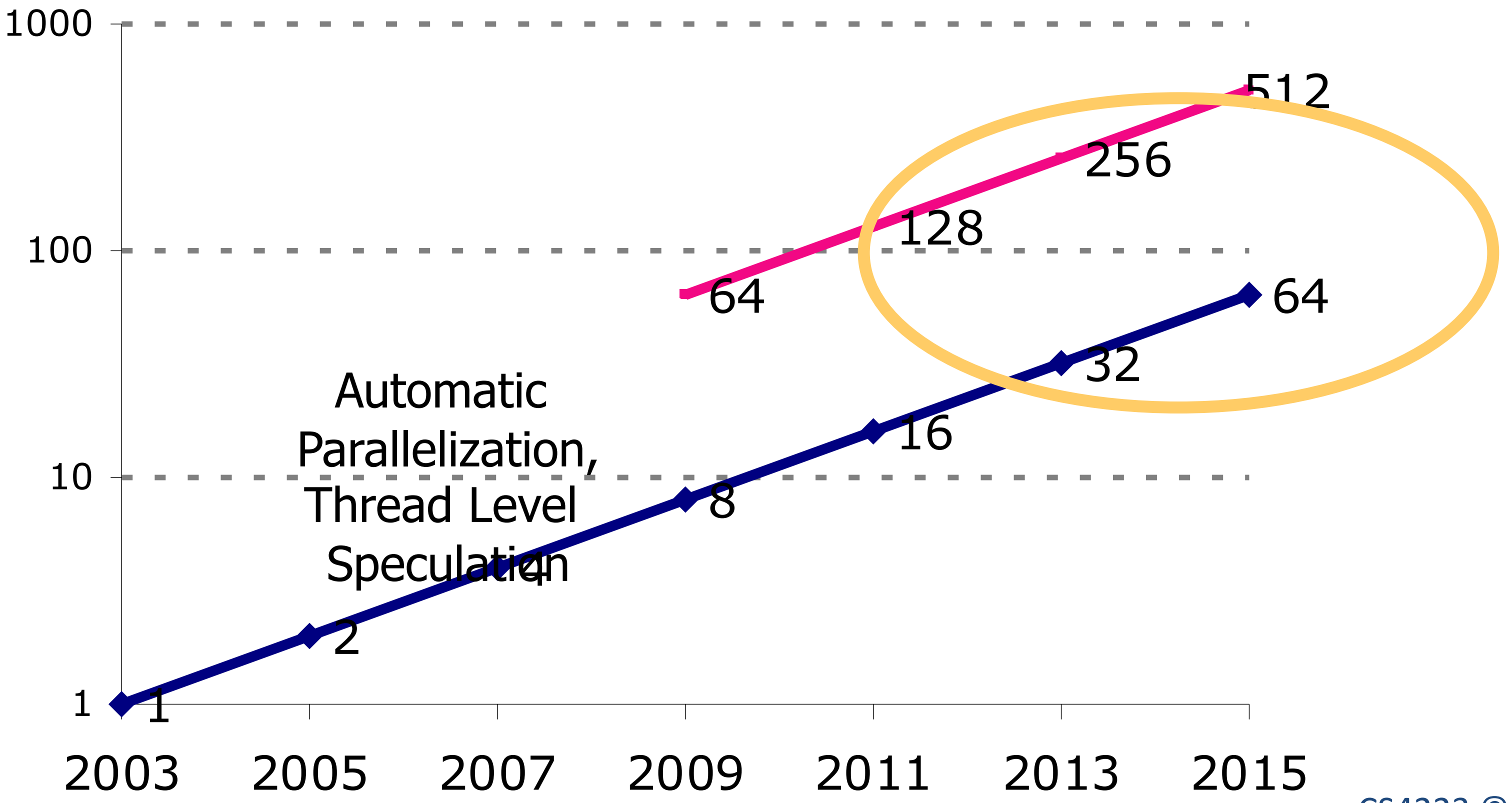
Multi-core Revolution

- Chip density is continuing to increase ~2x every 2 years
  - Clock speed is not
  - Number of processor cores may double instead

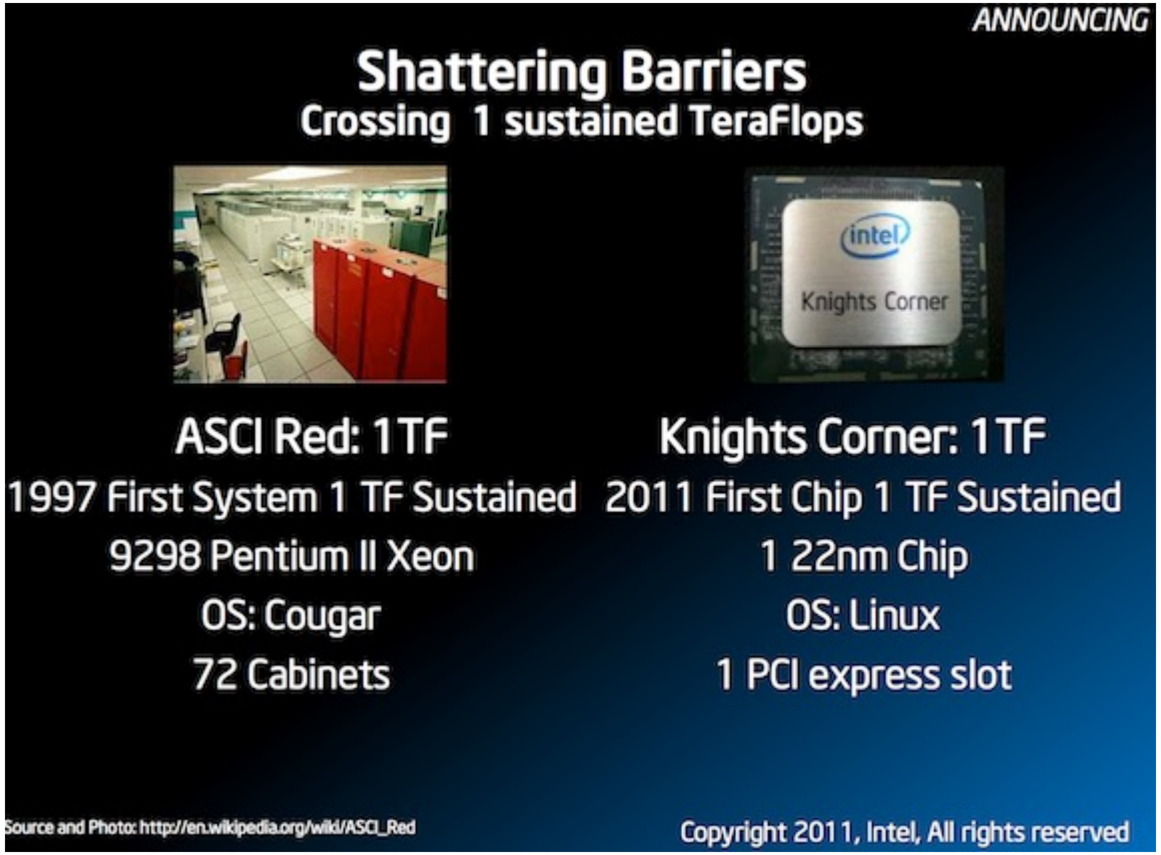


100+ Cores?

- Multi-core: 2X / 2 yrs ⇒ ≈ 64 cores in 8 years
- Many-core: 8X to 16X multi-core



