

Lecture 1

- Course Introduction
- Computer Architecture Overview

Course Introduction

- **Instructor:** Chia-Lin Yang
 - Email: yangc@csie.ntu.edu.tw
 - Office : CSIE 411
 - Office Hours: 10~ 11 AM Wednesday
 - Or make an appointment
 - Course Web Page:
 - http://eclab.csie.ntu.edu.tw/courses/ca2017_fall/
 - Lectures, homework, resources
 - FB Group : Discussion, posting interesting news on the IC/IT industries
- **TA : office hours (TBA)**
 - 林孟瑤 smt8922@gmail.com CSIE 308
 - 陳啟中 bryan830401@gmail.com CSIE 308
 - 柯志霖 j22491050@gmail.com CSIE 308
- **Textbook:**
 - Computer Organization & Design. **The Hardware/Software Interface.** 5nd Edition, David A. Patterson and John L. Hennessy

Course Introduction

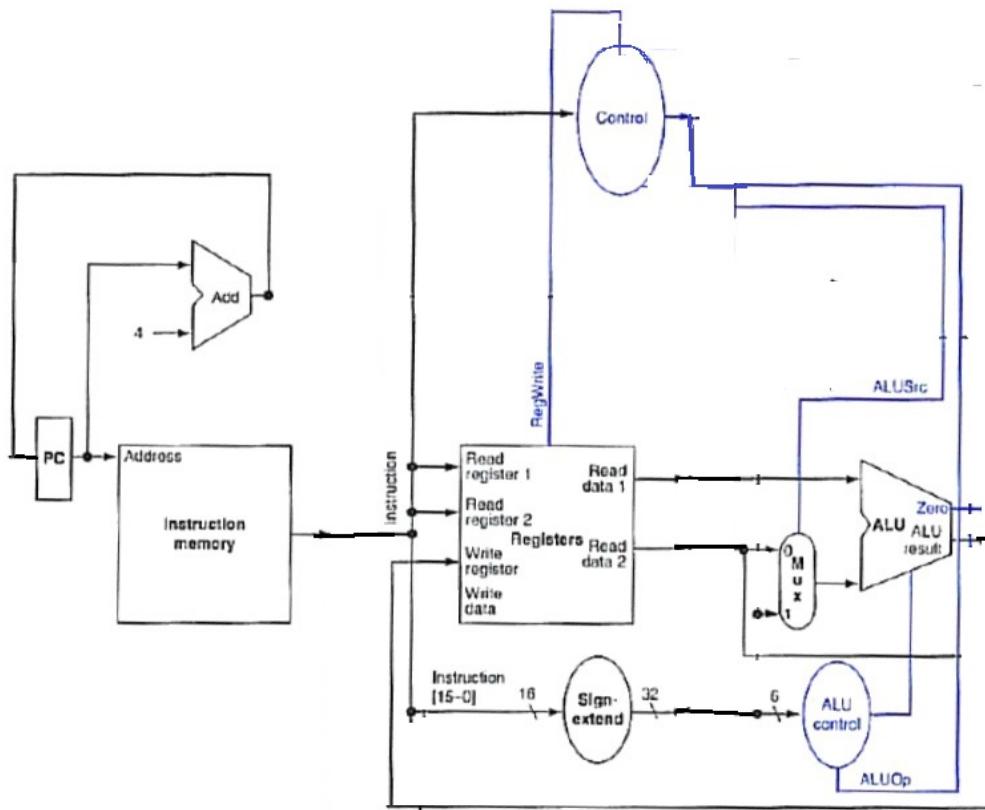
- Grading
 - 20% homework
 - Textbook exercises
 - MIPS assembly programming
 - Single cycle CPU (simple verilog exercise)
 - (20%) Two projects 3-person a team
 - Design a CPU pipeline in Verilog
 - Design CPU + Caches in Verilog
 - 60% exams
- Late homework policy
 - 10 pts reduction for each day late
- Honor code
 - **No cheating !!!**
- What is your duty?
 - Stop me when I talk too fast !!
 - Read the book, do the homework, attend the class

Processor Design

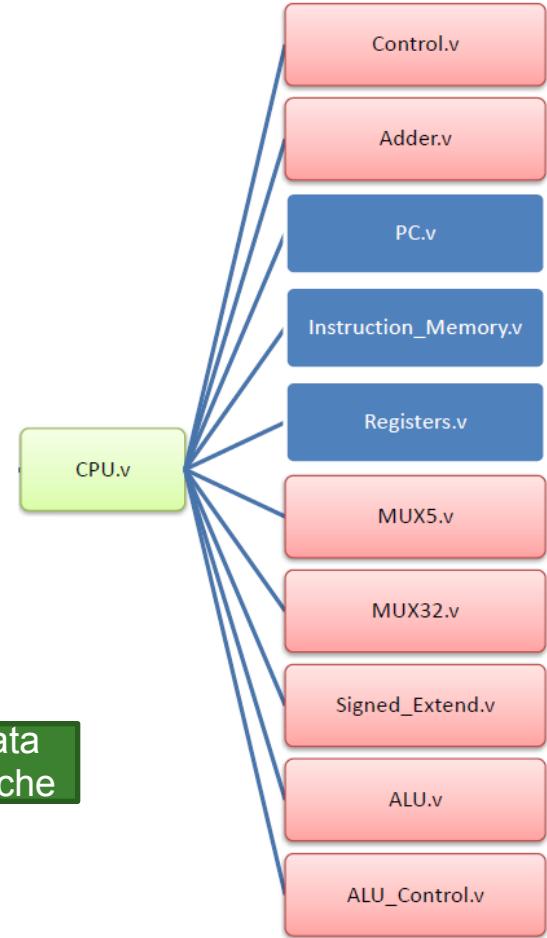
— Project 2 : Data Cache

Project 1 : Pipeline, FW and Hazard Detection modules

HW : Single cycle CPU module implementation (pink module)



Data
Cache



Ask questions when you do not understand !!

After explaining to a student, with various lessons and examples, that:

$$\lim_{x \rightarrow 8} \frac{1}{x-8} = \infty$$

I tried to check if he really understood that, so I gave him a different example

This was the result:

$$\lim_{x \rightarrow 5} \frac{1}{x-5} = \infty$$

當然,前提是,你需要知道“你不知道”!!

Why do I need to take this course?

You will be able to answer the following questions after taking this course.

Q1: How can I write a program with good performance? For example, why does code B perform better than code A?

```
for (i = 0; i < N; i = i+1)
  for (j = 0; j < N; j = j+1)
    {r = 0;
     for (k = 0; k < N; k = k+1)
     {
       r = r + y[i][k]*z[k][j];
     };
     x[i][j] = r;
   };
```

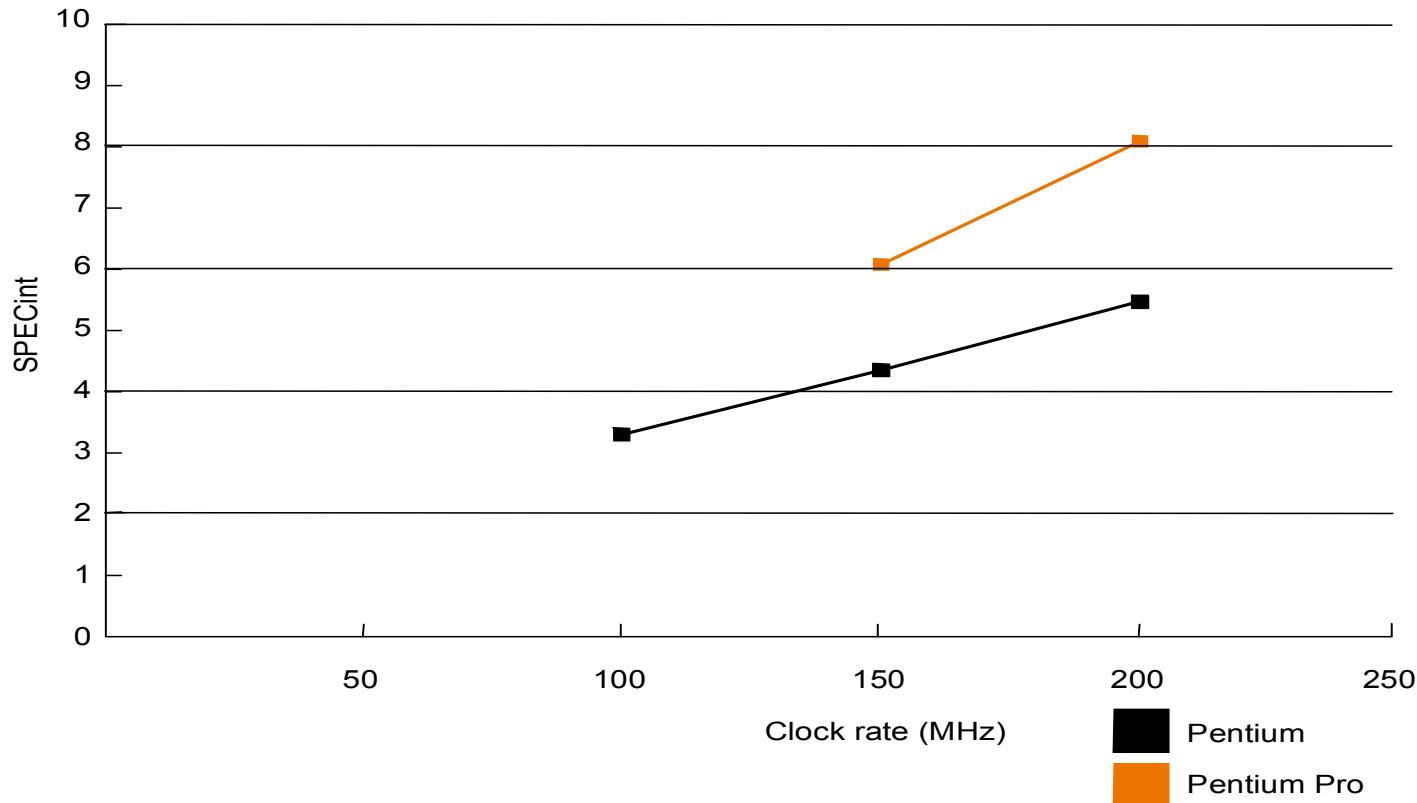
Code A

```
for (jj = 0; jj < N; jj = jj+B)
  for (kk = 0; kk < N; kk = kk+B)
    for (i = 0; i < N; i = i+1)
      for (j = jj; j < min(jj+B-1,N); j = j+1)
        {r = 0;
         for (k = kk; k < min(kk+B-1,N); k = k+1)
         {
           r = r + y[i][k]*z[k][j];
           x[i][j] = x[i][j] + r;
         };
       }
```

Code B

Why do I need to take this course ? (cont.)

- Q2: CPU frequency Performance

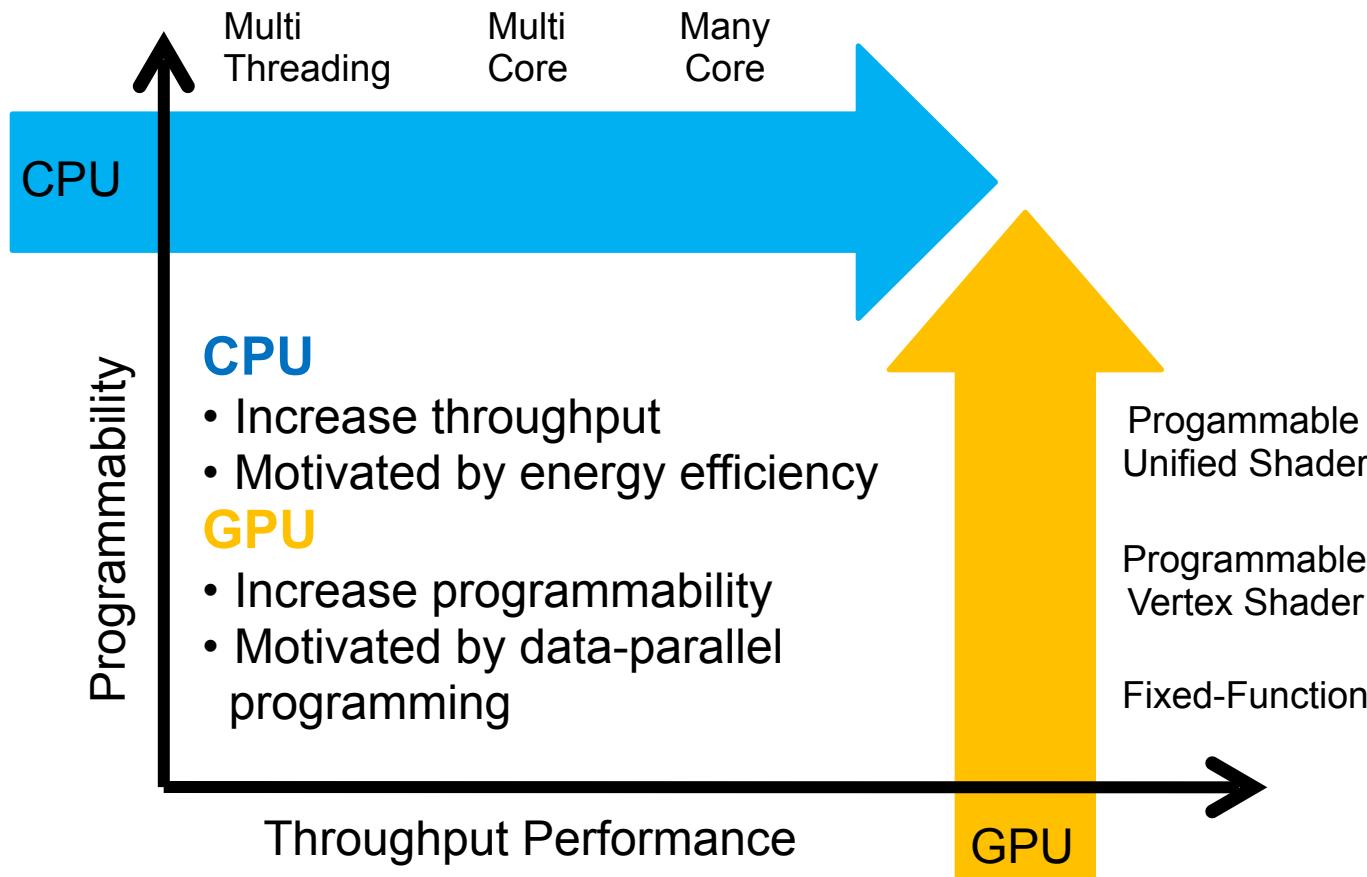


Why do I need to take this course ? (cont.)

Q3 : Why do Nvidia GPUs get so much attention today?

“The GPU is the Computer”

A general purpose computing engine, not just an accelerator”, GPU computing- To Exascale and Beyond, Bill Dally



Classes of Computing Applications

- **Desktop Computer**

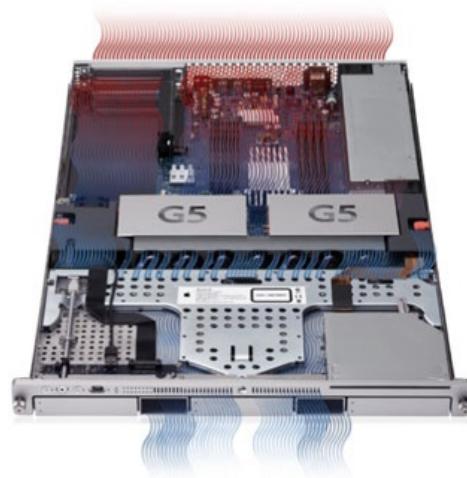
- A computer designed for use by an individual, usually incorporating a graphics display, keyboard, and mouse.
 - Design Emphasis
 - Performance & Cost (cost-effective solution)
 - Meeting performance requirement for a broad range of applications



Classes of Computing Applications (cont.)

■ Server

- A computer used for running larger programs for multiple users and typically accessed only via a network.
 - E.g.: file server, web server, supercomputer
- Design Emphasis
 - Dependability & Scalability
 - Throughput > Latency



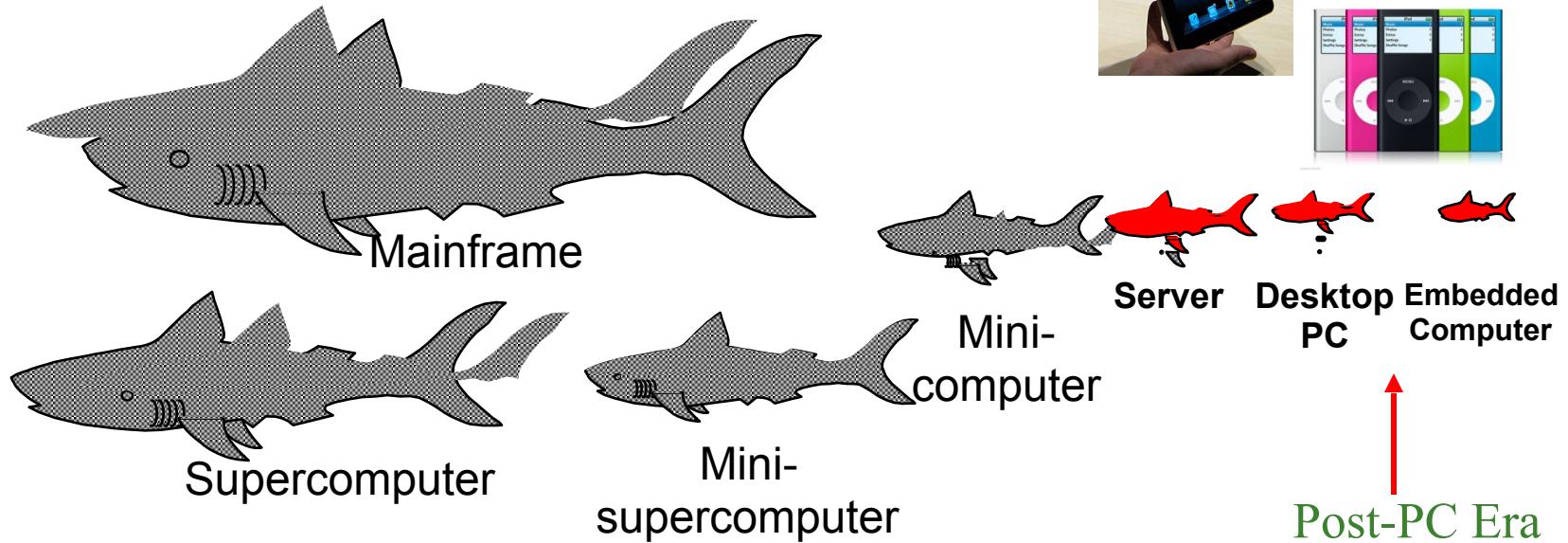
Classes of Computing Applications (cont.)

