EITF20: Computer Architecture

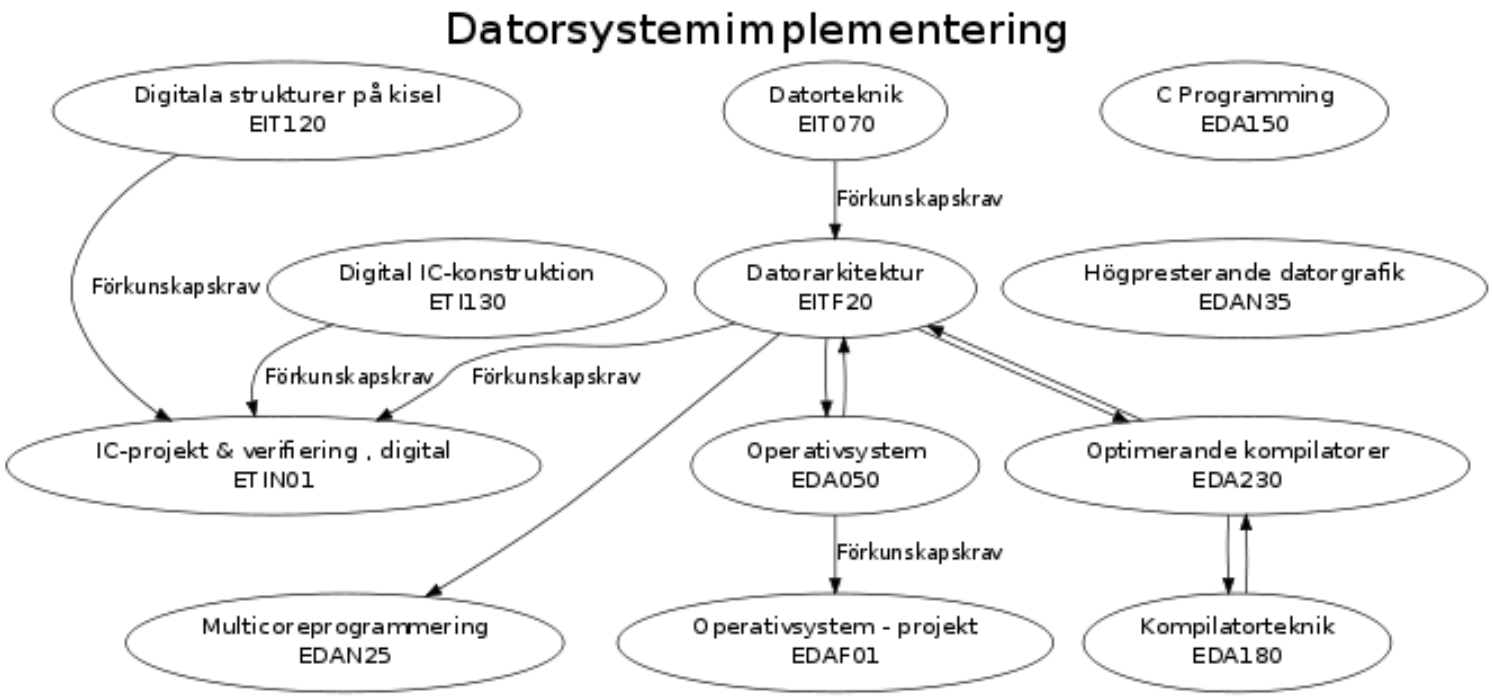
Part1.1.1: Introduction

Liang Liu
liang.liu@eit.lth.se

Course Factor

- **Computer Architecture (7.5HP)**

http://www.eit.lth.se/kurs/eitf20



Q1: How to upgrade with limited budget

Kingston
SSDNow UV400
599,00 kr

Corsair Vengeance DDR4 RAM
8 GB
769,00 kr

MSI CORE FROZR L
599,00 kr

Gigabyte GA-B250M-
DS3H, Socket-1151
699,00 kr

Question in exam

We are interested in comparing two proposed enhancements to a baseline processor to decide which design offers highest performance. All processors assume an average CPI of 1.5 together with cache. Assume instruction fetch has 100% cache hit rate (i.e., instruction cache is perfect). Each access between cache and main memory to read or write a cache block requires a setup time of 30 clock cycles, with each transfer of 32-bit data adding another 2 clock cycles penalty. The cache size for all 3 processor designs is 32,768 bytes, i.e. 32 kB.

Instruction mix	
instruction type	**% of all instructions**
load	24.55
store	10.45
uncond branch	3.50
cond branch	11.50
int computation	39.15
fp computation	10.85

1. The baseline processor runs at 3.0 GHz clock frequency, has a cache with 128 sets, associativity 1, uses Random replacement policy, write-back with an average of 20% dirty blocks, and no write-back buffer.

2. Based on the baseline processor, the first modification suggests to use the FIFO replacement policy and to add a write-back buffer (same percentage of dirty blocks as before) which can hold 80% of the write-back blocks (assume the content in the write-back buffer will not cause any CPU stalls). Moreover, the execution time for the integer computation has been decreased by 3 times (on average).

3. Based on the baseline processor, the second modified design suggests a cache with 32 sets, associativity 4 and 30% of dirty blocks for write-back, with a write buffer which can hold 20% of the write-back blocks. Moreover, due to improved implementation, all floating-point computations are executed 50% faster (on average).

Q2: More cores or higher frequency

Intel Core i3-7350K @ 4.20GHz + Compare

Description: *Intel HD Graphics 630*
Socket: *FCLGA1151*
Clockspeed: *4.2 GHz*
No of Cores: *2 (2 logical cores per physical)*
Typical TDP: *60 W*

Other names: *Intel(R) Core(TM) i3-7350K CPU @ 4.20GHz*
CPU First Seen on Charts: Q1 2017
CPUmark/$Price: 42.89 **Overall Rank:** 423
Last Price Change: $156.99 USD (2017-01-18)

Average CPU Mark

6734

Single Thread Rating: 2480
Samples: 48*
*Margin for error: Low

AMD Ryzen 5 1400 + Compare

Socket: *AM4*
Clockspeed: *3.2 GHz*
Turbo Speed: *3.4 GHz*
No of Cores: *4 (2 logical cores per physical)*
Typical TDP: *65 W*

Other names: *AMD Ryzen 5 1400 Quad-Core Processor*
CPU First Seen on Charts: Q2 2017
CPUmark/$Price: 52.73 **Overall Rank:** 297
Last Price Change: $159.99 USD (2017-04-07)

Average CPU Mark

8436

Single Thread Rating: 1712
Samples: 154*
*Margin for error: Low

Q3: Why not Intel for iphone?

ARM

Q4: Attacks?

Meltdown

Meltdown breaks the most fundamental isolation between user applications and the operating system. This attack allows a program to access the memory, and thus also the secrets, of other programs and the operating system.

If your computer has a vulnerable processor and runs an unpatched operating system, it is not safe to work with sensitive information without the chance of leaking the information. This applies both to personal computers as well as cloud infrastructure. Luckily, there are software patches against Meltdown.

Spectre

Spectre breaks the isolation between different applications. It allows an attacker to trick error-free programs, which follow best practices, into leaking their secrets. In fact, the safety checks of said best practices actually increase the attack surface and may make applications more susceptible to Spectre

Spectre is harder to exploit than Meltdown, but it is also harder to mitigate. However, it is possible to prevent specific known exploits based on Spectre through software patches.

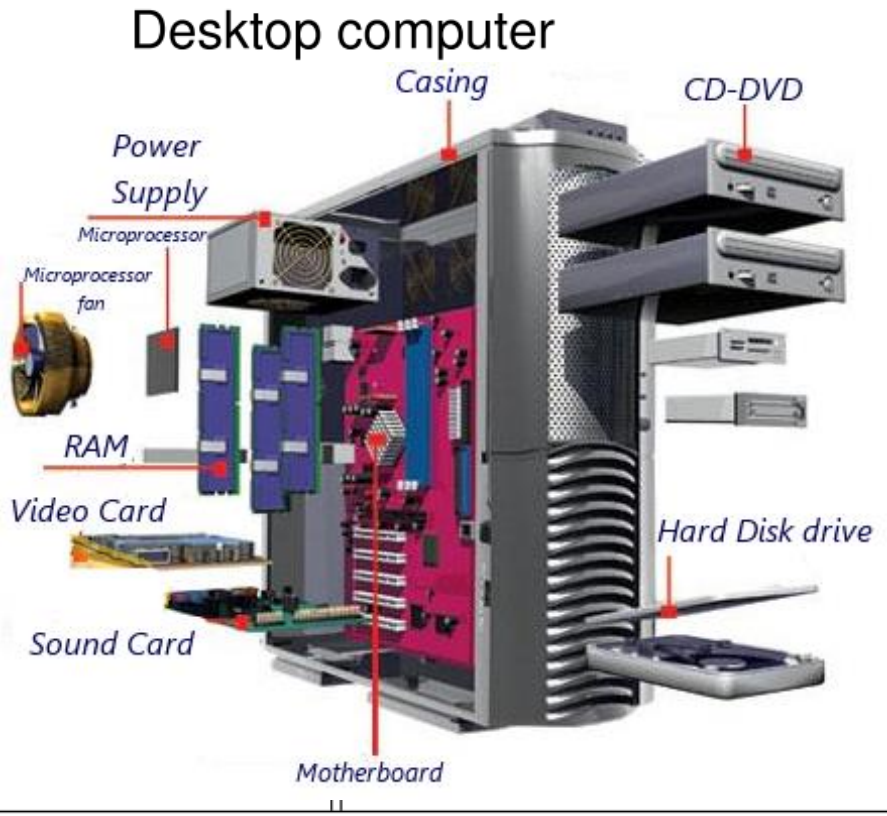
https://meltdownattack.com/

Outline

- Computers
- Computer Architecture
- This Course
- Trends
- Performance
- Quantitative Principles

Build a Computer…

Part		Price
Case		490
Power supply		399
Motherboard		790
CPU		2490
Memory		698
Disk		790
DVD/Blue-ray		590
Graphics		-
Sound, net, . . .		-
Keyboard, mouse, cables, . . .		?



$\sum$ = 3000 - 5000 - 10000 SEK

6247

Power Consumption: 65 to 250 watts

Build a Computer…

- A 1.2GHz 64-bit quad-core ARMv8 CPU
- 802.11n Wireless LAN
- Bluetooth 4.1
- Bluetooth Low Energy (BLE)

Like the Pi 2, it also has:

- 1GB RAM
- 4 USB ports
- 40 GPIO pins
- Full HDMI port
- Ethernet port
- Combined 3.5mm audio jack and composite video
- Camera interface (CSI)
- Display interface (DSI)
- Micro SD card slot (now push-pull rather than push-push)
- VideoCore IV 3D graphics core

3965 I lager för leverans inom 1 arbetsdagar

Tidigare 347,75 kr

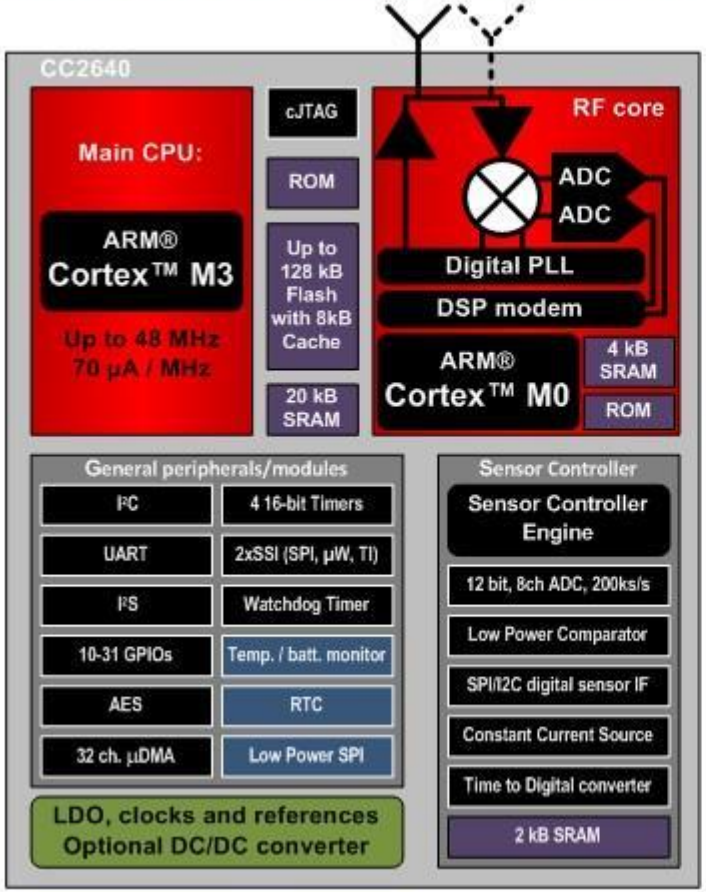
298,19 kr

Pris (ex. moms) Each

Power: 230 mA (1.2W)