Multi-core Revolution



ANNOUNCING

Shattering Barriers

Crossing 1 sustained TeraFlops

ASCI Red: 1 TF

1997 First System 1 TF Sustained

9298 Pentium II Xeon

OS: Cougar

72 Cabinets

Knights Corner: 1 TF

2011 First Chip 1 TF Sustained

1 22nm Chip

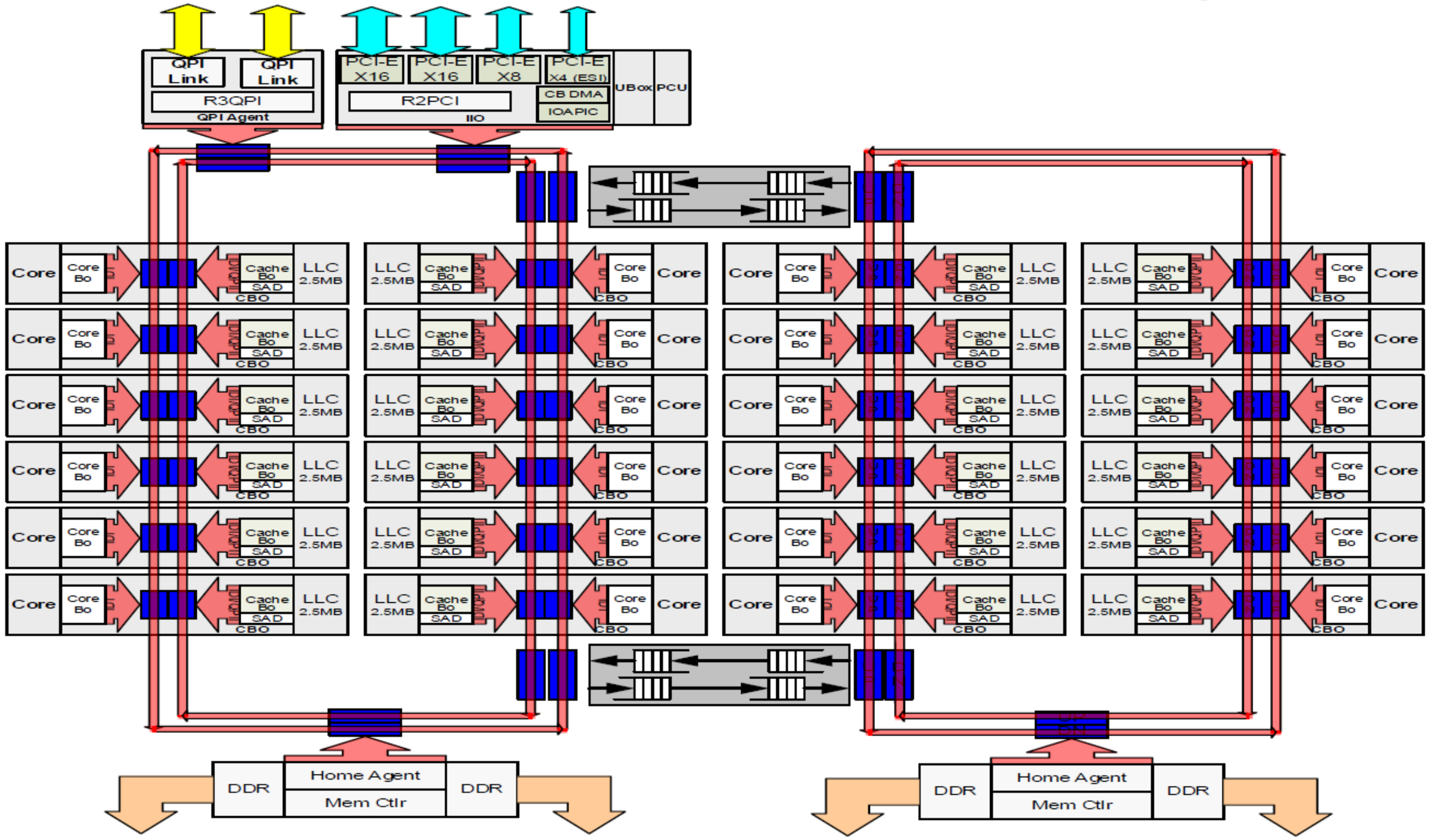
OS: Linux

1 PCI express slot

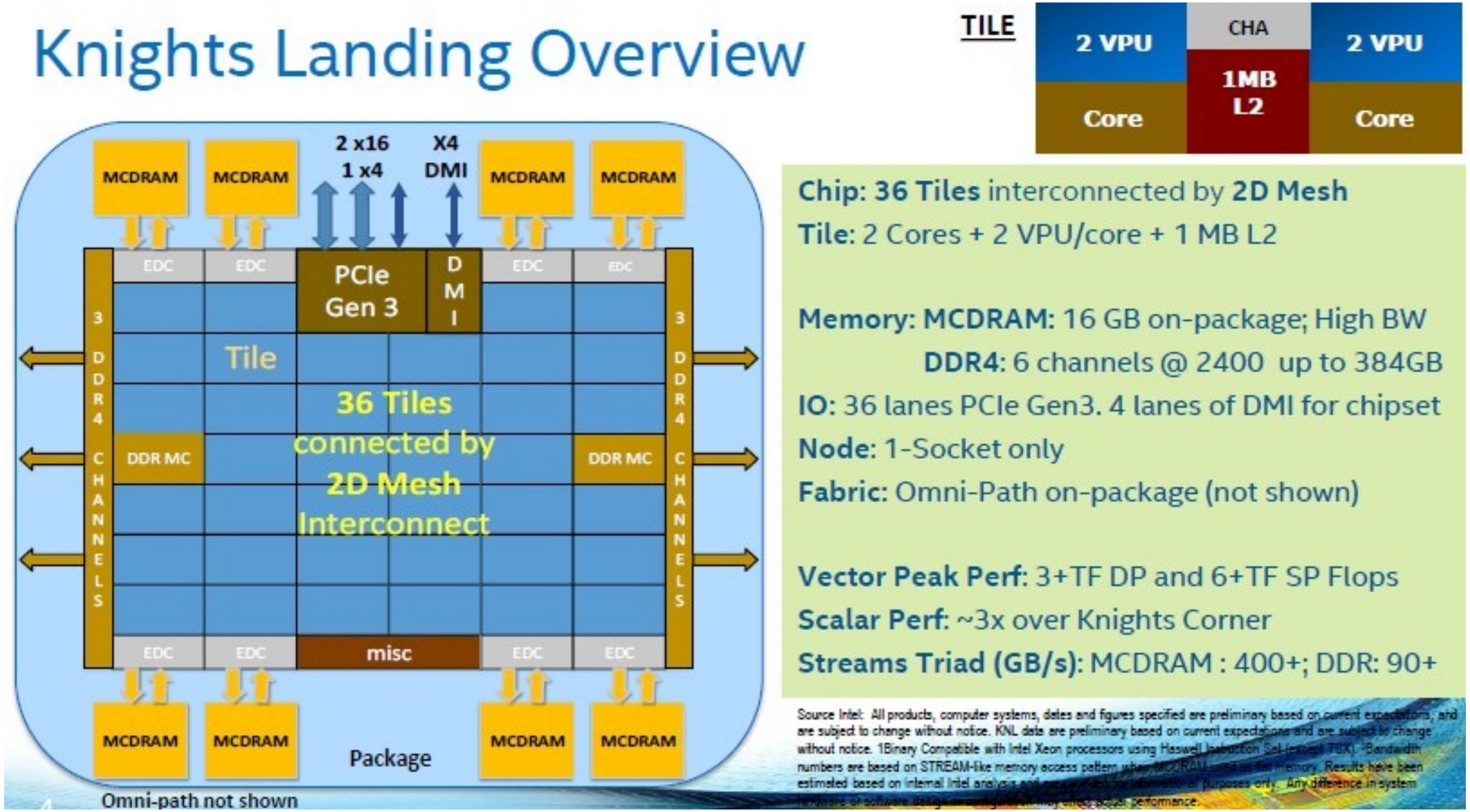
intel
Knights Corner

Source and Photo: http://en.wikipedia.org/wiki/ASCI_Red

Intel® Xeon® Processor E5 v4 Product Family HCC



- 24 cores (22 activated), 22x256KB L2, 55MB L3-cache
- 2.2 GHz, 145 Watt TDP (Thermal Design Power)



Knights Landing Overview

Chip: 36 Tiles interconnected by **2D Mesh**

Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW

DDR4: 6 channels @ 2400 up to 384GB

IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

Fabric: Omni-Path on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops

Scalar Perf: ~3x over Knights Corner

Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

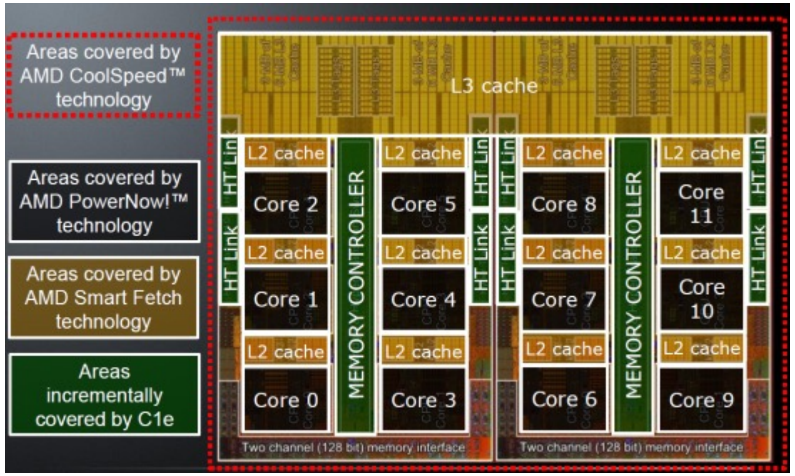
Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

- 1.5 GHz, 72 cores, 4 thread per core, 288 threads
- 3.46 TeraFlops performance at 245 Watts power