■ Embedded Computers

- ❑ A computer inside another device used for running one predetermined application or collection of software.
 - ❑ e.g., microcontroller for cars, phones, PDA, network processor
 - ❑ Design Emphasis
 - Minimize power & cost
 - Dependability



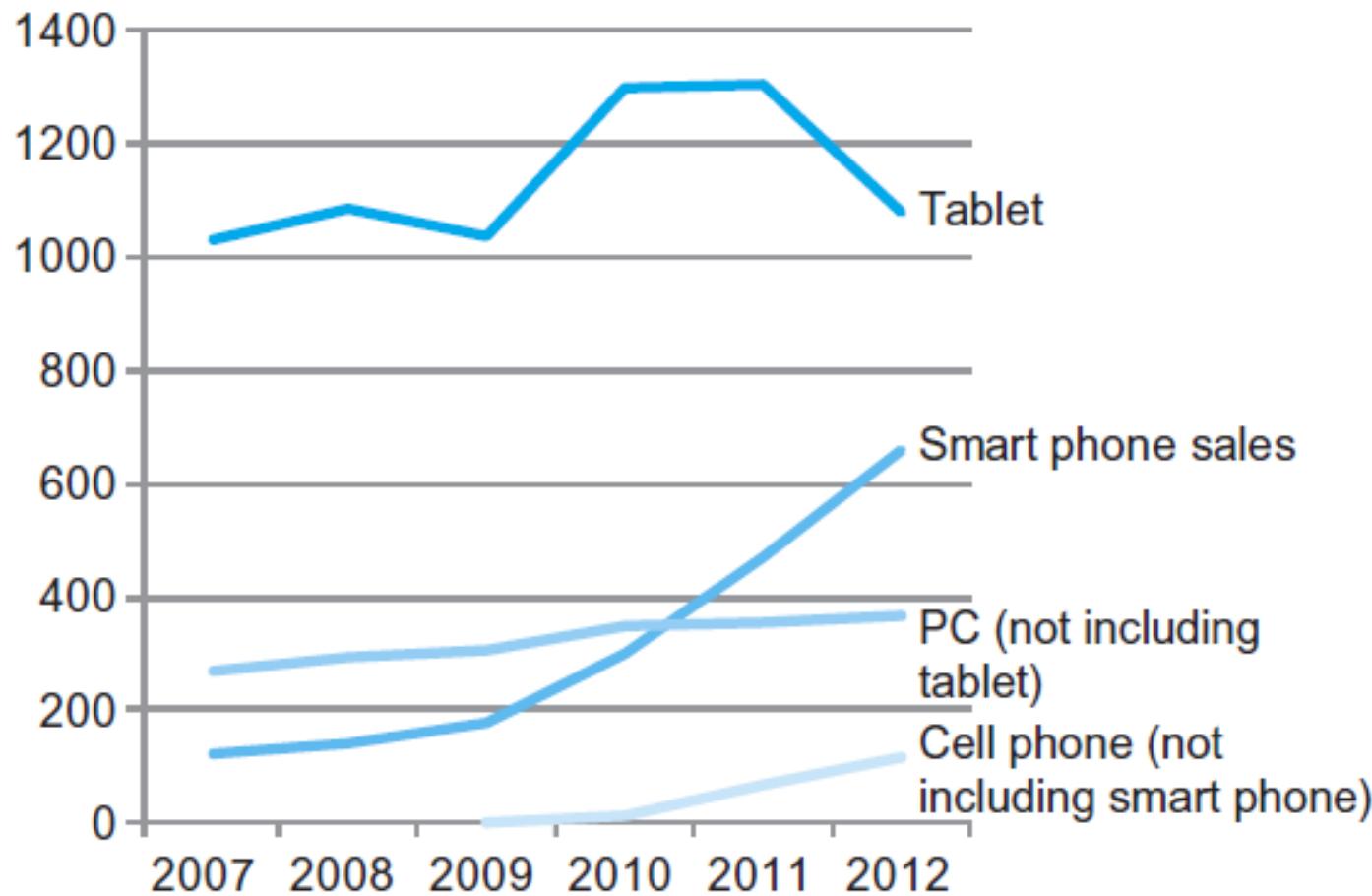
Computer Food Chain



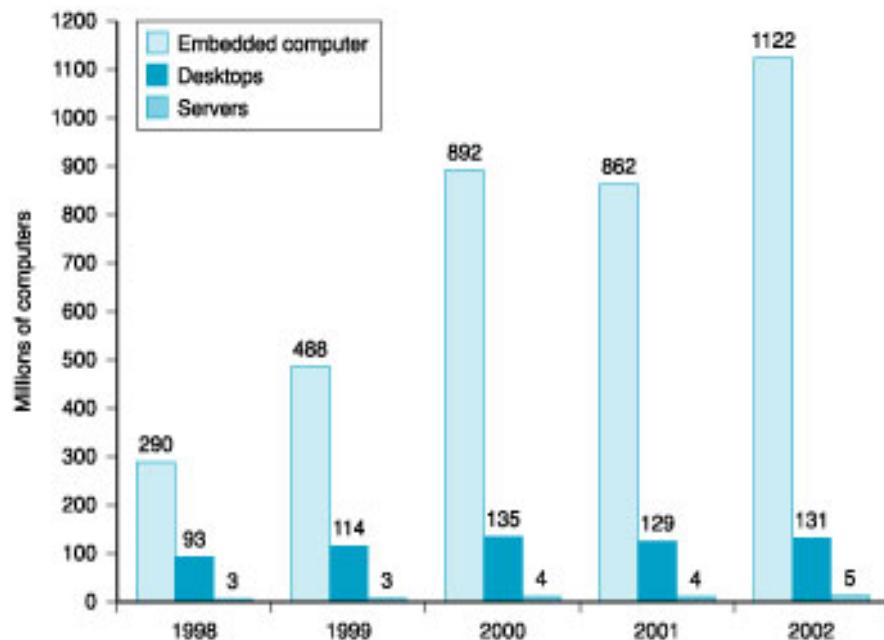
The PostPC Era

- Personal Mobile Device (PMD)
 - Battery operated
 - Connects to the Internet
 - Hundreds of dollars
 - Smart phones, tablets, electronic glasses
- Cloud computing
 - Warehouse Scale Computers (WSC)
 - Software as a Service (SaaS)
 - Portion of software run on a PMD and a portion run in the Cloud
 - Amazon and Google

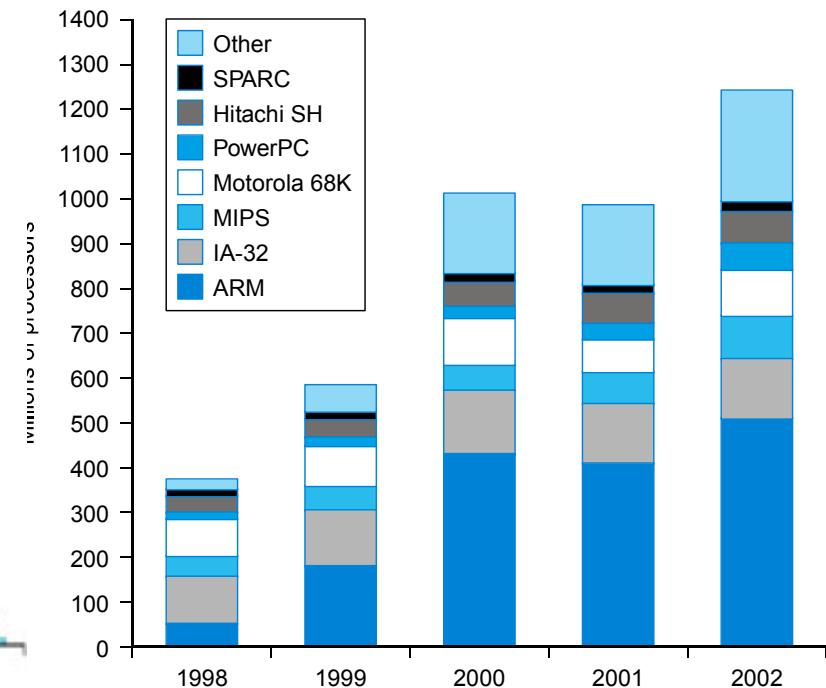
The PostPC Era



Who Wins the Market in PMD?



The number of distinct processors sold (1998~2002)



Sales of microprocessors by ISA (1998~2002)

ARM

Summaries from 商業周刊

Advanced RISC Machines Ltd. (ARM)

- Founded at the year of 1990, in Cambridge, England
 - Co-founded by Apple & Acron
- First low-power, 32 bit mobile processor



今昔比較



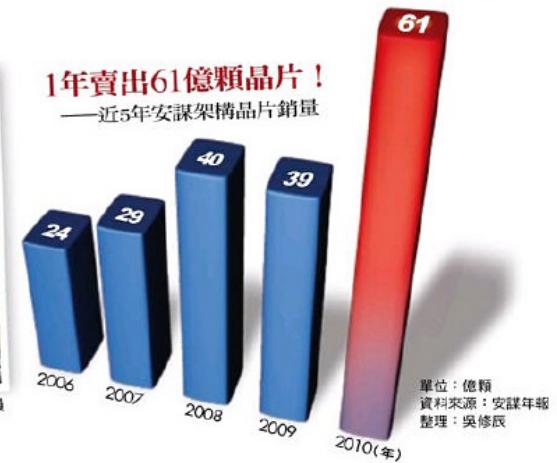
程傑拍攝



■安謀在劍橋總部（圖）的員工約900人，另有超過1,000名員工散布在海外12個國家。

1年賣出61億顆晶片！

—近5年安謀架構晶片銷量



Key factors for ARM to succeed

- Emphasize low-power since day one
- New business model
 - ARM sells designs not chips

▪ IP license (矽智財授權)
▪ ARM alliance: Samsung, Qualcomm, Mediatek, and more...



ARM Announces CPU Design Center in Taiwan, 2014



ARM was acquired by Softbank on 2016

[更新] 軟銀完成收購，ARM 正式成日資企業(加上軟銀、ARM 公開信)

作者 MoneyDJ | 發布日期 2016 年 09 月 06 日 9:15 | 分類 手機, 物聯網, 零組件 [Follow](#) [G+](#) [讚 16](#) [分享](#)



軟銀 (Softbank) 5 日發布新聞稿宣布，已完成對英國半導體巨擘安謀 (ARM Holdings) 的收購手續，已砸下 240 億英鎊 (約3.3兆日圓) 取得 ARM 已發行以及預定發行的所有股票，正式將 ARM 納為旗下完全子公司行列。軟銀對 ARM 的收購金額創下日本企業海外併購案的史上最大規模紀錄。

Chip War in Data Centers

- Small cores vs. Large cores



Microsoft's Quincy Data Center

Intel's Xeon

- Intel Xeon processors 7500 series
 - Designed for intelligent performance, smart energy efficiency
 - 8 cores / 16 threads
 - P6 architecture, 4-wide, out-of-order issue x86 CPU
 - 24MB of L3 shared cache
 - Quad-channel memory controller
 - Turbo boost technology
 - Intel virtualization technology
- NEC Express5800/A1080a-E server
 - CPU: 8 Xeon 7500 series CPUs
 - 64 cores / 128 threads
 - Memory: 128 DDR3-1066 DIMM slots (max: 2TB)
 - Storage: 12 hot plug slots



Intel Xeon 7500



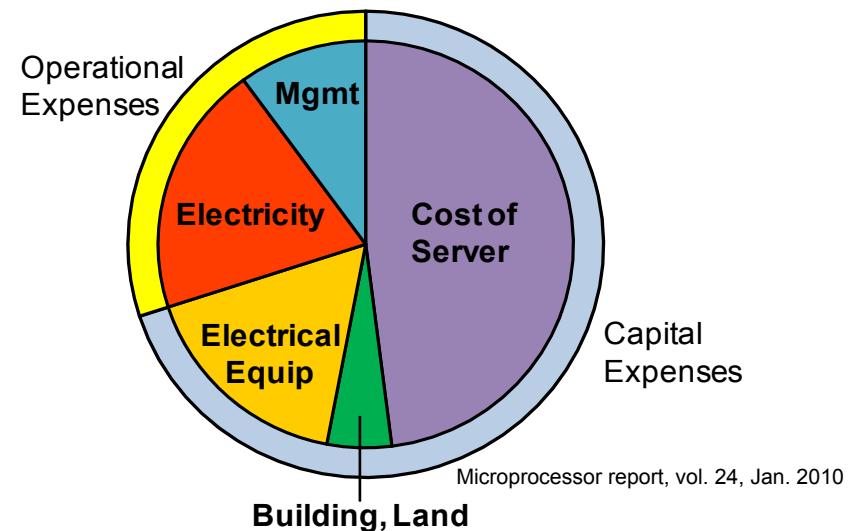
Powerful.
Intelligent.



NEC Express5800/A1080a server

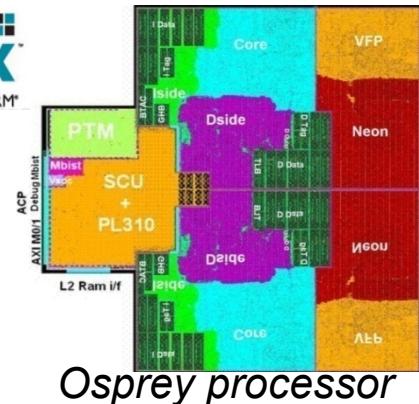
New opportunities for Small Cores

- Workload changes in data centers
 - Past: databases, financial services
 - Now: web service, cloud computing (accessing mails, photos, facebooks, etc)
 - Simple tasks & large amount of parallel tasks
- Servers must cut power

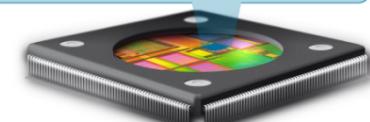
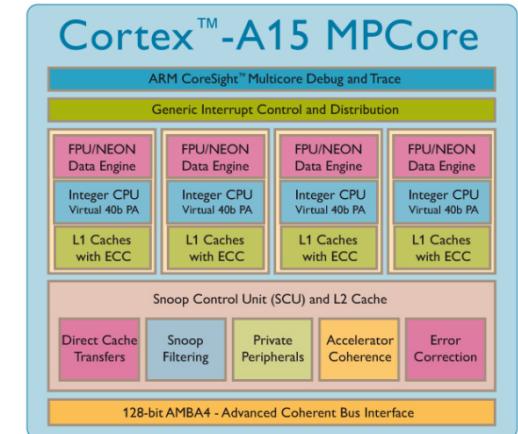


ARM's ambition

- Fall 2009, ARM announced its first multi-core server processor
 - Osprey: Dual Cortex-A9 cores
 - Cortex-A9 hard macros
 - TSMC 40G speed optimized
 - For enterprise servers
 - TSMC 40G power optimized
 - For mobile devices
- 2010 New processor: Eagle (Cortex-A15)
 - Aimed at the network infrastructure, server, and cloud-computing space
 - Hardware virtualization support
 - 4 CPUs sharing a L2 cache, linked by the AMBA-4 bus architecture.
- AMD announced ARM-based micro server
~ 2014



Osprey processor



Eagle processor (Cortex-A15)

Microsoft unveils new ARM server designs, threatening Intel's dominance

by James Vincent | @jjvincent | Mar 9, 2017, 7:45am EST

[f SHARE](#)

[t TWEET](#)

[in LINKEDIN](#)



NOW

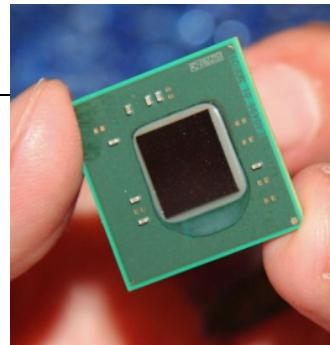


Chatbot lets
to \$25,000 wi

Intel's Atom

Intel Atom processor D525

- ❑ Designed for mobile internet devices, netbooks, and entry-level desktops.
 - 2 cores / 4 threads
 - Clock speed 1.8 GHz
 - 1MB of L2 cache
 - Voltage range: 0.8V-1.175V



Intel Atom processor

SeaMicro SM10000 (Atomized Server)

- ❑ 512 Intel Atom processors
 - 8 processors in a Compute Card
 - 64 Compute Cards in a system
- ❑ Uses 1/4 the power and 1/4 the space of today's best in class volume servers



SeaMicro SM10000



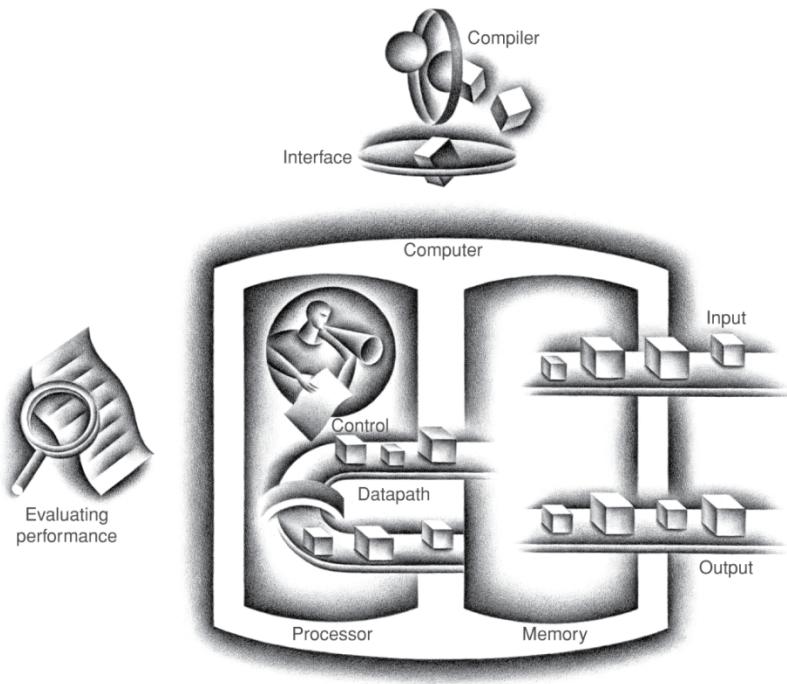
8 CPUs in a Compute Card

RISC-V

- RISC-V (pronounced "risk-five") is a new instruction set architecture (ISA) that was originally designed to support computer architecture research and education, which we now hope will become a standard open architecture for industry implementations. RISC-V was originally developed in the Computer Science Division of the EECS Department at the University of California, Berkeley
 - Lead by Prof. David Patterson and Prof. Krstye Asanoic