Build a Computer…

CC2640 *Bluetooth low energy*



Quick Facts

Ultra-low Power Consumption
- 65 µA/MHz ARM Cortex M3
- 8.2 µA/MHz Sensor Controller
- 0.7 µA sleep with retention and RTC
- 5.9 mA RX (single-ended)
- 6.5 mA TX (single-ended)

SoC Key Features
- Autonomous sensor controller engine
- 4x4 mm to 7x7 mm QFN
- 1.65 – 3.8 V supply range
- 128 kB Flash + 8 kB Cache
- 20 kB RAM

RF Key Features
- +5 dBm output power
- -97 dBm sensitivity
- 2360 MHz – 2500 MHz
- Pin compatible with CC13xx in 4x4 and 5x5 QFN (BLE + Sub 1GHz prop)

Power: 65 uA/MHz (32kHz-30MHz)

$ 2.98

Build a Computer…

Power: 80 Megawatts

Class of Computers

Feature	Personal mobile device (PMD)	Desktop	Server	Clusters/warehouse-scale computer	Embedded
Price of system	$100–$1000	$300–$2500	$5000–$10,000,000	$100,000–$200,000,000	$10–$100,000
Price of micro-processor	$10–$100	$50–$500	$200–$2000	$50–$250	$0.01–$100
Critical system design issues	Cost, energy, media performance, responsiveness	Price-performance, energy, graphics performance	Throughput, availability, scalability, energy	Price-performance, throughput, energy proportionality	Price, energy, application-specific performance

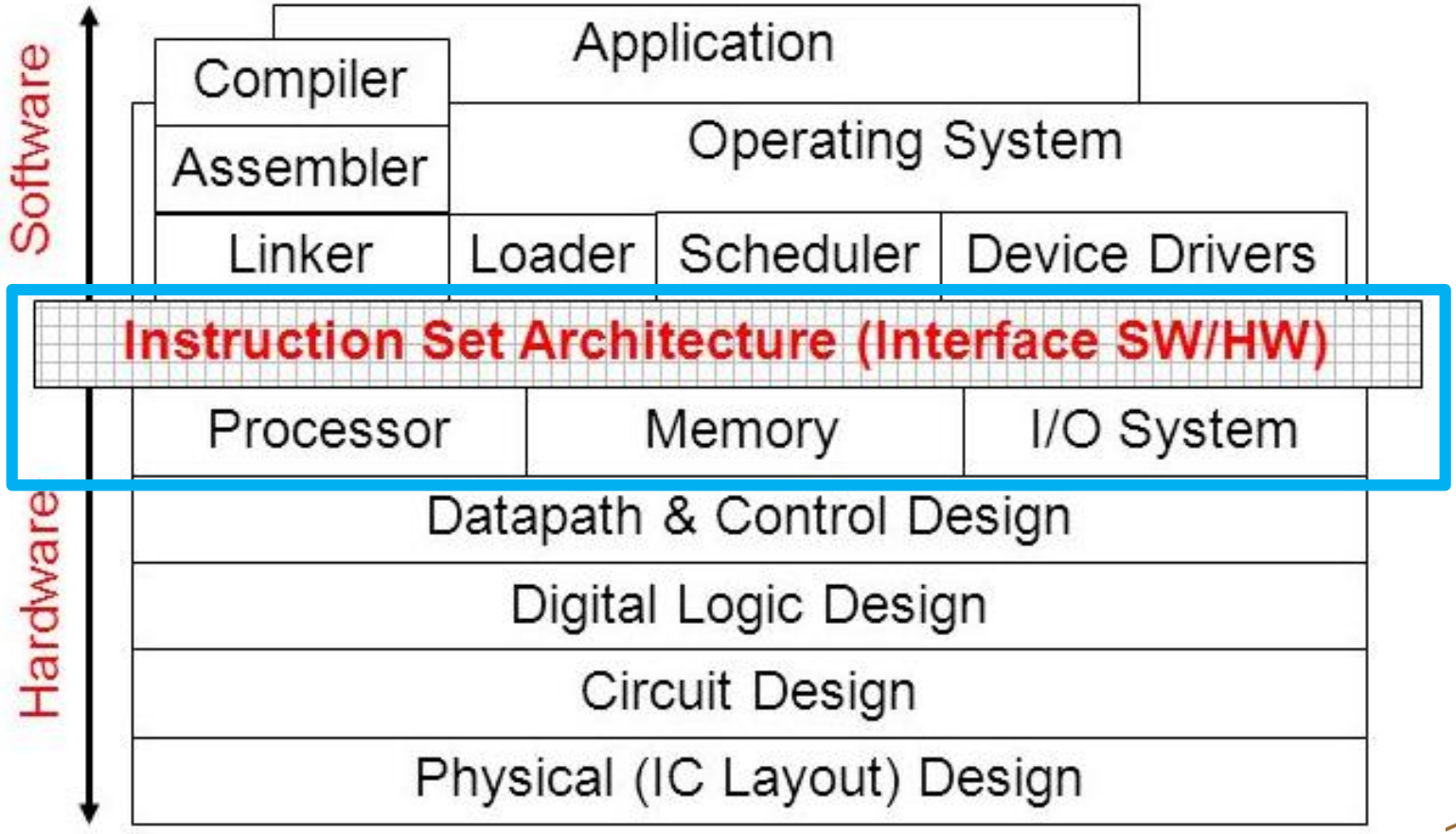
Do they have the same architecture? Why?

Outline

- Computers
- **Computer Architecture**
- This Course
- Trends
- Performance
- Quantitative Principles

The ART of designing computers is based on engineering principles and quantitative performance evaluation

Computer abstraction levels



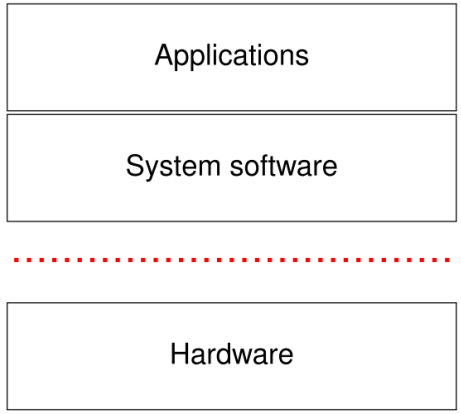
Computer architecture

Computer architecture is a set of disciplines that describe the functionality, organization and implementation of computer systems.

- **ISA: Instruction-set architecture**
- **Computer orginization: micro architecture**
- **Specific implementation**

Computer architecture

- **Design and analysis**
  - ISA
  - Orgnization (microarchitecture)
  - Implementation
- **To meet requirements of**
  - Functionality
  - Price
  - Performance (Execution Time)
  - Power
  - Reliability
  - Compatability
  - ..



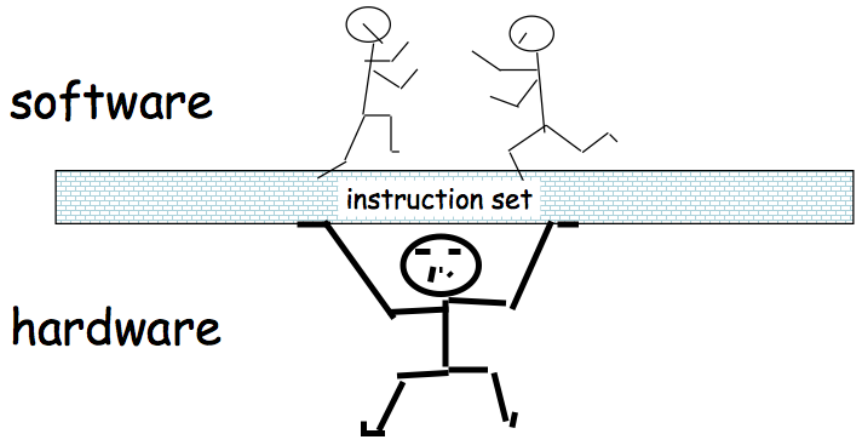
ISA

An instruction set architecture (ISA) is the interface between the computer's software and hardware and can be viewed as the programmer's view of the machine (functionality).

MIPS32 Add Immediate Instruction

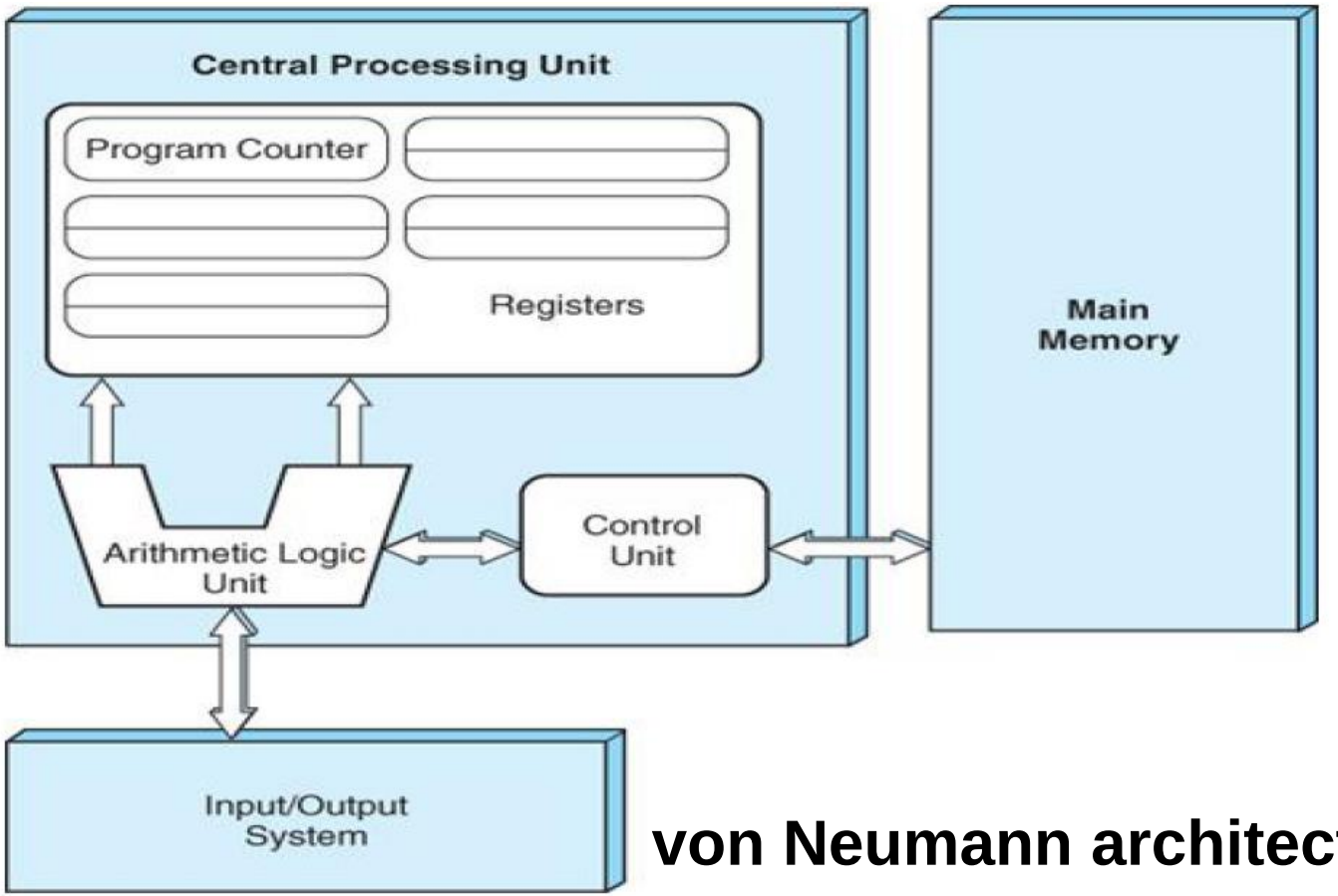
001000	00001	00010	0000000101011110
OP Code	Addr 1	Addr 2	Immediate value