AMD Multi-cores

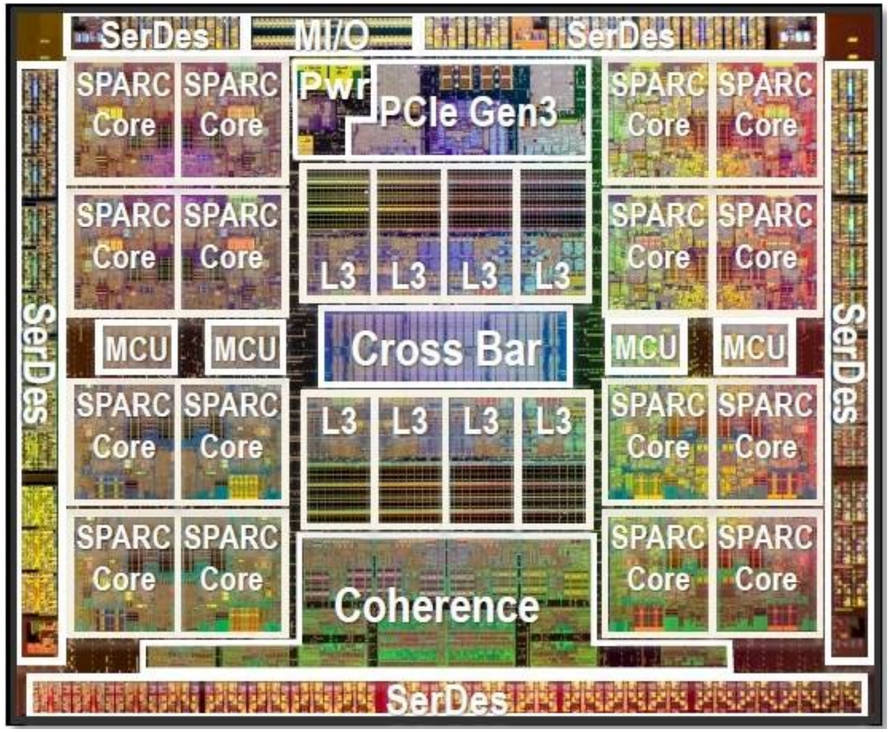
- 12-core Opteron Magny Cours
  - 12 cores at 2.2 GHz
  - Four channel DDR3

- 32-core Ryzen Threadripper 2
  - 32 cores at 3.0 – 4.2 GHz
  - Four channel DDR4


Areas covered by AMD CoolSpeed™ technology
Areas covered by AMD PowerNow!™ technology
Areas covered by AMD Smart Fetch technology
Areas incrementally covered by C1e
L3 cache
L2 cache
Core 2
Core 1
Core 0
Core 5
Core 4
Core 3
Core 8
Core 7
Core 6
Core 11
Core 10
Core 9
MEMORY CONTROLLER
HT Link
Two channel (128 bit) memory interface

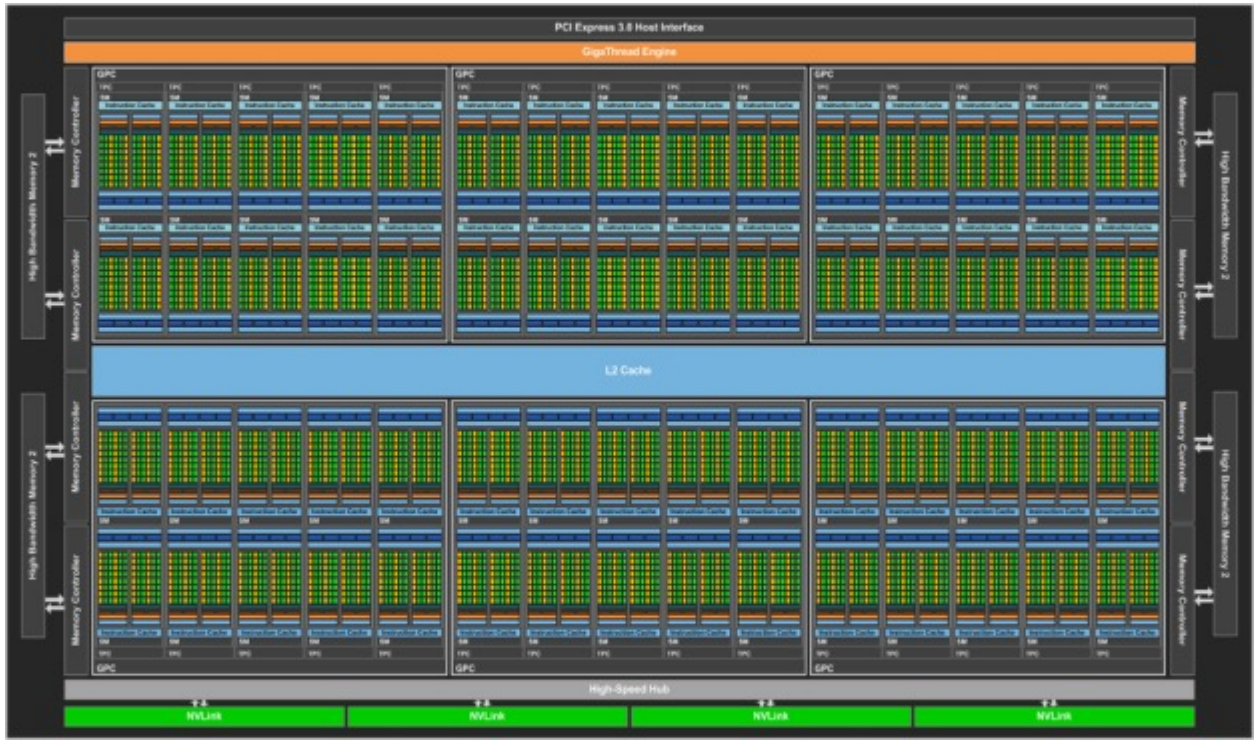
SPARC T5

- 16 CPU cores, 8 hardware threads per core → 128 threads per chip
- 8 MB Level 3 cache
- 3.6 GHz frequency
- 200-300 Watt TDP



nVidia Pascal

- 15.3 billion transistors
- 5.3 TFlop double-precision throughput, 300W TDP
- 56 SM units each with 64 FP32 and 32 FP64 CUDA cores
- 18 MB on-chip memory



One of the fastest supercomputers

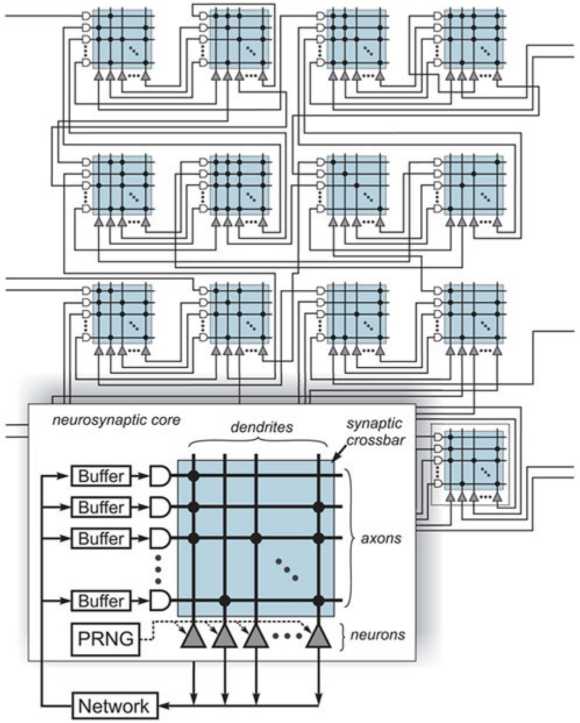
- Sunway TaihuLight from China’s National Research Center of Parallel Computer Engineering & Technology (NRCPC)
- Exclusively uses processor designed and made in China (not Intel)
- 40,960 compute nodes for a total of 10,649,600 computing cores
- 93 Petaflops performance
- Peak power consumption of 15.37 megawatts (6 Gflops/Watt) – top spot in Green500 in terms of Performance/Power metric

How does it compare to Human Brain?