Components of a Computer

The BIG Picture



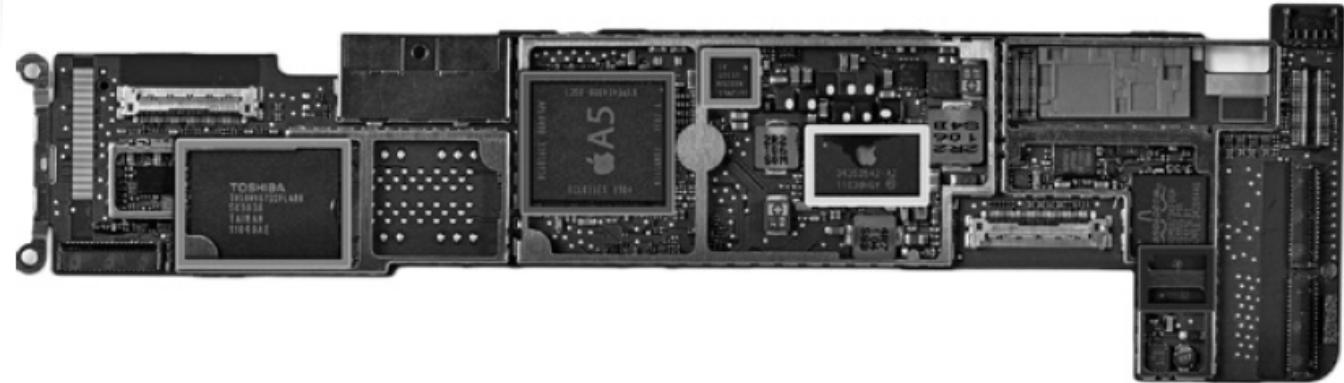
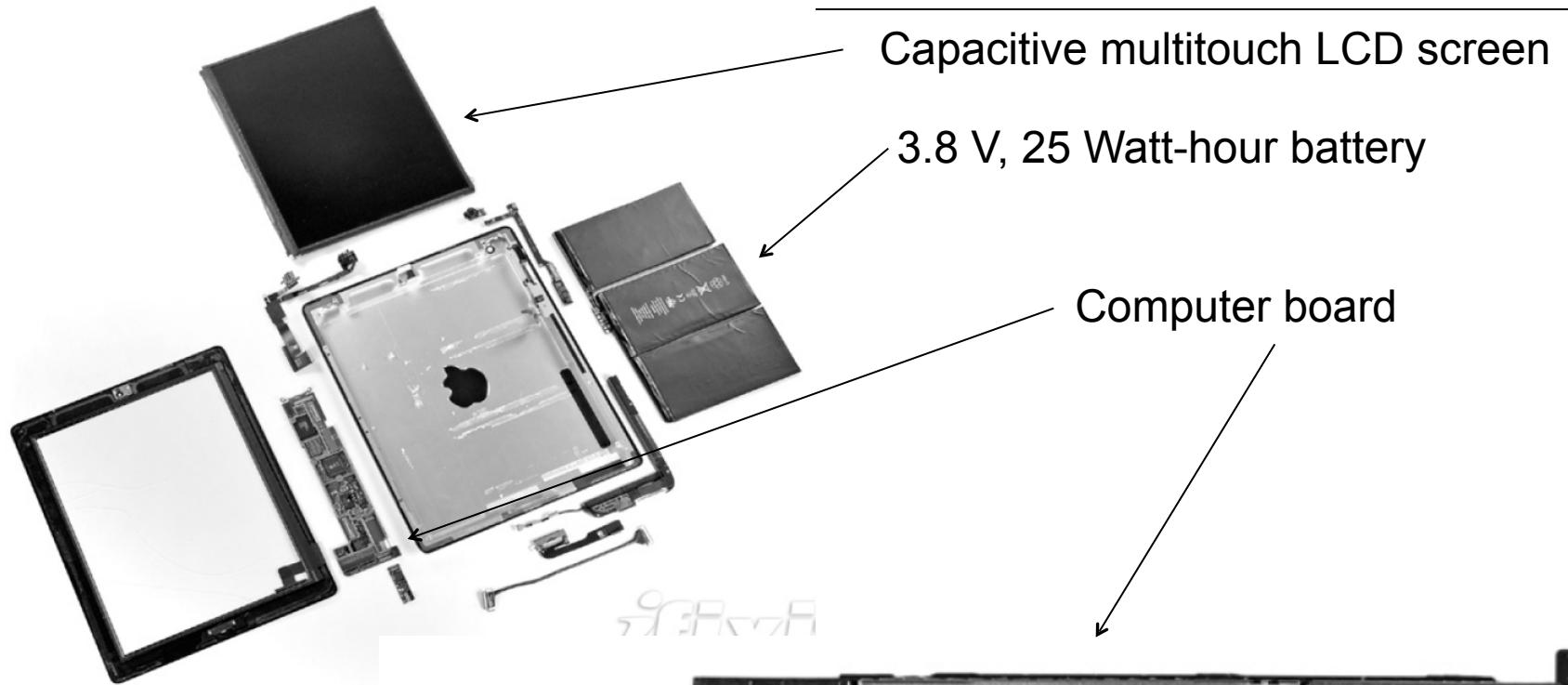
- Same components for all kinds of computer
 - Desktop, server, embedded
- Input/output includes
 - User-interface devices
 - Display, keyboard, mouse
 - Storage devices
 - Hard disk, CD/DVD, flash
 - Network adapters
 - For communicating with other computers

Touchscreen

- PostPC device
- Supersedes keyboard and mouse
- Resistive and Capacitive types
 - Most tablets, smart phones use capacitive
 - Capacitive allows multiple touches simultaneously



Opening the Box



Inside the Processor (CPU)

- Datapath: performs operations on data
- Control: sequences datapath, memory, ...
- Cache memory
 - Small fast SRAM memory for immediate access to data

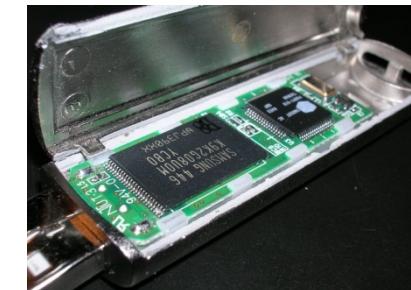
Inside the Processor

■ Apple A5



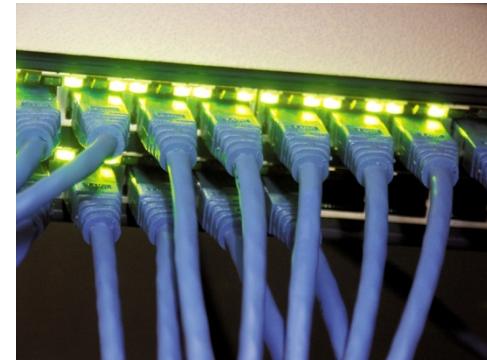
A Safe Place for Data

- Volatile main memory
 - Loses instructions and data when power off
- Non-volatile secondary memory
 - Magnetic disk
 - Flash memory
 - Optical disk (CDROM, DVD)

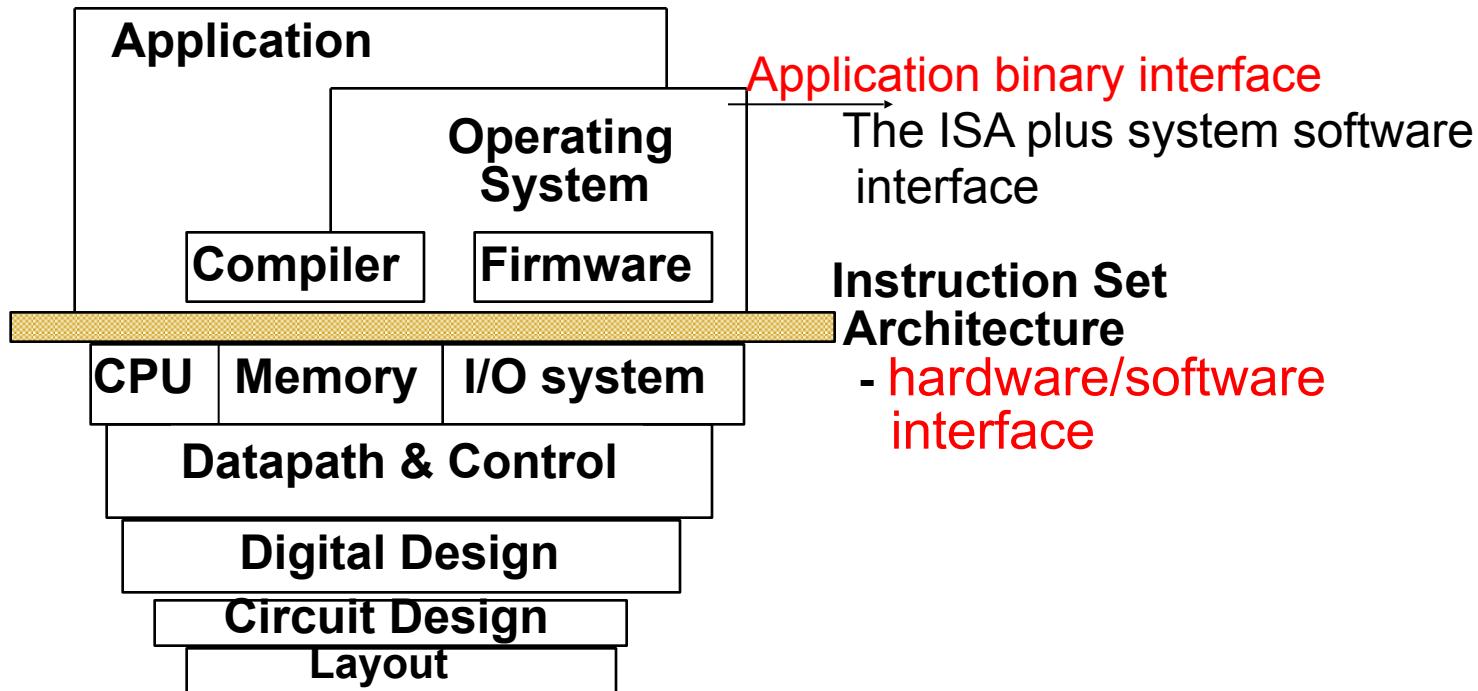


Networks

- Communication, resource sharing, nonlocal access
- Local area network (LAN): Ethernet
- Wide area network (WAN): the Internet
- Wireless network: WiFi, Bluetooth



Coordination of Layers of Abstraction



Abstraction helps us deal with complexity - hide lower-level details

Below Your Program

■ System software

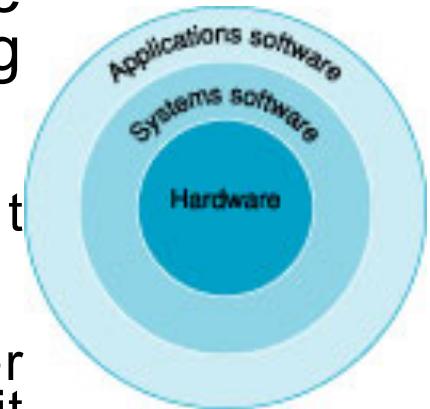
- Software that provides services that are commonly useful, including operating systems, compilers, and assemblers.

- Operating system

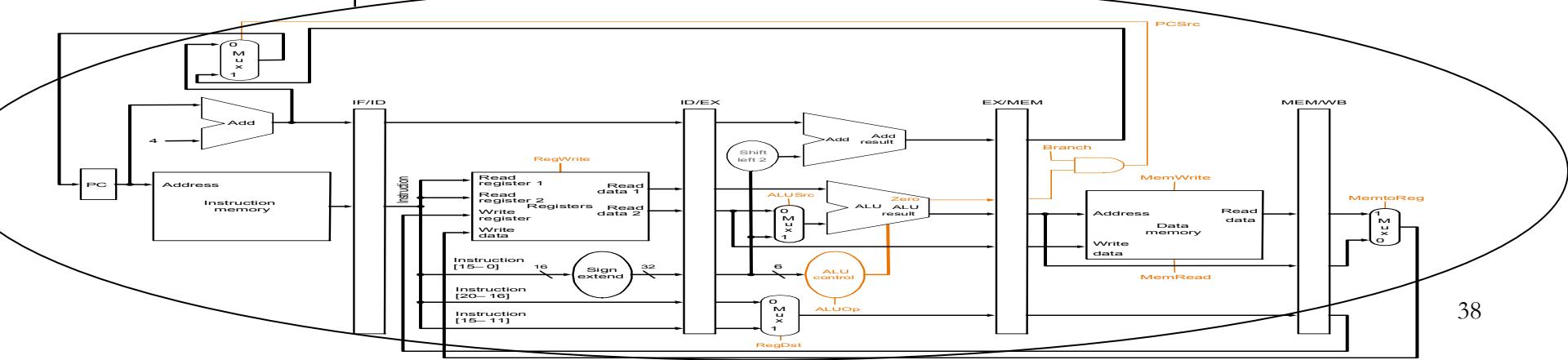
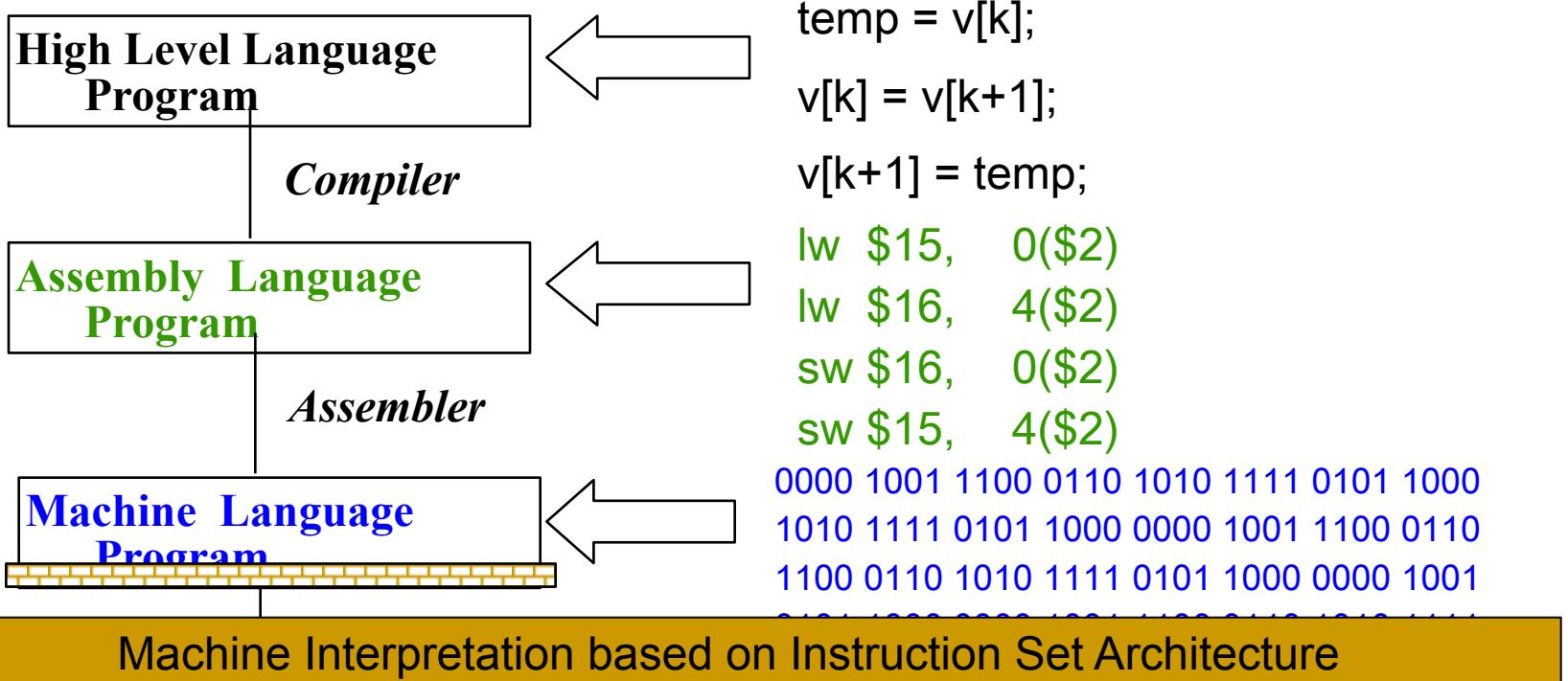
- Handling basic input and output operations
 - Allocating storage and memory
 - Providing for sharing the computer among multiple applications using it simultaneously

- Compiler

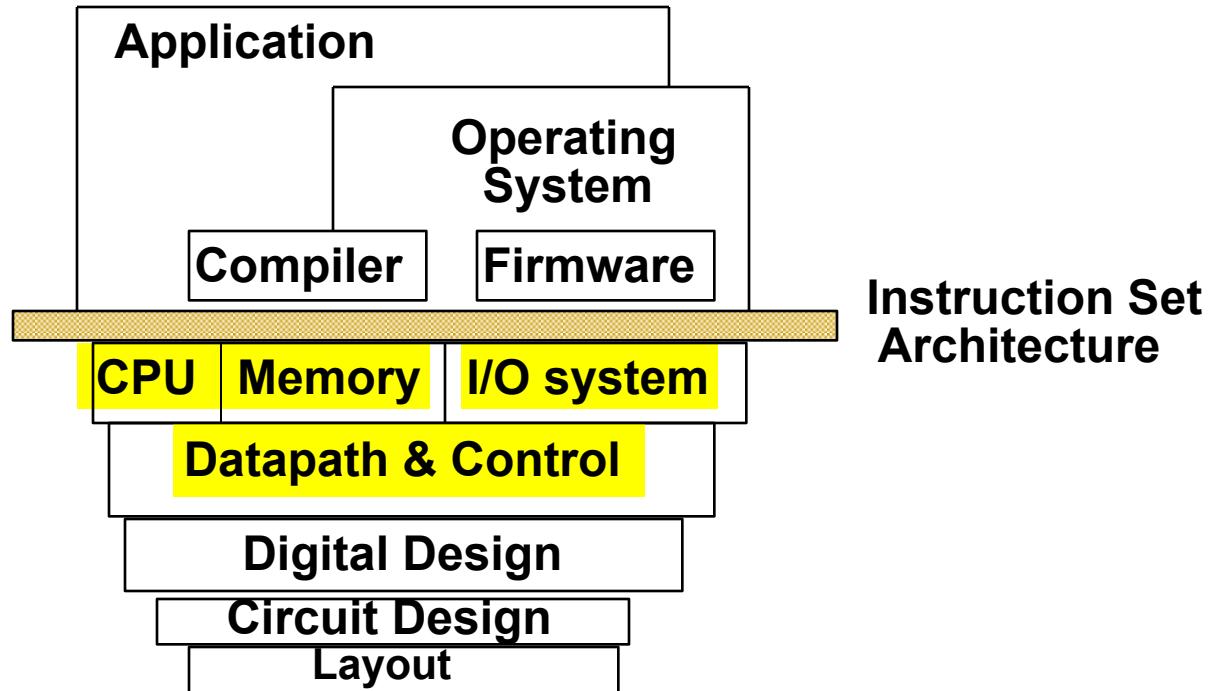
- A program that translates high-level language statements into assembly language statements.



Levels of Representation



What is “Computer Architecture” ?