Equivalent mnemonic: **addi** $r1, $r2, 350



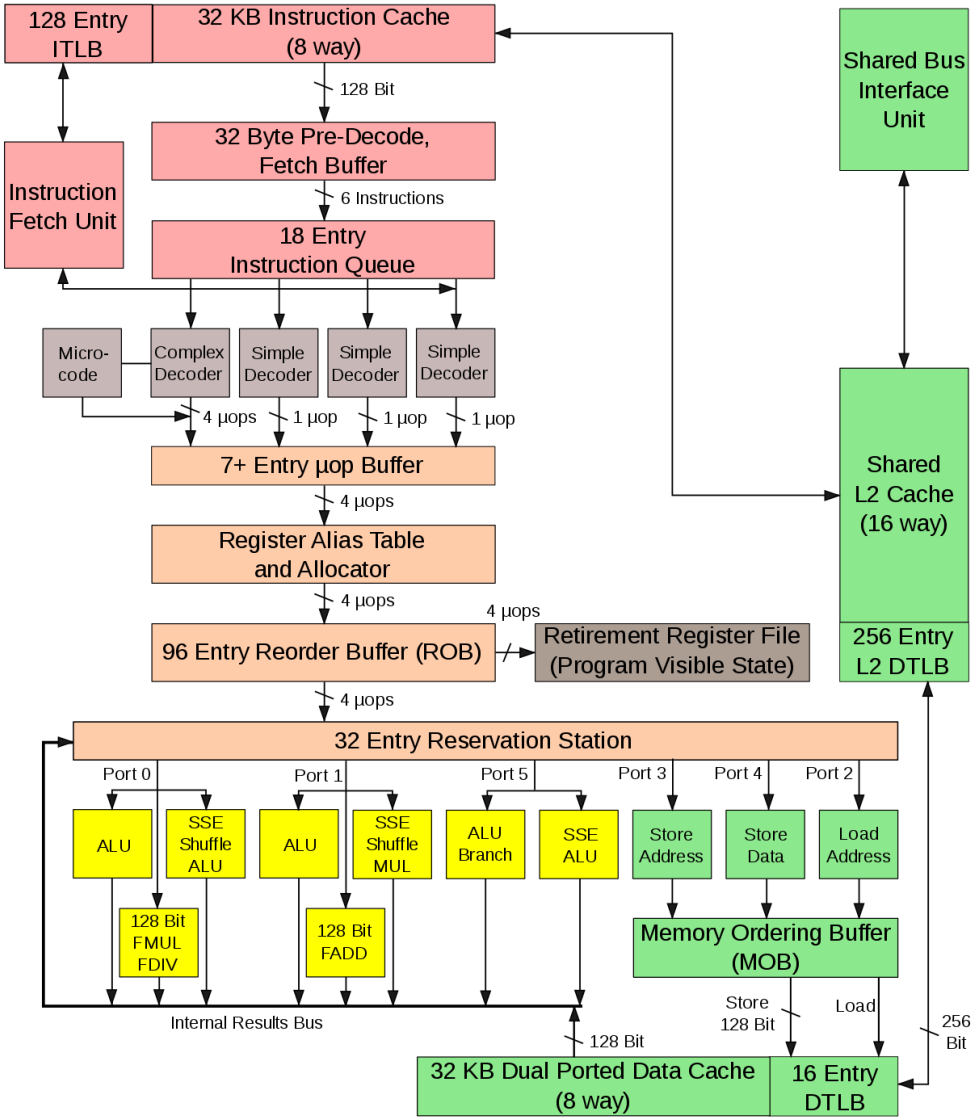
Microarchitecture

Microarchitecture is the way a given instruction set architecture (ISA) is implemented on a processor.



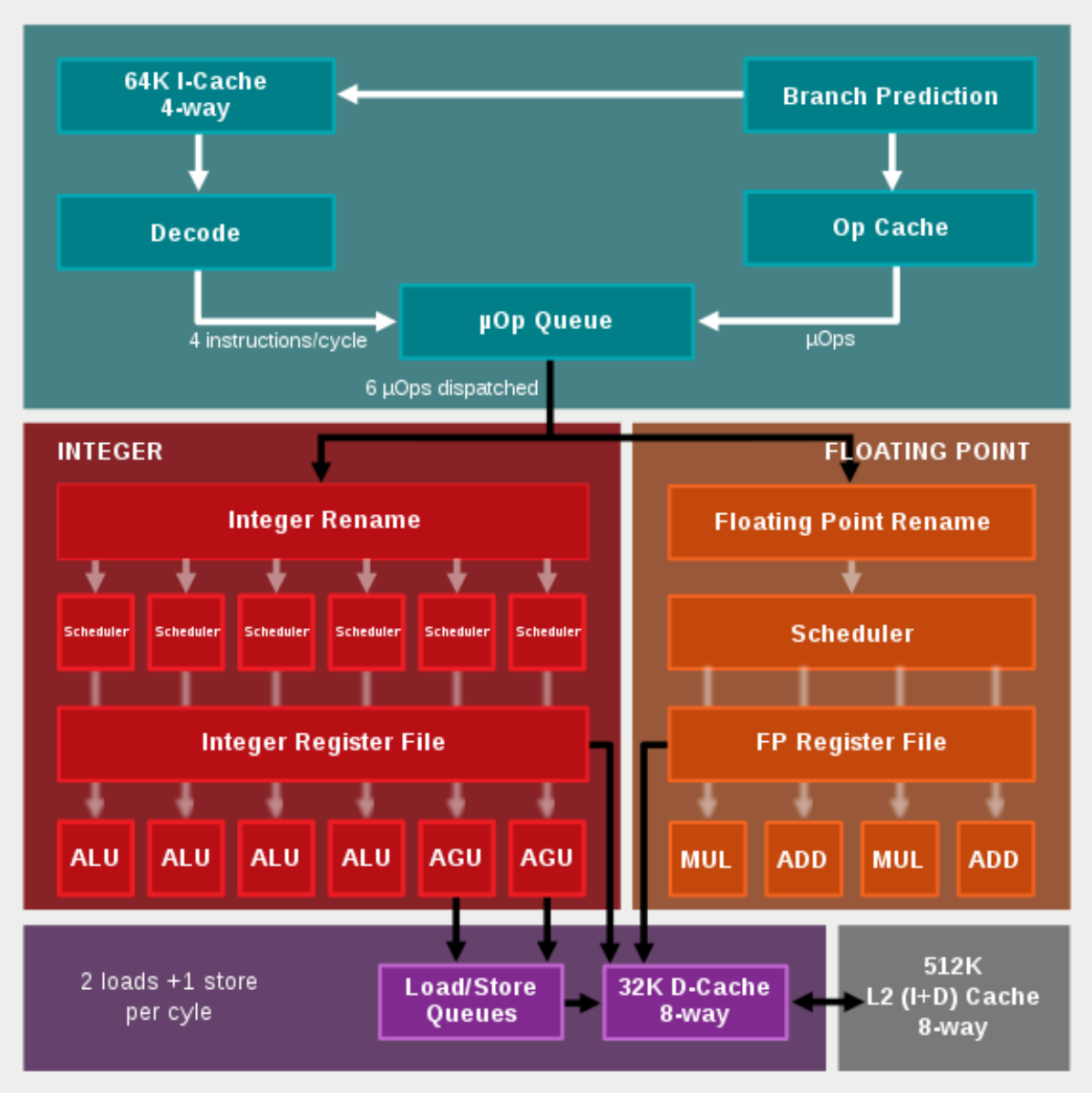
von Neumann architecture

Microarchitecture

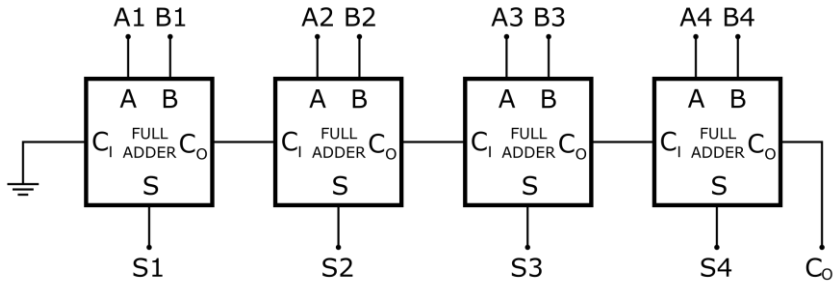


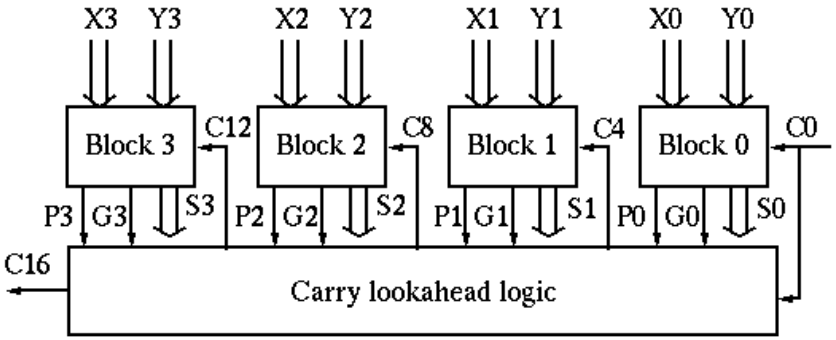
Intel Core 2 Architecture

Microarchitecture

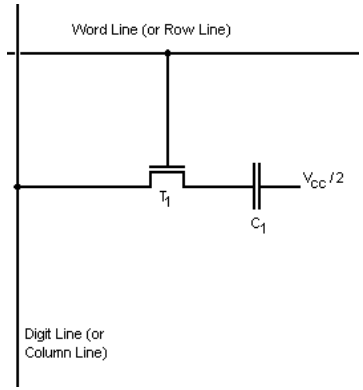


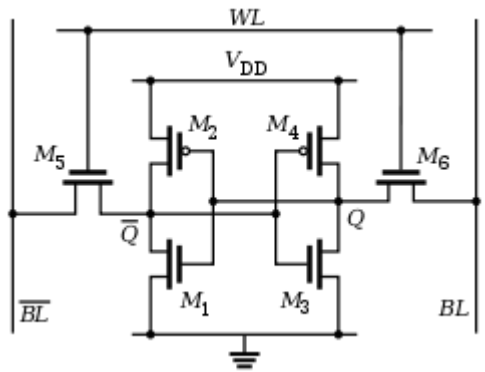
Implementation





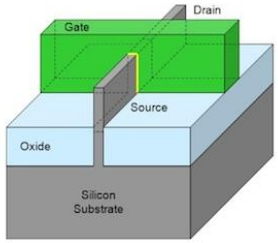
4-bit Adder





Memory Cell

What affects computer architecture?