- Supercomputer Sunway TaihuLight consumes 15,370,000 Watt power
- Human brain, which is at least several hundred times more complex, consumes 20 Watts of power
- Using 80,000 processors in a supercomputer, we could only mimic 1% of 1 sec worth of human brain activity
  - and even that took 40 minutes

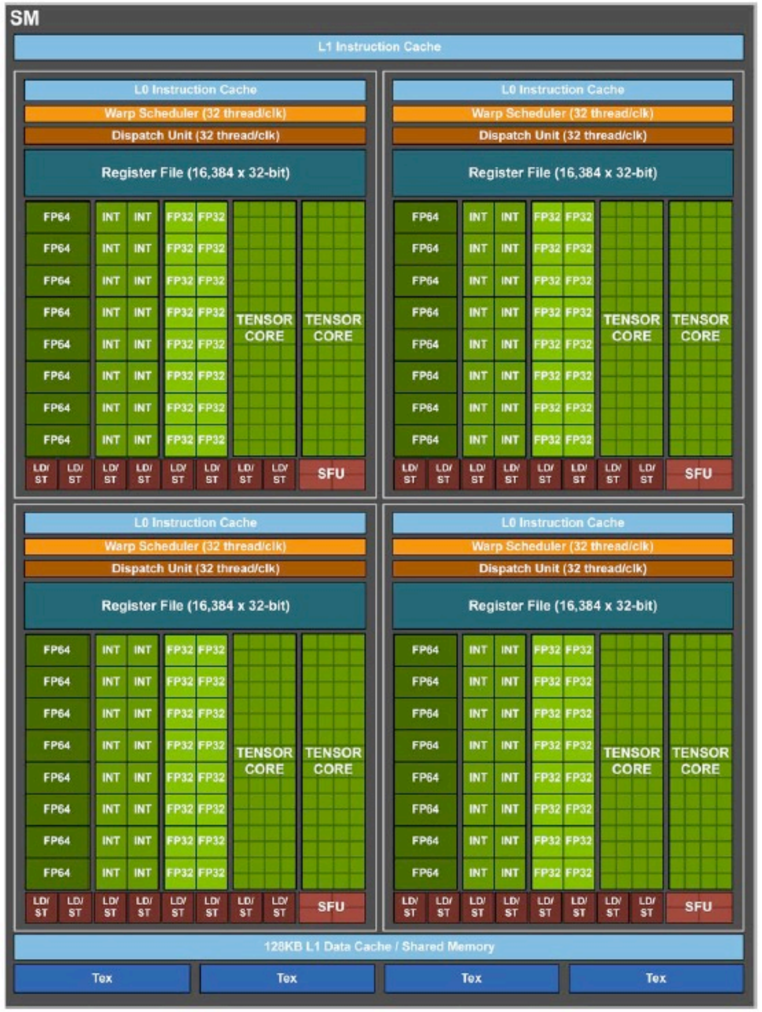
Accelerators: IBM True North

- Brain-inspired chip
- 1 million programmable neurons, 256 million programmable synapses, 4096 neurosynaptic cores
- 70mW powers chip; 1 trillion synapses need 4kW



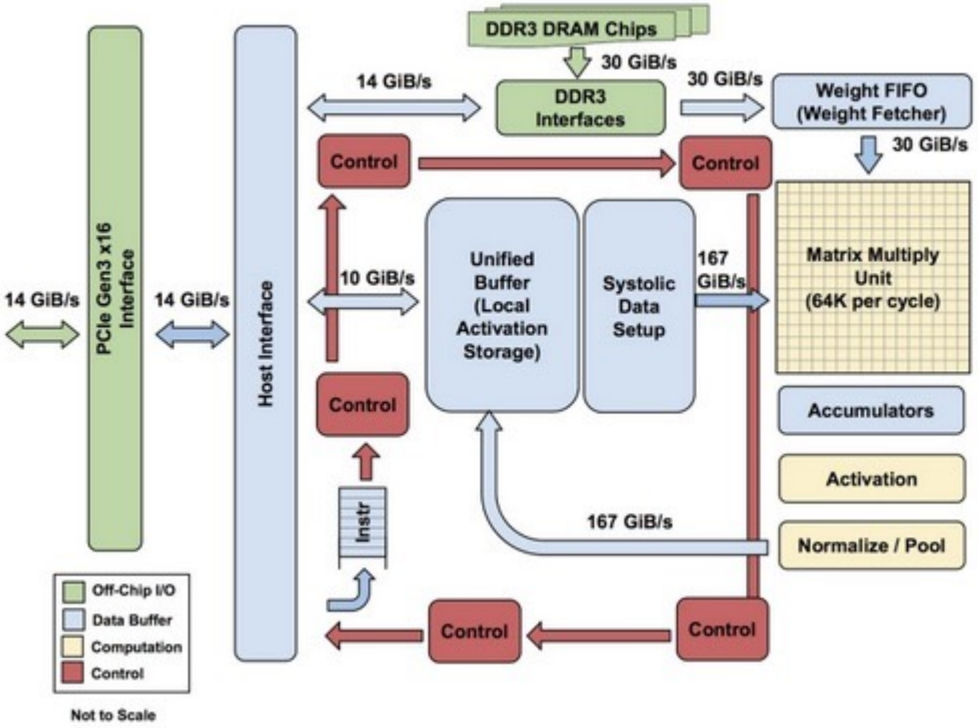
nVidia Volta

- 84 Volta Streaming Multiprocessor (SM) each with
  - 64 FP32 cores, 64 INT32 cores, 32 FP64 cores
  - 8 Tensor cores
- 21.1 billion transistors, 300 Watts, 815 mm^2 die size


SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

Accelerators: Google TPU

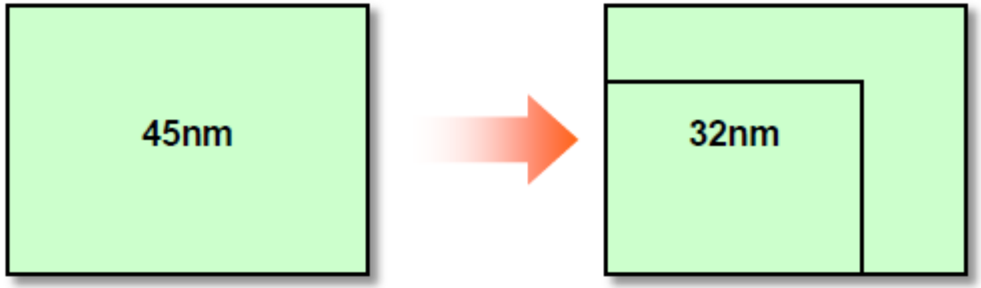
- Tensor Processing Unit (TPU): A programmable architecture for neural networks
- Use matrix as a primitive



Architecture Outlook

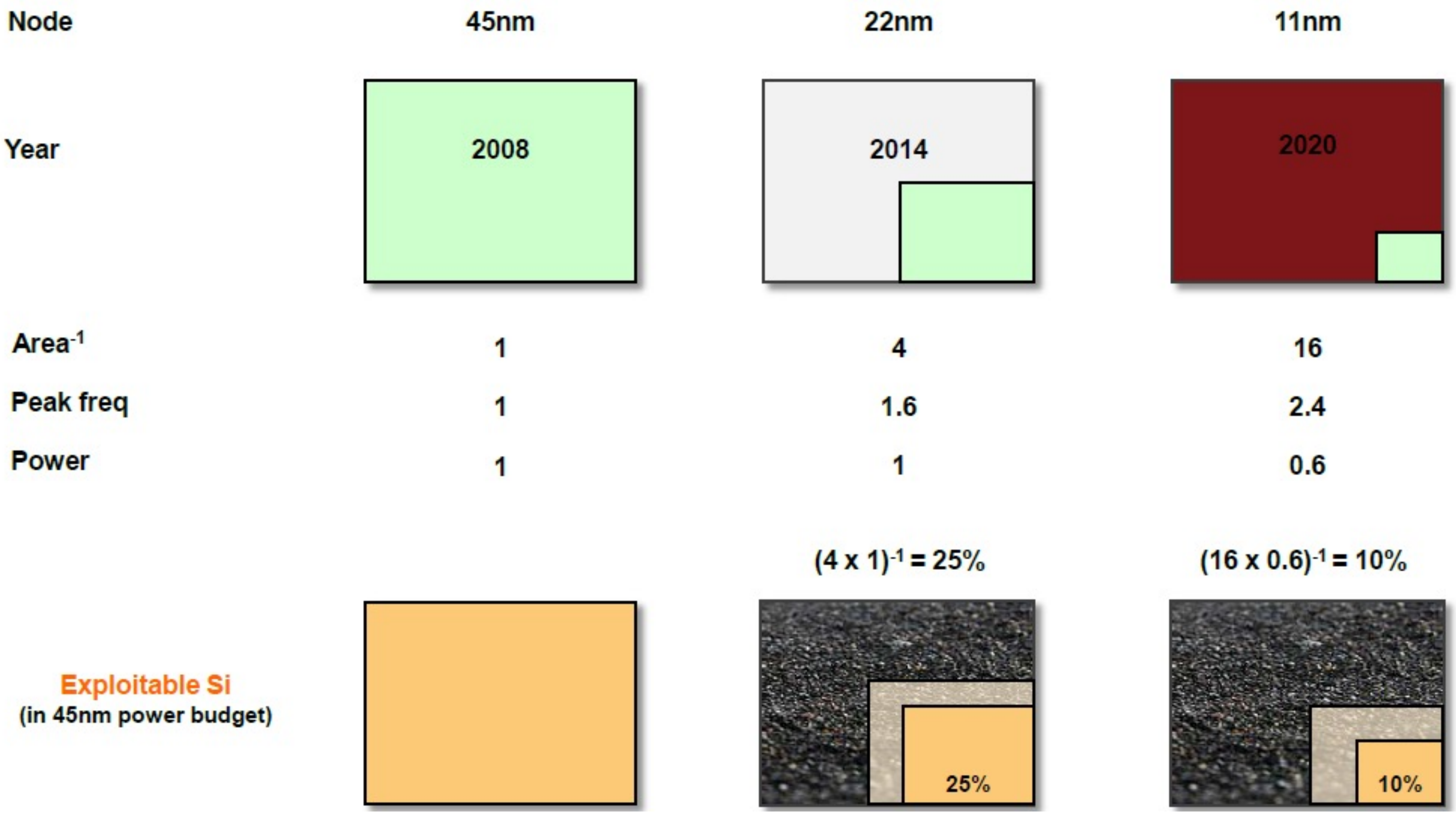
- Expect modestly pipelined processors
  - Reduce complexity for performance per watt
- Parallel is energy efficient path to performance
  - Lower threshold and supply voltages lowers energy per operation
- Redundant processors can improve chip yield
  - Cisco Metro 188 CPUs + 4 spares; Cell 8 out of 9 cores
- Small, regular processing elements easier to verify
- Compiler-controlled scratchpad memory rather than cache
- One size fits all?
  - Amdahl's Law $\Rightarrow$ Heterogeneous processors
  - Special function units to accelerate popular functions

