Κοινή λογική 1: Ίδια δουλειά τελειώνει στον ίδιο χρόνο

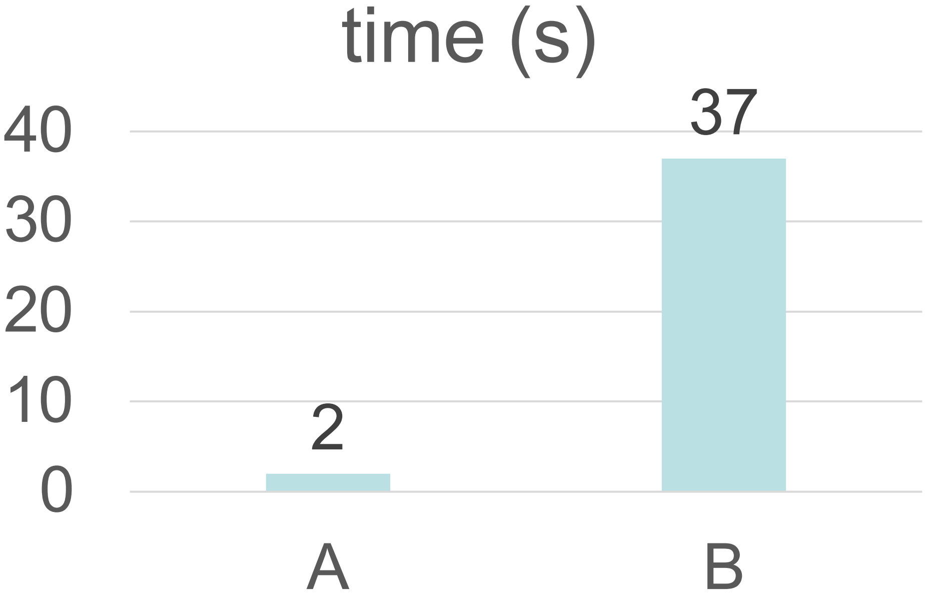
```
     (A)
for(i=0;i<SIZE;i++)
   for(j=0;j<SIZE;j++)
     s += a[i][j];
```

```
     (B)
for(j=0;j<SIZE;j++)
   for(i=0;i<SIZE;i++)
     s += a[i][j];
```

Ποιο είναι πιο γρήγορο;

a[SIZE][SIZE] δεδομένα που επεξεργάζεται το πρόγραμμα

SIZE=32000



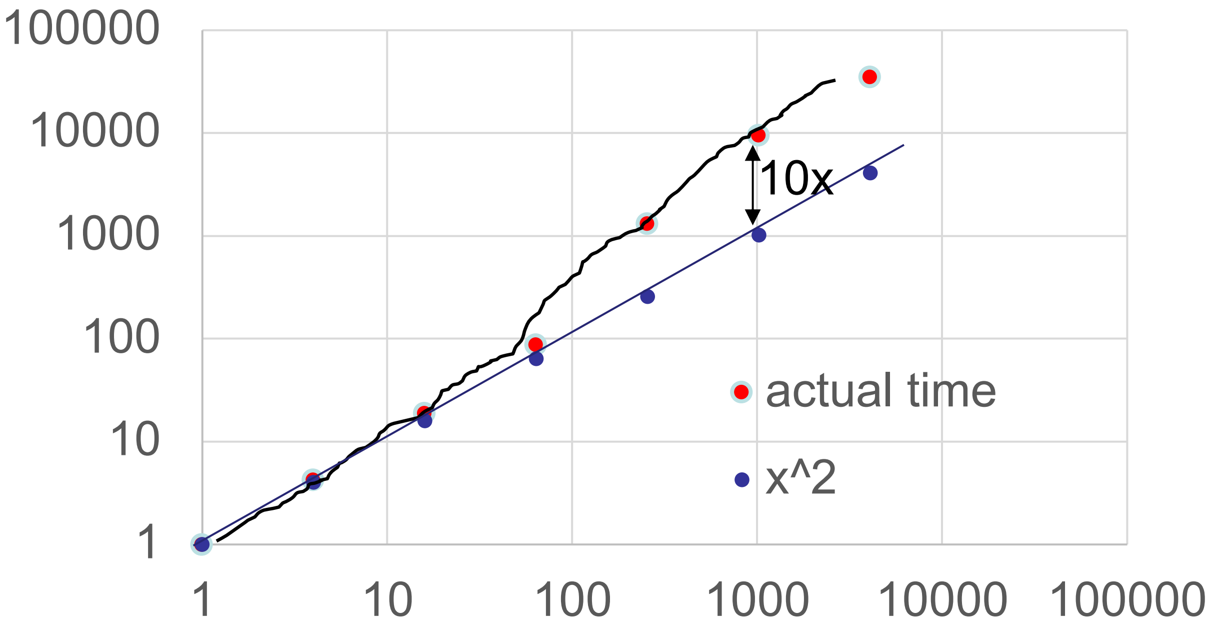
Κοινή λογική 2: x φορές περισσότερη δουλειά χρειάζεται x φορές περισσότερο χρόνο

(B)

```
for(j=0;j<SIZE;j++)
    for(i=0;i<SIZE;i++)
      s += a[i][j];
```

Εάν το SIZE αυξηθεί κατά x ο αριθμός των πράξεων αυξάνεται x^2. Ο χρόνος;

Πως μεγαλώνει ο χρόνος με μεγαλύτερο πρόβλημα



Κοινή λογική 3: Πιο λίγη δουλειά τελειώνει πιο γρήγορα

(A)

```
for(i=0;i<SIZE;i++)
   if (a[i]>1000000000)
     s++;
   else if (a[i] > 500000000)
     n++;
   else
     p++;
 }
printf("%d %d %d\n",s,n,p);
```

a[SIZE] πίνακας με μέγεθός 16000000000 ακεραίων (16εκ.)

(A)

```
 for(i=0;i<SIZE;i++)
     if (a[i]>1000000000)
       s++;
     else if (a[i] > 500000000)
       n++;
     else
       p++;
  }
 printf("%d %d %d\n",s,n,p);
8.6e9   3.4e9   4.0e9
```

1 x 8.6 + 2 x 3.4 + 3 x 4.0 = 27.4

(A)

```
 for(i=0;i<SIZE;i++)
      if (a[i]>1000000000)
        s++;
      else if (a[i] > 500000000)
        n++;
      else
        p++;
  }
 printf("%d %d %d\n",s,n,p);
8.6e9   3.4e9   4.0e9
```

1 x 8.6 + 2 x 3.4 + 3 x 4.0 = 27.4

(B)

```
 for(i=0;i<SIZE;i++)
      if (a[i]==0)
        s++;
      else if (a[i]==1)
        n++;
      else
        p++;
  }
  printf("%d %d %d\n",s,n,p);
0           0           16.0e9
```

3 x 16.0 = 48

(A)

```
for(i=0;i<SIZE;i++)
     if (a[i]>1000000000)
       s++;
     else if (a[i] > 500000000)
       n++;
     else
       p++;
 }
printf("%d %d %d\n",s,n,p);
```

```
8.6e9   3.4e9   4.0e9
```

1 x 8.6 + 2 x 3.4 + 3 x 4.0 = 27.4

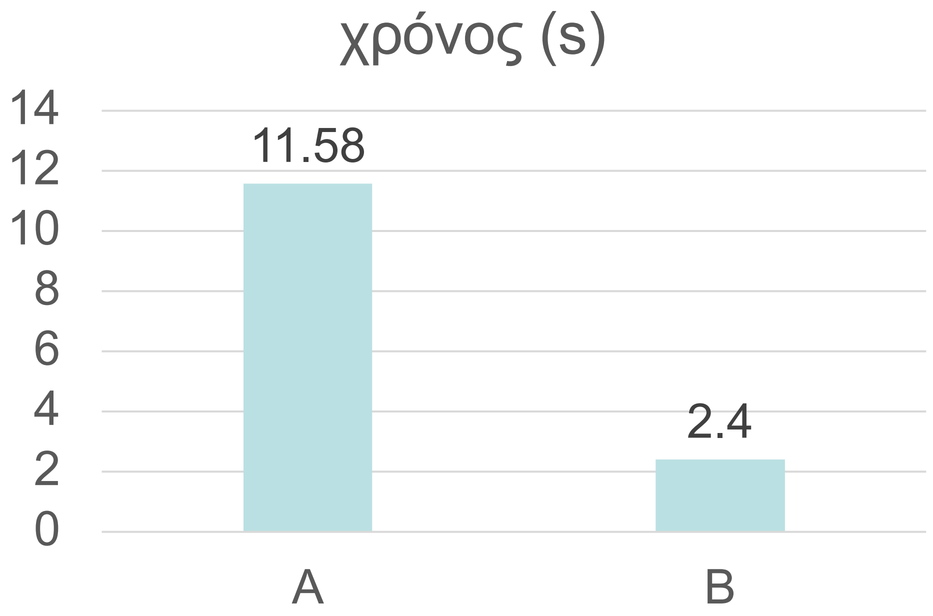
(B)

```
for(i=0;i<SIZE;i++)
     if (a[i]==0)
       s++;
     else if (a[i]==1)
       n++;
     else
       p++;
 }
 printf("%d %d %d\n",s,n,p);
```

```
0          0          16.0e9
```

3 x 16.0 = 48

48/27.4 = 1.75 περισσότερες συγκρίσεις το B

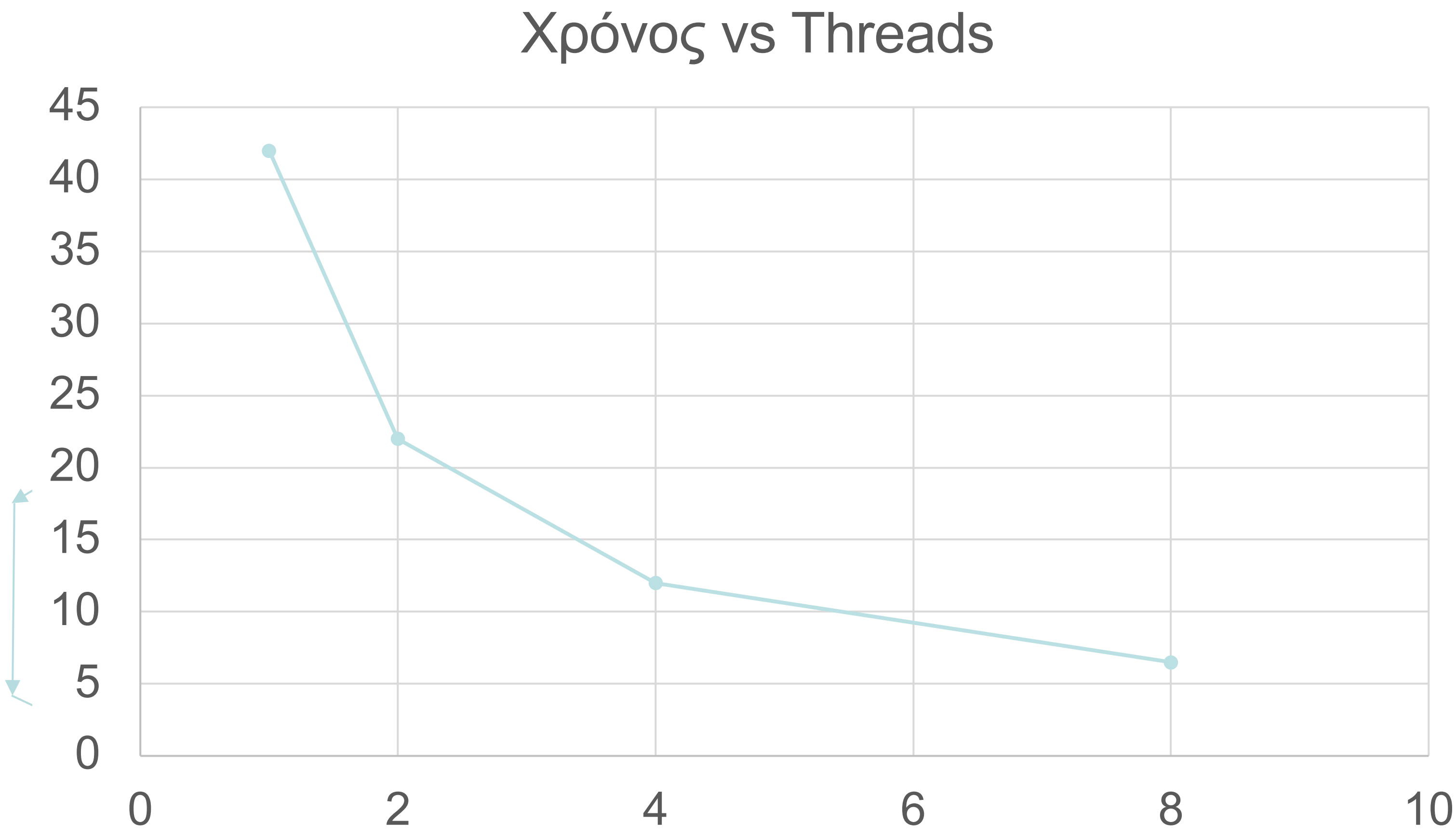

χρόνος (s)
14
12
10
8
6
4
2
0
11.58
2.4
A
B

Παράλληλη vs Σειριακή Επεξεργασία

```
void countf(int *a, int key, int begin, int end,int *count){
        int i;
        for(i=begin;i<end;i++)
                if (a[i]==key)
                        ++(*count);
}
```

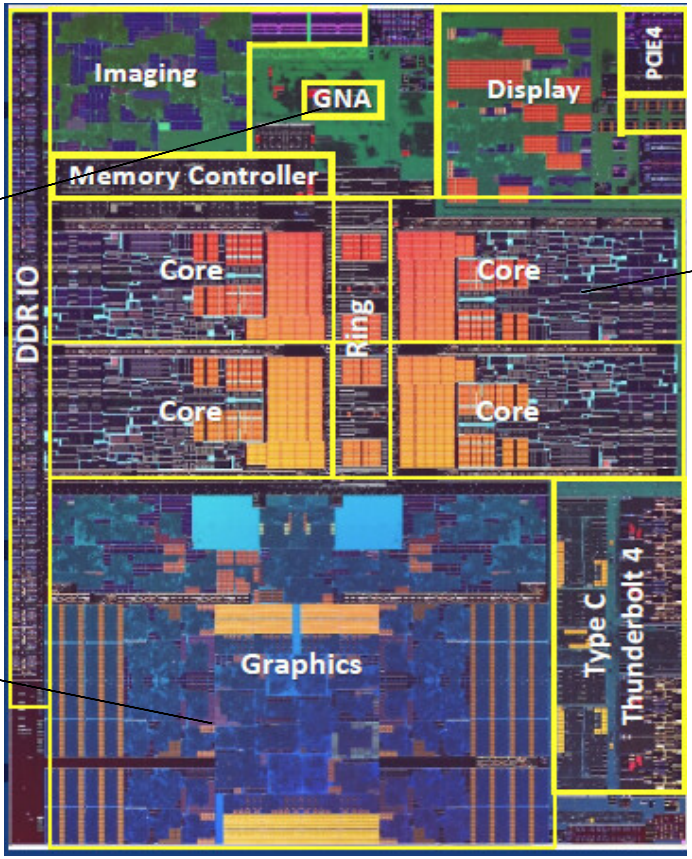
```
void countf(int *a, int key, int begin, int end,int *count){
            int i;
                                             // n is table size
}
```

```
a,
ey,
*n/p,
+1)*n/p,
counters[j]);
```



System on Chip (SoC)

- Why we need multiple type of cores?
- How this picture will look like in 5 years?