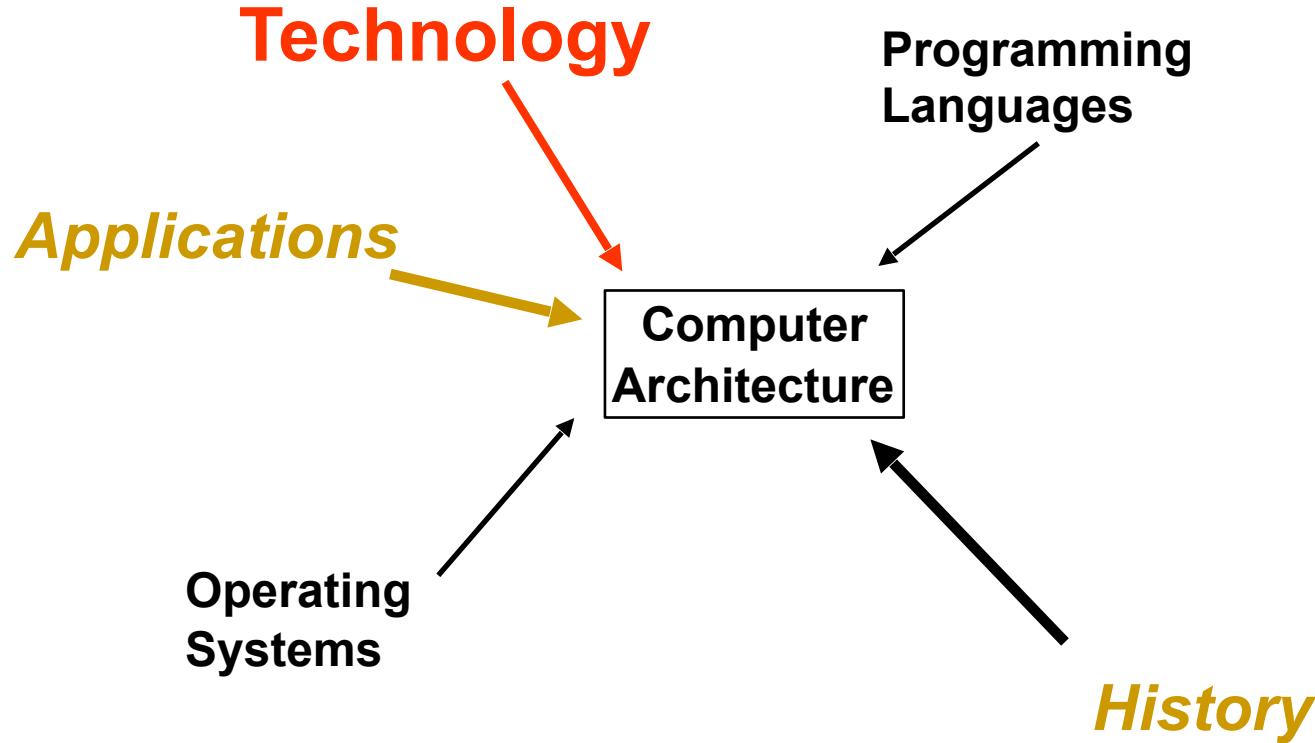


“What really matters is the functioning of the complete system, hardware, runtime system, compiler, operating system, and application”

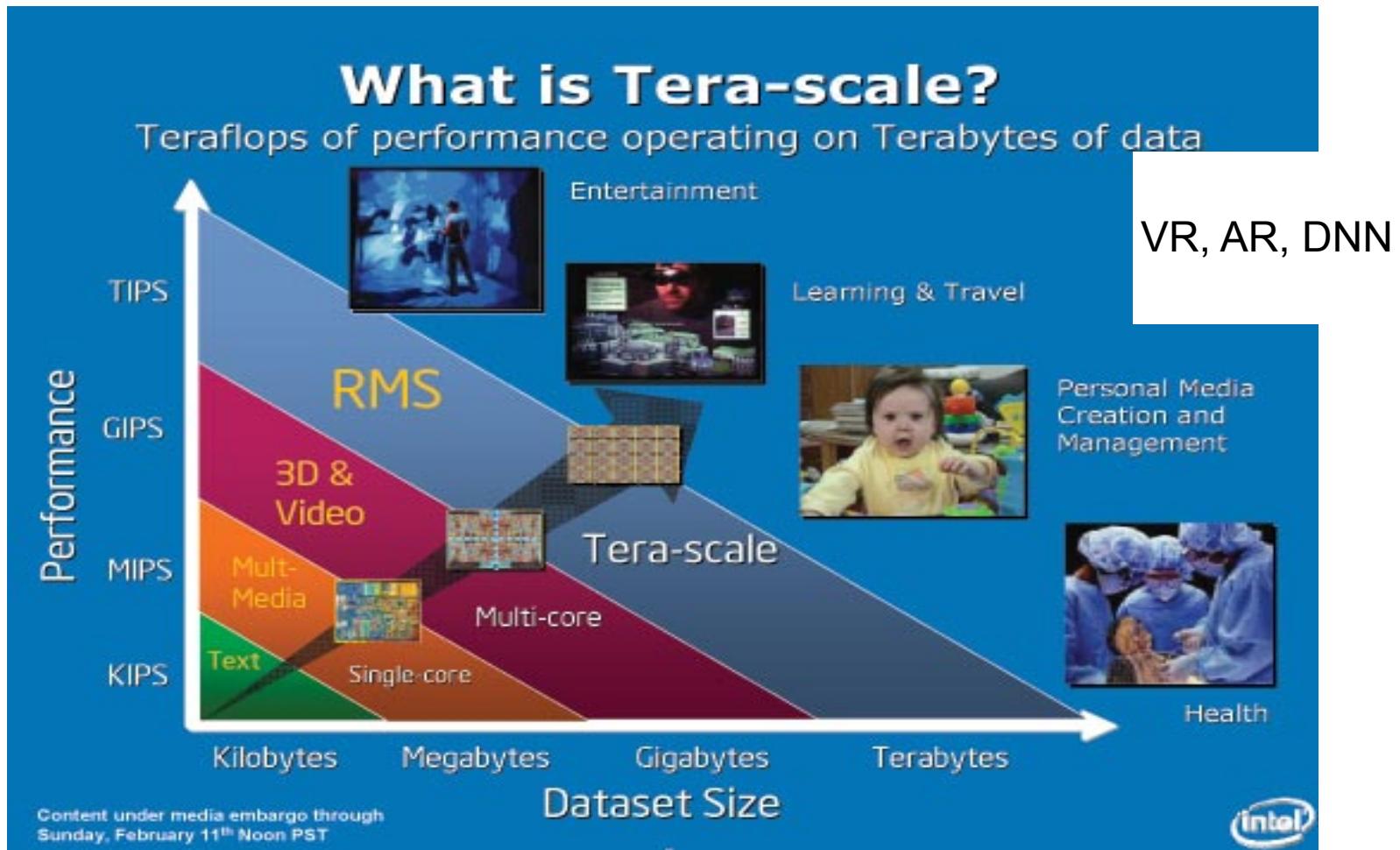
“In networking, this is called the “End to End argument”

--- H&P

Forces on Computer Architecture



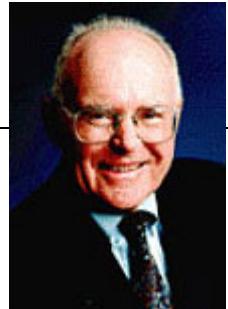
Application



Technologies Used in Computer

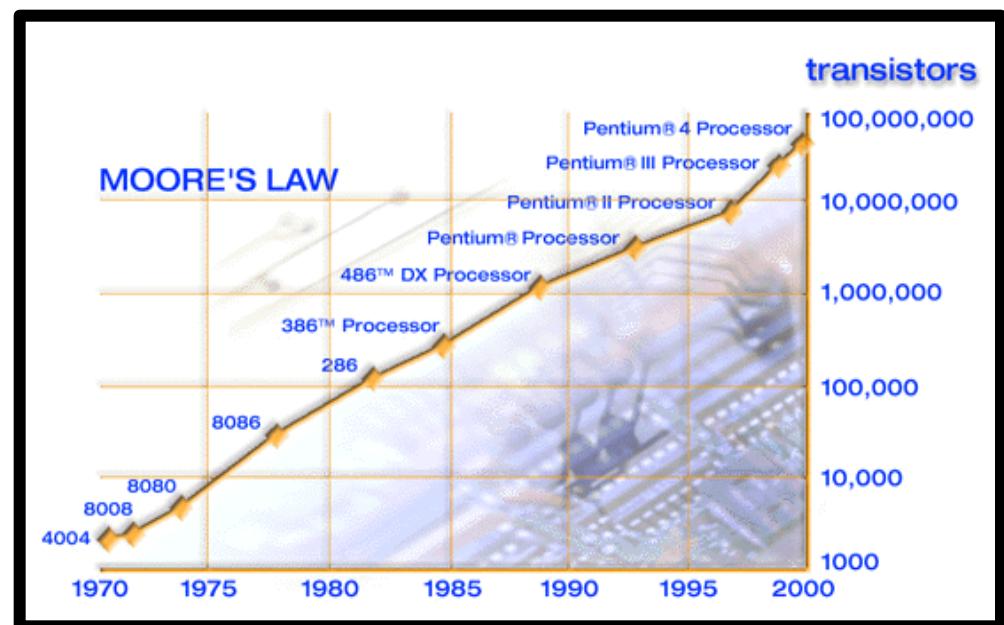
Year	Technology	Relative performance/cost
1951	Vacuum tube	1
1965	Transistor	35
1975	Integrated circuit (IC)	900
1995	Very large scale IC (VLSI)	2,400,000
2013	Ultra large scale IC	250,000,000,000

Moore's Law

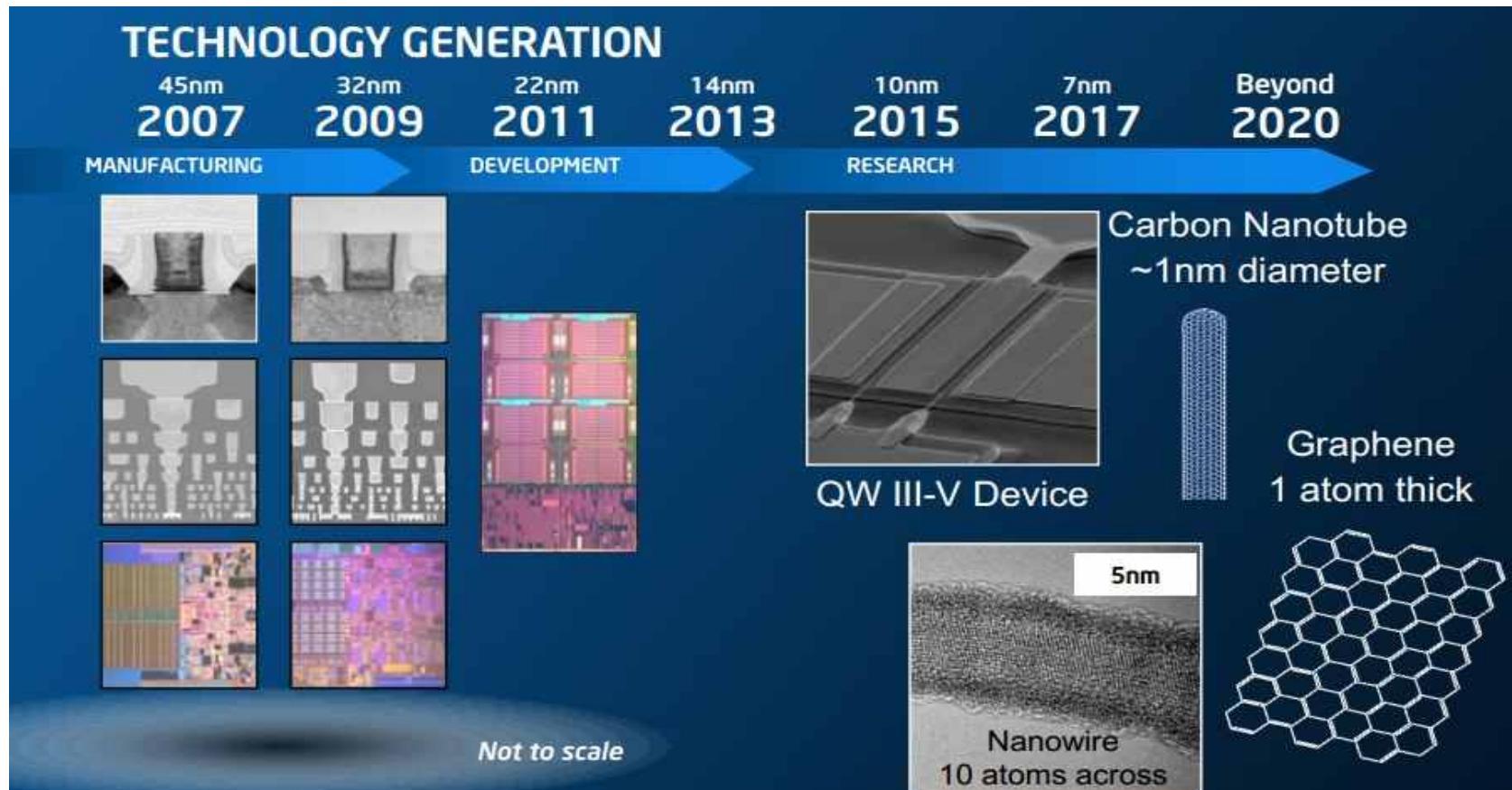


- Moore's Law (1965)
 - Gordon Moore, Intel founder
 - “The density of transistors in an integrated circuit will double every year.”

- Reality
 - “The density of silicon chips doubles every 18 months.”



Technology Outlook



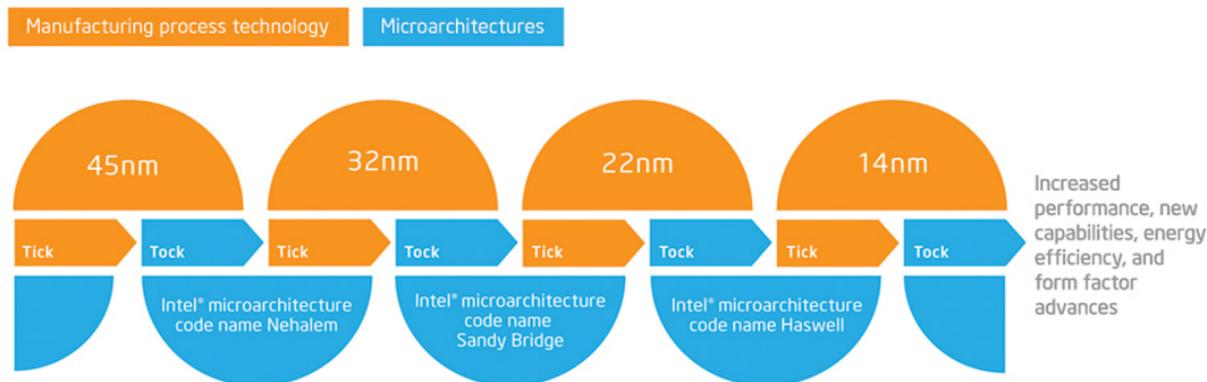
- Silicon lattice is ~ 0.5nm, hard to imagine good devices smaller than 10 lattices across - reached in 2020



Monolithic 3D is now on the roadmap for 2019, 08/01/2013

<http://www.electroiq.com/articles/sst/2013/08/monolithic-3d-is-now-on-the-roadmap-for-2019.html>