Silicon Generations: Shrink and Add



Expectation: shrink and add new functionality in about same area

The Creation of Dark Silicon



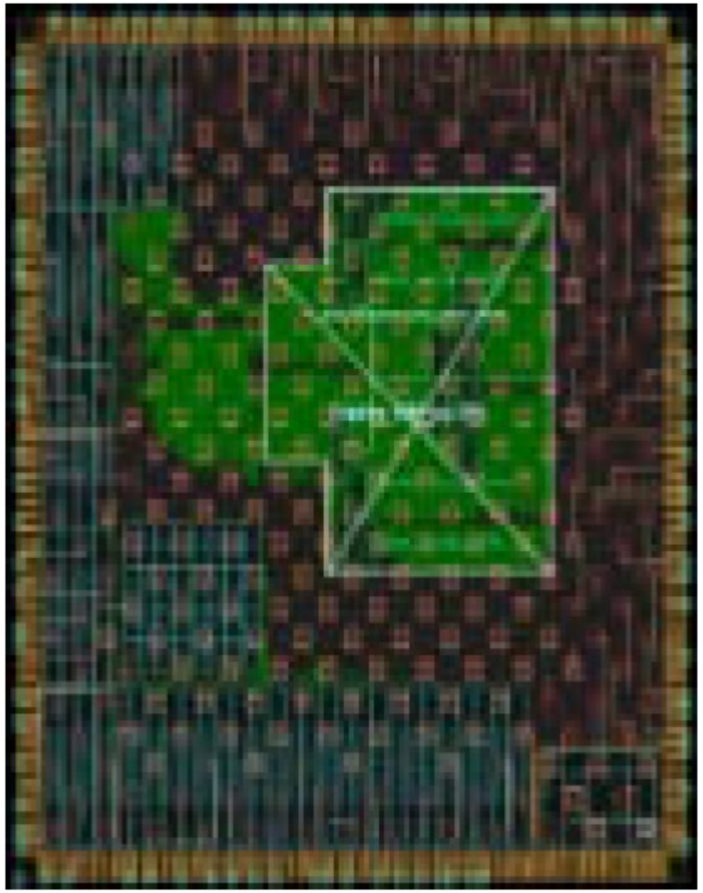
Node	45nm	22nm	11nm
Year	2008	2014	2020
Area^{-1}	1	4	16
Peak freq	1	1.6	2.4
Power	1	1	0.6
		$(4 \times 1)^{-1} = 25\%$	$(16 \times 0.6)^{-1} = 10\%$
Exploitable Si (in 45nm power budget)		25%	10%

Source: Jem Davies, ARM

Dark Silicon

- We can have more transistors and cores

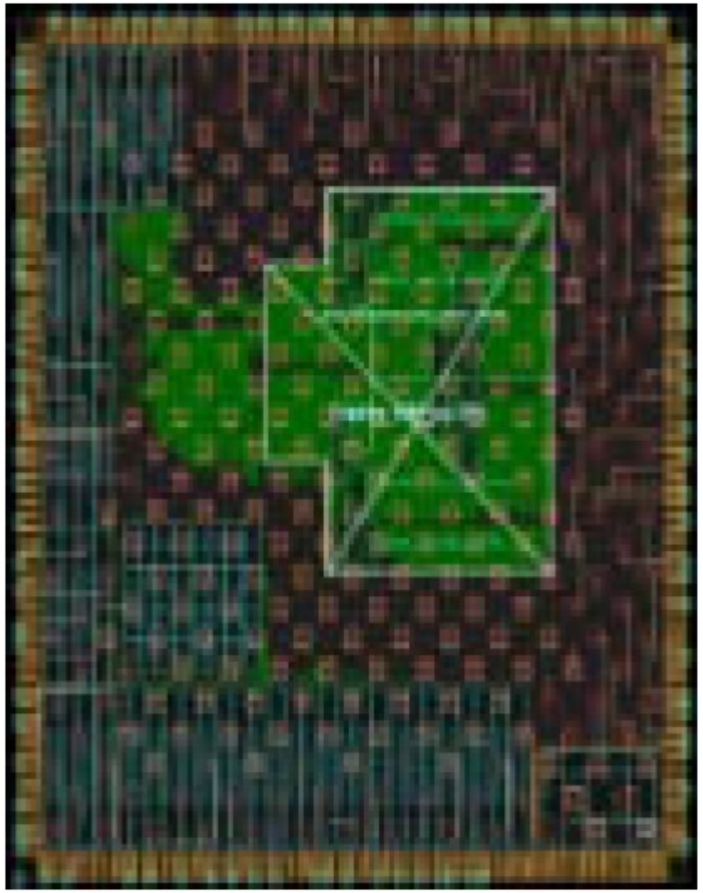
- We just cannot power them all at the same time



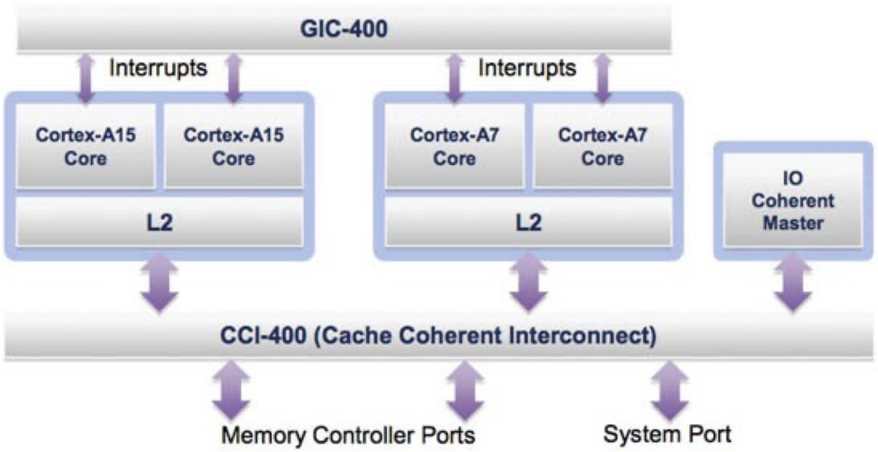
CS4223 © Mitra 2017

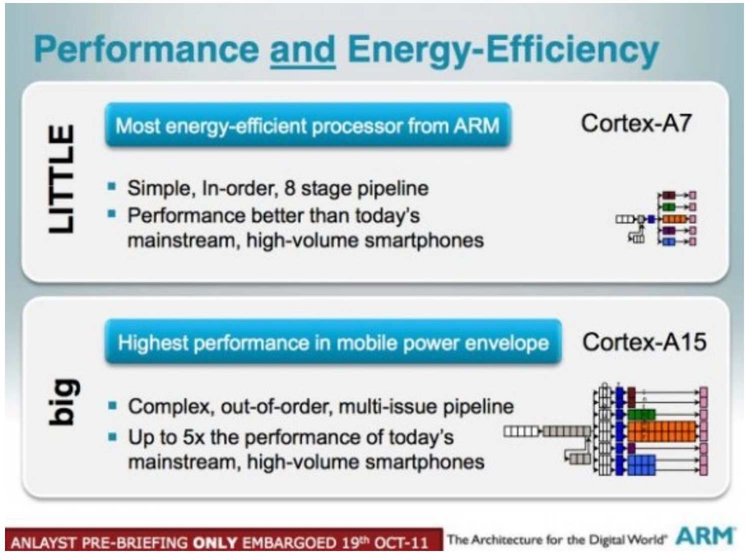
So what can we do?