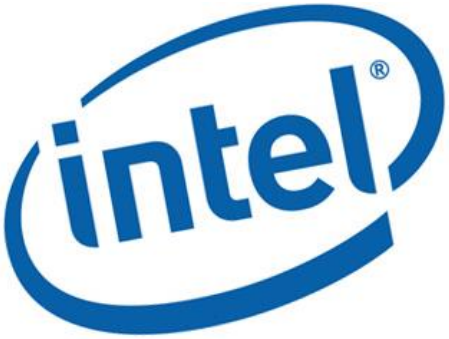
Applications

Technology

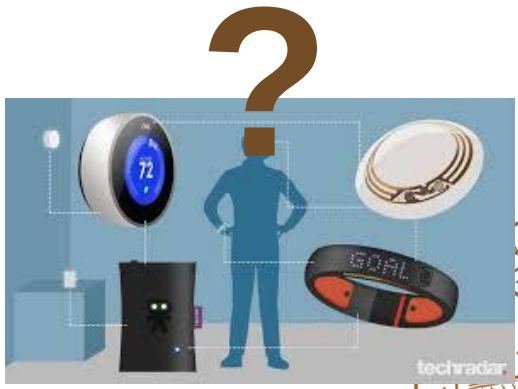
Software

Computers

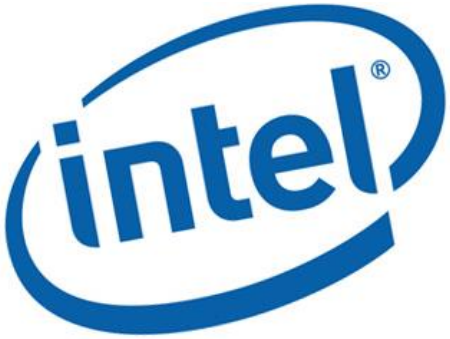
Architecture change due to new applications

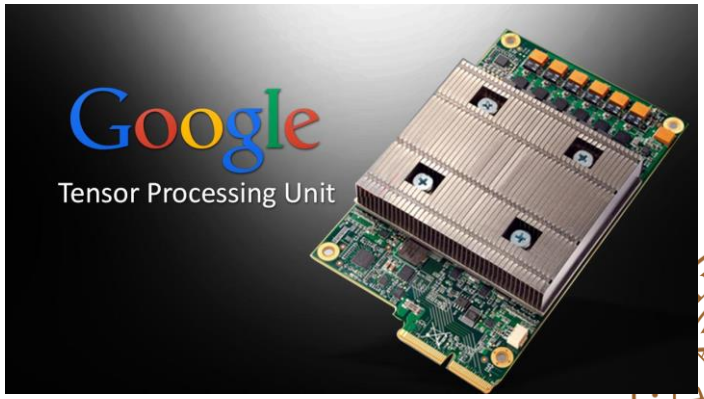


ARM

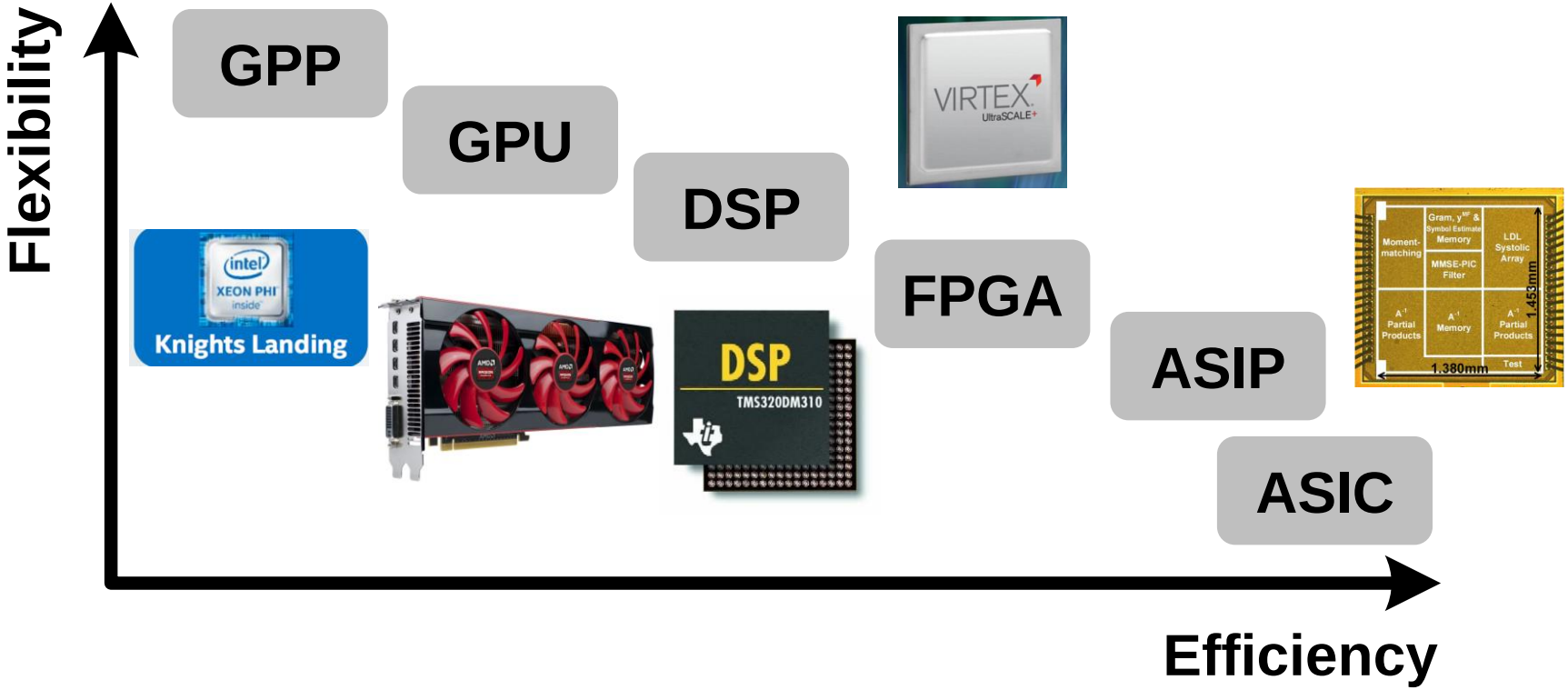


Architecture change due to new applications





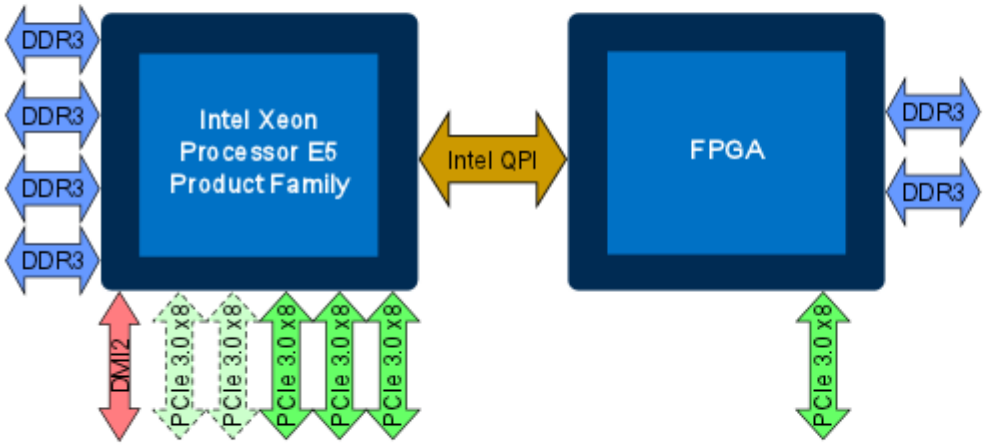
Devices


Flexibility
GPP
GPU
DSP
FPGA
ASIP
ASIC
Knights Landing
DSP
TMS320DM310
VIRTEX
UltraSCALE+
Efficiency

Devices (Hybrid Computing, which I believe is the future)

Intel® Xeon® Processor + Field Programmable Gate Array Software Development Platform (SDP) Shipping Today

Software Development for Accelerating Workloads using Intel® Xeon® processors and coherently attached FPGA in-socket

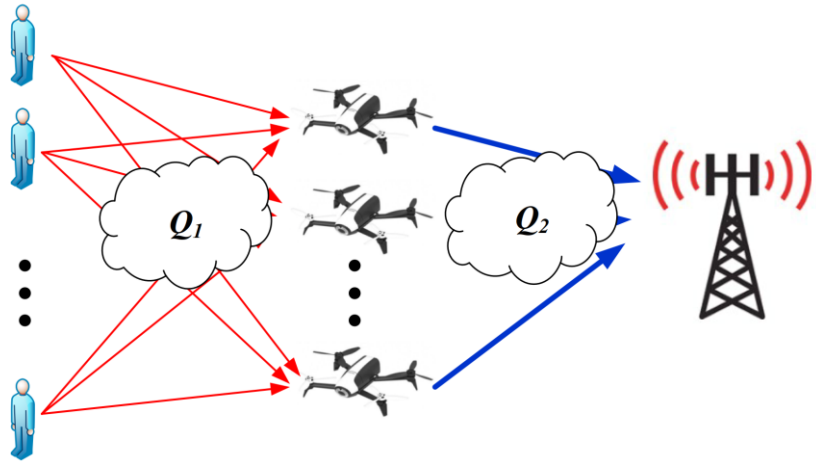


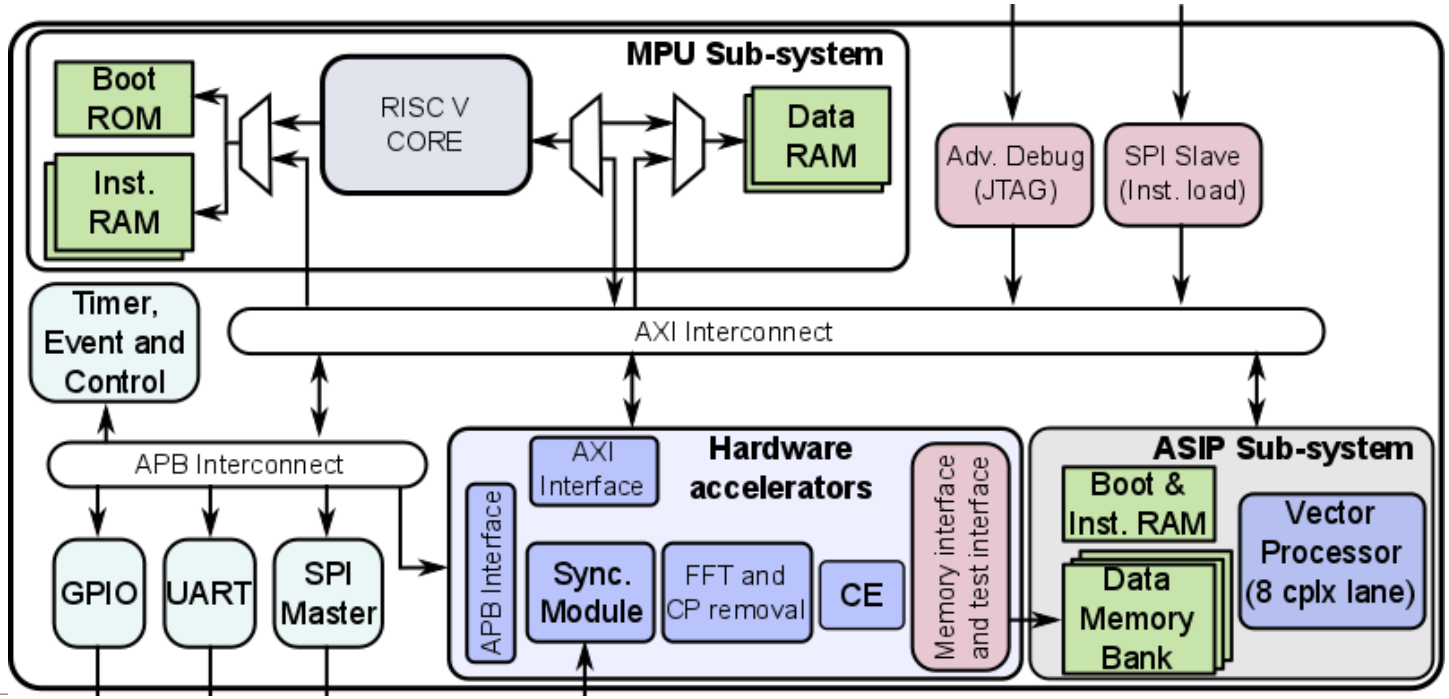
Processor	Intel® Xeon® Processor E5
FPGA Module	Altera® Stratix® V
QPI Speed	6.4 GT/s full width (target 8.0 GT/s at full width)
Memory to FPGA Module	2 channels of DDR3 (up to 64 GB)
Expansion connector to FPGA Module	PCI Express® (PCIe) 3.0 x8 lanes - maybe used for direct I/O e.g. Ethernet
Features	Configuration Agent, Caching Agent, (optional) Memory Controller
Software	Accelerator Abstraction Layer (AAL) runtime, drivers, sample