The Tick-Tock model through the years

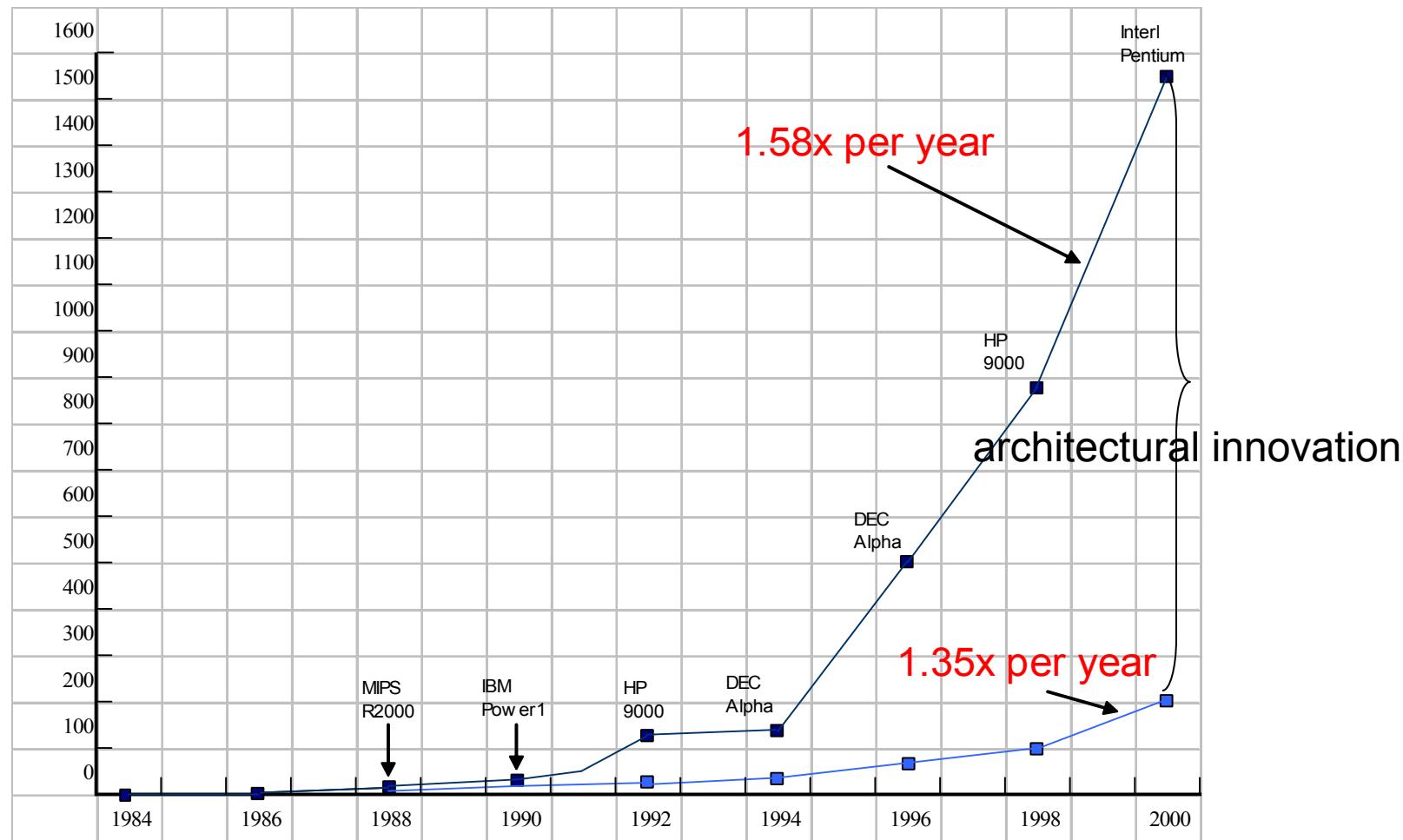


PAO: Process-Architecture-Optimization

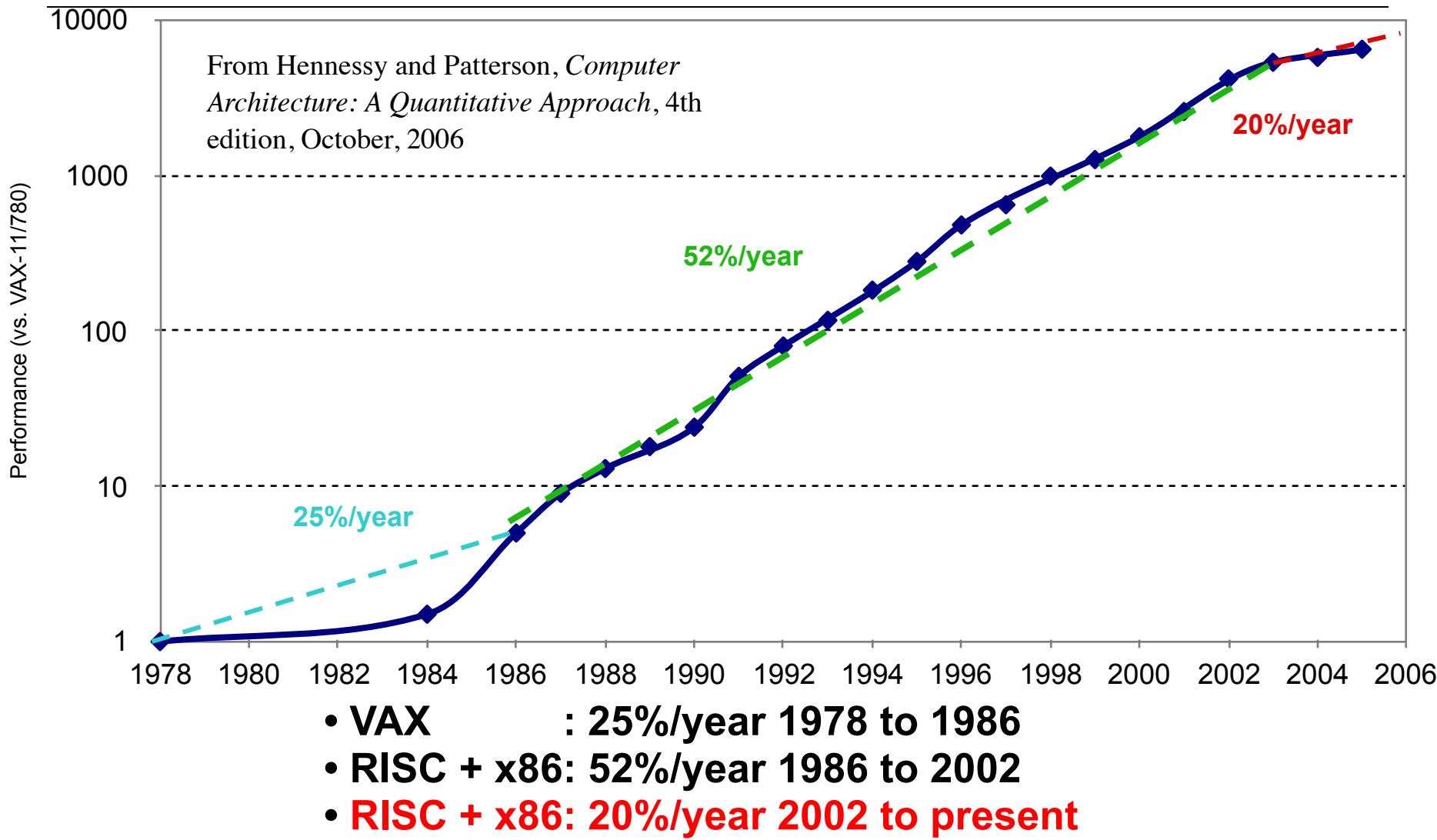
Intel PAO Schedule			
Cycle	<u>Process</u>	Introduction	Microarchitecture
Process	<u>14 nm</u>	2014	<u>Broadwell</u>
Architecture	<u>14 nm</u>	2015	<u>Skylake</u>
Optimization	<u>14 nm</u>	2016	<u>Kaby Lake</u>
Optimization	<u>14 nm</u>	2017	<u>Coffee Lake</u>
Process	<u>10 nm</u>	2017	<u>Cannonlake</u>
Architecture	<u>10 nm</u>	2018	<u>Icelake</u>
Optimization	<u>10 nm</u>	2019	<u>Tigerlake</u>

→ Expected in Oct

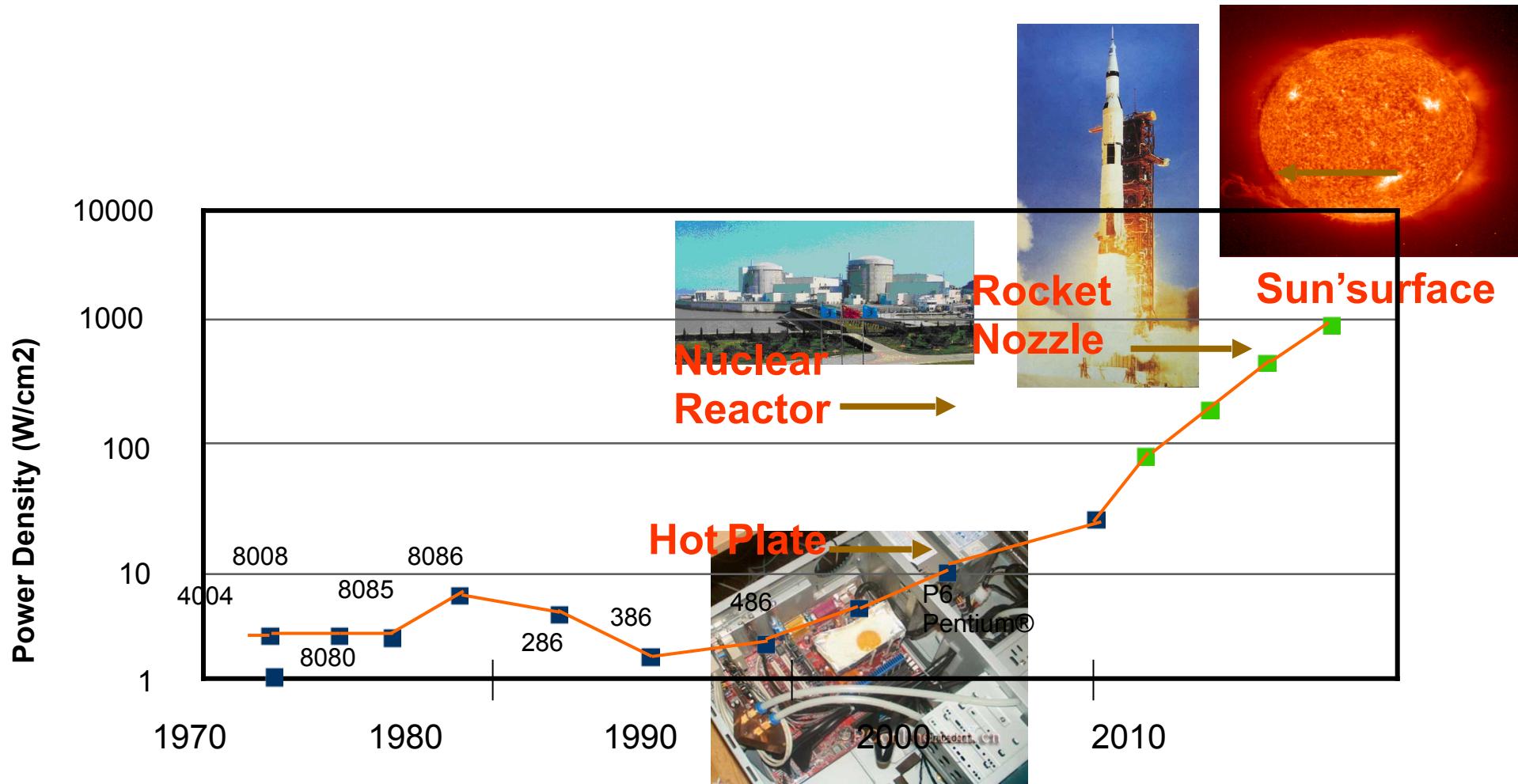
Processor Performance



Crossroads: Uniprocessor Performance

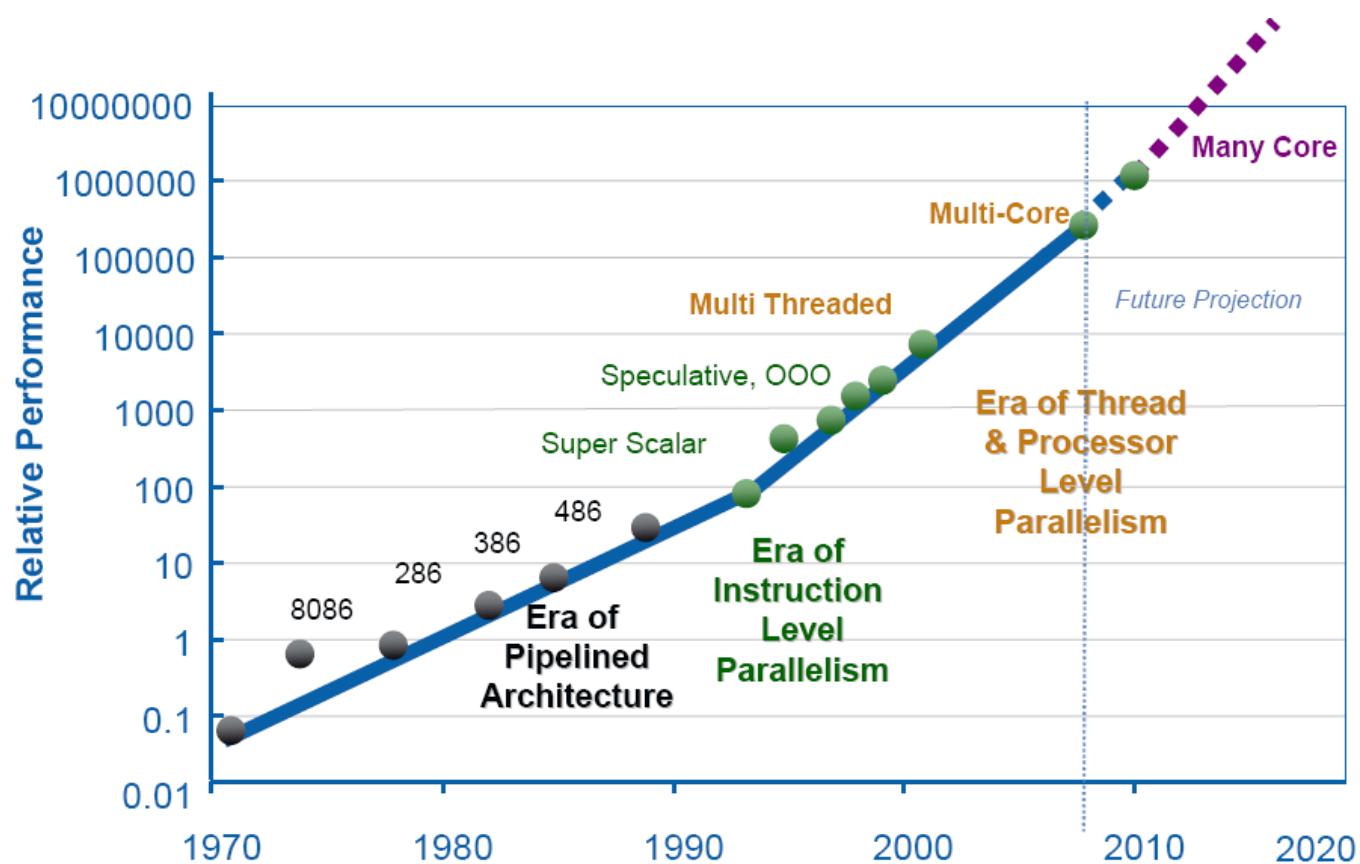


Power Dissipation Increases



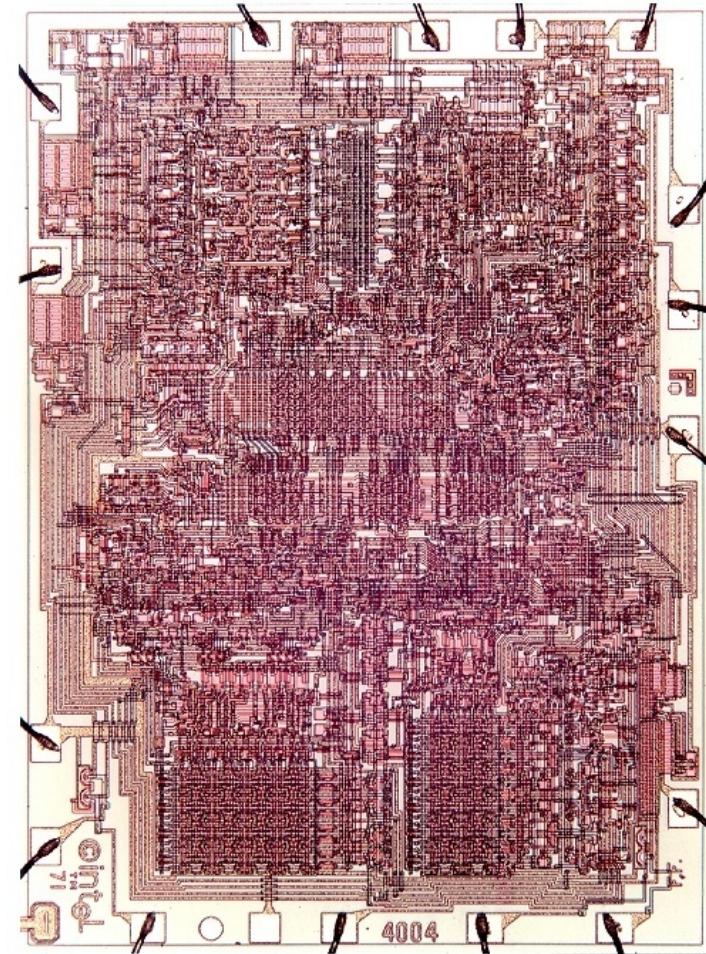
Parallelism for Energy Efficiency

Present and Future



Evolution of Intel Microprocessors : 4004

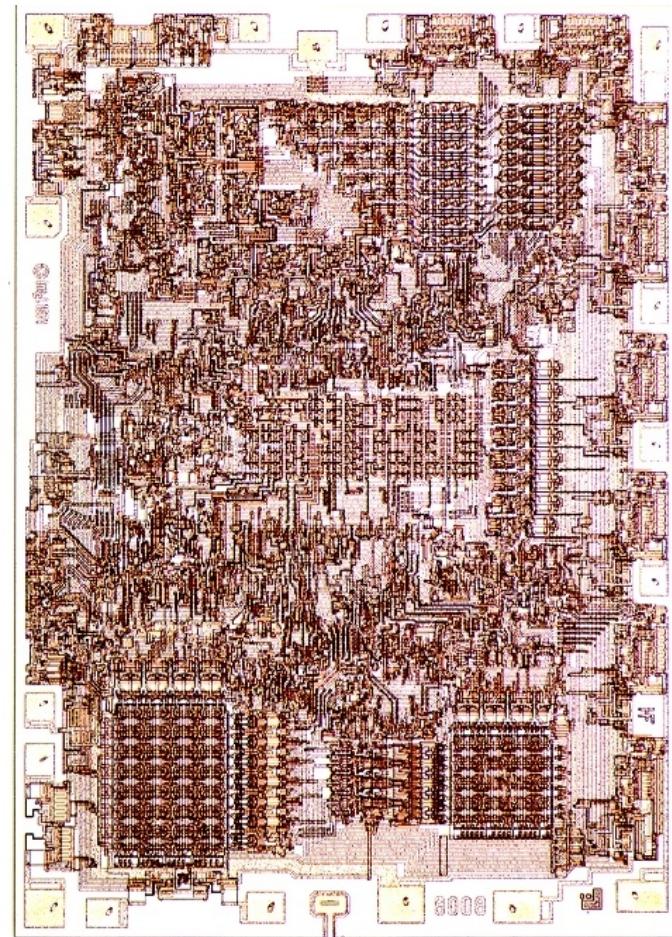
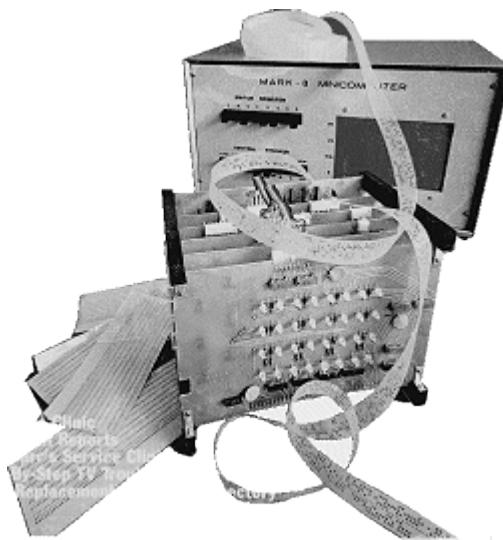
- First microprocessor (1971)
 - For Busicom calculator
- Characteristics
 - 10 μm process
 - 2300 transistors
 - 400 – 800 kHz
 - 4-bit word size



Courtesy of Intel Museum

8008

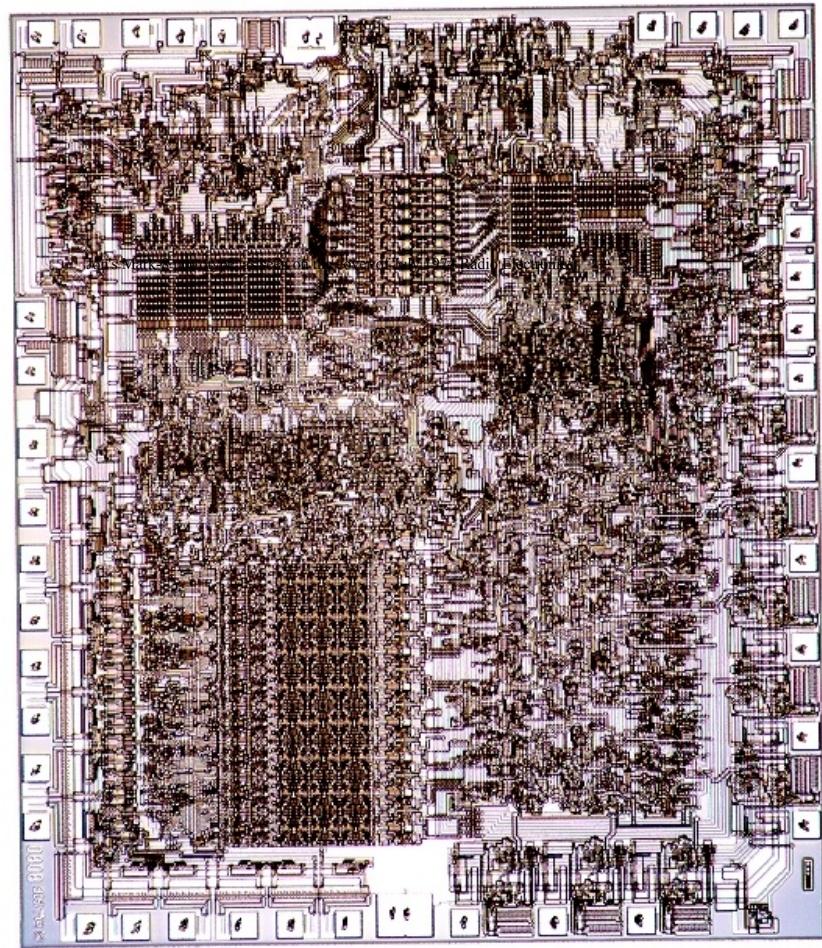
- 8-bit follow-on (1972)
 - Mark 8: Dumb terminals
- Characteristics
 - 10 μm process
 - 3500 transistors
 - 500 – 800 kHz
 - 8-bit word size