- Need heterogeneous cores
- Power on only the most appropriate cores
- Power-efficient computation



ARM big.LITTLE asymmetric multi-core





Performance and Energy-Efficiency

LITTLE

Most energy-efficient processor from ARM — Cortex-A7

- Simple, In-order, 8 stage pipeline
- Performance better than today's mainstream, high-volume smartphones

big

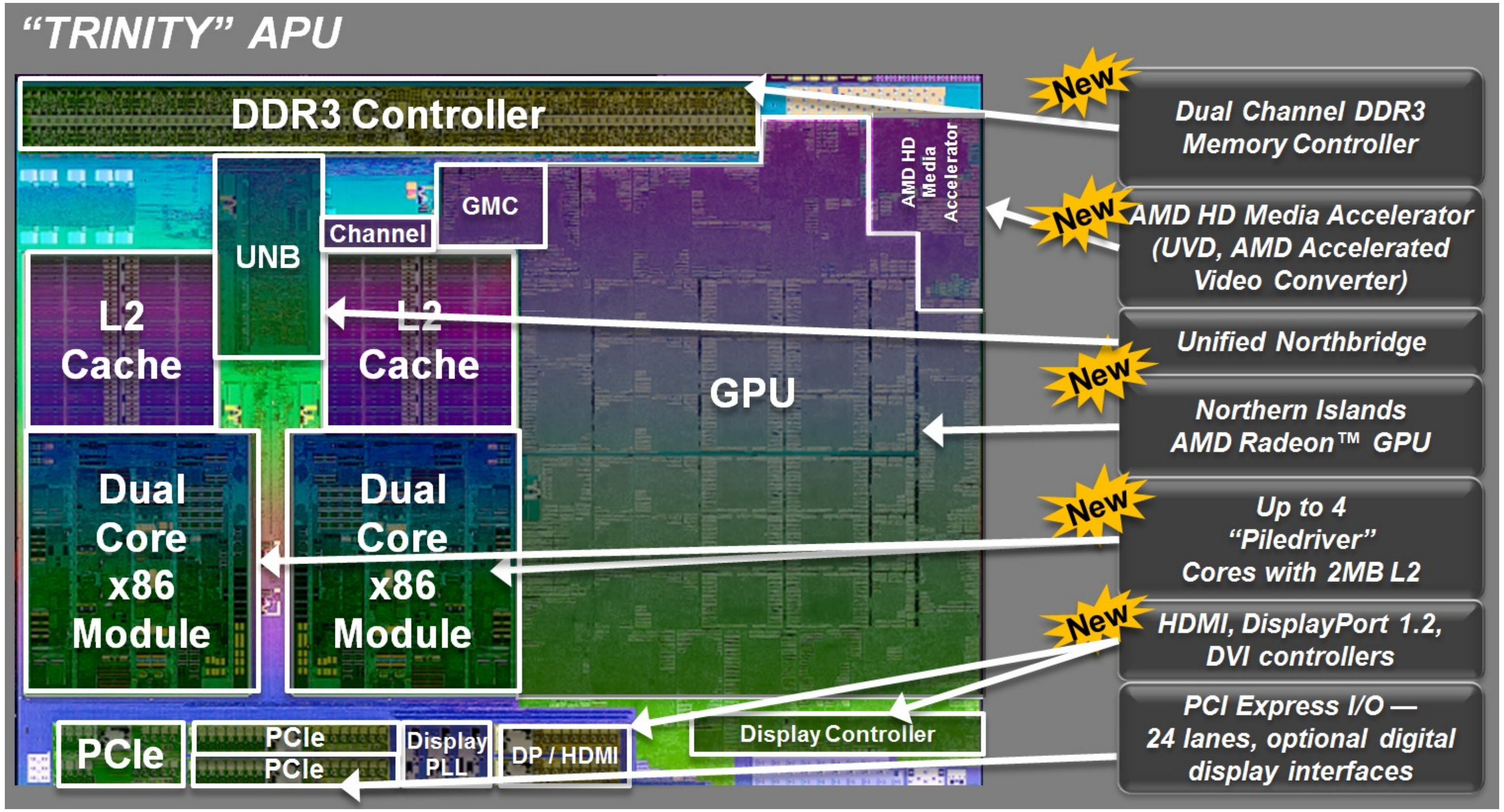
Highest performance in mobile power envelope — Cortex-A15

- Complex, out-of-order, multi-issue pipeline
- Up to 5x the performance of today's mainstream, high-volume smartphones

ANLAYST PRE-BRIEFING ONLY EMBARGOED 19th OCT-11 The Architecture for the Digital World ARM

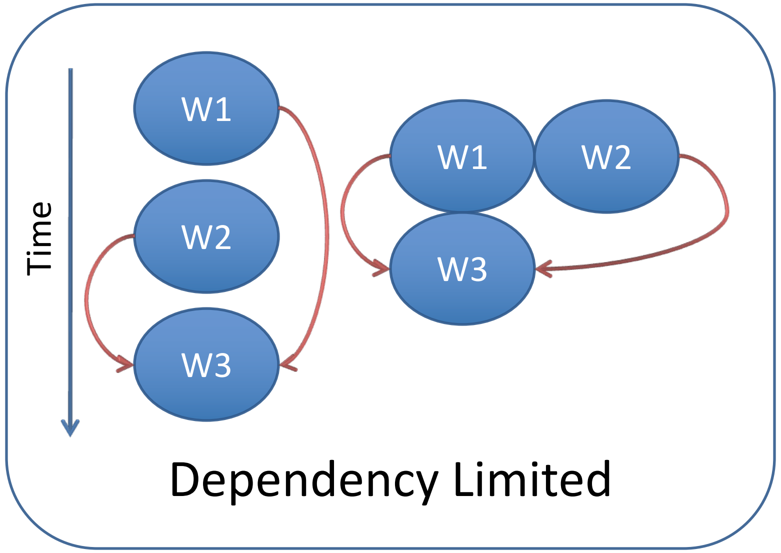
Cortex-A7 manages always-on, always connected workloads
#ARM
SMS
Social Media
Cortex-A15 handles peak performance tasks
Video

AMD Trinity APU

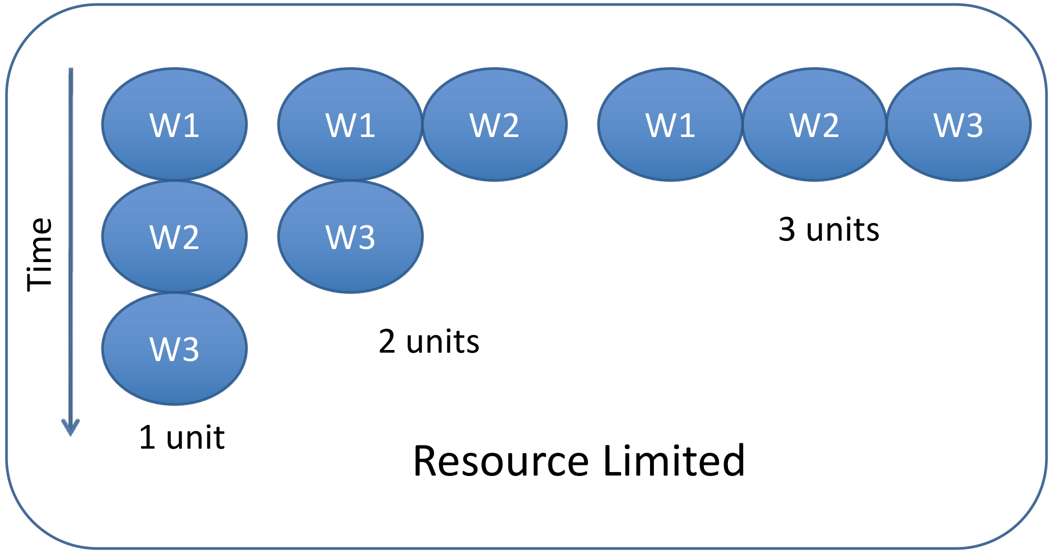

"TRINITY" APU
DDR3 Controller
AMD HD Media Accelerator
GMC
Channel
UNB
L2 Cache
L2 Cache
GPU
Dual Core x86 Module
Dual Core x86 Module
PCIe
PCIe
PCIe
Display PLL
DP / HDMI
Display Controller
New
Dual Channel DDR3 Memory Controller
New
AMD HD Media Accelerator (UVD, AMD Accelerated Video Converter)
Unified Northbridge
New
Northern Islands AMD Radeon™ GPU
New
Up to 4 "Piledriver" Cores with 2MB L2
New
HDMI, DisplayPort 1.2, DVI controllers
PCI Express I/O — 24 lanes, optional digital display interfaces

What is Parallelism?