ML Accelerator

CPU cores

Graphics cores
(igpu,Integrated GPU)

ΕΠΛ605 Προχωρημένη Αρχιτεκτονική Υπολογιστών

Εισαγωγή: Τάσεις Τεχνολογίας Υπολογιστών

1η Εργασία Ημερομηνία παράδοσης 26/1/21 πριν την αρχή του μαθήματος ηλεκτρονικά (επανάληψη για βασικές έννοιες). Reviews for readings

READING

Chapter 1 Όλα τα Sections

Some slides based on Elsevier Material

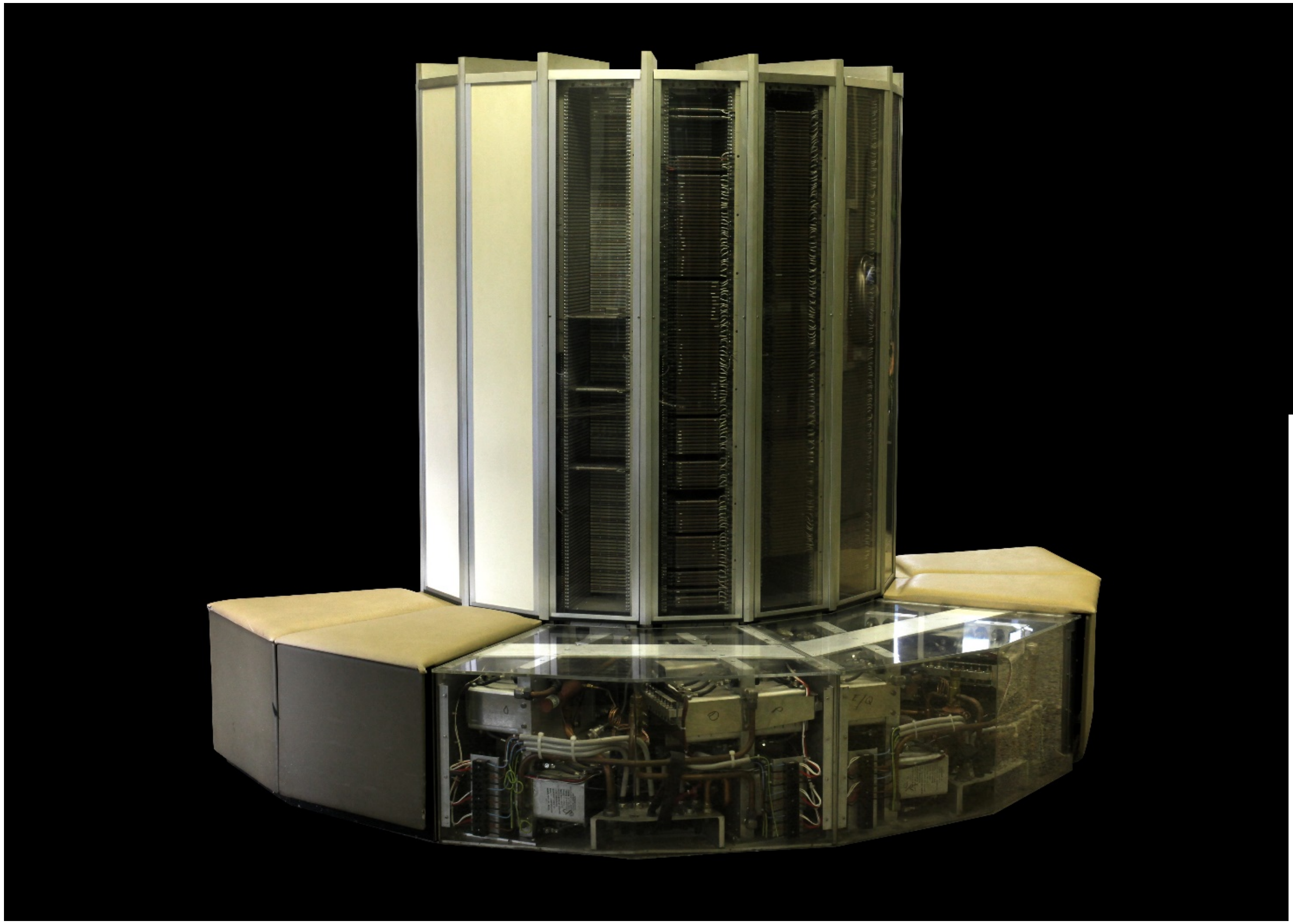
Computer Technology

- Amazing improvements:

today can buy for $500 a mobile phone that has superior performance, more main memory and disk space than a $9million computer in 1975

- 9million/500 = 18000 times cost reduction
- Inflation: $1 1975 = $4.56 today
- 41million/500 = 80000 times cost reduction!

1975: Cray-1 (Wikipedia)



1975: Cray-1 (Wikipedia)

- Supercomputer
- Build by Cray Research
- Architect: Seymour Cray
- 64 bit machine: scalar and vector units
- Clock: 80MHz
- Technology: ECL, 2.5M transistors (1000s of small chips)
- FLOPS: 160 MFLOPS
- Memory: 1M words (word: 64 bits, 8MB)
- Weight: 5.5 tons
- Cost: ~$9million
- Power: 115KW
- Sales Volume: ~80 units

2015: Galaxy S6 Edge

- Smartphone
- Build by Samsung
- 64 bit Exynos System on Chip
  - CPU: octa core, 4 A57, 4 A53
  - GPU: Mali 3D rendering engine
- Clock:
  - CPU: 2.1GHz, 1.5GHz
  - GPU: 600-700 MHz
- Technology 14nm FinFET, ~billion of transistors
- GPU FLOPS: >100s GFLOPS
- Memory: 3GB
- Storage: 32-128GB
- Weight: 132 g
- Cost: ~$500
- Battery: 2600 mAh
  - Battery Life depends on applications running
  - 10hrs @4V => 1W
- Sales Volume: 6million units in first month

Comparison

	1975	2015	Ratio
Transistors	2.5million	Billions	1000
Main Memory	1MB	3GB	3000
Clock	80MHz (12.5ns)	2.1GHz (0.48ns)	25
Cores	1 CPU+ Vector	8 CPU+GPU	>8
FLOPS	160M	100G	1000
Cost	$9million ($41mil)	$500	18000 (41000)
Weight/Dimen	5.5 tons	132 g	~40000
Power	115KW	1W	115000
Sales	10s	millions	million
Applications	Few Scientific	Many Client	100-1000

Comparison

	1975	2015	2021 (G S21)
Transistors	2.5million	Billions	Billions
Main Memory	1MB	3GB	16GB
Clock	80MHz (12.5ns)	2.1GHz (0.48ns)	2.2-2.9GHz
Cores	1CPU+vector	8CPU+GPU	8CPU+GPU+3NPU
FLOPS	160M	100G	
Cost	$9million ($41mil)	$500	
Weight/Dimen	5.5 tons	132 g	227 g
Power	115KW	1W	?
Sales	10s	millions	?
Applications	Scientific	Client	Client+AI

Observations

- Amazing density/bandwidth/performance improvements
  - Processor, memory, (storage, network)…
- Weight/Dimensions reduction
- New applications
- More sales (sw/hw)
- Lower cost

- Diversification in type of cores

- (Latency) Clock rate improves at a much slower rate

- Source of improvements/trends?

Compute Stack Abstractions

Applications (scientific, client, cloud,IoT)

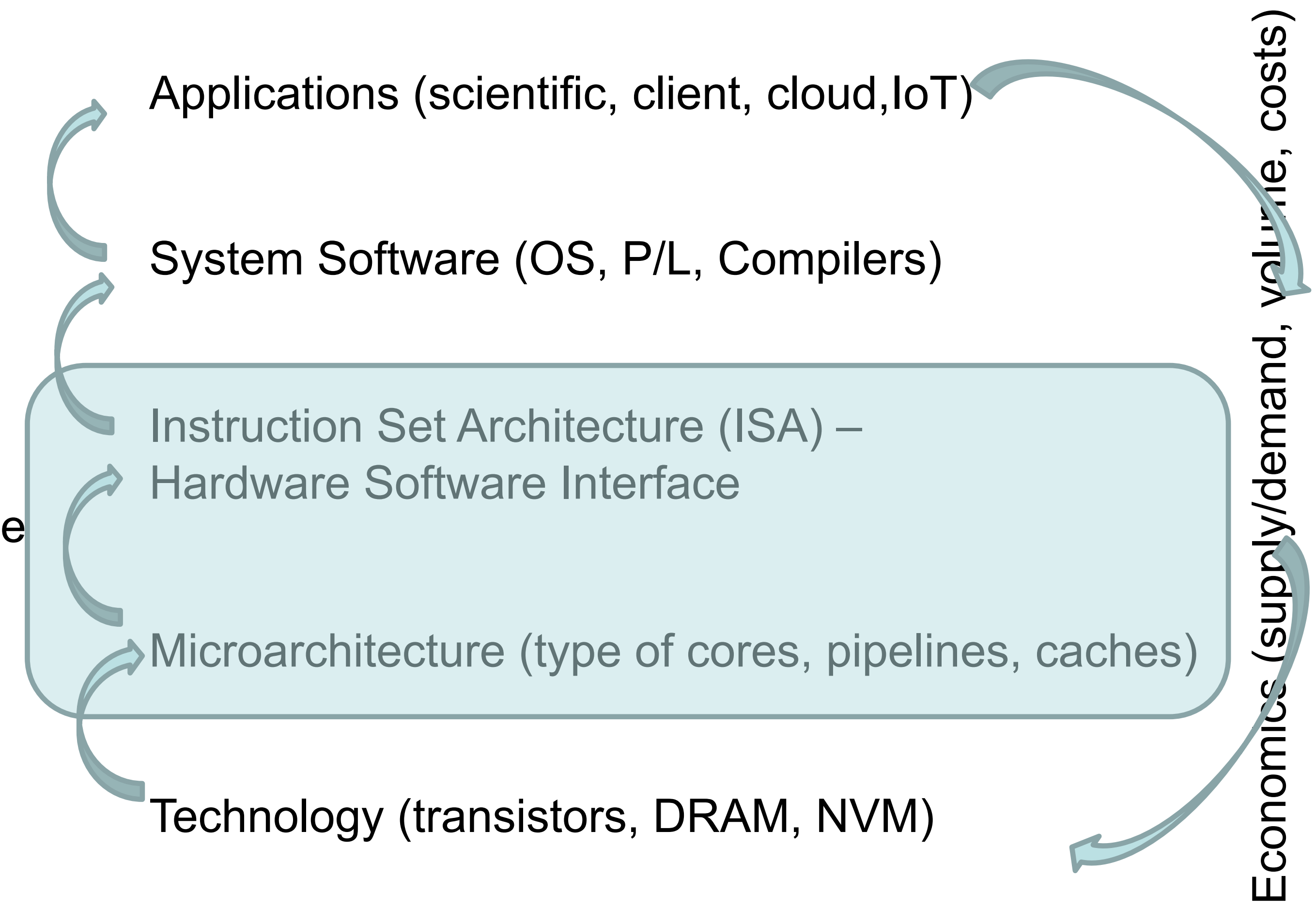
System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

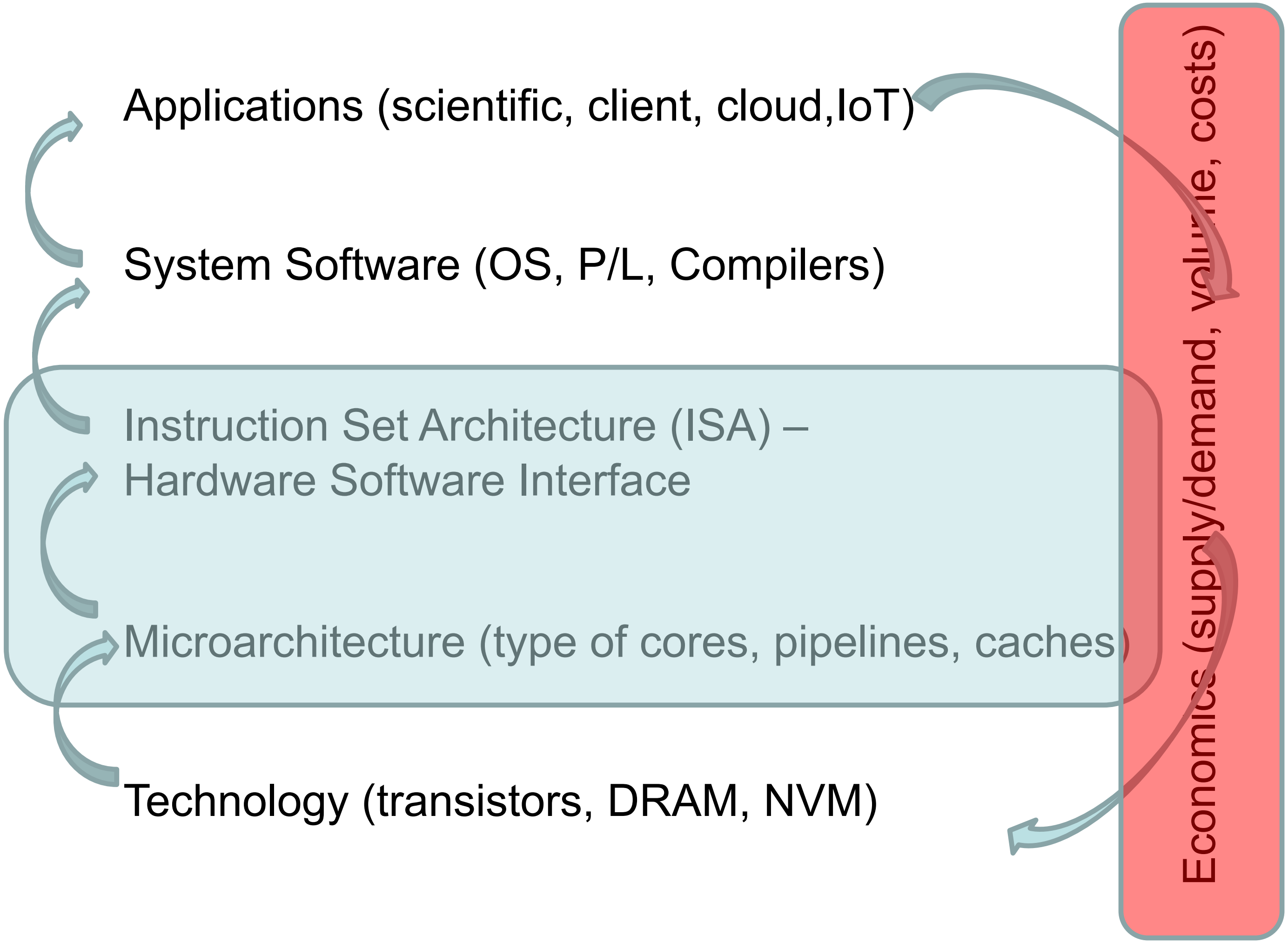
Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

Economics (supply/demand, volume, costs)