Courtesy of Intel Museum

8080

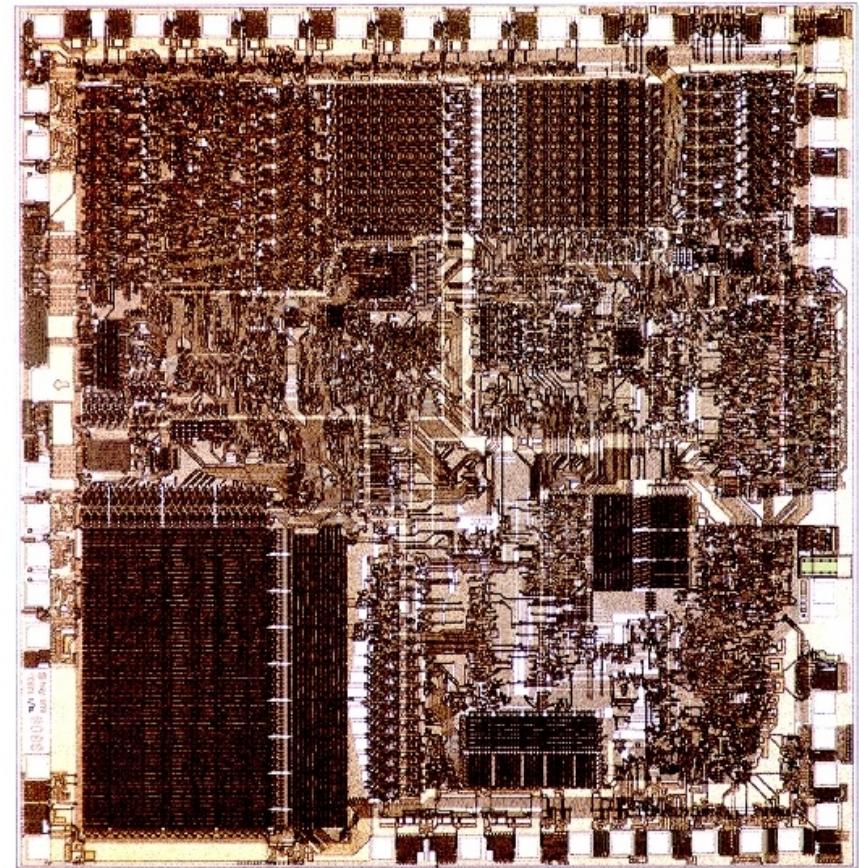
- 16-bit address bus (1974)
 - Altair : first personal computer
- Characteristics
 - 6 μm process
 - 4500 transistors
 - 2 MHz
 - 8-bit word size



Courtesy of Intel Museum

8086 / 8088

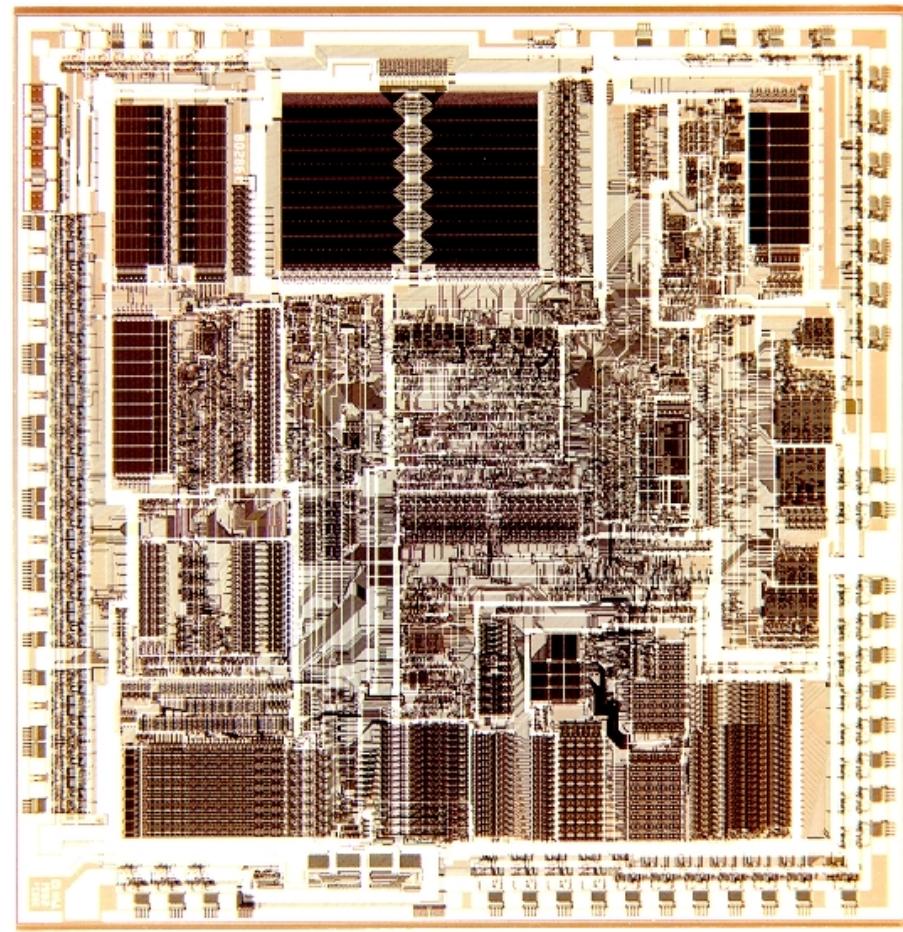
- 16-bit processor (1978-9)
 - IBM PC and PC XT
 - Revolutionary products
 - Introduced x86 ISA
- Characteristics
 - 3 μm process
 - 29k transistors
 - 5-10 MHz
 - 16-bit word size



Courtesy of Intel Museum

80286

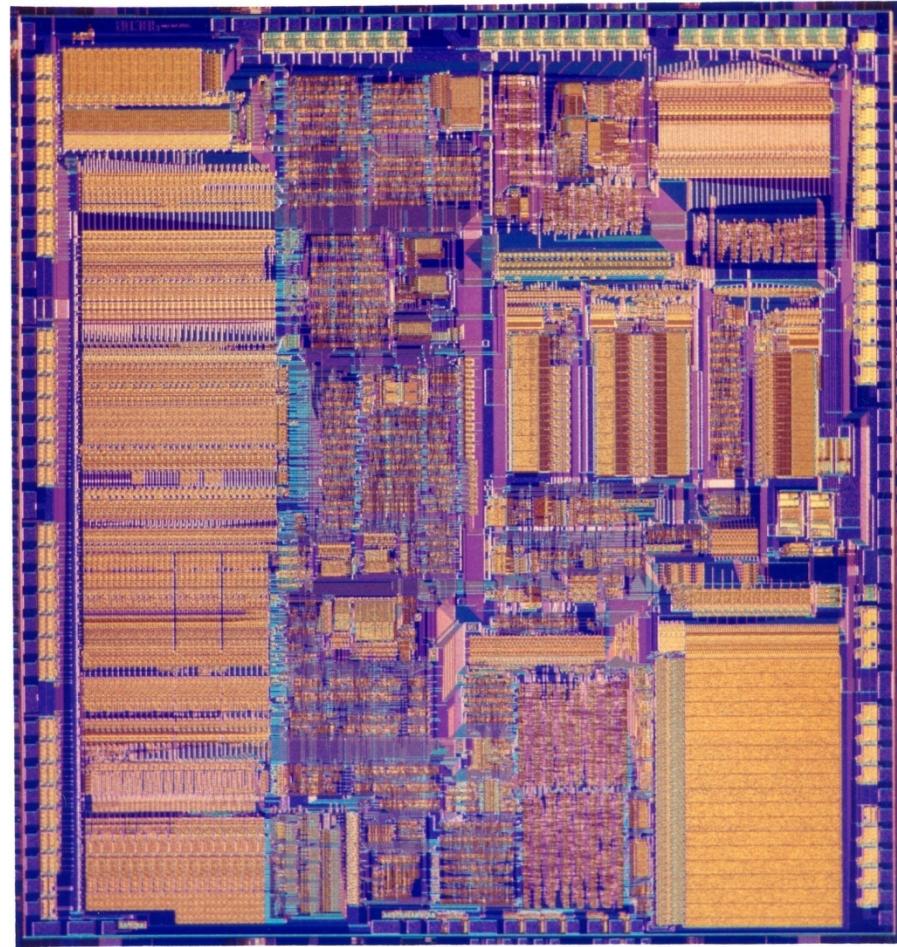
- **Virtual memory (1982)**
 - IBM PC AT
- **Characteristics**
 - 1.5 μ m process
 - 134k transistors
 - 6-12 MHz
 - 16-bit word size



Courtesy of Intel Museum

80386

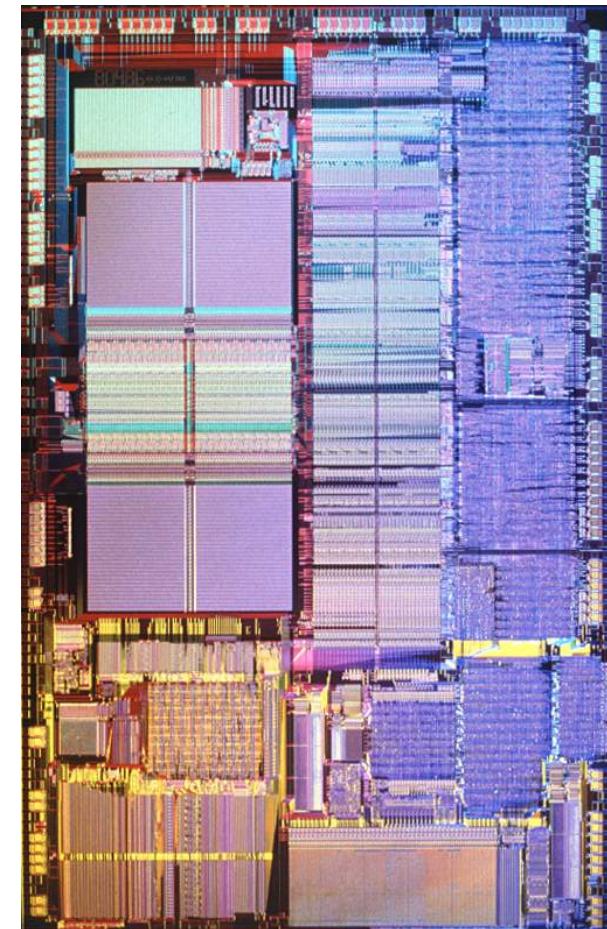
- 32-bit processor (1985)
 - Modern x86 ISA
- Characteristics
 - 1.5-1 μm process
 - 275k transistors
 - 16-33 MHz
 - 32-bit word size



Courtesy of Intel Museum

80486

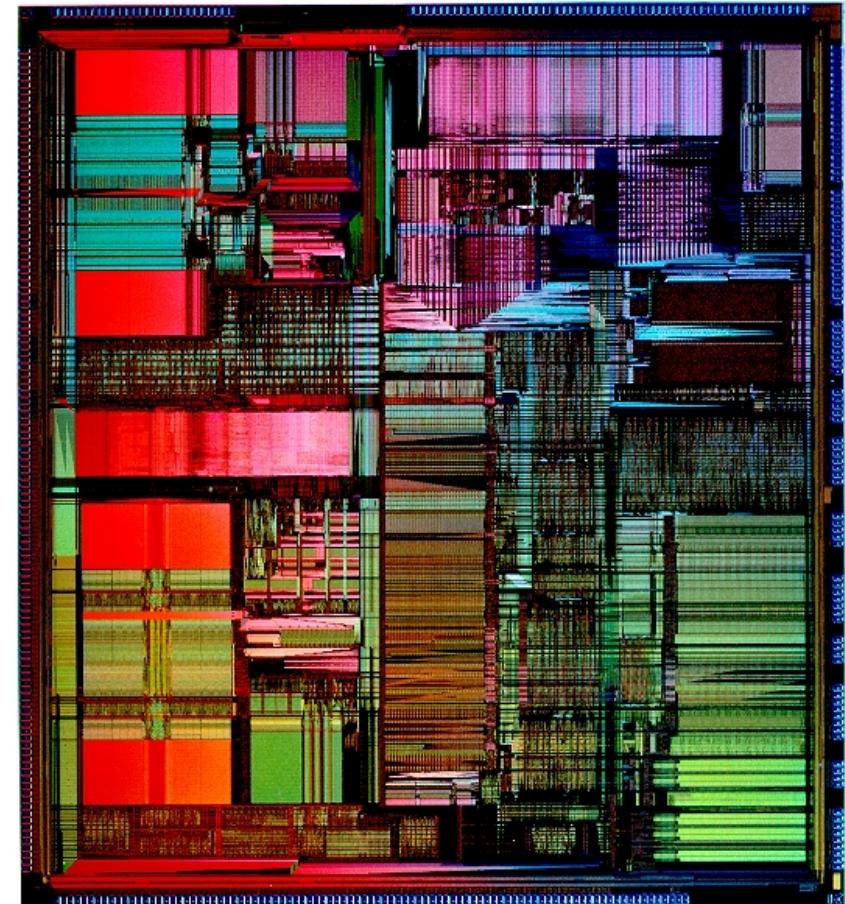
- **Pipelining (1989)**
 - Floating point unit
 - **8 KB cache**
- **Characteristics**
 - 1-0.6 μm process
 - 1.2M transistors
 - 25-100 MHz
 - 32-bit word size



Courtesy of Intel Museum

Pentium

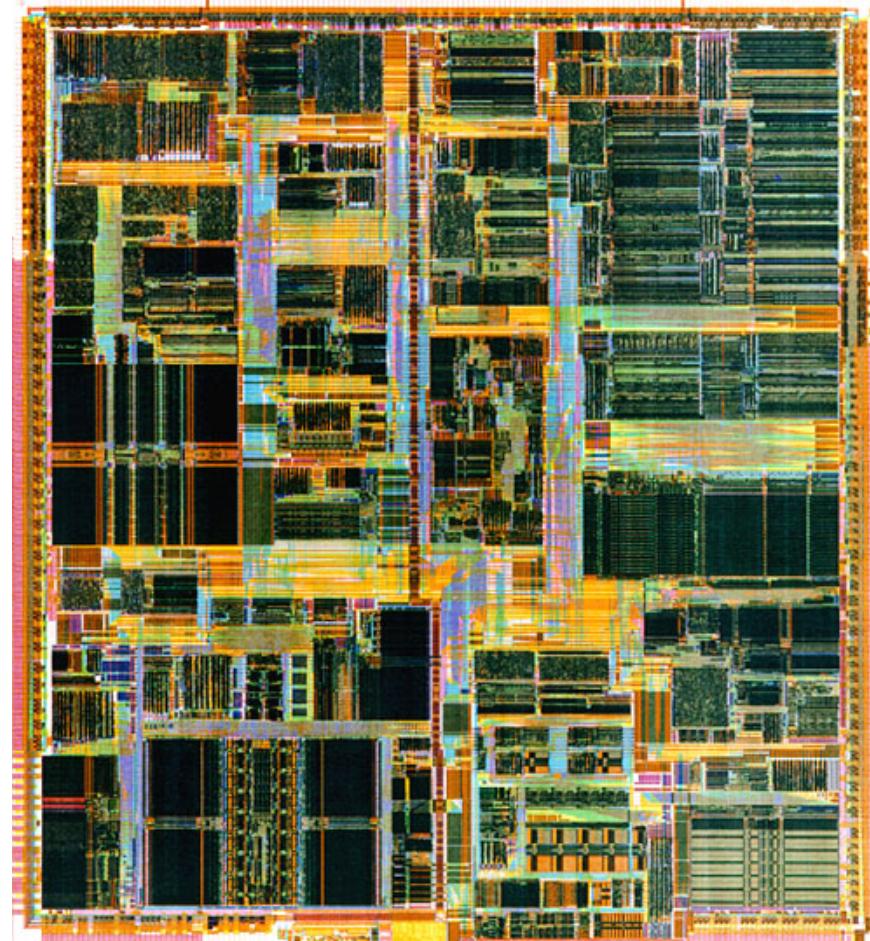
- **Superscalar (1993)**
 - 2 instructions per cycle
 - **Separate 8KB I\$ & D\$**
- Characteristics
 - 0.8-0.35 μm process
 - 3.2M transistors
 - 60-300 MHz
 - 32-bit word size



Courtesy of Intel Museum

Pentium Pro / II / III

- **Dynamic execution (1995-9)**
 - ❑ 3 micro-ops / cycle
 - ❑ Out of order execution
 - ❑ 16-32 KB I\$ & D\$
 - ❑ **Multimedia instructions**
 - ❑ PIII adds 256+ KB L2\$
- **Characteristics**
 - ❑ 0.6-0.18 μm process
 - ❑ 5.5M-28M transistors
 - ❑ 166-1000 MHz
 - ❑ 32-bit word size