- Independent units of work that can execute concurrently if sufficient resources exist



Dependency Limited



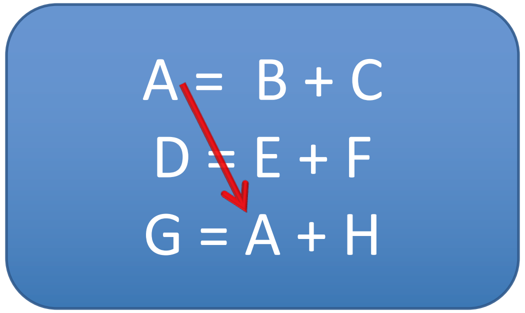
Resource Limited

Where to find parallelism?

- Parallelism can be found/exists at different granularities
  - Instruction Level
  - Data Level
  - Thread Level
  - Task Level

Instruction-Level Parallelism (ILP)

- A measure of how many operations in a sequential program can be performed simultaneously



A = B + C

D = E + F

G = A + H

- Micro-architectural and compiler techniques are employed to extract ILP
  - Instruction pipelining, superscalar execution, out-of-order execution, VLIW, dataflow architecture

Data Level Parallelism (DLP)

- DLP is parallelism inherent in program loops where similar operation sequences are performed on elements of a large data structure
- Need compiler and programmer's help in extracting DLP

```
for (i=0; i<N; i++)
A[i] = C x B[i];
```

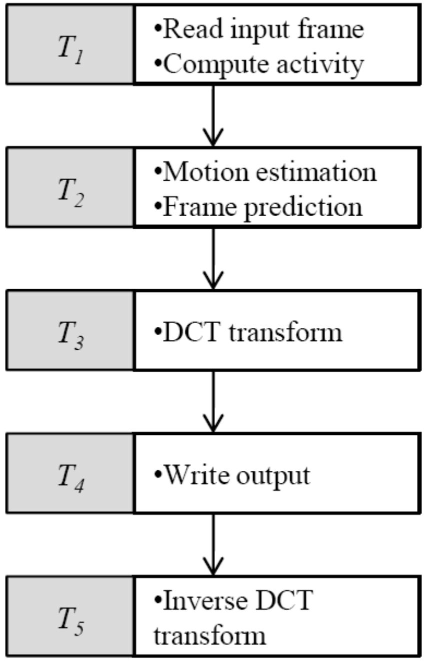
Thread Level Parallelism (TLP)

- Higher level of parallelism available as multiple threads of control within a process
- Need compiler and programmer’s help in extracting TLP

```
for (i=0; i<200; i++)
     for (j=1; j<20000; j++)
                 val [i,j] = val[i,j-1] +1;
```

Task Level Parallelism

- Higher level of parallelism where different processes execute on the same or different data
- Need user and programmer’s help in indentifying task parallelism



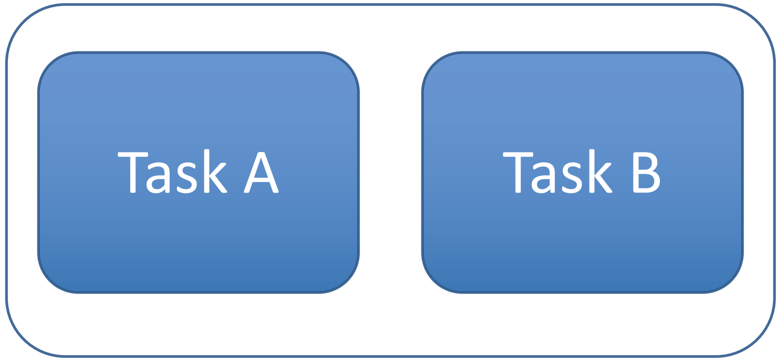
T_1 •Read input frame •Compute activity

T_2 •Motion estimation •Frame prediction

T_3 •DCT transform

T_4 •Write output

T_5 •Inverse DCT transform



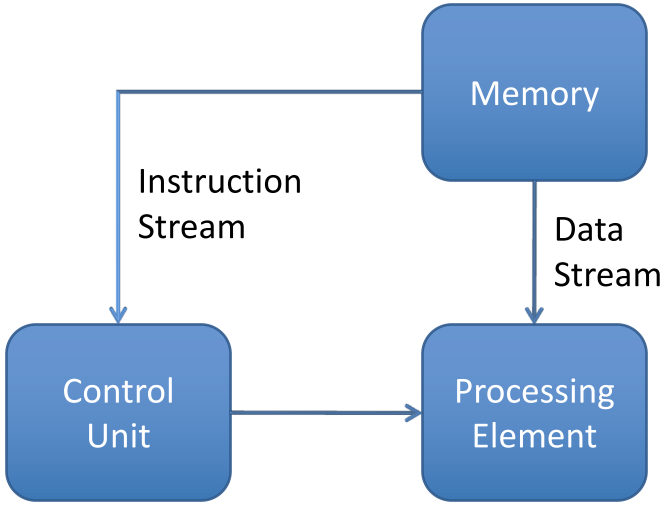
Task A

Task B

Flynn’s Taxonomy of Parallel Computers

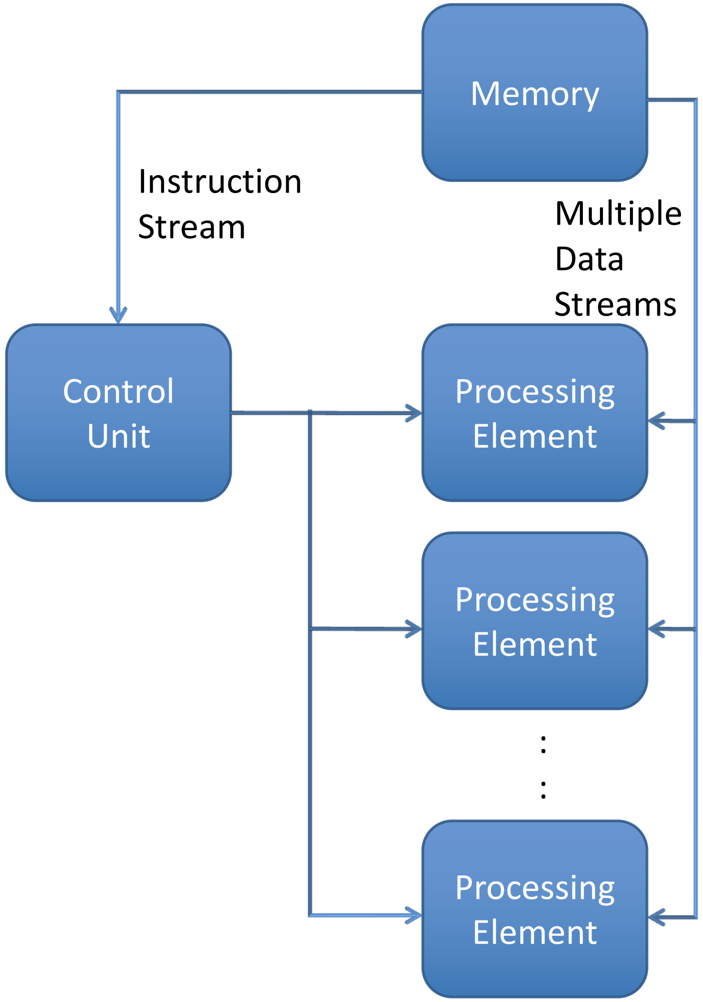
		Number of Data Streams	
		Single	Multiple
Number of instruction streams	Single	SISD	SIMD
	Multiple	MISD	MIMD