Our focus…
but interdependence
with other layers

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

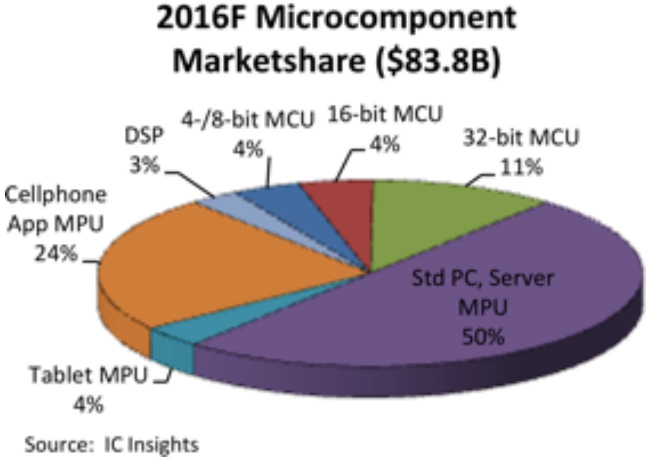
Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

Economics (supply/demand, volume, costs)

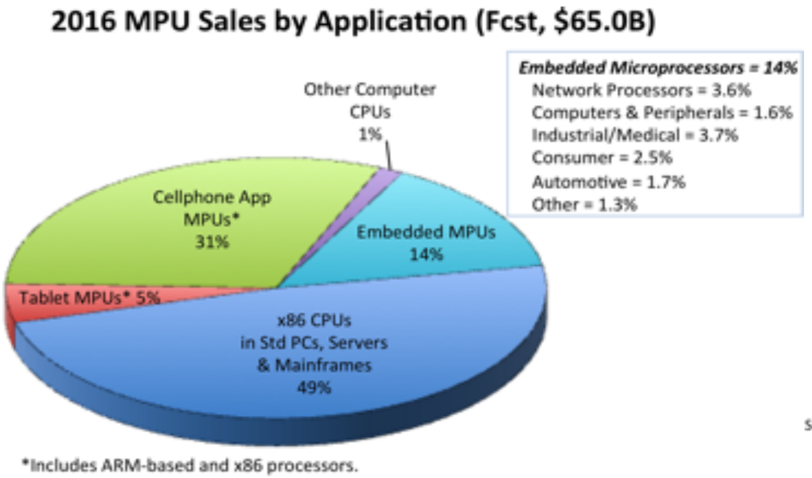
IC Sales

- 267 billion Integrated circuits expected to ship in 2016 (McClean Report)
- $203 billion sales (IC Insights)
  - 36.7 Billion Memory/Flash Units
  - 300 million pc/laptop processors (2015)
  - 23 million server processors – mostly Intel x86 (2015)
  - 14 billion ARM based chips (2015)



2016F Microcomponent Marketshare ($83.8B)

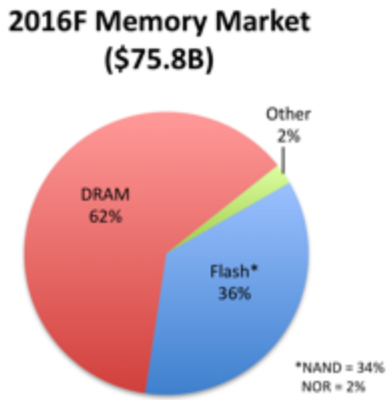
Source: IC Insights



2016 MPU Sales by Application (Fcst, $65.0B)

*Includes ARM-based and x86 processors.

Source: IC Insights



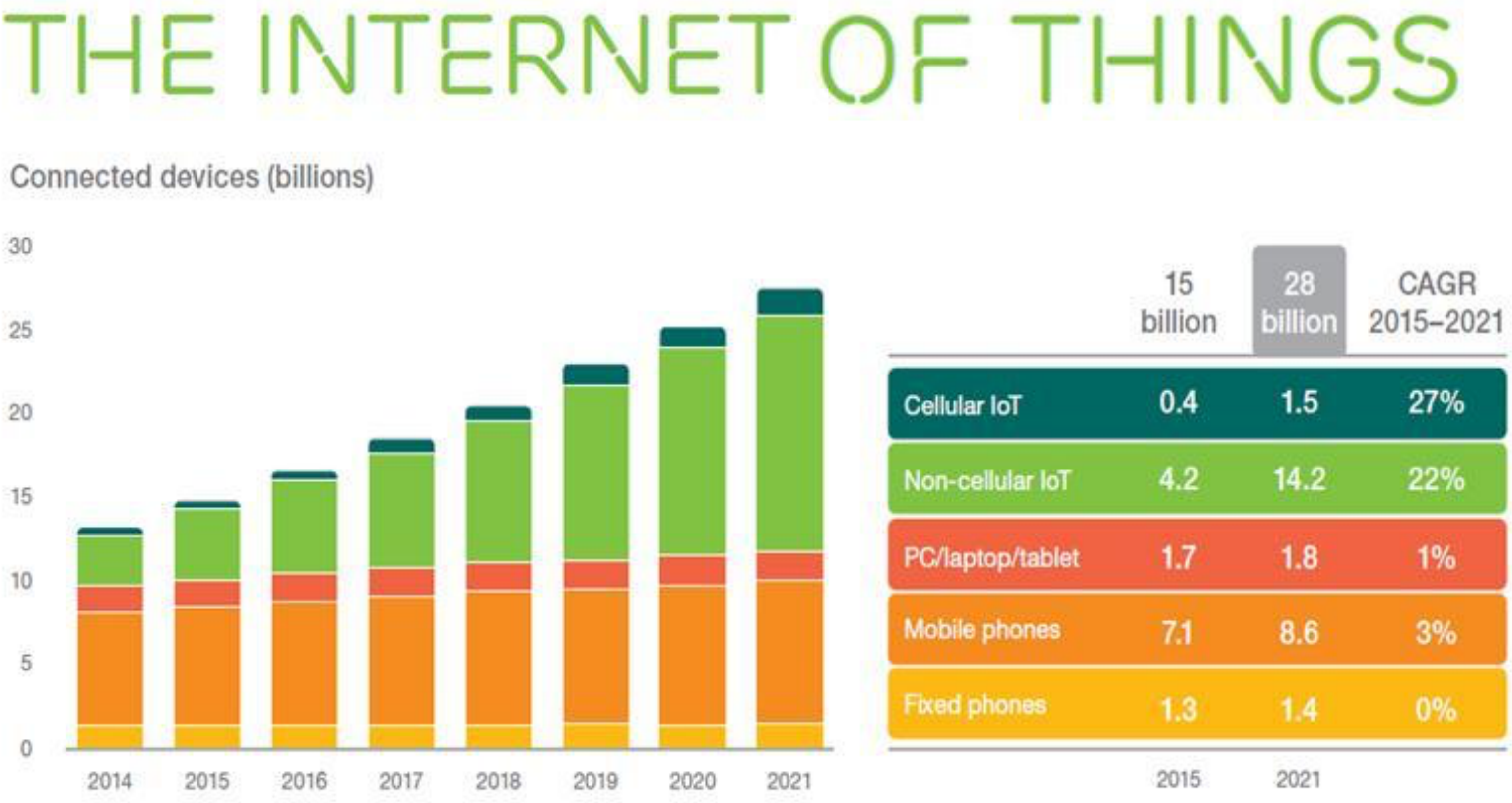
2016F Memory Market ($75.8B)

Source: IC Insights

Classes of Computers

- Embedded and Internet of Things:
  - Emphasis: price (and other depending on application/requirements, e.g. safety, battery autonomy)
  - Embedded/IoT computer can be simple or quite sophisticated (e.g., temperature vs autonomous driving processor)
- Personal Mobile/Laptops
  - e.g. smart phones, tablet computers
  - Emphasis on cost, video quality, real-time, battery life
- Desktop
  - Emphasis on cost-performance, graphics
- Servers
  - Emphasis on availability, scalability, throughput, quality of service
- Clusters/Warehouse Scale Computers
  - Used for “Software/Infra/Platform as a Service (XaaS)”
  - Emphasis on availability and price-performance
  - Based on commodity parts (typically server grade)

The IoT Era



THE INTERNET OF THINGS

Connected devices (billions)

	15 billion	28 billion	CAGR 2015–2021
Cellular IoT	0.4	1.5	27%
Non-cellular IoT	4.2	14.2	22%
PC/laptop/tablet	1.7	1.8	1%
Mobile phones	7.1	8.6	3%
Fixed phones	1.3	1.4	0%
	2015	2021	

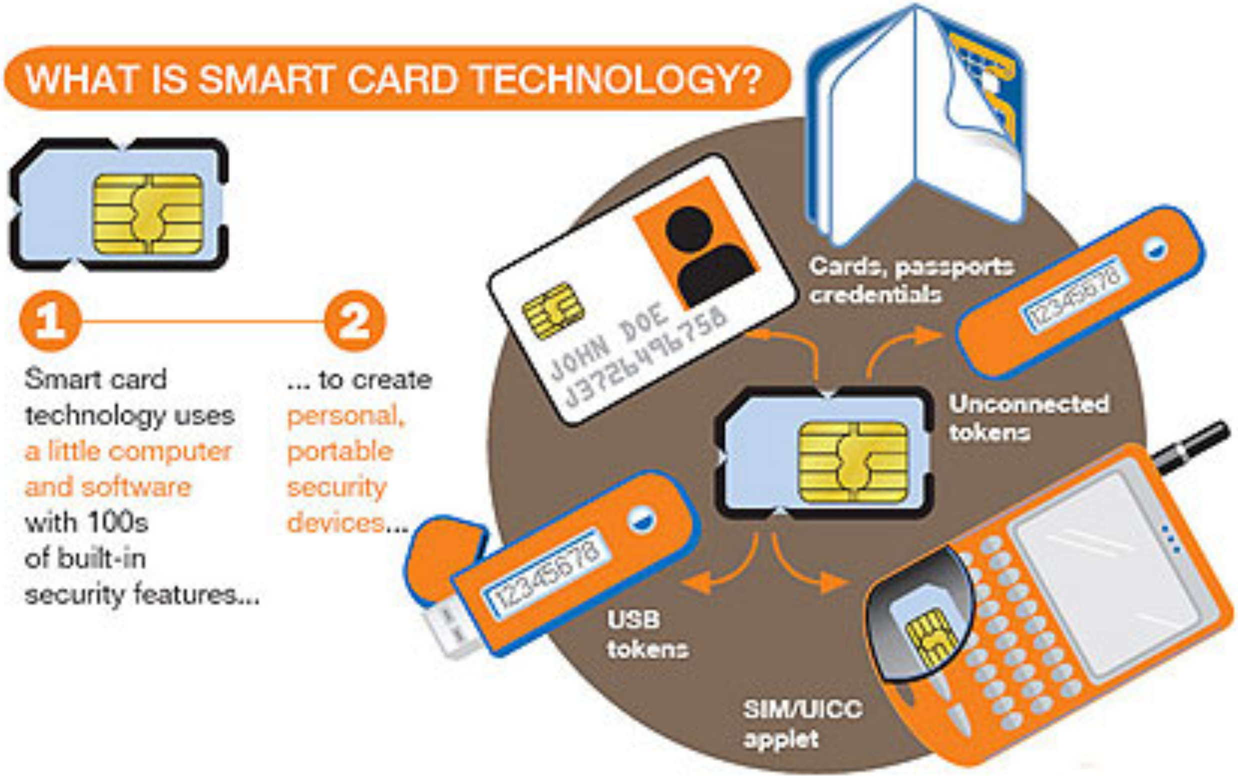
Many sensory devices that collect (pre-process) and transmit data (directly or through other devices) to data centers

Data centers servers analyze (data analytics) and take decisions

E.g. Autonomous Cars

Missing in this picture is the server growth to support IoT

Embedded: Smart Card

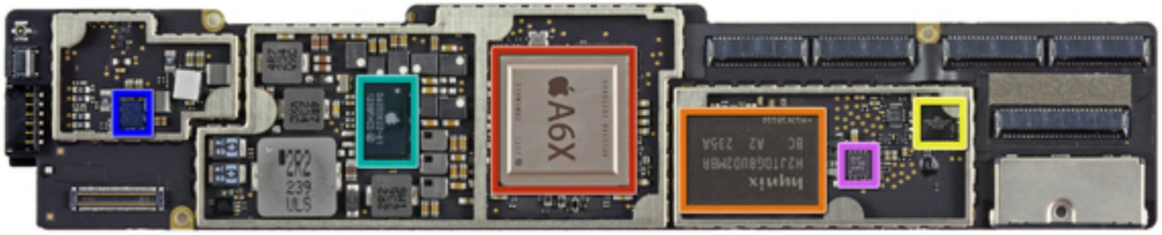


WHAT IS SMART CARD TECHNOLOGY?

1 Smart card technology uses a little computer and software with 100s of built-in security features...

2 ... to create personal, portable security devices...

Personal Mobile: iPad

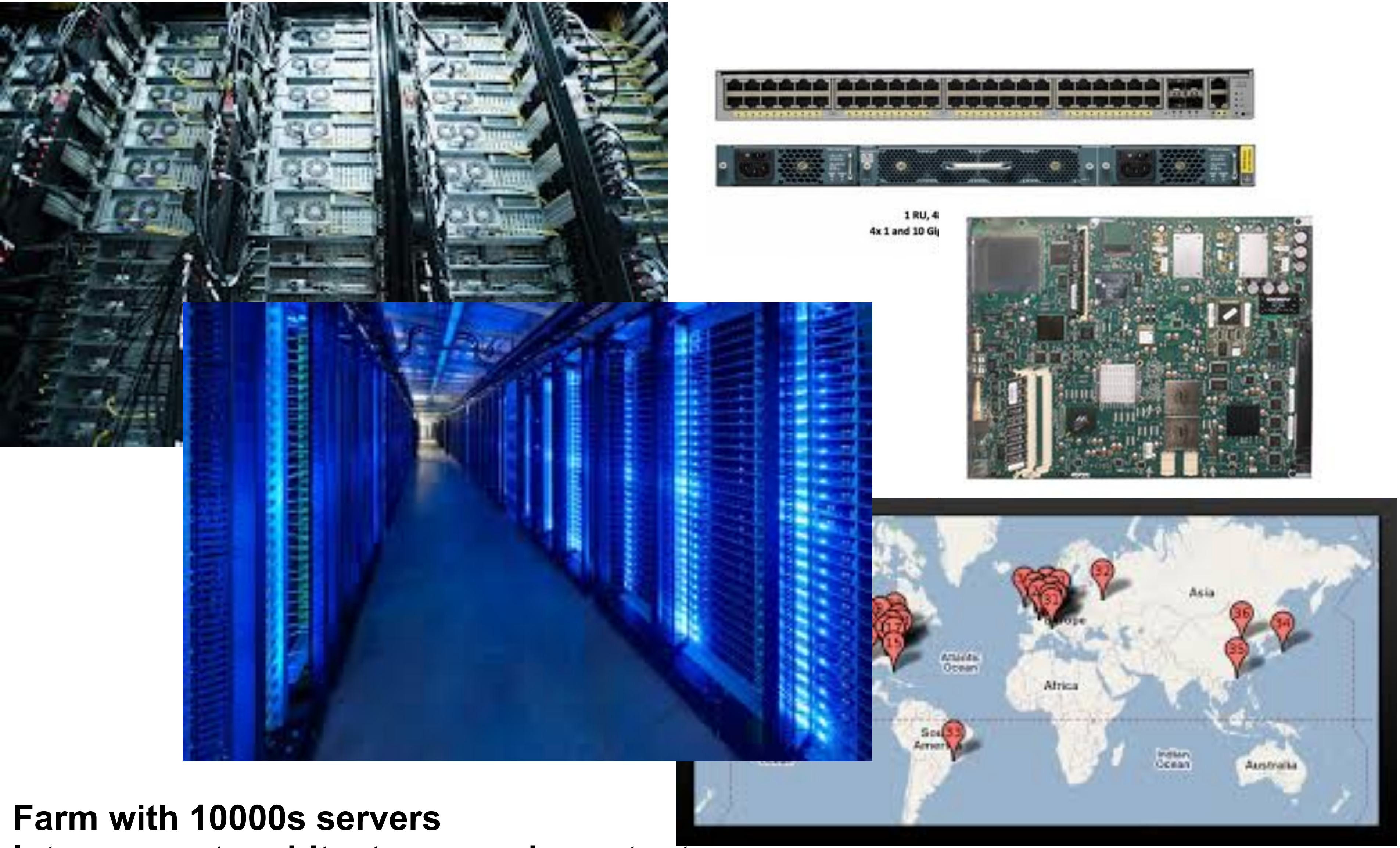


- 🔴 Apple A6X Processor
- 🟠 Hynix H2JTDG8UD2MBR 16 GB NAND Flash
- 🟡 Apple 338S1116 Cirrus Logic Audio Codec
- 🟢 343S0622-A1 Dialog Semi PMIC
- 🔵 Apple 338S1077 Cirrus Logic Class D Amplifier
- 🟣 QVP TI 261 A9P2