Available as part of Intel & Altera co-sponsored Hardware Accelerator Research Program

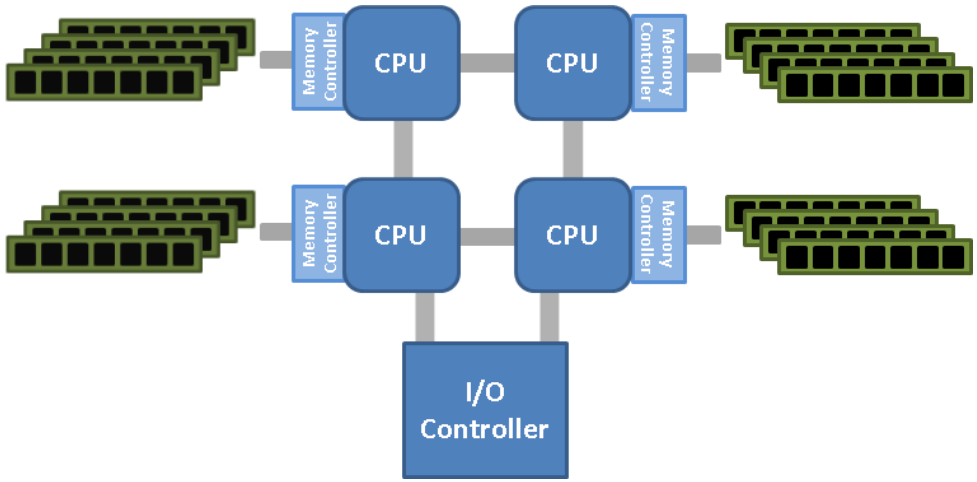
IDF15 INTEL DEVELOPER FORUM

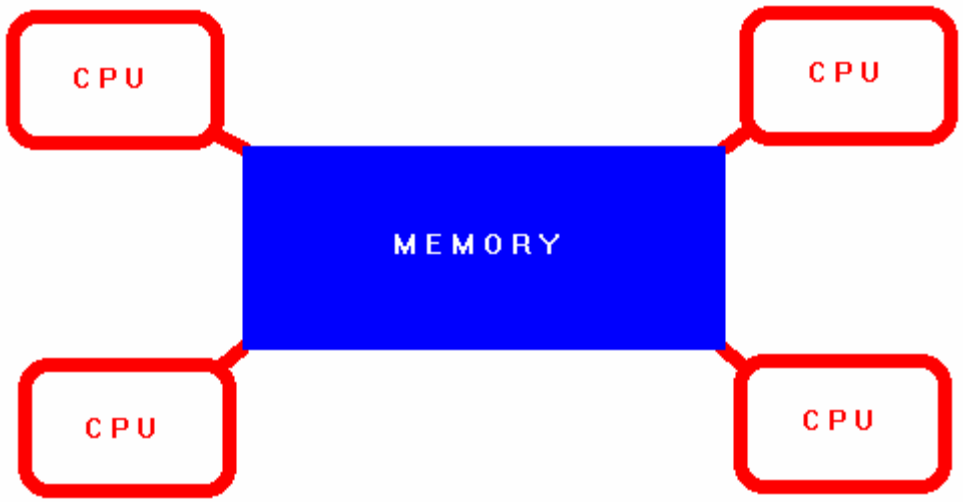
Devices (Hybrid Computing, which I believe is the future)





Memory centric architecture (e.g., in-memory computing)





Outline

- Computers
- Computer Architecture
- **This Course**
- Trends
- Performance
- Quantitative Principles

Course Objectives

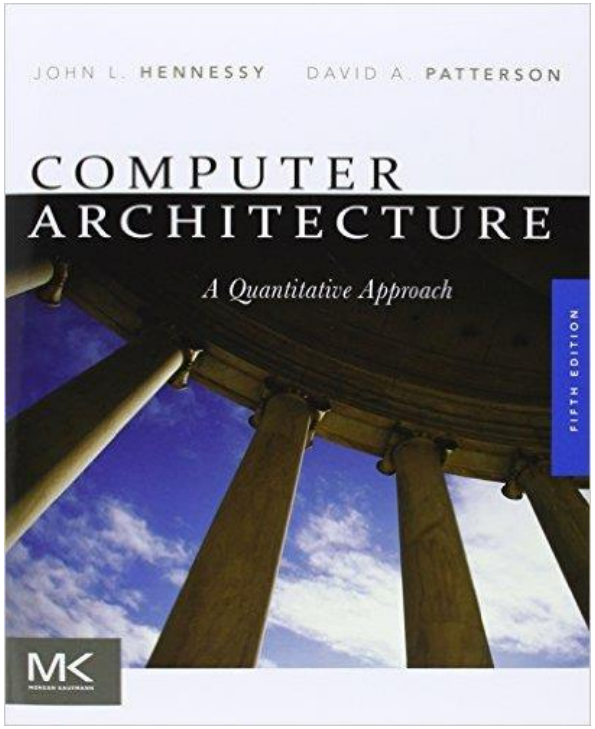
After this course, (hope) you will...

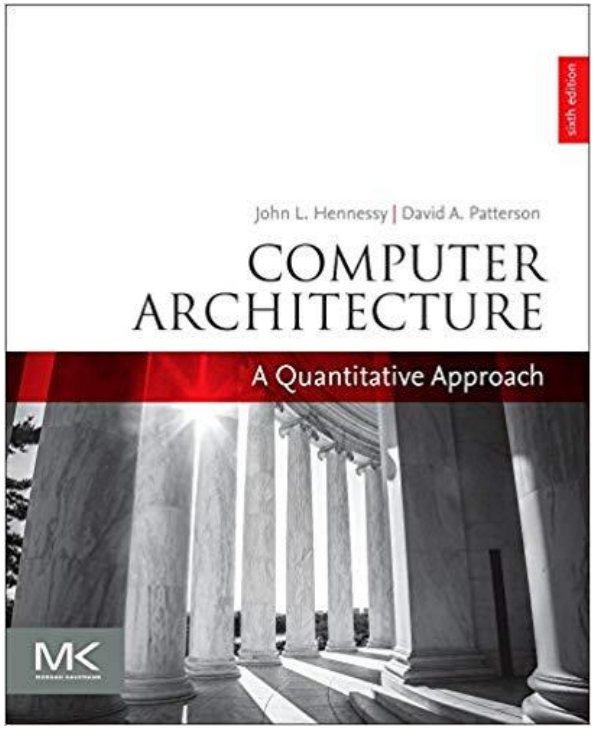
- **Have a thorough knowledge about the design principles for modern computer systems**
- **Be able to evaluate design alternatives towards design goals using quantitative evaluation methods**
- **”Side” effects...**
  - Better digital IC designer
  - Better understanding of compiler, operating system, high-performance programming

Book Recommendation

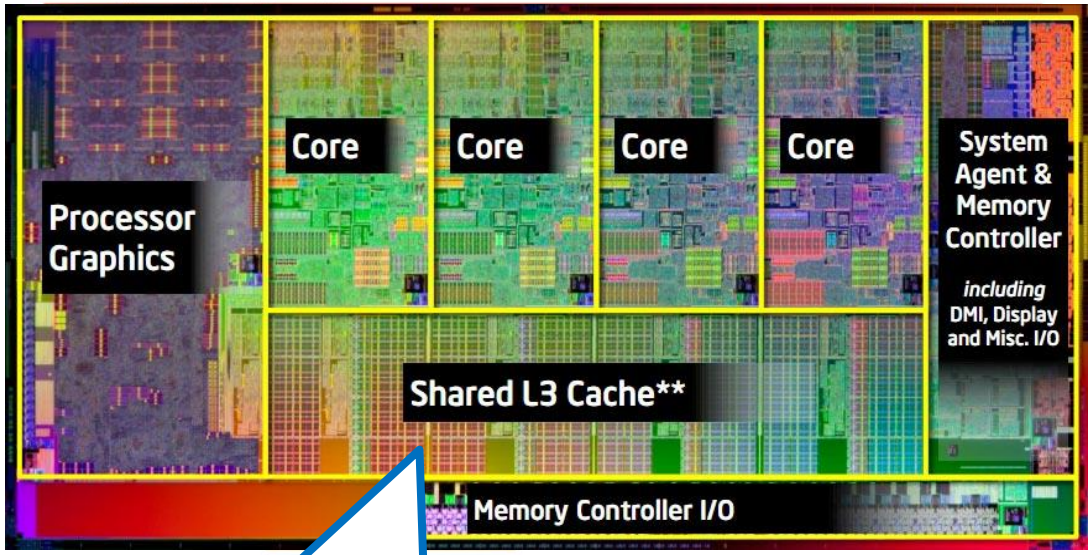
Computer Architecture – A Quantitative Approach

- HP: Hennessy, Patterson
- 5th Edition (6th Edition available)





Course Content & Schedule



- Concept &Theory
- Assignment & Project
- Exams

- Overview
- Instruction set architecture
- Pipeline
- Memory System
- Storage System
- Multiprocessor
- Application-specific Processor
- …

Teachers

Lecture

- Liang Liu, Associate Professor
- Email: liang.liu@eit.lth.se
- Room: E2342
- Homepage: http://www.eit.lth.se/staff/Liang.Liu

Course secretary

- Erik Göthe, Room 3152B
- e-mail: erik.gothe@eit.lth.se

Teaching Assistants

- Mohammad Attari
- Lucas Ferreira

Lucas Ferreira

Mohammad Attari

Invited Lectures

- Sven Karlsson, Ericsson Research, Lund
- AI/Machine Learning Processors & Accelerators
- 2019-12-19

Lectures and Labs

Lectures (10 or more)

- Tuesday : 13:15-15:00 **E:B/**Thursday: 15:15-17:00 **E:B**
- Covers design principles and analysis methodology
- Read the literature **before** each lecture
- Does **NOT** cover **all** the literature

Seminar/Exercise (2)

- Problems in previous exams

Labs (4)

- Tuesday: 08:15-12:00 **E:4118-E:4119**
- **OR** Friday: 13:15-17:00 **E:4118-E:4119**
- 2 students/group
- Read manual and literature **before** the lab
- Do Home Assignments **before** lab
- Experiment and **discuss** with assistants
- **Understand** what you have done (**or FAIL the exam…**)
- Finish Lab before **DEADLINE**

https://cloud.timeedit.net/lu/web/lth1/ri14566600000YQQ45Z5587007y5Y4313gQ7g5X4Y55ZQ076.html

Examination (Written)

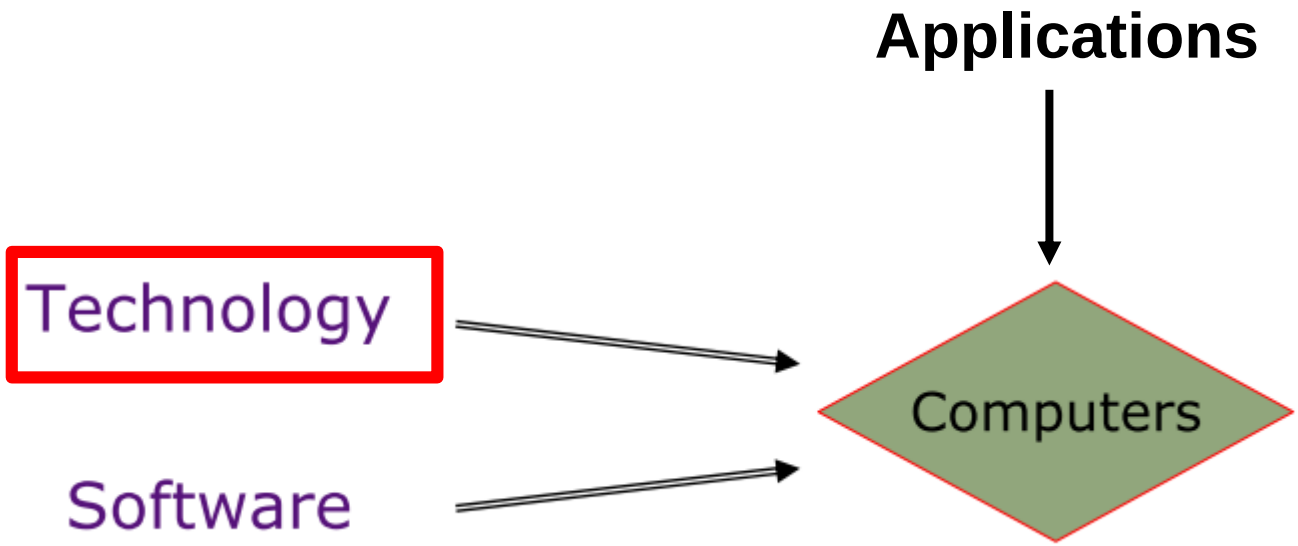
- **Anonymous exam**
- **Pass all labs to be able to attend written exam**
- **~Five problems**
  - Highly lab related
  - Problem solving
  - Descriptive nature
- **Oral (and written) exam before Christmas (only) for exchange students**

Questions?