SISD: Single Instruction Single Data



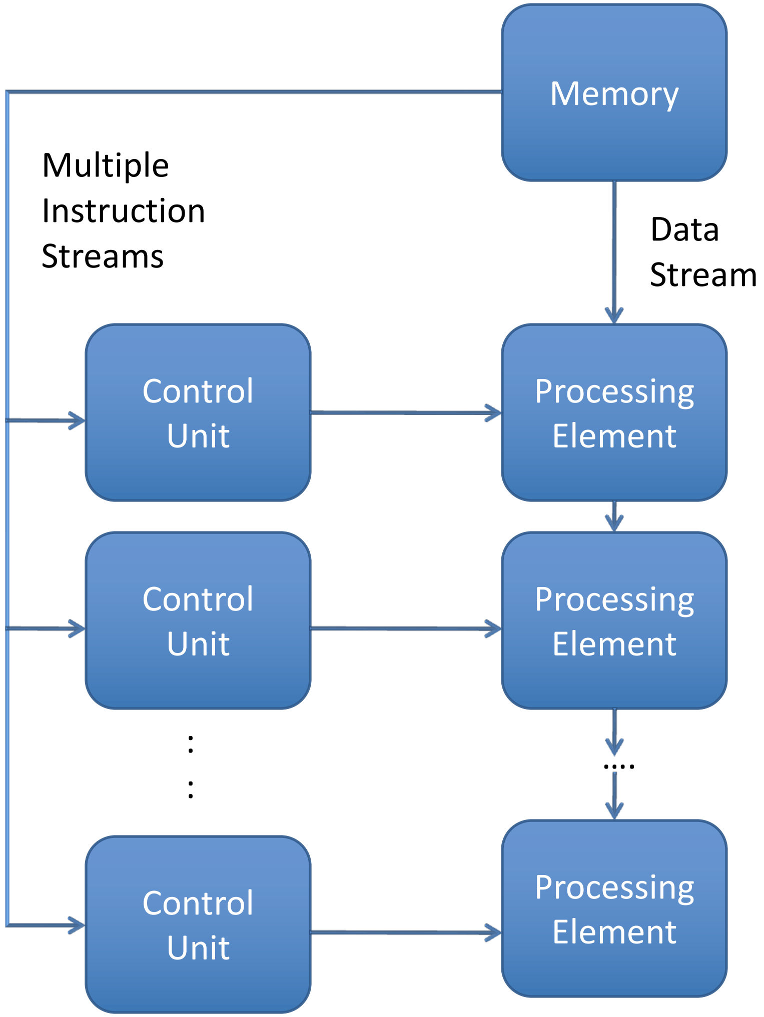
- SISD is not considered as a parallel architecture
- SISD exploits parallelism at the instruction level
- Pipelined, Superscalar, and VLIW architectures are examples of SISD architecture

SIMD: Single Instruction Multiple Data



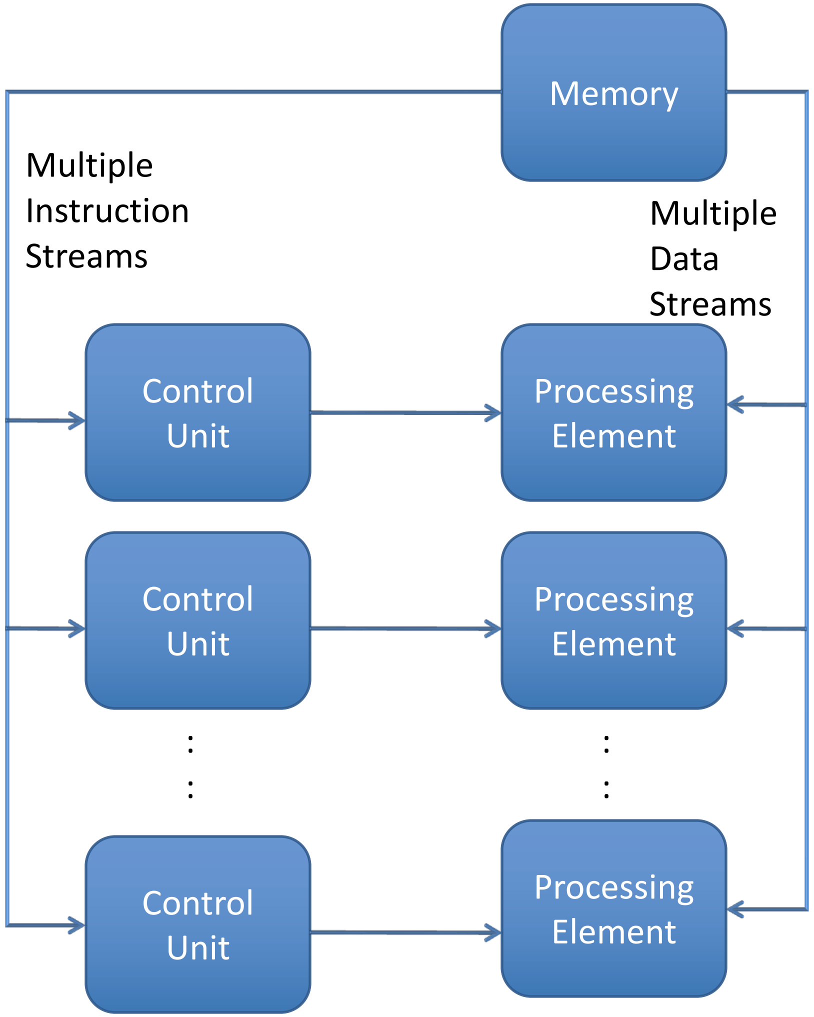
- A single instruction operates on multiple data to exploit data parallelism
- Vector processors and GPUs are excellent examples of SIMD architecture
- More efficient in terms of instruction count and loop control overhead

MISD: Multiple Instruction Single Data



- Multiple processing elements execute from different instruction streams and data is passed from one processing element to the next
- Example: Systolic array such as CMU iWrap
- Data passing restriction is quite severe --- hard to generalize

MIMD: Multiple Instruction Multiple Data



- Most flexible architecture
- Used in most parallel computers today

Syllabus

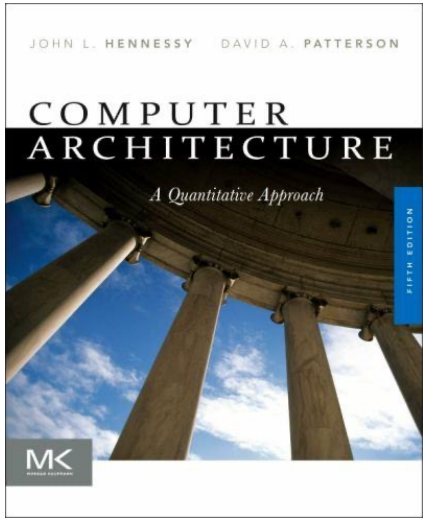
- ILP exploitation via superscalar and VLIW
- DLP exploitation via vector processors (GPU)
- TLP exploitation via multi-core
  - Cache Coherence
  - Memory Consistency
  - Synchronization
- Power/Energy issues

Know your lecturer

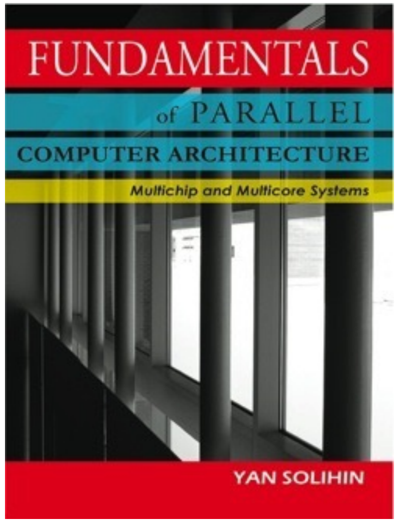
- Trevor E. Carlson (you can call me Trevor)
- Office: COM3-02-10
- Extension: x7997
- Email: tcarlson@comp.nus.edu.sg
- Please note that this is my primary email account
- I will only respond to your email if it is sent from SoC or NUSNET account (i.e., do not use yahoo, gmail etc.)
- Consultation: By appointment through email

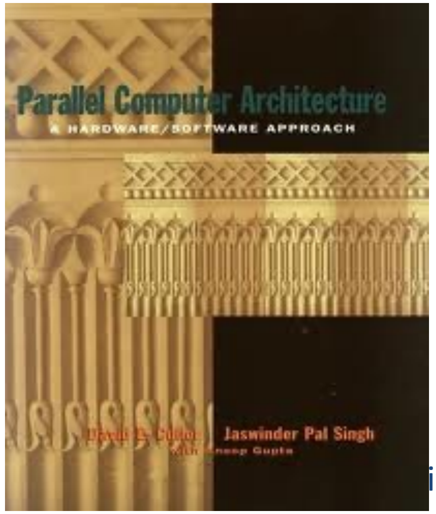
Textbooks

- Helpful textbook



- Multiple reference books
  - Relevant chapters will be distributed
- Multiple eBooks


FUNDAMENTALS of PARALLEL COMPUTER ARCHITECTURE
Multichip and Multicore Systems
YAN SOLIHIN



Resources

- Primary information source is Luminus

- Files: Lecture notes, assignments, lab exercises
- Lesson Plan: Very important; describes in detail the schedule for each week
- Forum: Ask course-related technical questions in the forum. Email is only for your personal concerns.

Other useful resources

- Subscribe to Canvas Announcements for SMS/email notification
- Anonymous feedback: Let me have your feedback during the semester without disclosing your identity. We do listen to your feedback ☺

Assessment

~~Final exam~~	~~40%~~
Tests and Midterm (tentative)	55%*
Project/Assignment	40%
In-class Quizzes	5%

*Tentative breakdown

Quizzes and Paper Reviews

- Short in-class quizzes to review recent topics

- Papers / Supplemental Material
  - Allows us to connect fundamentals to modern research.
  - Answer a number of questions
    - What are the key novel components? Which is the most important? Why?
    - How does work relate to the foundational work from the textbook?
    - Why does this work matter? Why is it important? (Justification)
  - More details to be posted on Luminus