2 x 4Gb Elpida LP DDR2

Desktop/Cloud Server

Racks with Servers: Data Center



Farm with 10000s servers
Interconnect architecture very important

Στόχοι Προχωρημένης Αρχιτεκτονικής Υπολογιστών (605) και Προχωρημένης Παράλληλης Επεξεργασία (655) (221,325,420)

- Να γράφετε προγράμματα με καλύτερη απόδοση
- Να αναλύετε που πάει ο χρόνος κατά την διάρκεια της εκτέλεσης ενός προγράμματος
- Τις τάσεις τεχνολογίας υπολογιστών
  - Περιορισμοί και δυνατότητες
- Να μάθετε τις αρχιτεκτονικές και τις τεχνικές που χρησιμοποιούνται για βελτίωση της επίδοσης των μοντέρνων υπολογιστικών συστημάτων και πως αλληλοεπιδρούν με το λογισμικό και τα δεδομένα
  - speculation, hyperthreading, multi-cores, prefetching, turbo-mode, DVFS, accelerators (SIMD, GPUs, Deep-Neural-Nets)
  - Sensors, tablets, smartphones, laptops, servers, data centers, IoT
- Εμπειρία στον σχεδιασμό συστημάτων
- Παράλληλος Προγραμματισμός
- Γνώση εργαλείων για ανάλυση επίδοσης προγραμμάτων

Πληροφορίες για το ΕΠΛ605

- Διδάσκων: Γιάννος Σαζεϊδης
- Διαλέξεις: Τρίτη - Παρασκευή 16:30-18:00 (Online)
- Φροντιστήριο: Παρασκευή 18:00-19:00 (Online)
- Βοηθός Διδασκαλίας: Παναγιώτα Νικολάου
- Εργαστηριο: Τετάρτη 16:30-18:00 (Online)
- www.cs.ucy.ac.cy/courses/EPL605

- Βιβλιογραφία: Computer Architecture: A Quantitative Approach, Henessy & Patterson 6th edition
- Επιλεγμένα Άρθρα

Πληροφορίες για το ΕΠΛ605

- Προαπαιτούμενα
  - Οργάνωση Υπολογιστών και Συμβολικός Προγραμματισμός (ΕΠΛ221)
    - Αρχιτεκτονική συνόλου εντολών (ISA): πχ MIPS, x86 κτλ
    - Διάδρομος δεδομένων (datapath)
    - Μονάδα ελέγχου (control unit)
    - Σχεδιασμός ενός ψηφιακού συστήματος/απλού επεξεργαστή
    - Τα βασικά σε σχέση με κρυφή μνήμη (caches) και διασωλήνωση (pipelining)
    - Διαχείριση Μνήμης (εικονίκη/πραγματική μνήμη, caches)
  - Καλή Γνώση C και UNIX

Πληροφορίες για ΕΠΛ605

- Εργαστήριο:
  - εισαγωγή εργαλείων (πχ scripting, pin, perf,simulators…)
  - αξιολόγηση εργασιών
  - συζητήσεις, παρουσιάσεις
- Αξιολόγηση

– Εργασίες (4-5)	15-20 %
– Περίληψη/παρουσίαση άρθρων/θεμάτων	12%
– Συμμετοχή	10%
– Τελική Εργασία (project)	15-20%
– Τελική Εξέταση	40-45%

- Για επιτυχία στο μάθημα τελική εξέταση ≥ 50%

Reading Summaries

- Expect to read 25-30 papers (2-3 every week)
- Some are presentations or videos

- 3 παράγραφοι
  - Σύντομη περίληψη για το πρόβλημα, στόχο, συνεισφορά
  - Περίληψη της μεθοδολογίας και κυριότερων αποτελεσμάτων (εάν υπάρχουν)
  - Την γνώμη σας για δυνατά σημεία και αδυναμίες
- 400 λέξεις το πολύ για κάθε περίληψη

- Need to submit before the beginning of the 1st lecture every week (blackboard or teams I will announce it)
- Βαθμολόγηση: 0-3 (όχι υποβολή 0, μέτρια 1, ικανοποιητική 2, εξαιρετική 3)

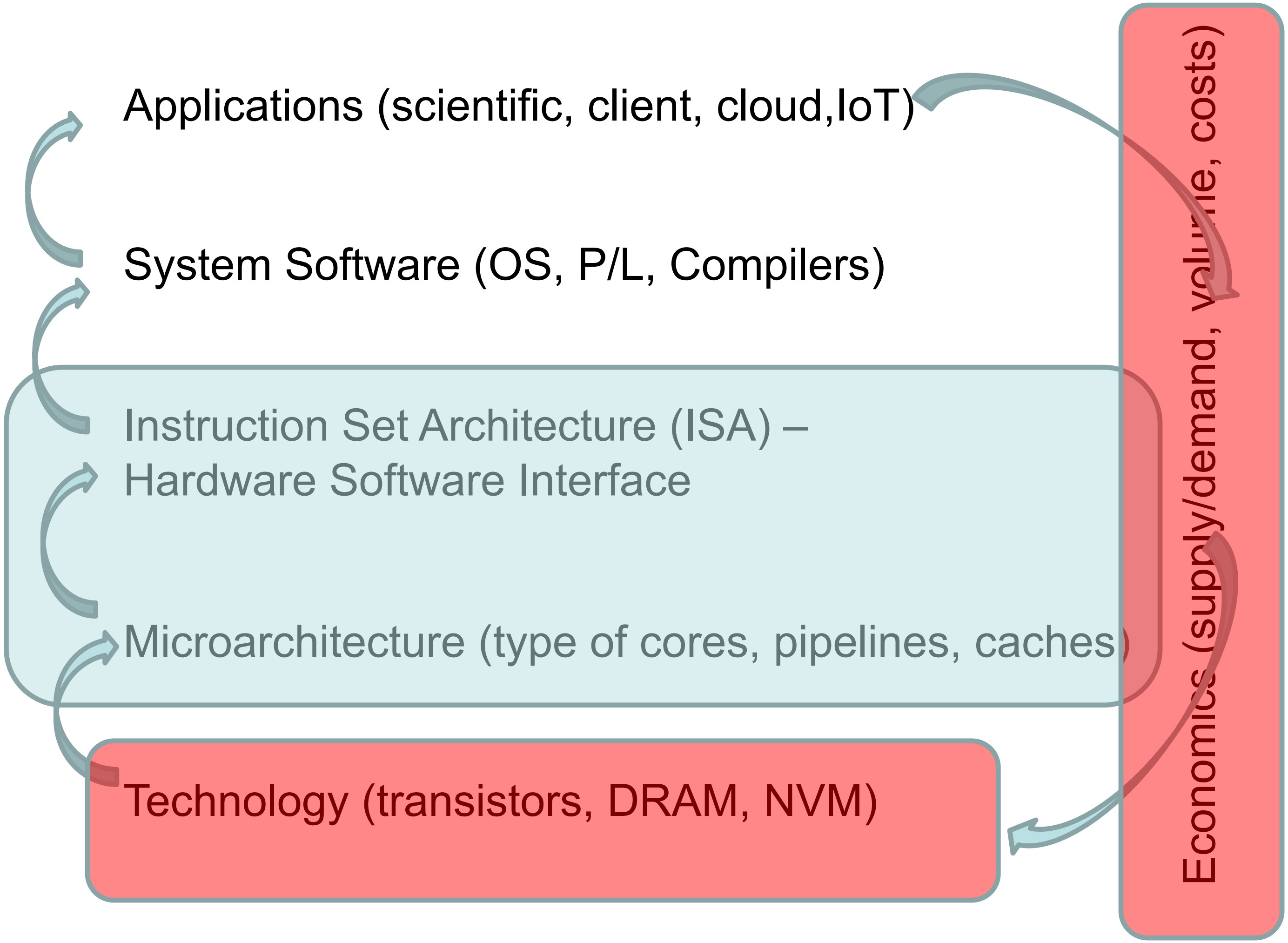
Projects

- A survey in an area of interest
  - Read technical papers >10
  - Other info on the web
  - TPUs, ML, GPUs, Optane Memory, AI for architecture
- Experimental
  - Confirm something in a published work
  - Extend an idea in a paper
  - Initial evaluation of an idea using simulator or performance counters
- Report and Presentation

Readings for Summary

- Moore’s Law
  - Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114.
- Dennard Scaling
  - A 30 Year Retrospective on Dennard’s MOSFET Scaling Paper
- Power Wall
  - Power: A First-Class Architectural Design Constraint

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

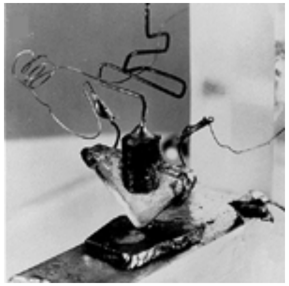
Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

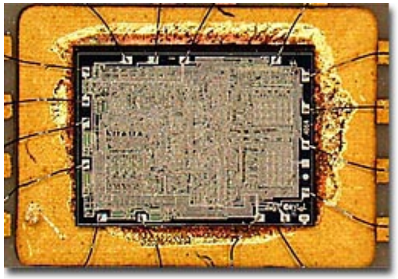
Economics (supply/demand, volume, costs)

Βασική Τεχνολογία Υλοποίησης

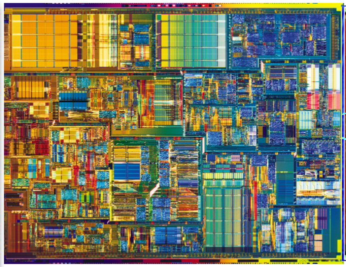
- Processor logic and memory arrays implemented with silicon based transistors
- DRAM with transistors and capacitors
- Metal wires for connectivity
- Feature Size:.,45nm,32nm,28nm,22nm,14nm,10nm, 7nm, 5nm…
- Συνέπεια σμίκρυνσης στον ίδιο χώρο περισσότερους πόρους (transistors, functional units, memory cells, cores)