Courtesy of Intel Museum

Pentium 4

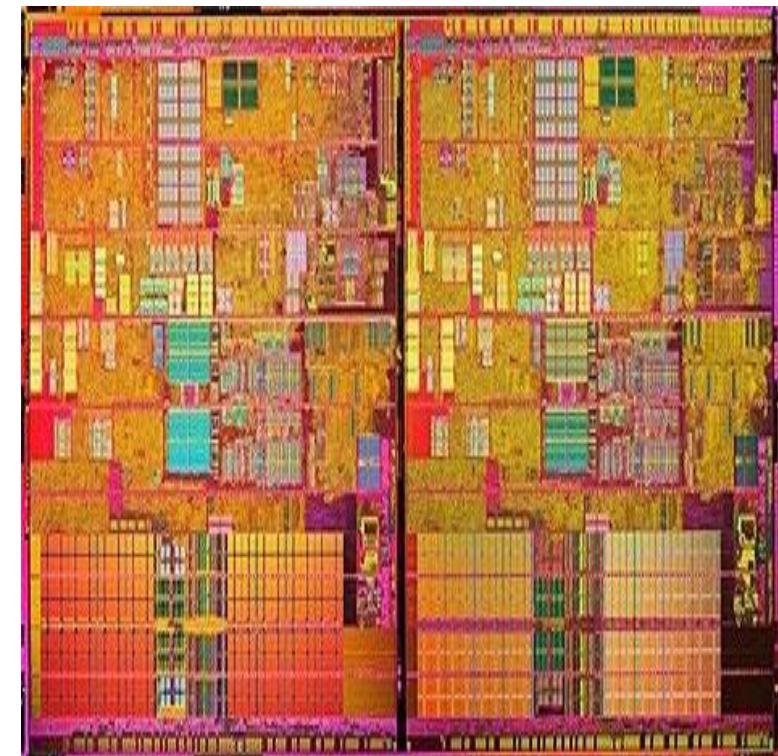
- Deep pipeline (2001)
 - Very fast clock
 - 256-1024 KB L2\$
- Characteristics
 - 180 – 90 nm process
 - 42-125M transistors
 - 1.4-3.4 GHz
 - Extended Memory 64 Technology
 - HyperThreading



Courtesy of Intel Museum

Pentium D

- Dual core (2005)
 - 2 Pentium 4 cores
 - 1 M L2 cache each core
- Characteristics
 - 90nm process technology.
 - 230 million transistors.
 - 3.2 GHz, 3 GHz, 2.8 GHz, 2.66 GHz
 - Extended Memory 64 Technology
 - HyperThreading

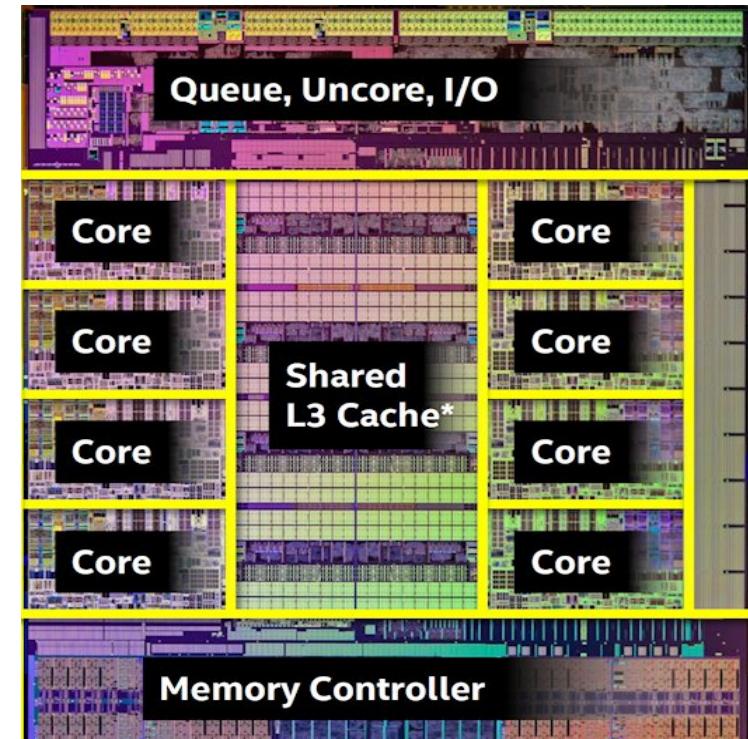


Courtesy of Intel Museum

Intel® Core i7 (2008 ~ core # 4 ~ 10)

▪ Intel® Core™ i7-5960X Processor

# of Cores	8
# of Threads	16
Clock Speed	3 GHz
Max Turbo Frequency	3.5 GHz
Intel® Smart Cache	20 MB
Intel® QPI Speed	0 GT/s
# of QPI Links	0
Instruction Set	64-bit
Instruction Set Extensions	SSE4.2, AVX 2.0, AES
Max Memory Size (dependent on memory type)	64 GB
Memory Types	DDR4-1333/1600/2133
# of Memory Channels	4
Max Memory Bandwidth	68 GB/s
ECC Memory Supported ‡	 No



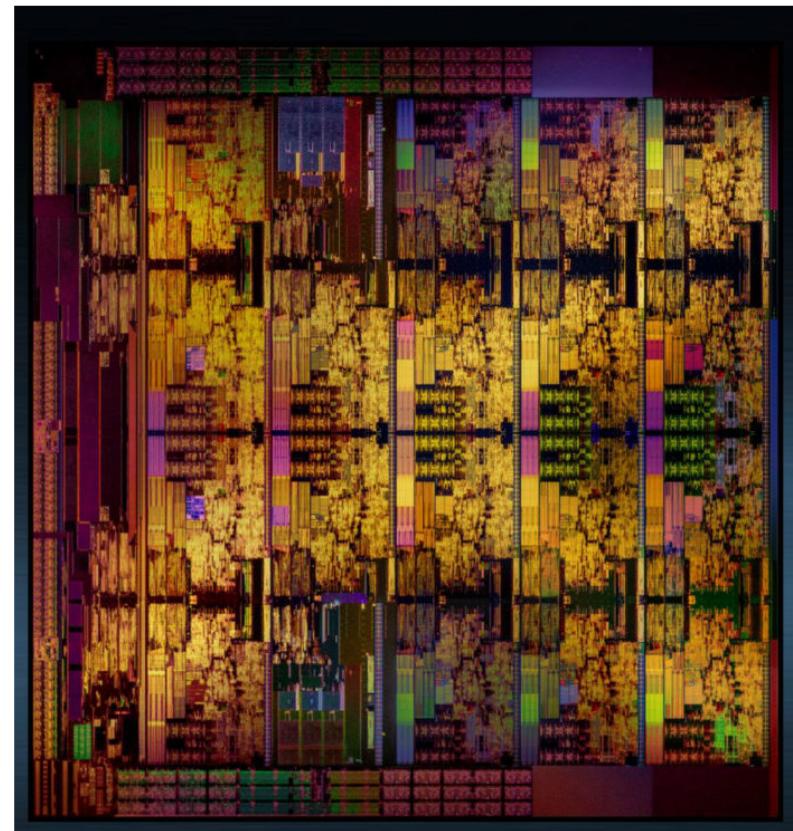
http://ark.intel.com/products/82930/Intel-Core-i7-5960X-Processor-Extreme-Edition-20M-Cache-up-to-3_50-GHz#@specifications

Intel® Core™ i9-7980XE X-series Processor (2017)

Performance

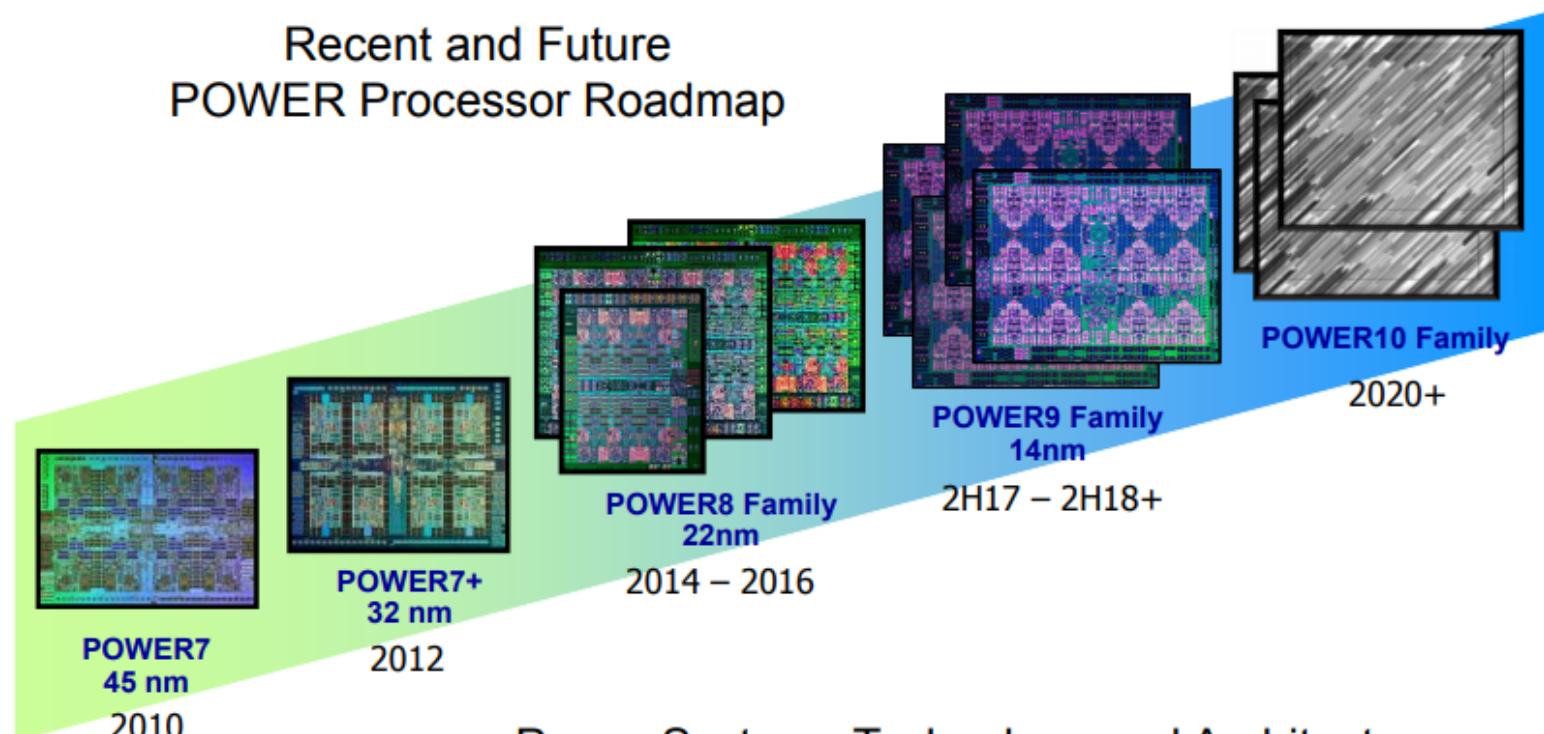
# of Cores	18
# of Threads	36
Processor Base Frequency	2.60 GHz
Max Turbo Frequency	4.20 GHz
Cache	24.75 MB
Bus Speed	8 GT/s D
# of QPI Links	0
Intel® Turbo Boost Max Technology 3.0 Frequency †	4.40 GHz
TDP	165 W

Intel Core i9-7980XE Die Shot



http://ark.intel.com/products/126699/Intel-Core-i9-7980XE-X-series-Processor-24_75M-Cache-up-to-4_20-GHz

IBM Power



Power Systems Technology and Architecture
Leveraging the economics of the New Era

Power9

New Core Microarchitecture

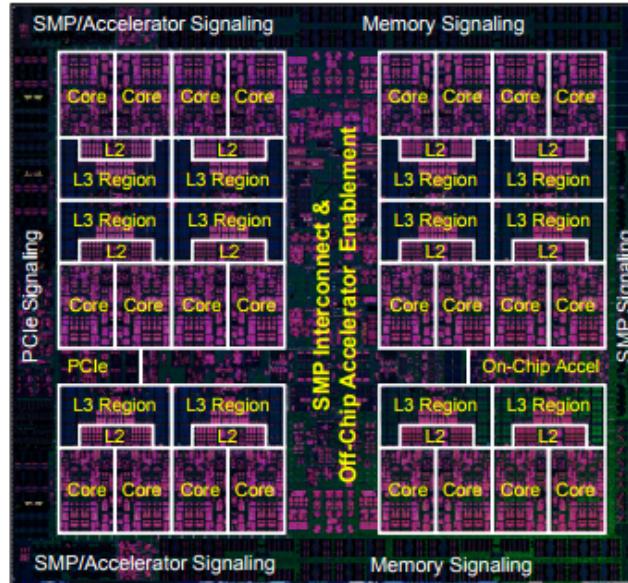
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy

- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation

- Quality of service assists
- New interrupt architecture
- Workload optimized frequency
- Hardware enforced trusted execution



14nm finFET Semiconductor Process

- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors

Leadership Hardware Acceleration Platform

- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth and advanced new features (25G)
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- New CAPI: Improved latency and bandwidth, open interface (25G)

State of the Art I/O Subsystem

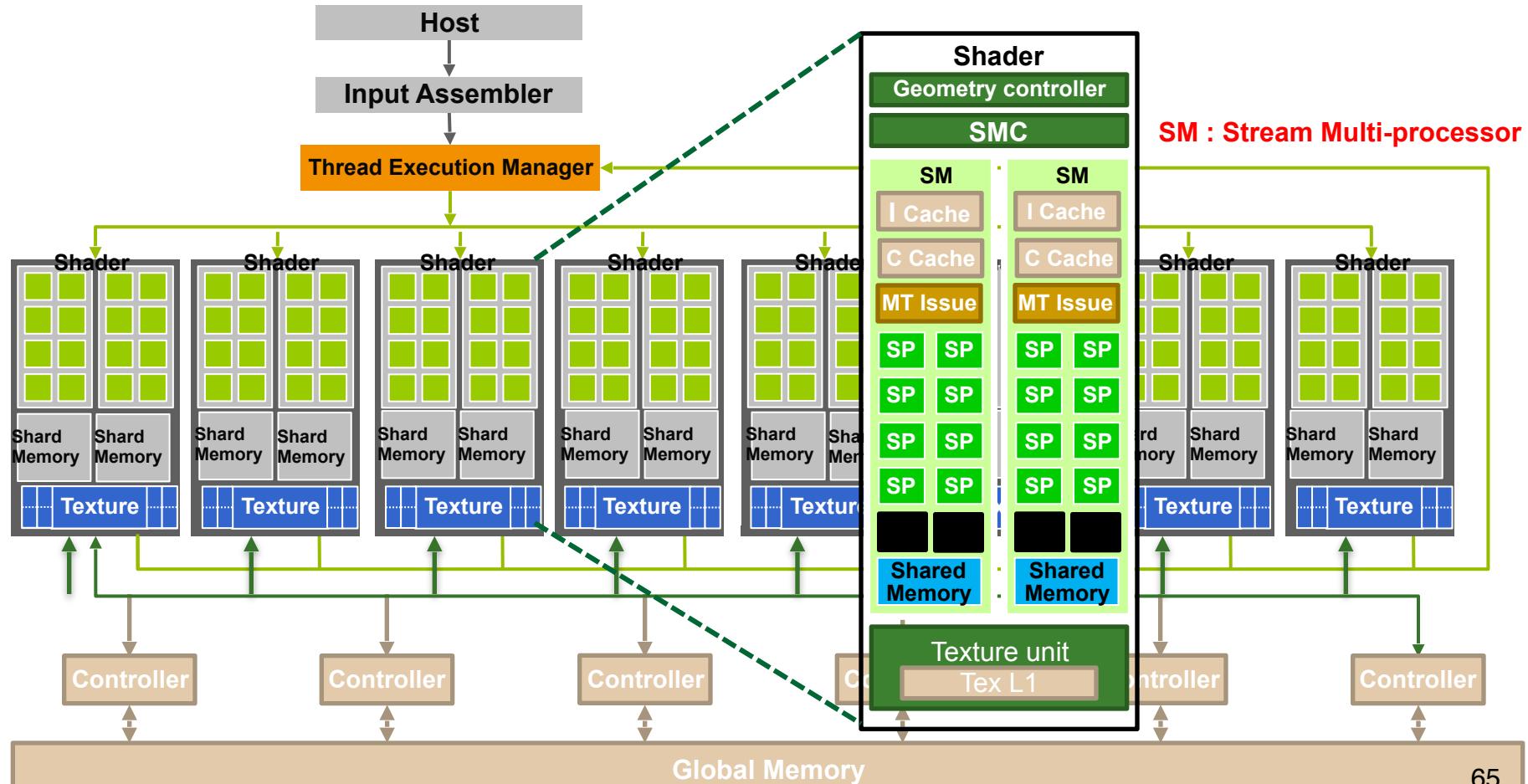
- PCIe Gen4 – 48 lanes

High Bandwidth Signaling Technology

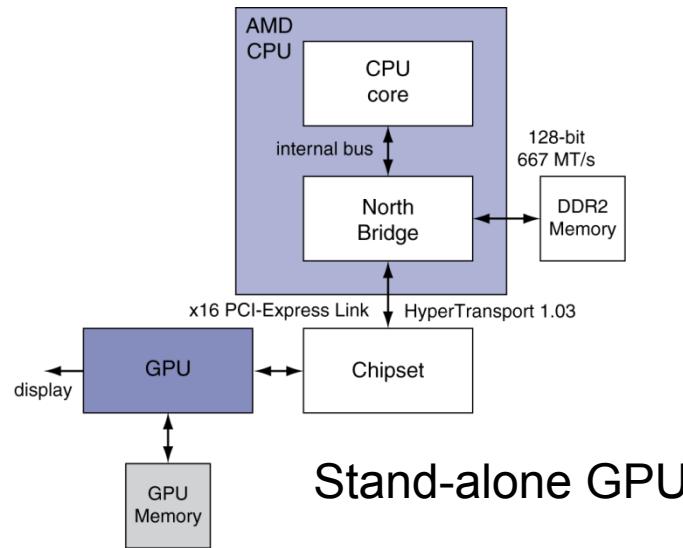
- 16 Gb/s interface
 - Local SMP
- 25 Gb/s Common Link interface
 - Accelerator, remote SMP

GPU Hardware Architecture

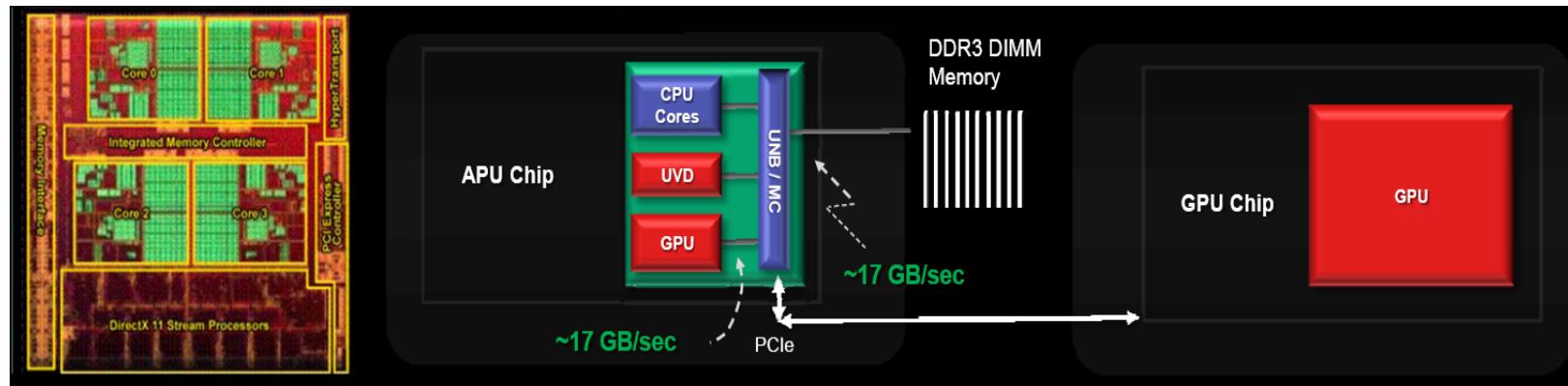
- Baseline: NVIDIA Tesla Architecture[1]