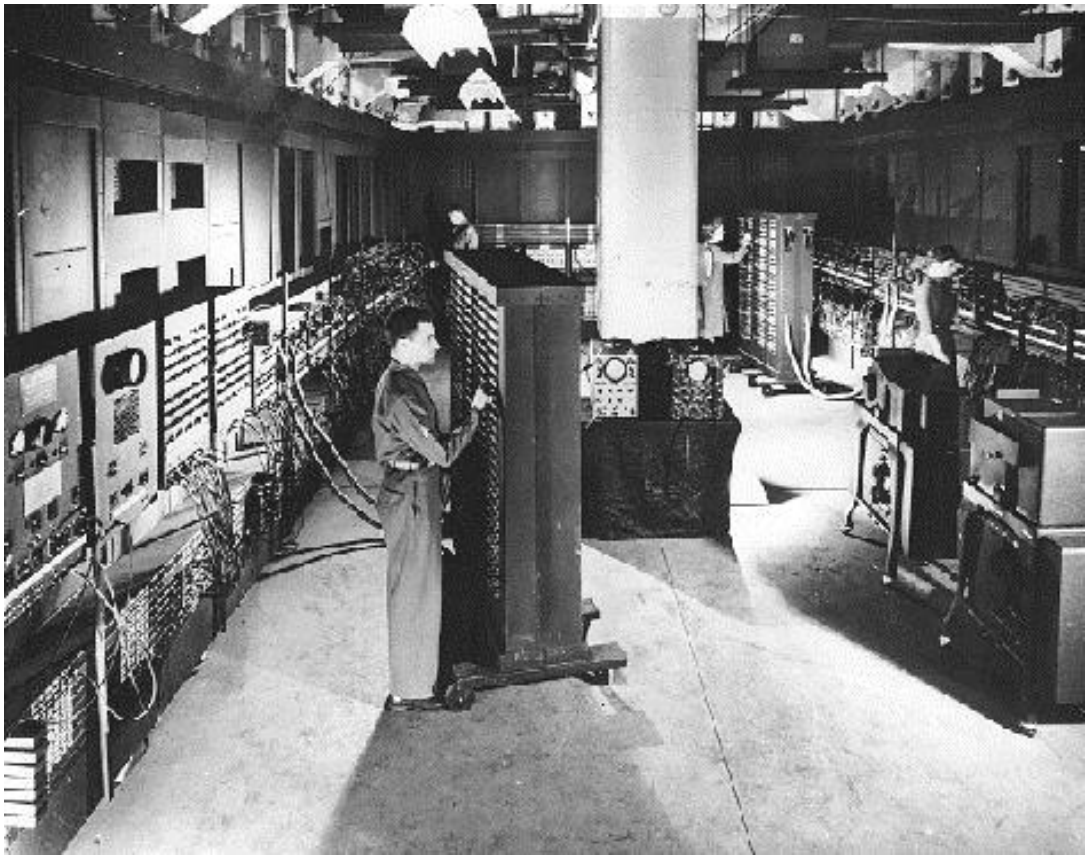
Outline

- Computers
- Computer Architecture
- This Course
- **Trends**
- Performance
- Quantitative Principles

What affects computer architecture?



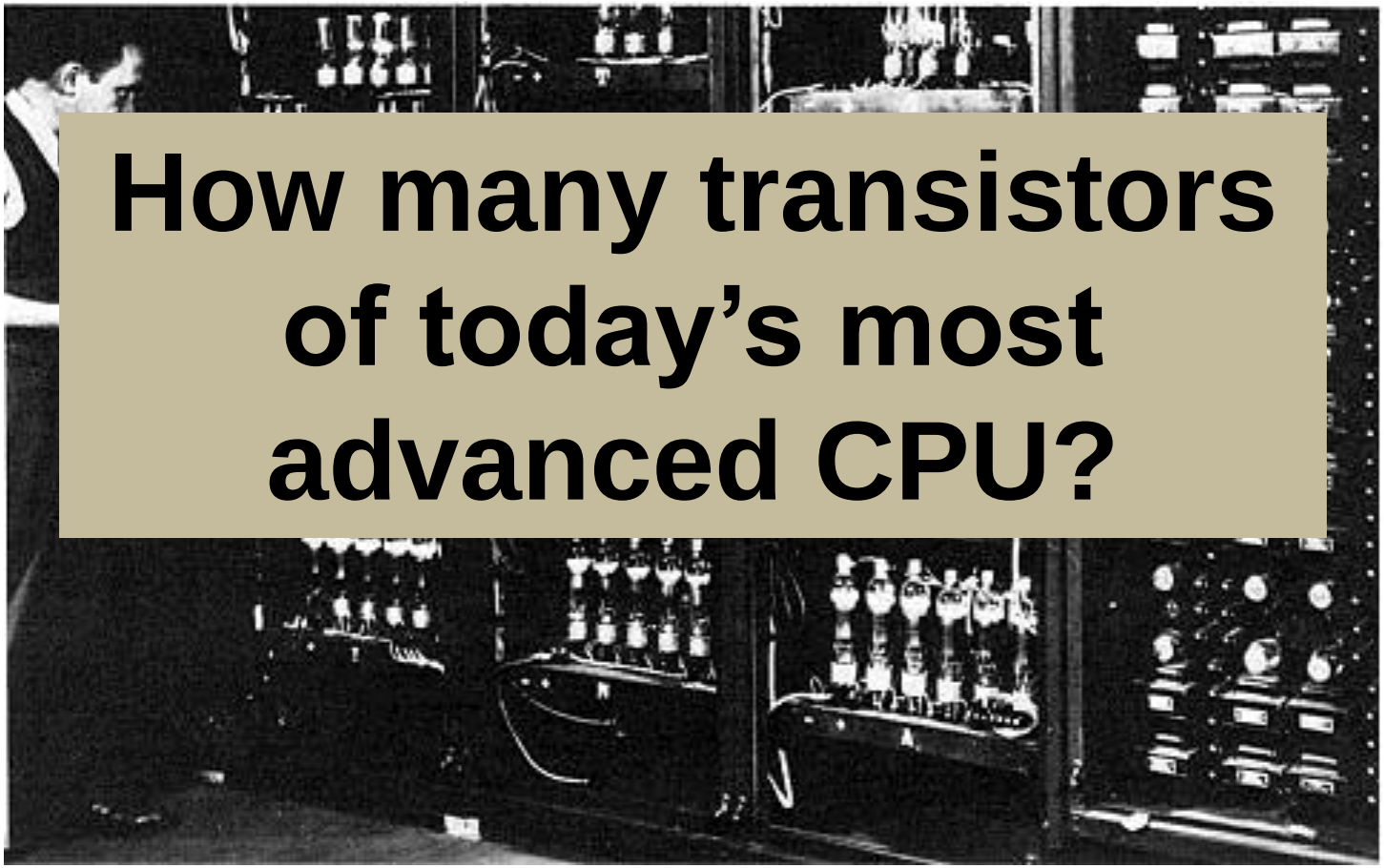
The first electronic computer



ENIAC-1946

18 000 vacuum tubes, 30 ton, 150m^2 ,140kW

The first electronic computer

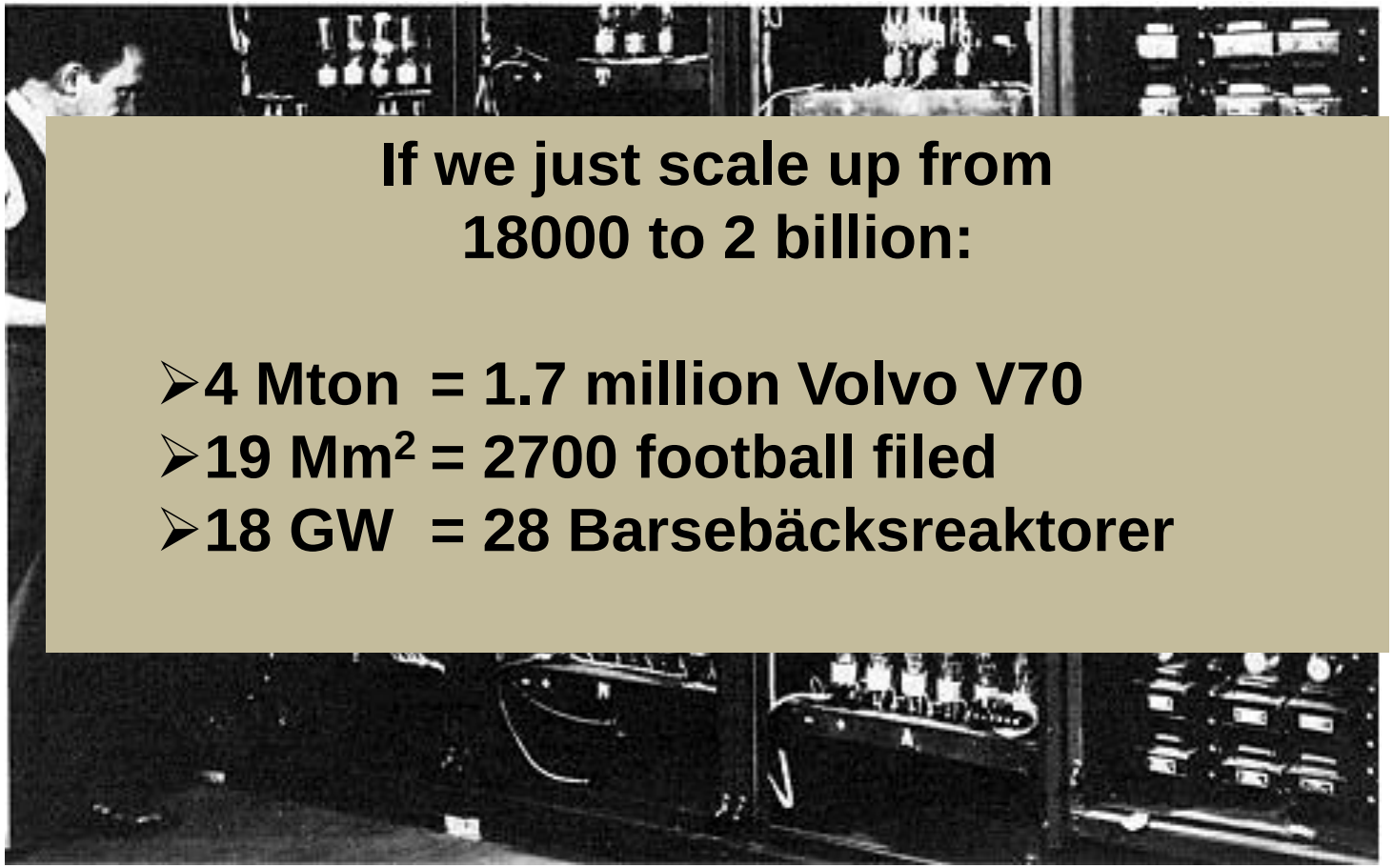


How many transistors of today's most advanced CPU?