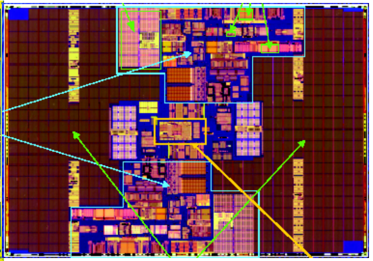
1947-10^0



1971-10^3

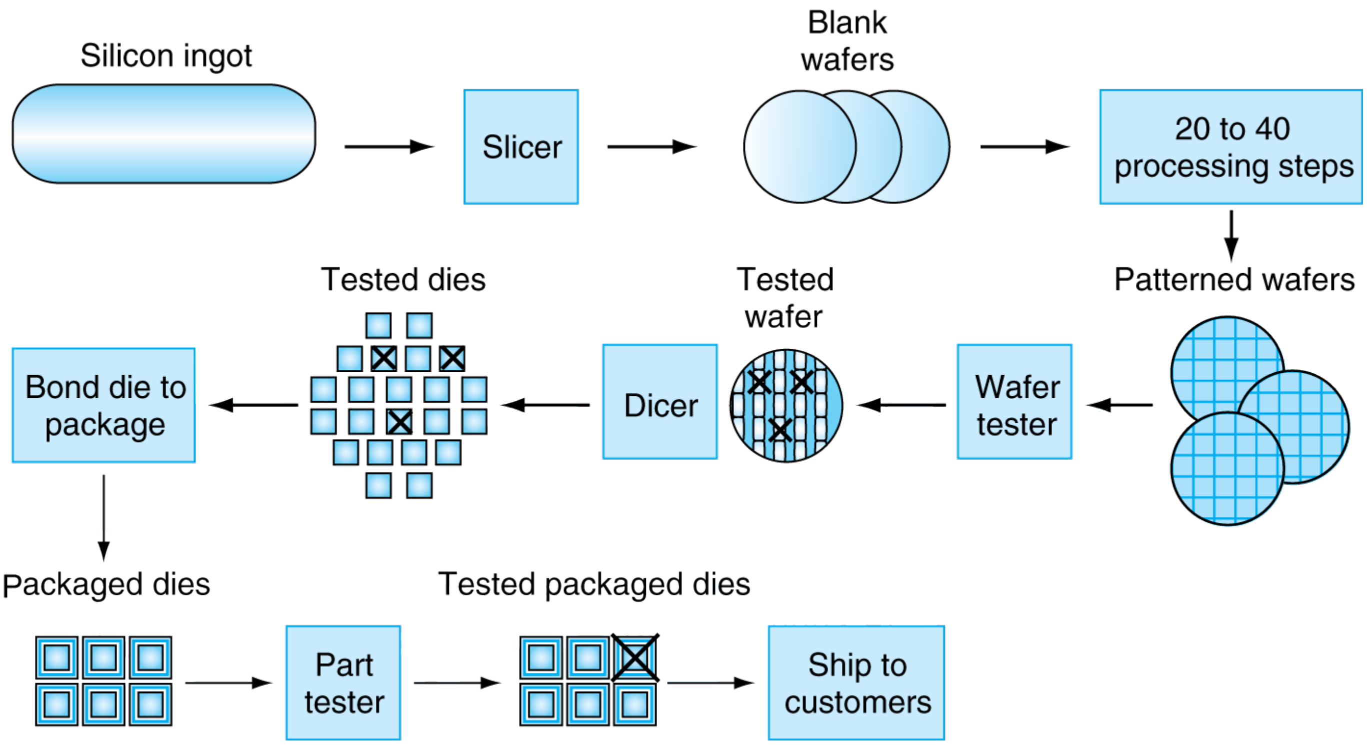


2002-10^7



2005-10^9

Manufacturing ICs



Silicon ingot

Slicer

Blank wafers

20 to 40 processing steps

Patterned wafers

Wafer tester

Tested wafer

Dicer

Tested dies

Bond die to package

Packaged dies

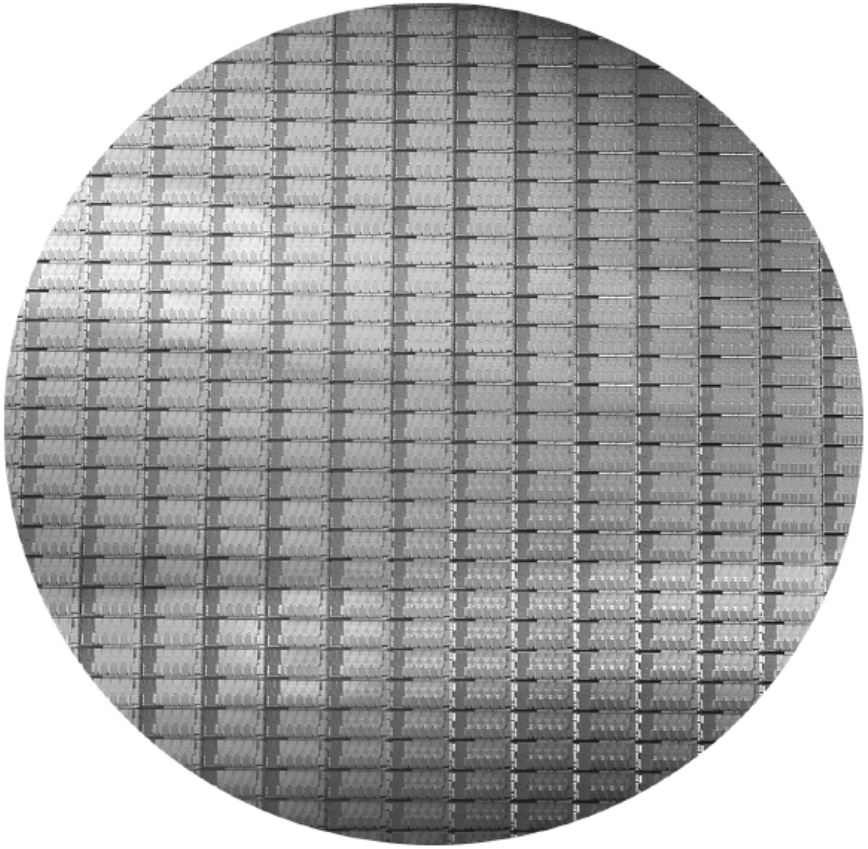
Part tester

Tested packaged dies

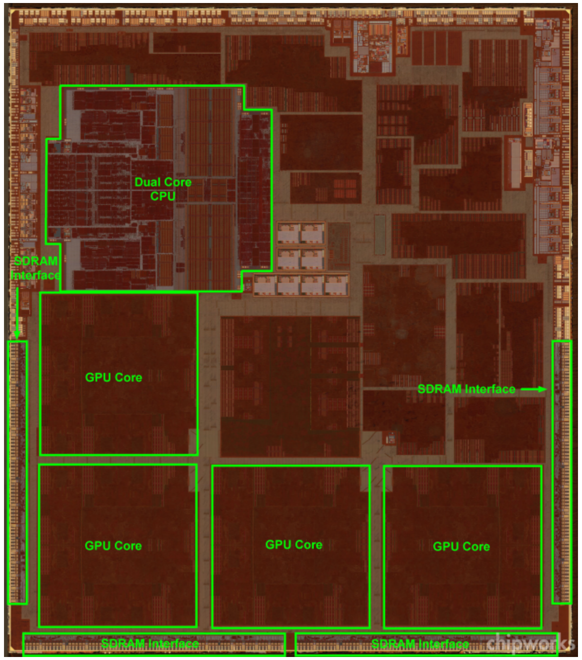
Ship to customers

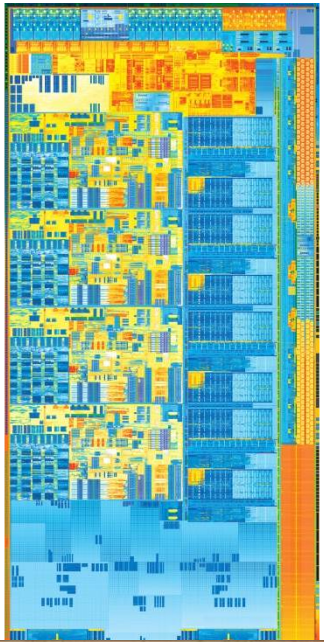
- Yield: proportion of working dies per wafer
- Bigger die (cost due to area and lower yield)

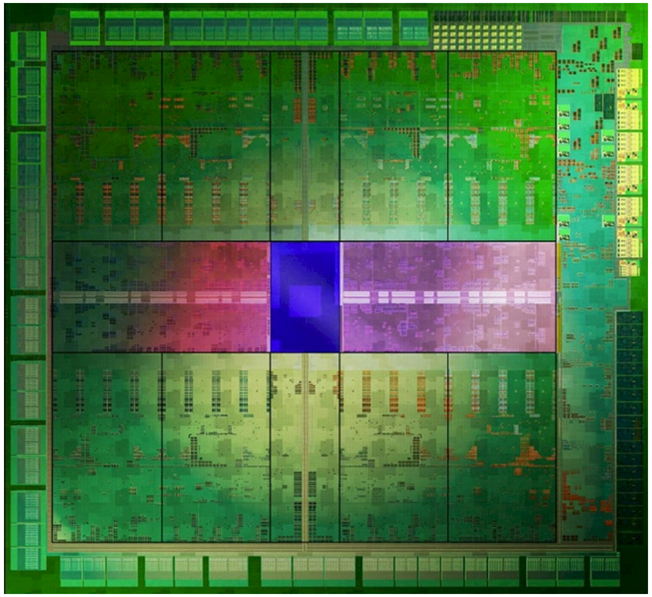
Intel Core i7 Wafer



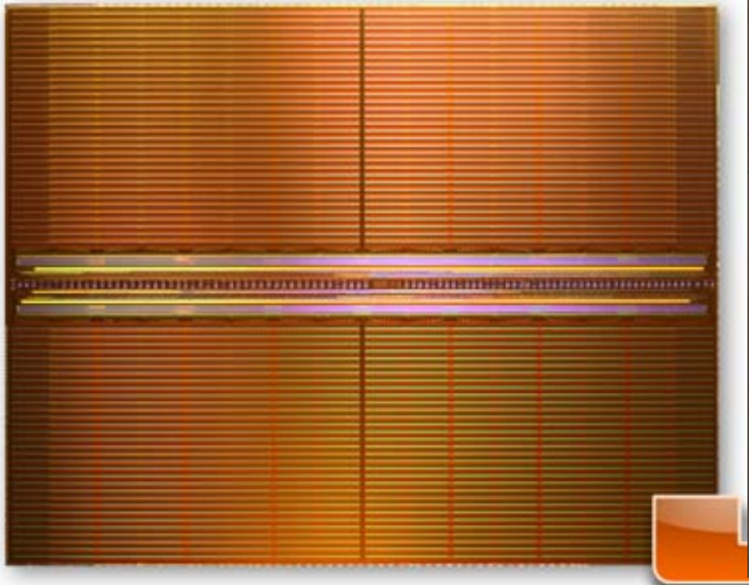
- 300mm wafer, 280 chips, 32nm technology
- Each chip is 20.7 x 10.5 mm







https://www.youtube.com/watch?v=aCOyq4YzBtY

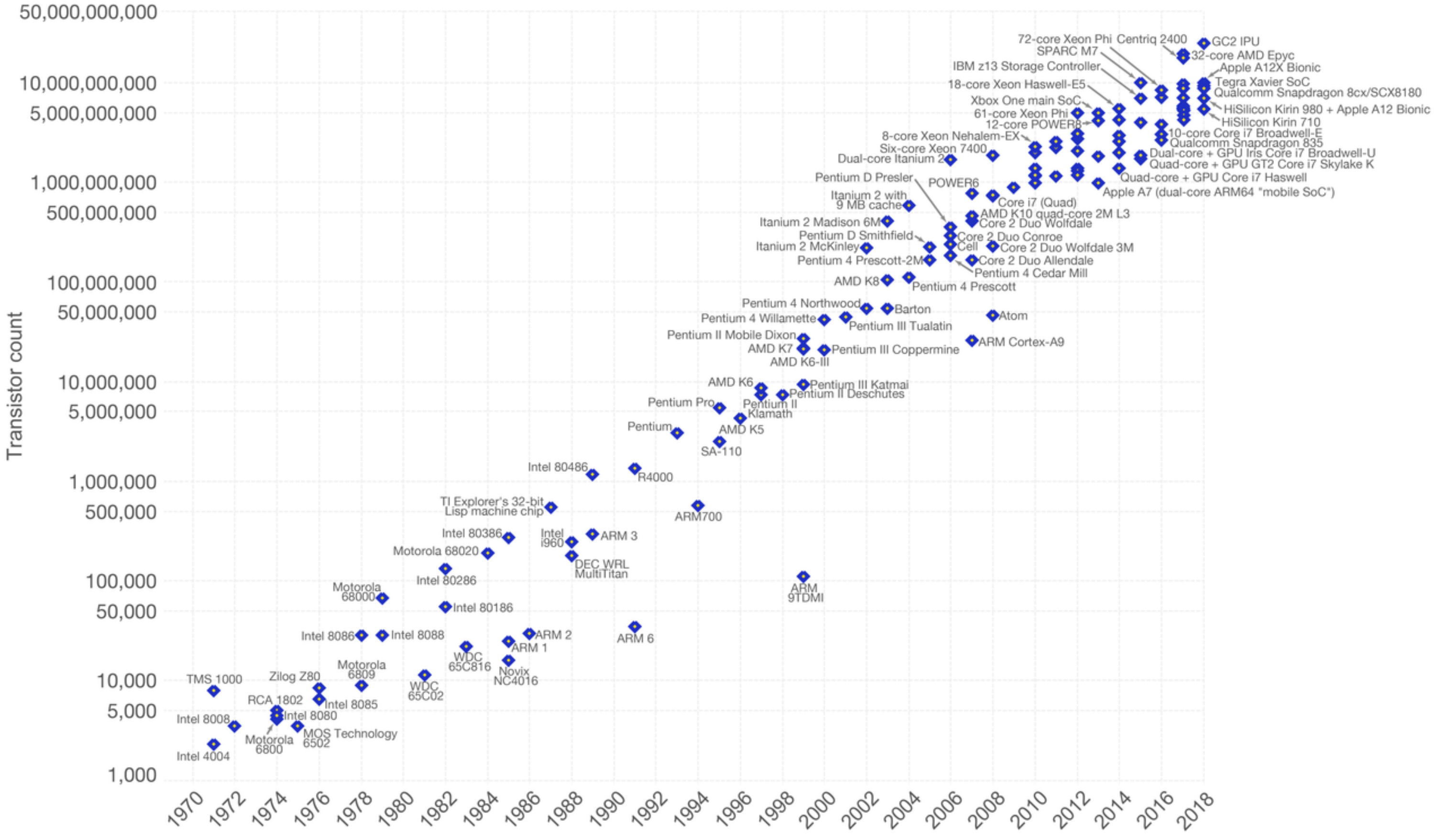


Technology Scaling: transistors getting smaller (by 2x every 2-3yrs)

Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.

Our World in Data



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Trends in Technology: Growth

- Integrated circuit technology (slowing)
  - Transistor density: 35%/year (Moore's law)
  - Die size: 10-20%/year
  - Integration overall: 40-55%/year

Amazing Density Trends but slowdown

- DRAM capacity: 25-40%/year (slowing)

NVM (e.g., Optane)

Foundries <u>≤</u>10nm:
Intel
TSMC
Samsung

- Flash capacity: 50-60%/year
  - 8-10X cheaper/bit than DRAM
- Magnetic disk technology: 40%/year (slowing)
  - 8-10X cheaper/bit than Flash
  - 200-300X cheaper/bit than DRAM
- Network Technology (routers and links)

Basic Technology Economics

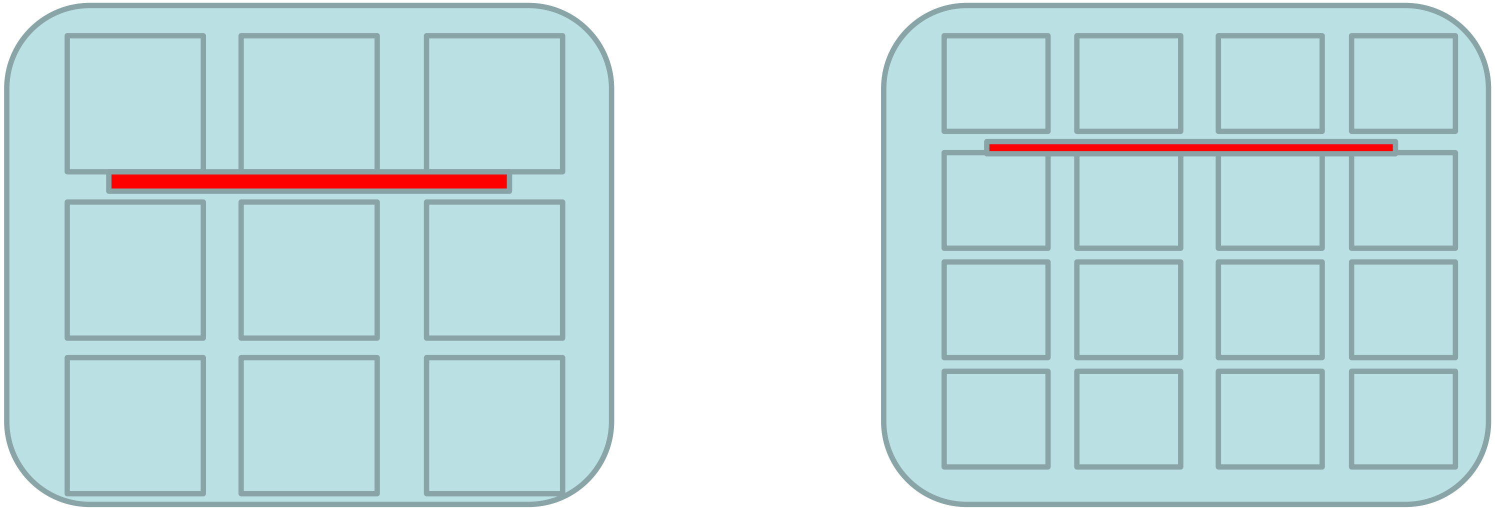
- Making feature size smaller is very expensive: multi-billion cost for a factory (Intel, TSMC, Global Foundries)
- Such cost is affordable if it can be amortized in few years with millions/billions of IC sales
- Larger volume helps lower non-recurring expenses (NRE) cost paid by each customer

E.g. $5billion new factory, expect to sell 20million chips/year
NRE Cost/chip for 5 yrs => 5e9/(5x20e6) = $50

If yield is 50% => (5e9/(5x20e6))/yield = $100
(i.e., you manufacture 2xC dies to be able to cell C dies)

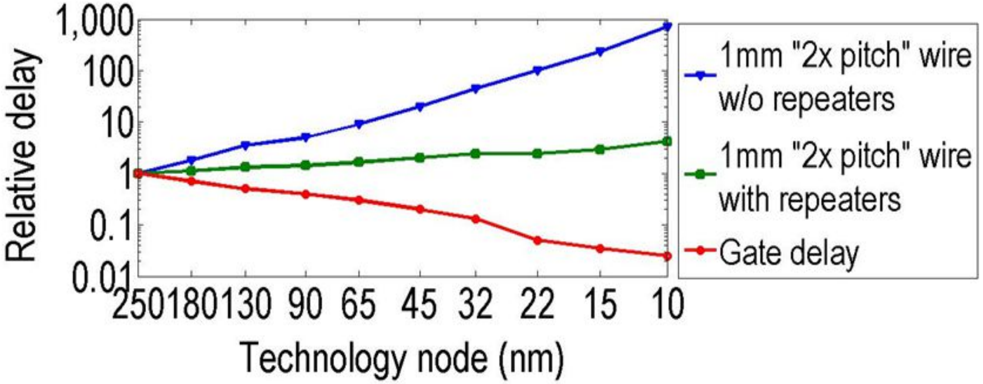
- Key: need combinations of reasonable costs and high volume and good yield technology