Heterogeneous Computing : Integrated CPU/GPU

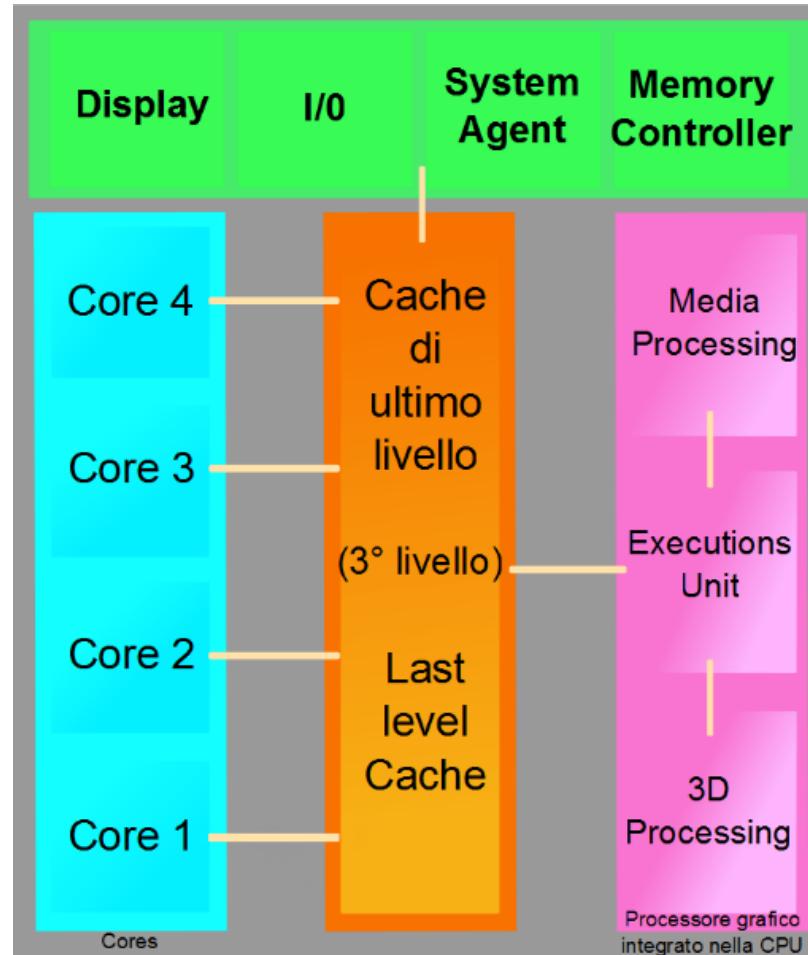


Stand-alone GPU



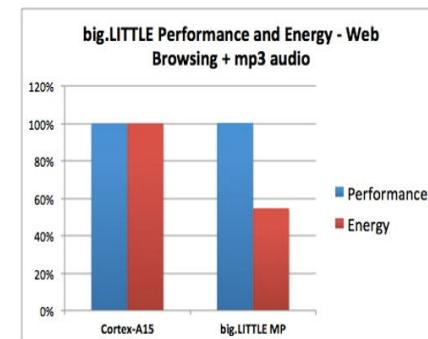
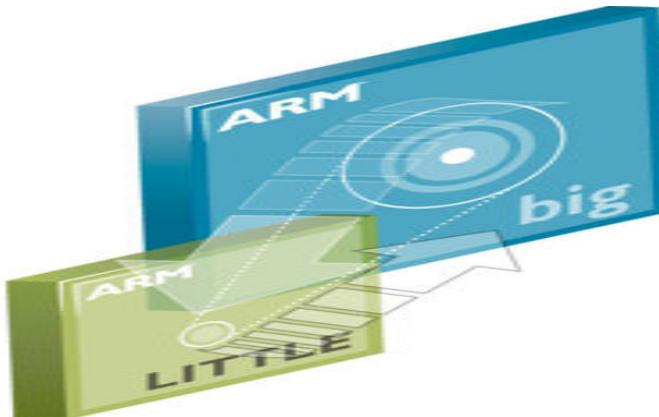
AMD Fusion

Intel Sandy-bridge, Ivy-bridge



ARM Big.Little Technology – 2011

- ARM big.LITTLE processing is designed to deliver the vision of the right processor for the right job. In current big.LITTLE system implementations a ‘big’ ARM Cortex™-A15 processor is paired with a ‘LITTLE’ Cortex™-A7 processor to create a system that can accomplish both high intensity and low intensity tasks in the most energy efficient manner. For example, the performance capabilities of the Cortex-A15 processor can be utilized for heavy workloads, while the Cortex-A7 can take over to process most efficiently majority of smartphone workloads. These include operating system activities, user interface and other always on, always connected tasks.



Qualcomm Snapdragon 835 First to 10 nm

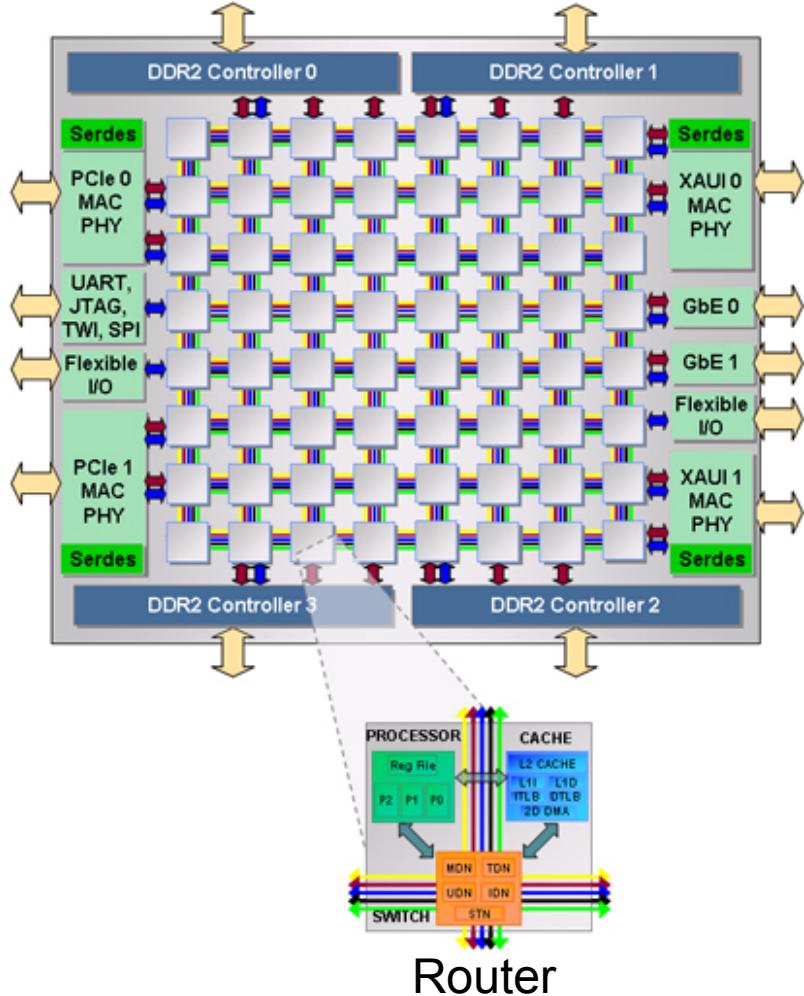
Contributed by Andy Wei

Posted: April 24, 2017

With the first set of [Samsung Galaxy S8 teardowns](#), we have access now to the first SoCs produced on "10 nm" class technology. First to the 10 nm productization finish line is the Qualcomm Snapdragon 835, built on Samsung LSI Foundry's 10 nm LPE technology. In parallel, Samsung's own Exynos 8895 was released concurrently, and we expect to find these in our second set of Samsung Galaxy S8 teardowns.

The Qualcomm Snapdragon 835 has a die size of 72.3 mm². Compared to the die size of the Snapdragon 820 at 113.7 mm², this represents a 36% die size shrink. The Qualcomm Snapdragon 835 appears to be mainly a shrink of the Snapdragon 820/821 family, with upgrades in similar IP blocks, but no new major IP blocks have been added. There is, however, a major change in the CPUs. The Snapdragon 820 family used a very large area 2+2 big-little implementation which seemed out of place in a mobile applications processor. We suspect the Kyro was a re-use of cores designed for Qualcomm's ARM server ambitions. The Snapdragon 835 uses a more ARM-like 4+4 big-little implementation in the Kyro 280, which are a derivative of the ARM Cortex-A73/A53 implementation we have already seen in the HiSilicon Kirin 960.

Tilera 64-core processor

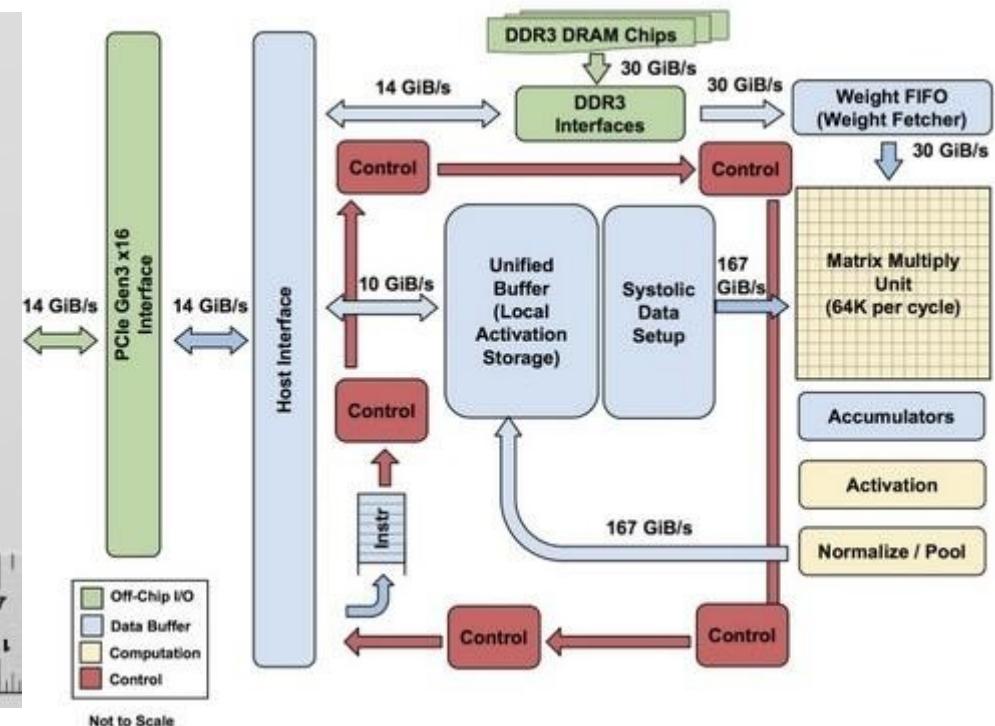
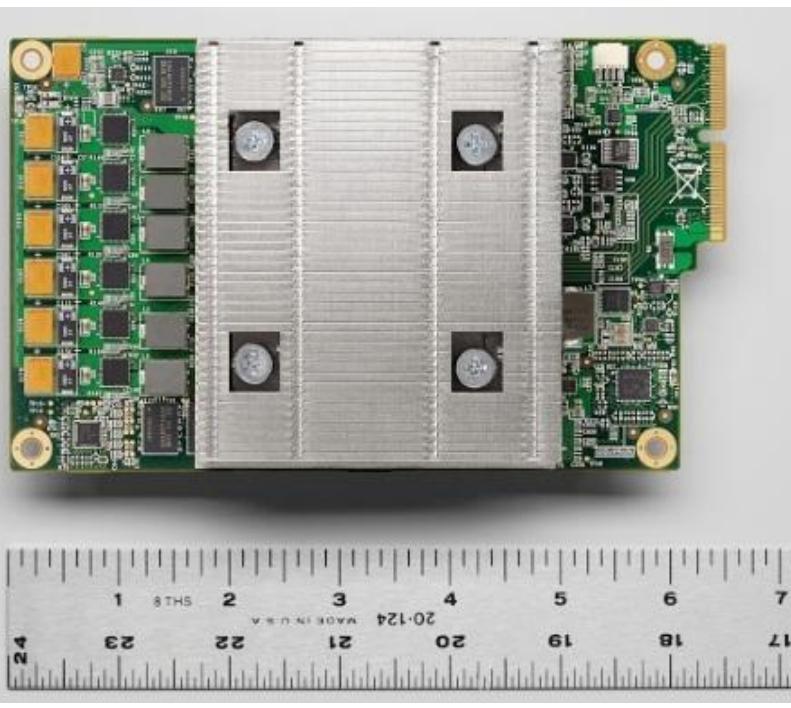


“Tilera® Corporation provides multicore processors that deliver the world's highest performance computing for a broad range of applications such as networking, wireless infrastructure, digital multimedia, and cloud computing. Tilera's processors are based on an intelligent mesh (iMesh™) architecture and provide far greater scalability than any other processor on the market.”

***“The Processor is
the new Transistor”
[Rowen]***

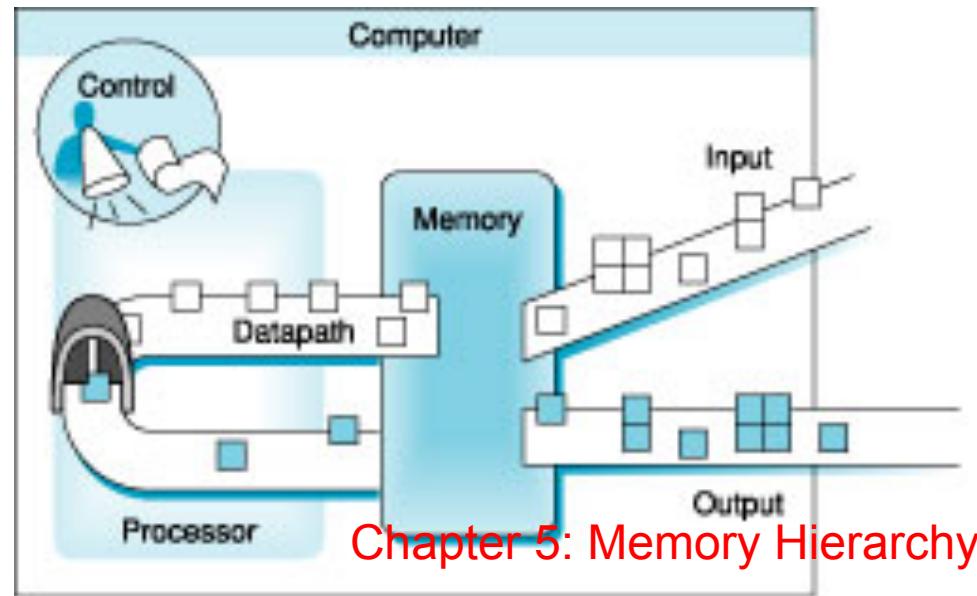
Google TPU

- For high energy-efficiency deep learning inference



Course Outline

Chapter 4: Processor



Chapter 5: Memory Hierarchy

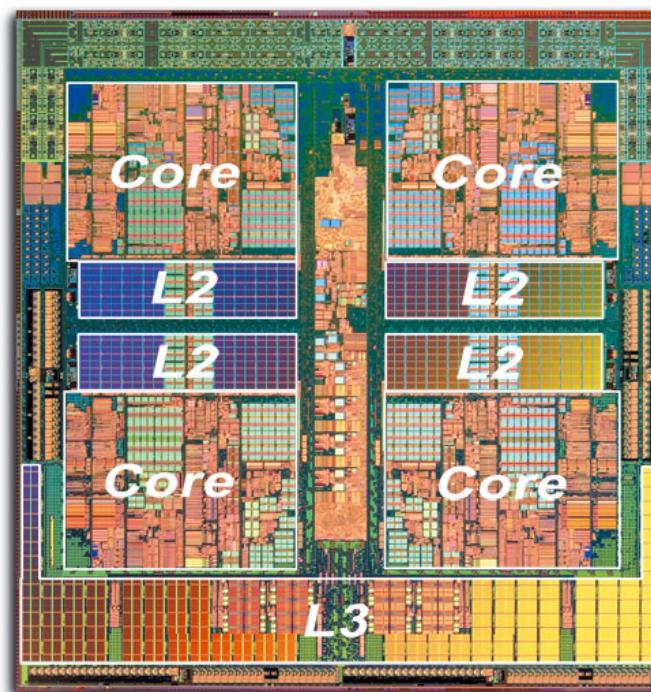
Compiler

Chapter 2: Instruction set Architecture

Interface

Course Outline (cont.)

Chapter 6: Parallel Processor - from Client to Cloud



cache policy to
affic from snoops

Eight Design Principle for Computer Architecture/System

- Design for ***Moore's Law***
- Use ***abstraction*** to simplify design
- Make the ***common case fast***
- Performance via ***parallelism***
- Performance via ***pipelining***
- Performance via ***prediction***
- ***Hierarchy*** of memories
- ***Dependability*** via redundancy

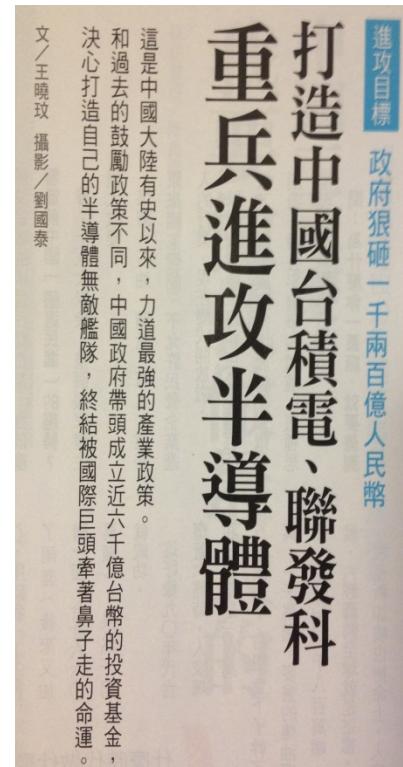


What You Can Learn in This Course

- How are programs written in a high-level language finally executed on hardware?
 - How can I write an efficient program?
 - How hardware and software work well together?
 - What techniques can be used by hardware designers to improve performance
-
- This is the course to help you cross the barrier between software and hardware !!!
 - A foundation course to train you to become an expert in hw/sw co-design
- 台灣產業升級, 急需之人才 !!

台灣IC產業的危機

- 大陸投入一千兩百億人民幣(約五千九百億台幣)



來源: 1. 新華網, 6/24, 2014; 2. 天下雜誌, 554期, 8/20~9/2, 2014