ENIAC-1946

18 000 tubes, 30 ton, 150m^2 ,140kW

The first electronic computer



If we just scale up from 18000 to 2 billion:

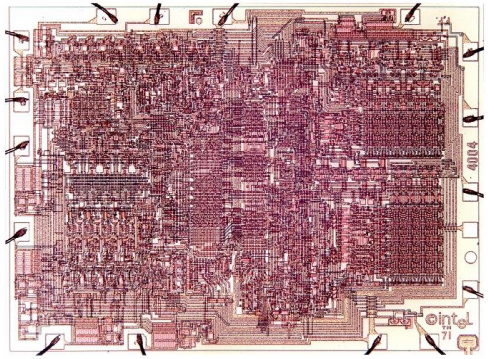
- **4 Mton = 1.7 million Volvo V70**
- **19 Mm^2 = 2700 football filed**
- **18 GW = 28 Barsebäcksreaktorer**

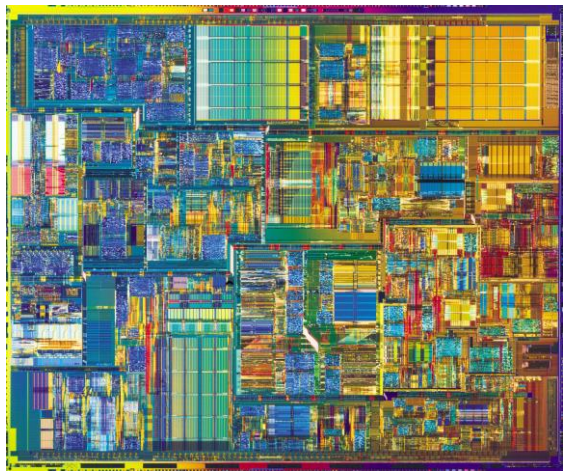
ENIAC-1946

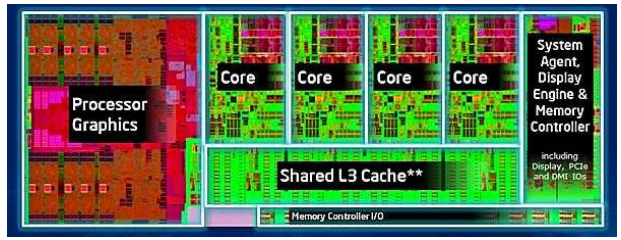
18 000 tubes, 30 ton, 150m^2 ,140kW

Development of Microprocessor

	Year	Transistors	Frequency	cores	Cache
Intel4004	1971	2300	108 kHz	“1”	None
Z80	1976	8500	2.5 MHz	1	None
Intel386	1985	280 000	16 MHz	1	None
Intel486	1989	1 185 000	20 - 50 MHz	1	8 kB
Pentium 4	2000	44 000 000	1 - 2 GHz	1	256 kB
Nehalem	2008	731 000 000	> 3.6 GHz	4	8 MB
Sandy Bridge	2011	995 000 000	3.8 GHz	4+	8 + 1 MB
Haswell	2013	1 860 000 000	> 3.6 GHz	6	15 + 1.5 M
Itanium 9560	2012	3 100 000 000	2.5 GHz	8	32 + 6 MB







Moore's Law

The experts look ahead

Cramming more components onto integrated circuits

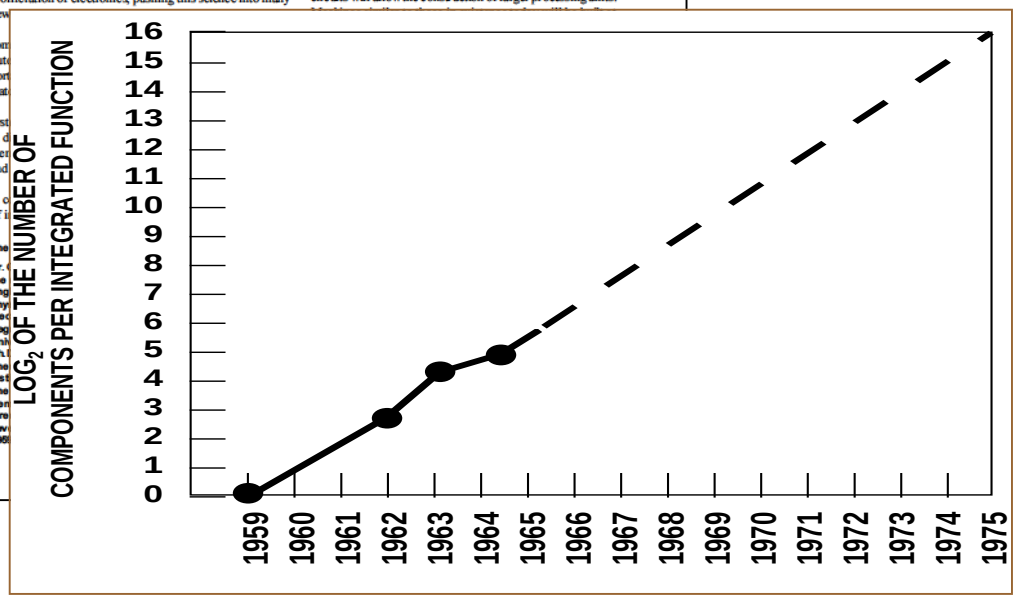
With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

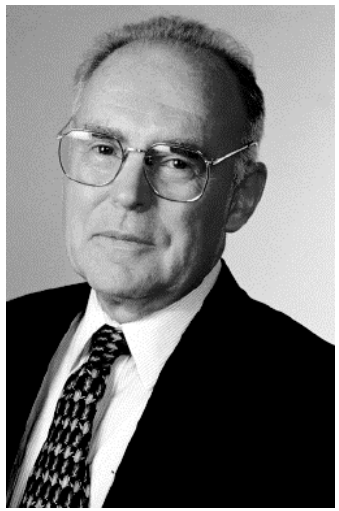
The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many

machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units.

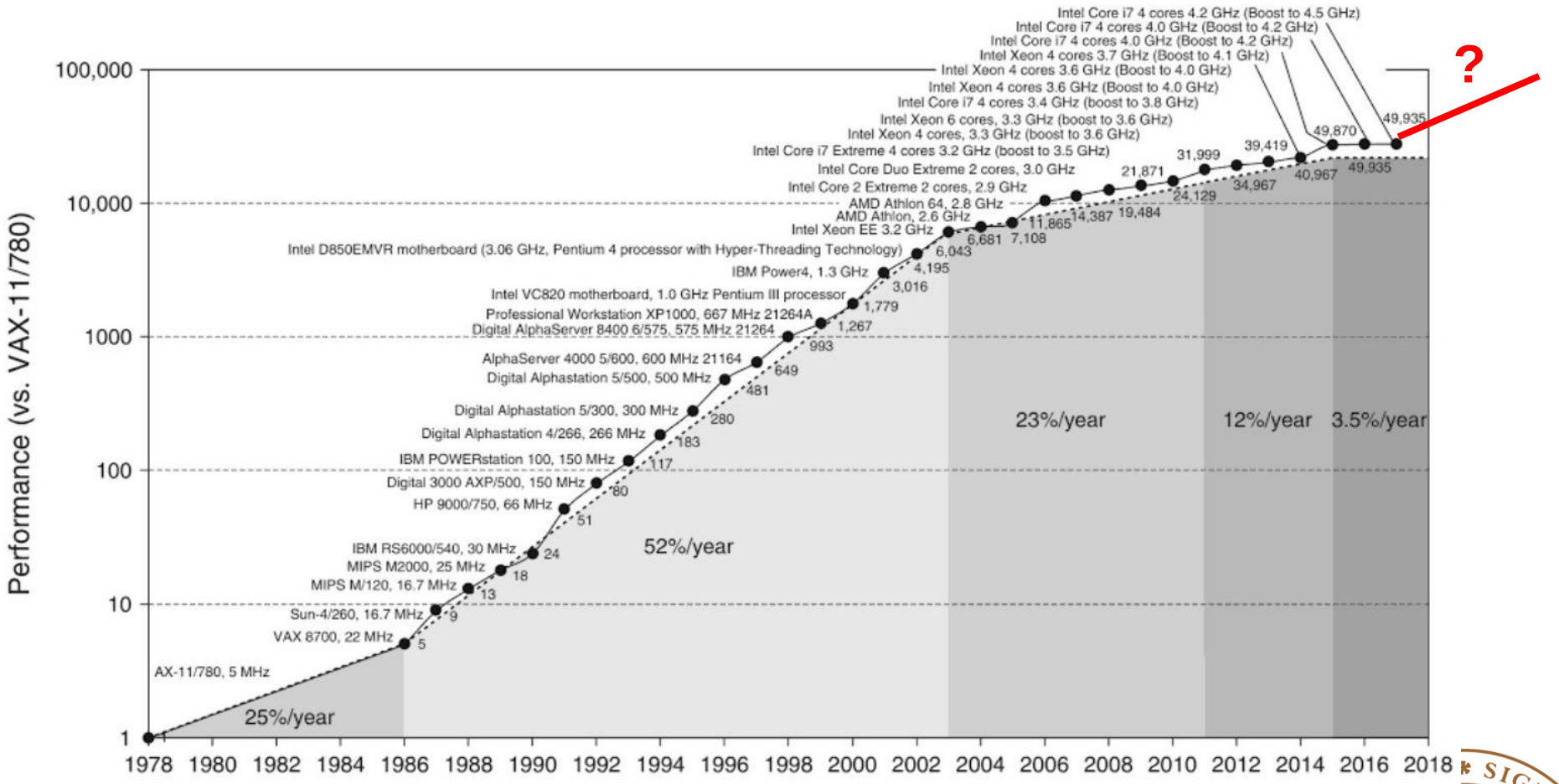


- **Electronics, Apr. 19, 1965**

Gordon Moore (co-founder of Intel) described a doubling every year in the number of components per integrated circuit