Transistor vs Wire Scaling



Transistors and Wires

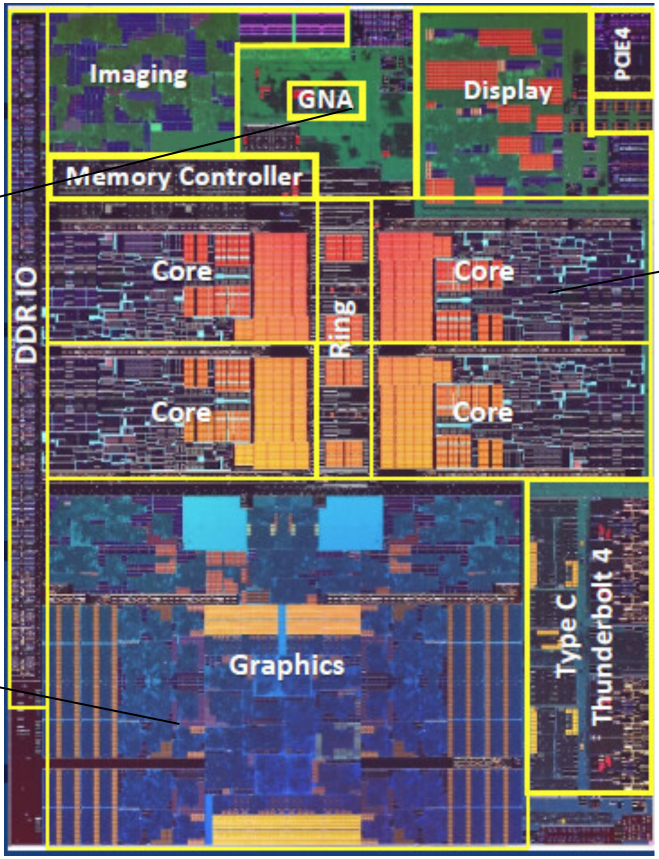
- Feature size
  - Minimum size of transistor or wire in x or y dimension
  - 10 microns in 1971 to 0.005 micron (5nm) in 2021 (10/0.005~ 2000x reduction in 1D and ~4000000x reduction in 2D)
  - Scaling of speed of transistors vs wires not-uniform



- ***Non-uniform scaling supports localized activity, caching, and communication***

System on Chip (SoC)

- A single system contain multiple type of cores

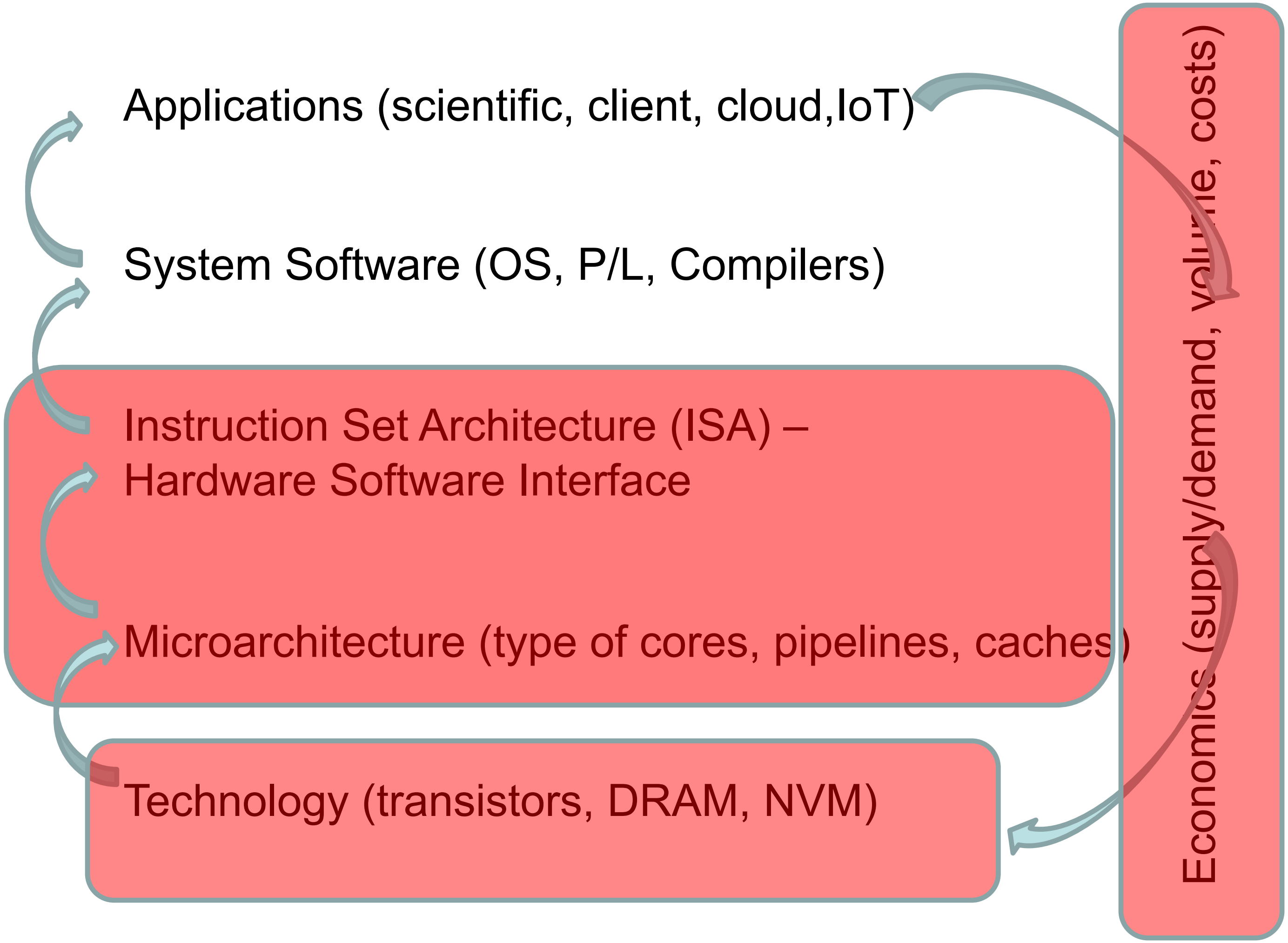


ML Accelerator

CPU cores

Graphics cores (igpu, Integrated GPU)

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

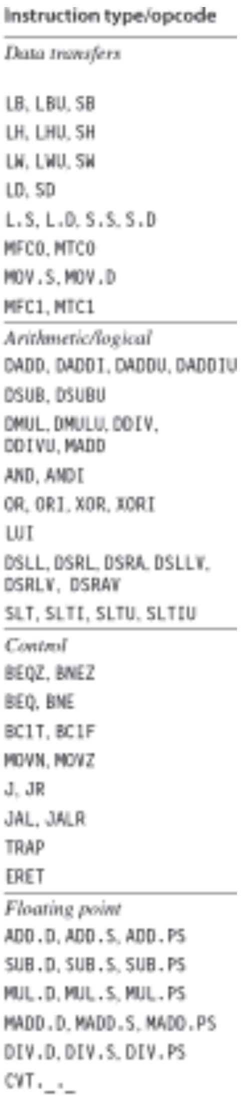
Economics (supply/demand, volume, costs)

What do we do the transistors?

- Computer architects job
    - Pipelines, caches, number of cores, core types, interconnect
- Also define instructions to interface to the hardware resources and manage them
- Architecting system: processors, processor types, memory architecture, interconnections, storage
- Considering goals and constraints: performance, power, temperature, power delivery, battery life, reliability, form factor, cost

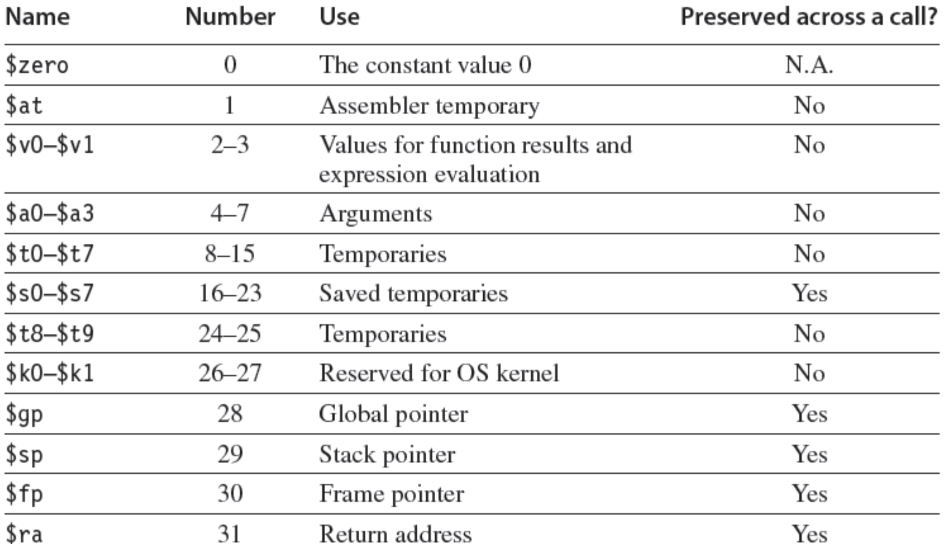
ΑΣΕ (ISA): η διασύνδεση στο υλικό

operations



Instruction type/opcode
Data transfers
LB, LBU, SB
LH, LHU, SH
LW, LWU, SW
LD, SD
L.S, L.D, S.S, S.D
MFC0, MTC0
MOV.S, MOV.D
MFC1, MTC1
Arithmetic/logical
DADD, DADDI, DADDU, DADDIU
DSUB, DSUBU
DMUL, DMULU, DDIV, DDIVU, MADD
AND, ANDI
OR, ORI, XOR, XORI
LUI
DSLL, DSRL, DSRA, DSLLV, DSRLV, DSRAV
SLT, SLTI, SLTU, SLTIU
Control
BEQZ, BNEZ
BEQ, BNE
BC1T, BC1F
MOVN, MOVZ
J, JR
JAL, JALR
TRAP
ERET
Floating point
ADD.D, ADD.S, ADD.PS
SUB.D, SUB.S, SUB.PS
MUL.D, MUL.S, MUL.PS
MADD.D, MADD.S, MADD.PS
DIV.D, DIV.S, DIV.PS
CVT._._

registers



Name	Number	Use	Preserved across a call?
$zero	0	The constant value 0	N.A.
$at	1	Assembler temporary	No
$v0–$v1	2–3	Values for function results and expression evaluation	No
$a0–$a3	4–7	Arguments	No
$t0–$t7	8–15	Temporaries	No
$s0–$s7	16–23	Saved temporaries	Yes
$t8–$t9	24–25	Temporaries	No
$k0–$k1	26–27	Reserved for OS kernel	No
$gp	28	Global pointer	Yes
$sp	29	Stack pointer	Yes
$fp	30	Frame pointer	Yes
$ra	31	Return address	Yes

Encoding of operations
Other registers (FP, vector)
Other operands (immediate)
Syscalls
Address Space
Endianess
…

Trends in Computer Architecture

PAST/PRESENT:

- For several years improve performance with more Instruction Level Parallelism (ILP)
  - Pipelining
  - Out-of-order execution
  - Superscalar execution
  - Prediction and Speculative execution
  - Clock Rate Increase
- Improvements in Memory Hierarchy: capacity and access time
  - Multi-level Caches,
  - Replacement Policies
  - Prefetching
  - Different memory technology (SRAM, DRAM, NVM, Flash, Disk)
- Support for running multiple-threads in many processors

Emergence of Applications that benefit from Parallelism and Domain Specific Architectures

- Classes of architectural parallelism:
  - Instruction-Level Parallelism (ILP)
  - Multiprogram (multicores,multiprocessors)
  - Data Level Parallelism (Vector architectures, Graphic Processor Units, GPUs)
  - Thread-Level Parallelism (multi-core CPU)
  - Request-Level Parallelism (multi-core CPU)
- Domain Specific Architectures for ML/AI (TPUs)

Trends in Computer Architecture

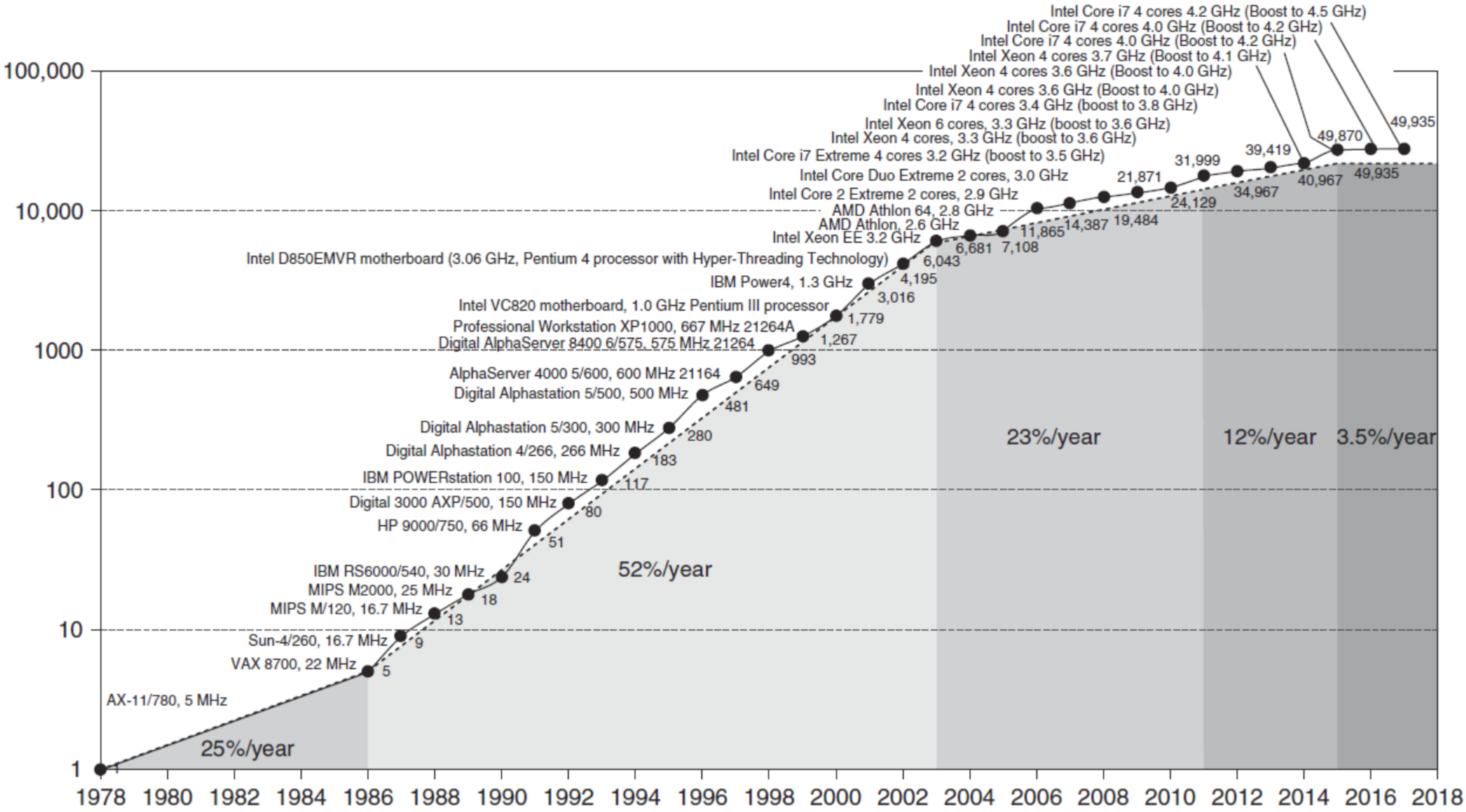
PRESENT/FUTURE:

- Continue to improve ILP and memory hierarchy
  - Single processor performance improvement key business requirement
  - technology and program properties make it challenging
  - Amdahl's Law makes it essential

- Support from HW/ISA
  - for other types of parallelism DLP, TLP, RLP
  - DSA – e.g. deep neural nets for classification

- DLP,TLP,DSA: require explicit restructuring of the applications (e.g. parallel programming 325,655) for some cases auto-vectorization/parallelization

Single Thread Performance



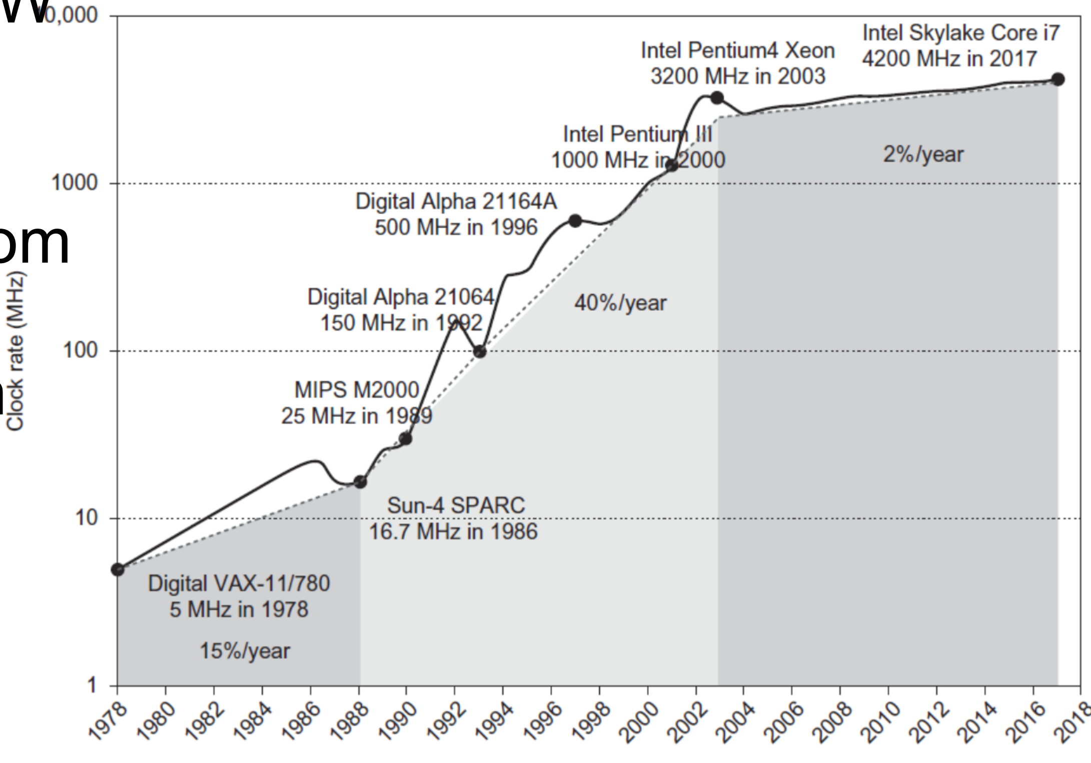
Why ST Performance improves slower?

- Difficult to improve further
  - Can’t find more cheap ILP
  - Can improve predictors/replacement policies /prefetchers (?)

- Power limited => Frequency Limited
- Power more cores for parallel execution

Power and Thermal Limitations (TDP)

- Intel 80386 consumed ~ 2 W
- 3.3 GHz Intel Core i7 consumes 130 W
- Heat must be dissipated from 1.5 x 1.5 cm chip
- This is the limit of what can be cooled by air

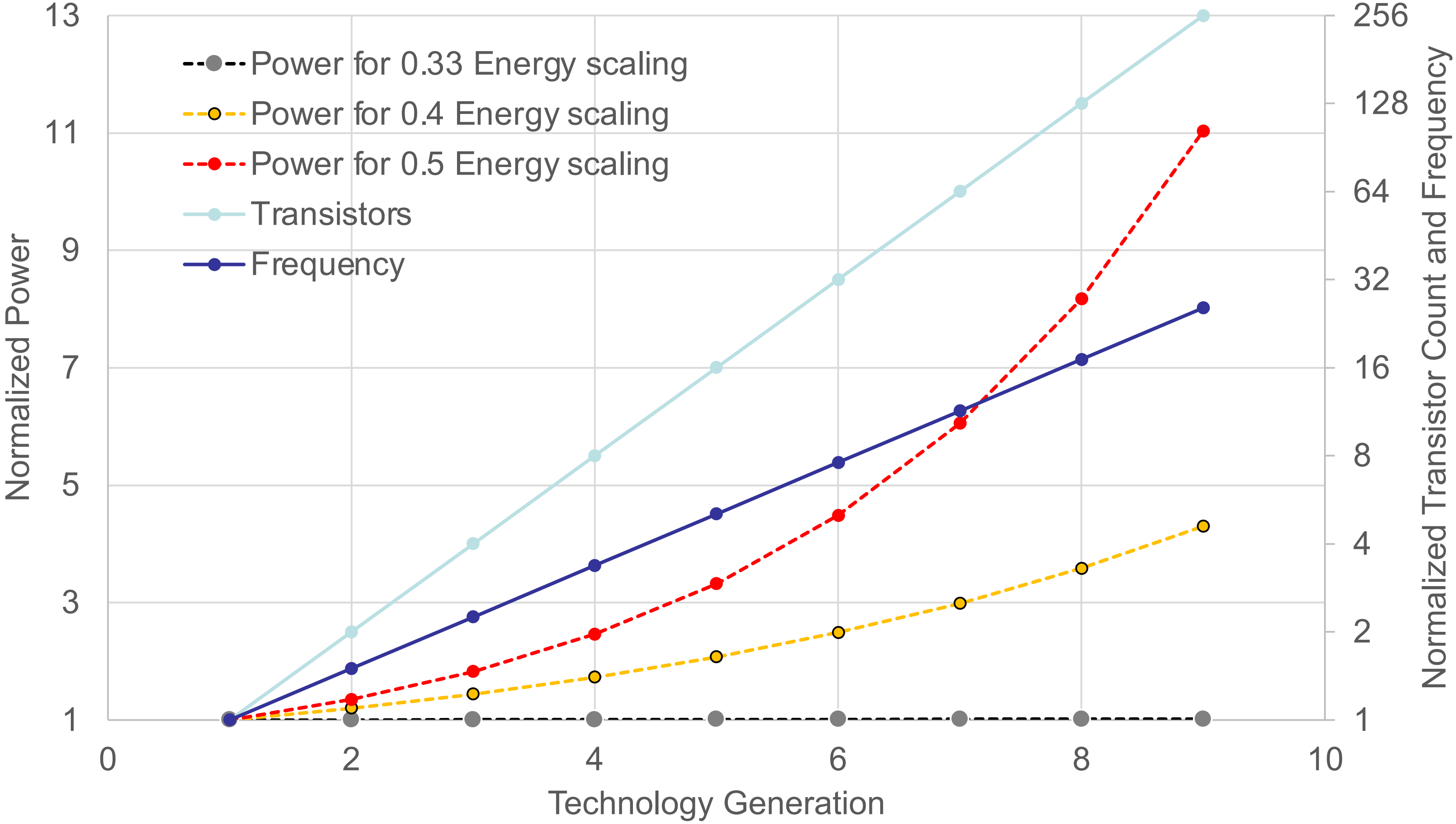


- Different systems (e.g., mobiles, laptops, desktops, server) have fixed TDP envelopes (e.g., 1W,15W,95W,150W)
  - Depends on cooling technology used

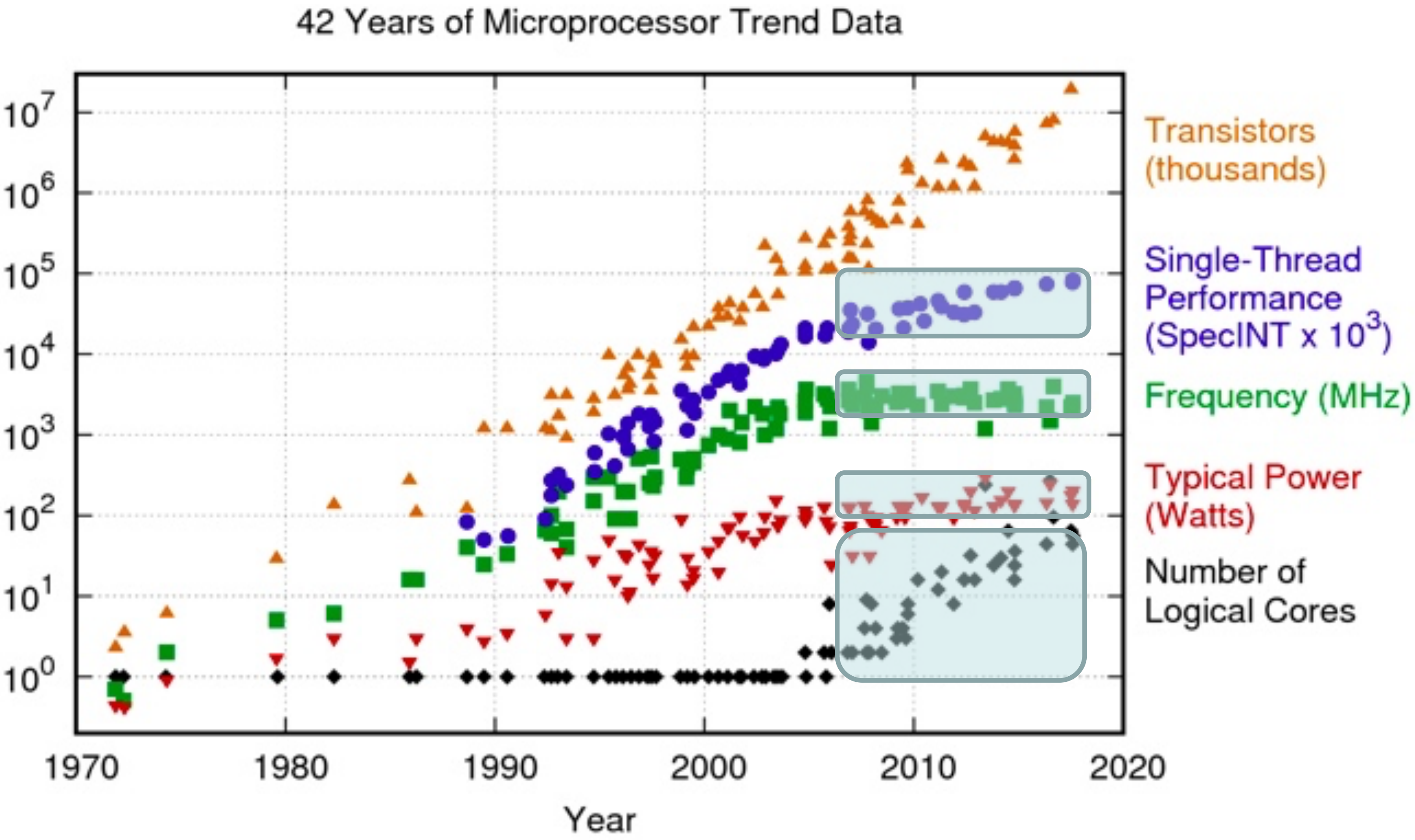
Clock, Energy, Power (simplified example)

- Power = Energy/Cycle Time => Power = Energy x F
- Transistors keep getting smaller and faster
  - Moore's "Law": doubling transistor density every 24-36 months
- Transistors burn energy
  - Smaller transistor burns less energy (ideal need 35% of previous generation)
  - if 2x transistors in next generation chip (from Moore)
  - For same power
    - new Frequency = old Power /(old Energy x 0.35 x 2)
    - New Frequency = 1.5 x old Frequency
    - Assume same fraction of transistors active and smaller transistor speed 1.5 faster
- Fact: energy per transistor is scaling slower (>35%) with smaller feature size but industry still increased frequency by 1.5
  - Supply voltage is not scaling (Dennard Scaling)
- So power kept increasing for man years…

And hit the power wall….



Microprocessor Trends



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

and different type of cores CPU, GPU, TPU

https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

Bandwidth/Density vs Latency

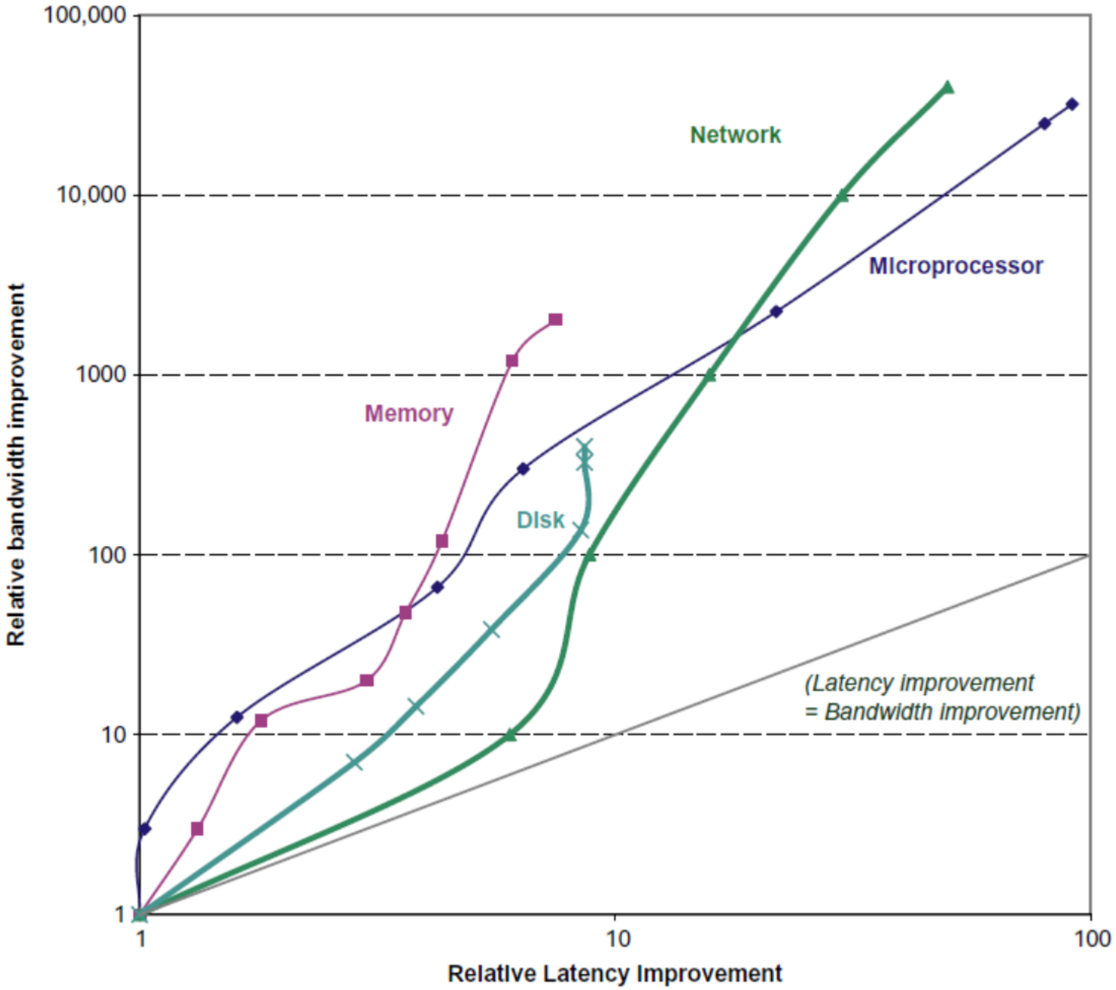
- Bandwidth or throughput – grows faster
  - Total work done in a given time
  - 10,000-25,000X improvement for processors
  - 300-1200X improvement for memory and disks