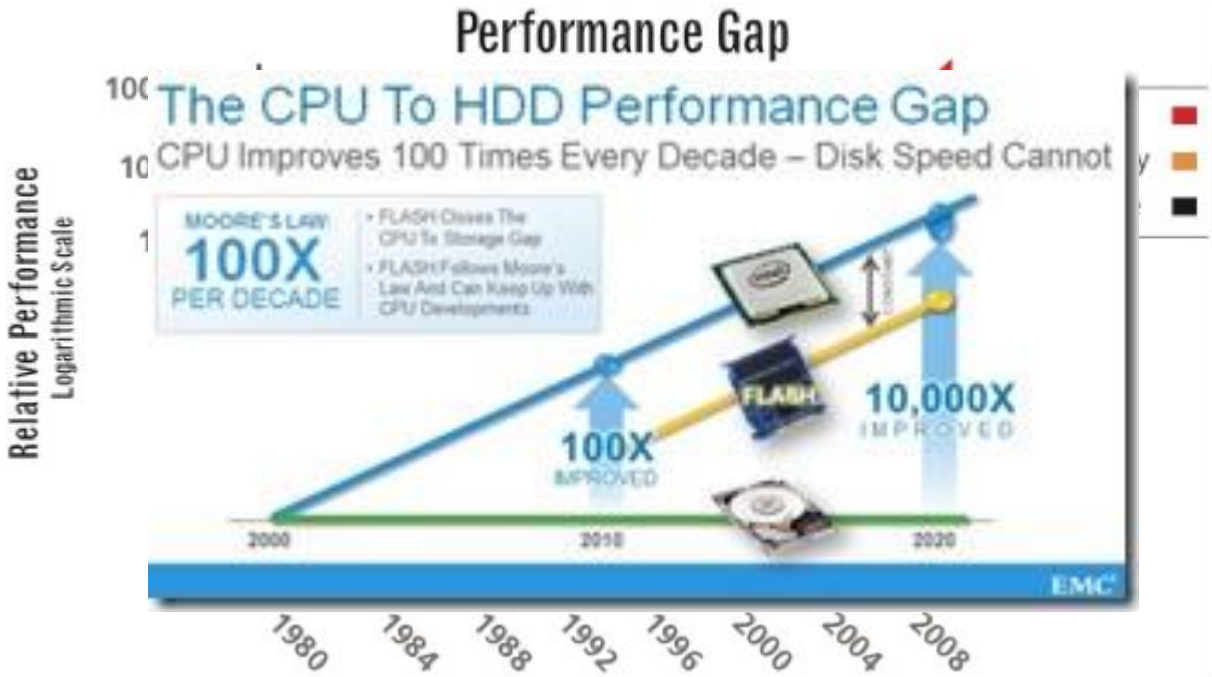


Performance of Microprocessor

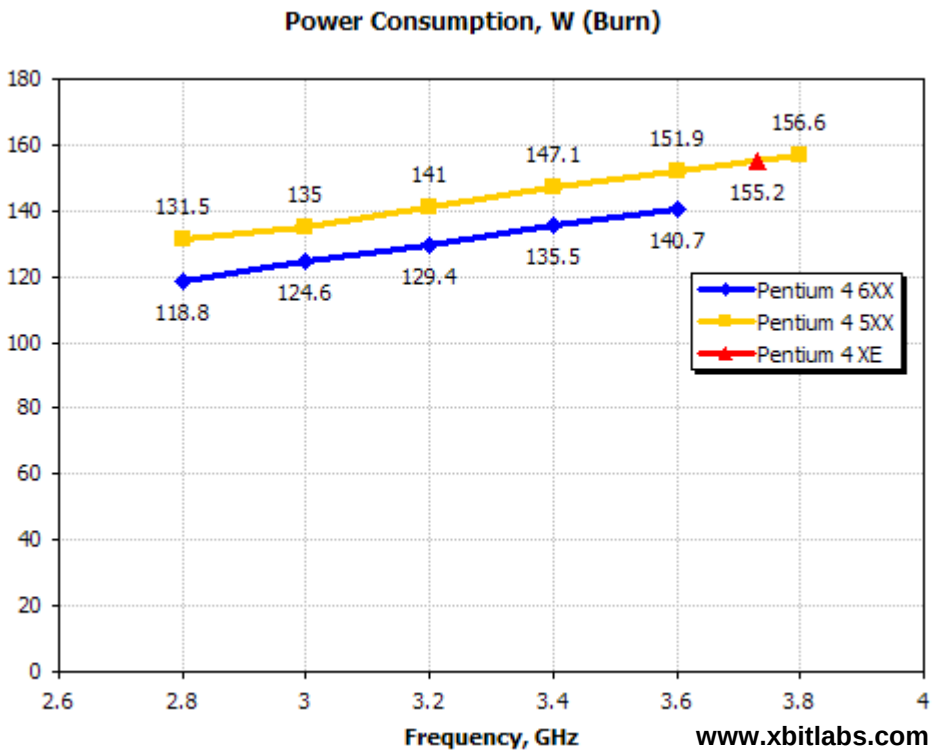


Does not Apply to All

- Processing power doubles every 18 months
- Memory size doubles every 18 months
- Disk capacity doubles every 18 months
- Disk positioning rate (seek & rotate) doubles every ten years!
- Speed of DRAM and disk improves a few % per year



Moore's Law: power density



Pentium IV chip area (in 130 nm technology) **1.3 cm^2**

This gives about 100 W/cm^2 that needs to be transported away (cooling)

Comparison: This little thing operates at about 10 W/cm^2.

Heating (power) is an issue

Samsung has reportedly stopped Galaxy Note 7 production

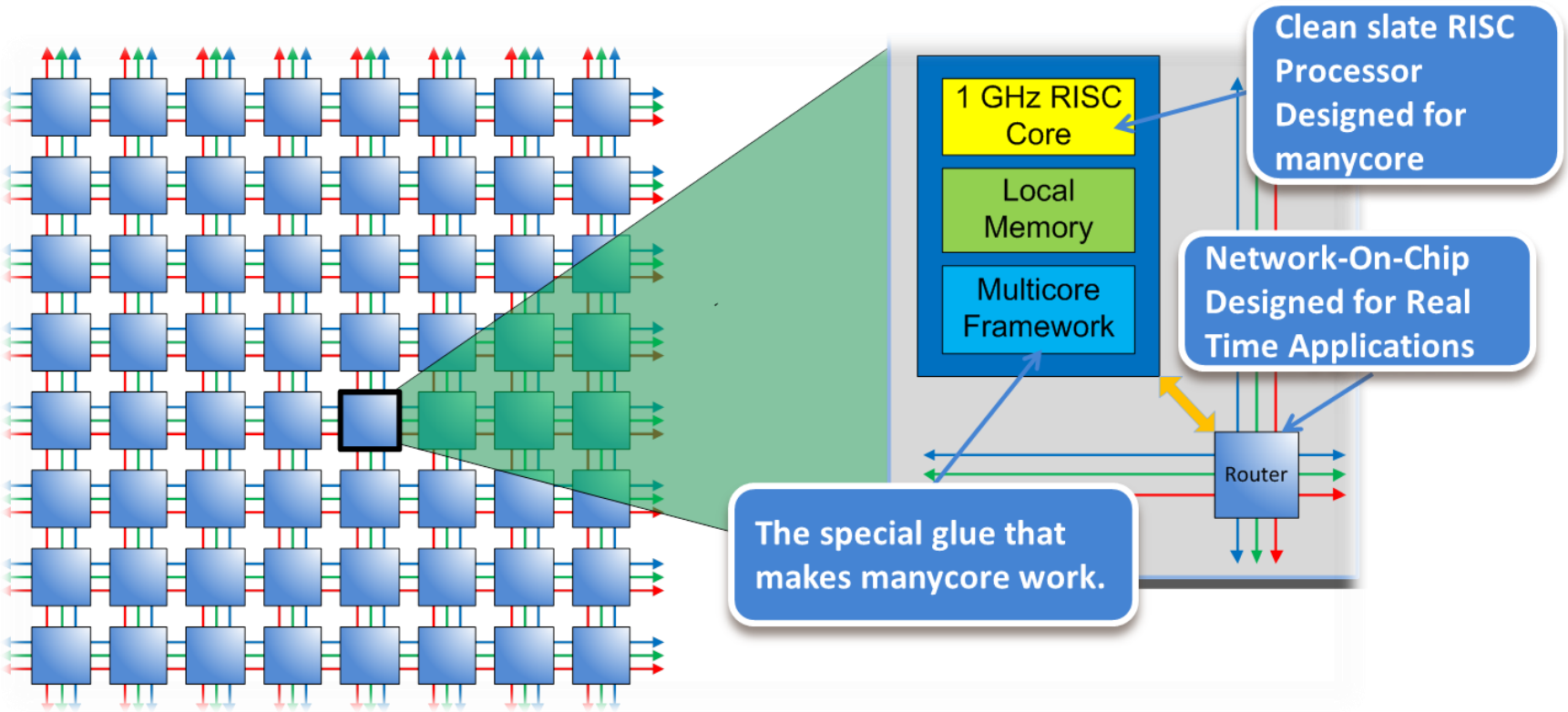
by: **NIRAVE GONDHIA**
6 HOURS AGO

3.1K shares

f 3022 | 69

Samsung has had a hard couple of months and the Galaxy Note 7 recall looks set to cause even more trouble for the company, with a new report suggesting that Samsung has temporarily halted production of the Galaxy Note 7.

Architecture change?



Coprocessor to ARM/Intel CPU

25mW per core

Ease To Use

Outline

- Computers
- Computer Architecture
- This Course
- Trends
- **Performance**
- Quantitative Principles

What is Performance?

Plane	DC to Paris	Speed
Boeing 747	6.5 h	980 km/h
Concorde	3 h	2160 km/h

- **Time to complete a task (T_{exe})**
  - Execution time, response time, latency
- **Task per day, hour…**
  - Total amount of tasks for given time
  - Thoughput, bandwidth
- **Speed of Concorde vs Boeing 747**
- **Throughput of Boeing 747 vs Concorde**

Performance

$$Performance(X) = \frac{1}{T_{exe}(X)}$$

"X is n times faster than Y" means:

$$\frac{T_{exe}(Y)}{T_{exe}(X)} = \frac{Performance(X)}{Performance(Y)} = n$$

How to define execution time?

Aspect of CPU performance

$$\text{CPUtime} = \text{Execution time} = \textit{seconds}/\textit{program} =$$

$$\underbrace{(\textit{executed})\textit{instr.}/\textit{program}}_{IC} * \underbrace{\textit{cycles}/\textit{instr.}}_{CPI} * \underbrace{\textit{seconds}/\textit{cycle}}_{T_c}$$

	IC	CPI	T_c
Program	X		
Compiler	X	(X)	
Instr. Set	X	X	
Organization		X	X
Technology			X

Instructions are not created equally

"Average Cycles per Instruction"

CPI_{op} = Cycles per Instruction of type *op*

IC_{op} = Number of executed instructions of type *op*

$$CPUtime = T_c * \sum (CPI_{op} * IC_{op})$$

"Instruction frequency"

$$\overline{CPI} = \sum (CPI_{op} * F_{op}) \text{ where } F_{op} = IC_{op}/IC$$

Average CPI: example

Op	F_{op}	CPI_{op}	$F_{op} * CPI_{op}$	% time
ALU	50 %	1	0.5	(33 %)
Load	20 %	2	0.4	(27 %)
Store	10 %	2	0.2	(13 %)
Branch	20 %	2	0.4	(27 %)

$\overline{CPI} = 1.5$

Invest resources where time is spent!

Outline

- Computers
- Computer Architecture
- This Course
- Trends
- Performance
- **Quantitative Principles**

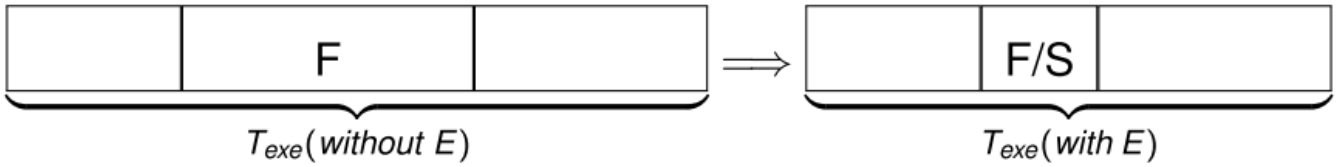
Quantitative Principles

This is intro to design and analysis

- Take advantage of parallelism
  - ILP, DLP, TLP, ...
- Principle of locality
  - 90% of execution time in only 10% of the code
- Focus on the common case
  - In making a design trade-off, favor the frequent case ove the infrequent case
- Amdahl's Law
  - The performance improvement gained from uisng faster mode is limited by the fraction of the time the faster mode can be used
- The Processor Performance Equation

Amdahl's Law

Enhancement E accelerates a fraction F of a program by a factor S



Speedup due to enhancement E:

$$Speedup(E) = \frac{T_{exe}(without\ E)}{T_{exe}(with\ E)} = \frac{Performance(with\ E)}{Performance(without\ E)}$$

$$T_{exe}(with\ E) = T_{exe}(without\ E) * [(1 - F) + F/S]$$

$$Speedup(E) = \frac{T_{exe}(without\ E)}{T_{exe}(with\ E)} = \frac{1}{(1-F)+F/S}$$

Best you could ever hope to do:

$$Speedup_{maximum} = \frac{1}{(1 - Fraction_{enhanced})}$$

Amdahl's Law: example

- New CPU is 10 times faster!
- 60% for I/O which remains almost the same…

$$\text{Speedup}_{\text{overall}} = \frac{1}{(1 - \text{Fraction}_{\text{enhanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}}}$$

$$= \frac{1}{(1 - 0.4) + \frac{0.4}{10}} = \frac{1}{0.64} = 1.56$$

Apparently, its human nature to be attracted by 10X faster, vs. keeping in perspective its just 1.6X faster