- Latency – grows slower
  - Time between start and completion of an event
  - 30-80X improvement for processors
  - 6-8X improvement for memory and disks

Trends in Technology

Bandwidth vs Latency

1000-100000
improvement



10-100
improvement

Log-log plot of bandwidth and latency milestones

Historic trend: bandwidth grows faster than latency improvement

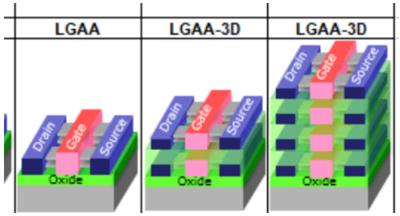
Bandwidth vs Latency

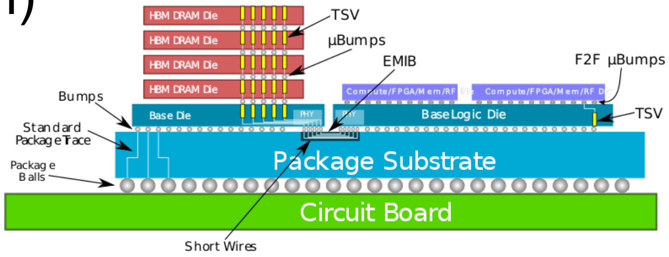
- Τι είναι πιο εύκολο: να καλύψω μια απόσταση Χ φορές πιο γρήγορα ή να την καλύψουν Χ παράλληλα

- Τι είναι πιο εύκολο: Να επιταχύνω Χ φορές μια πράξη ή να εκτελώ Χ πράξεις παράλληλα

Μέλλον του Moore’s Law

- Roadmaps Until ~2028 (7nm, 5nm,3nm,2.1nm,1.5nm…)
  - Will continue but is getting harder
  - Costs will increase (need larger volumes)
- Options
  - 3D transistors
  - 2.5D,3D stacking (chiplets,die integration)
  - Specialization
    - Accelerators
    - In-memory computation
    - ASICs
    - Human brain (spiking nets)
  - May need new technology
    - Non-cmos
    - quantum – small steps but for some class of applications
    - DNA based – storage





Security

- Architectural support to facilitate software security
  - Tainting
  - Secure enclaves to protect even from malicious OS
- Secure microarchitectures
  - Performance/power features should not leak information

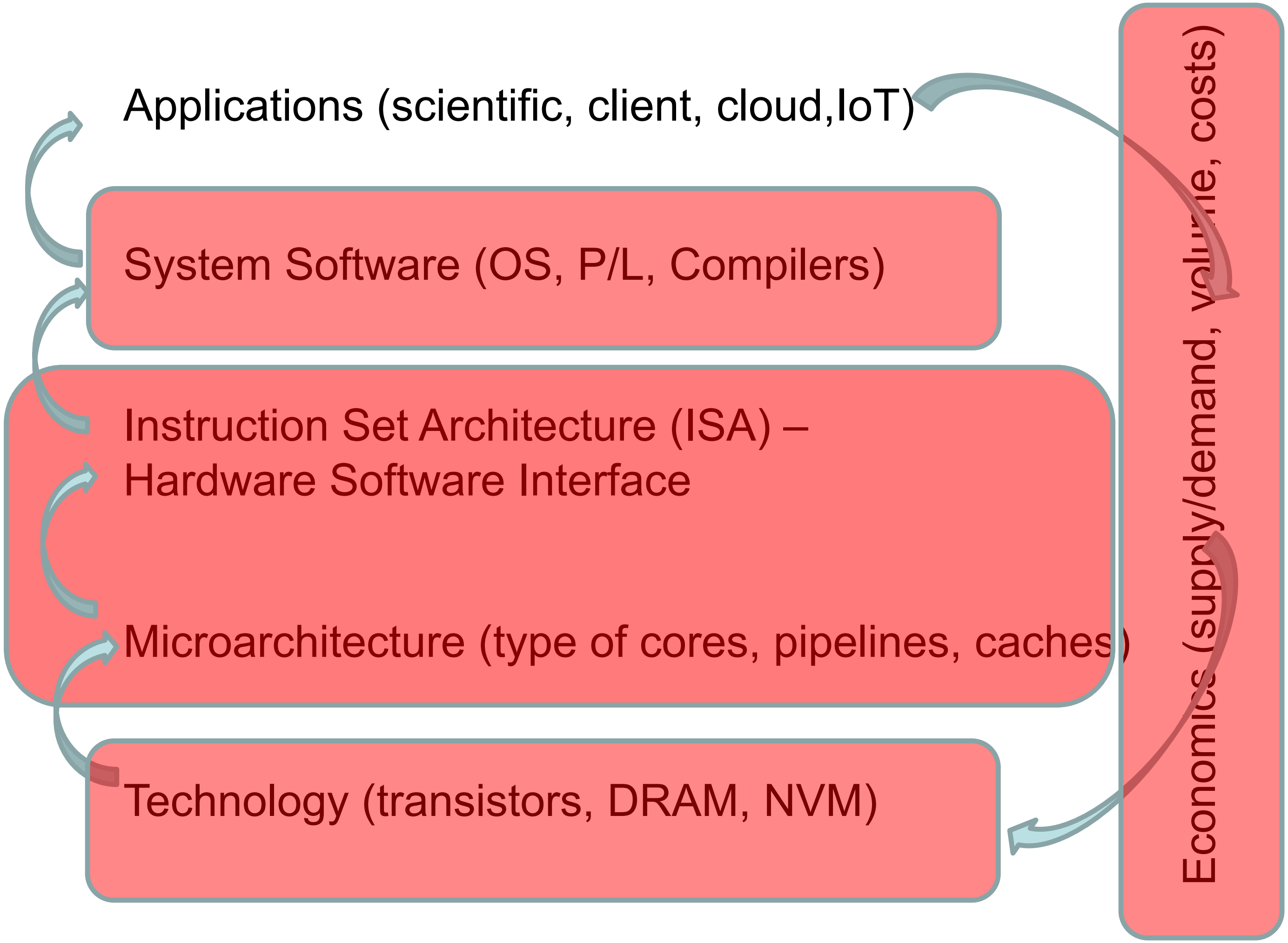
Programmability and Performance Monitoring

- Architectural features to facilitate debugging
- Counters that facilitate monitoring performance and performance bottlenecks as well as power, temperature, errors

Dependability

- Scaling helps improve performance but component reliability should also improve
- Otherwise computers will fail more often
- Failures
  - Crash (error detected than can not be recovered from)
  - Silent Data Corruption (not knowing there is error)
- Servers used for Mission Critical and Functional Safety applications employ reliability and availability features to deal with errors

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

Microarchitecture (type of cores, pipelines, caches)

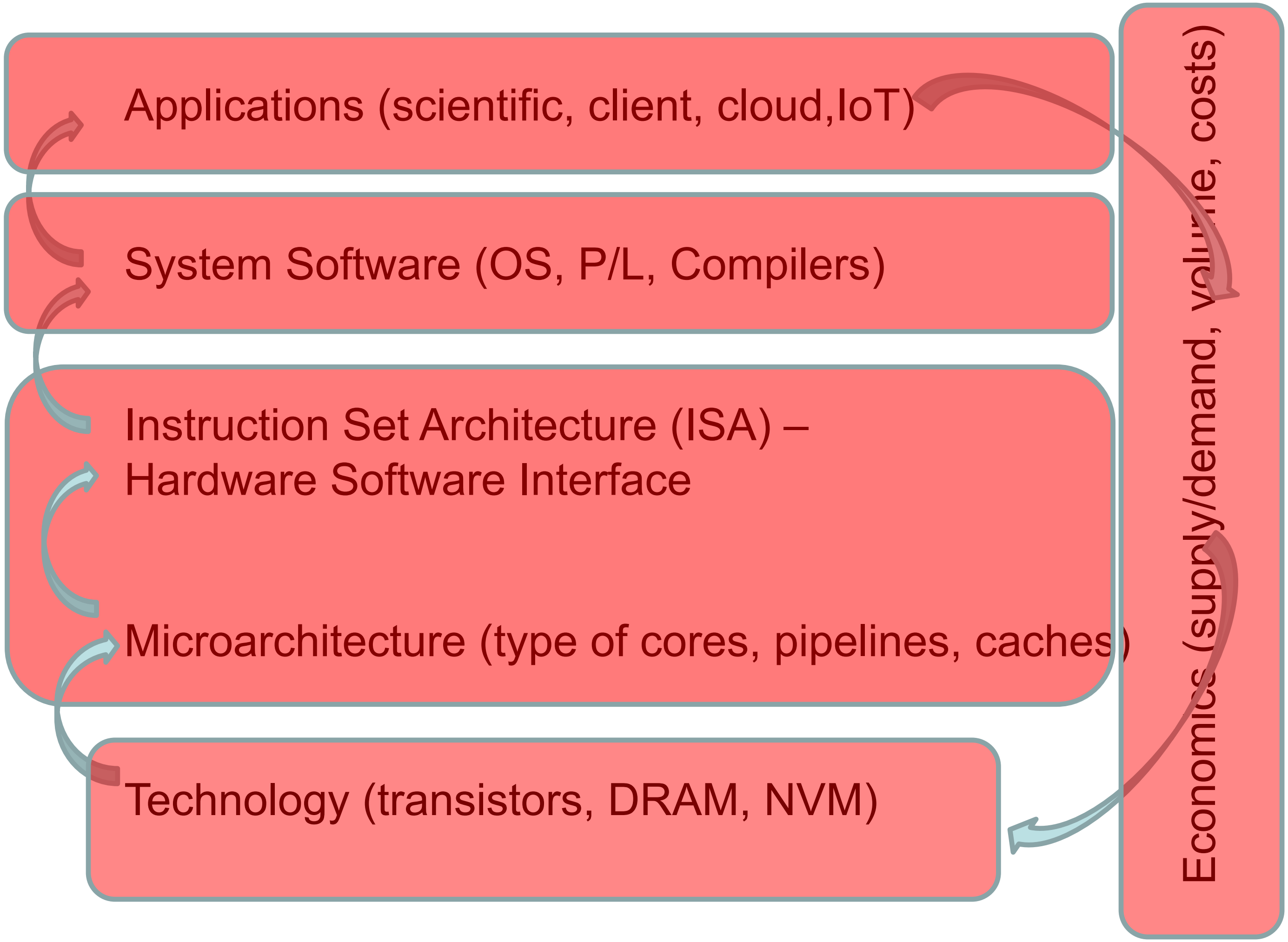
Technology (transistors, DRAM, NVM)

Economics (supply/demand, volume, costs)

Software and System Software

- Compilers key for exploiting new hardware capabilities through the ISA
  - Complex cost functions for making good choices
- Operating Systems critical for managing the various hardware resources (various cores, core types, power, memory)
- Programming Languages facilitate productivity but without hurting performance
  - Architectural features to facilitate productivity
- No security holes

What drives the improvements/trends?



Applications (scientific, client, cloud,IoT)

System Software (OS, P/L, Compilers)

Instruction Set Architecture (ISA) –
Hardware Software Interface

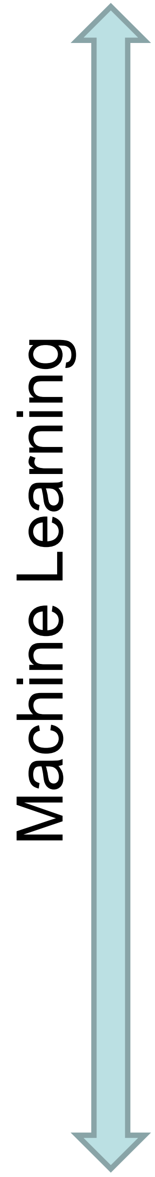
Microarchitecture (type of cores, pipelines, caches)

Technology (transistors, DRAM, NVM)

Economics (supply/demand, volume, costs)

Applications Keep Growing

- **Scientific: weather prediction, genome sequencing**
  - Need: large memory, heavy-duty floating point, availability (hw)
- **Cloud services and Social Media: amazon, google, Microsoft facebook, liknkedin,**
  - Need: throughput, quality of service (tail latency), availability (hw+sw)
- **Commercial: database/web serving,**
  - Need: data movement, high memory + I/O bandwidth, availability
- **Home office, multimedia, games**
  - Need: integer, memory bandwidth, graphics
- **Mobile apps:**
  - Need: low power, integer performance, integrated graphics, QoS
- **IoT and Embedded Applications**
  - Smart Cars (functional safety): Autonomous driving
  - Internet of Things: RFID on everything, temperature, air, water monitoring, traffic, smart houses, survailance(security)
  - Deeply Embedded: disposable “smart dust” sensors
  - Need: low power, low cost (not always) – e.g. safety applications
  - Large collection of embedded/IoT systems needs backing up from servers to do processing (data analytics), aggregation, coordination etc



Machine Learning

Applications Characteristics

Properties:

- Control Flow:
  - Irregular, Regular, Mix
- Instruction Locality
  - Temporal, Spatial, Mix
- Data Locality
  - Temporal, Spatial, Streaming, Irregular, Mix
- Data Types:
  - Integer, FP, Vectors, Matrices, Mix
- Data and Memory Dependencies between instructions
  - Dependent: not much instruction parallelism
  - Independent: instructions can execute in parallel
  - Regular, Irregular
  - Mix

Concluding Remarks

- Hardware properties challenge common sense
  - For better program performance (analysis) need to understand computer architecture
- Technology trends
  - Cost/performance is improving
    - Bandwidth more than latency
  - Due to underlying technology development
  - New applications
  - Though rate of reduction is slowing down
    - Moore’s law and Dennard’s Law
- Power limitations
  - shift to parallelism and domain specialization

What is new

- No silver bullet (perfect solution) for all applications
- Specialization and heterogeneity
  - Small cores, big cores, gpus, accelerators, neural processors etc

- Στο 605 θα μελετήσουμε τις διάφορες τάσεις και αρχιτεκτονικές για single-thread, multi-thread, data-level, domain specific accelerators…