Lecture 1
BASIC CONCEPTS

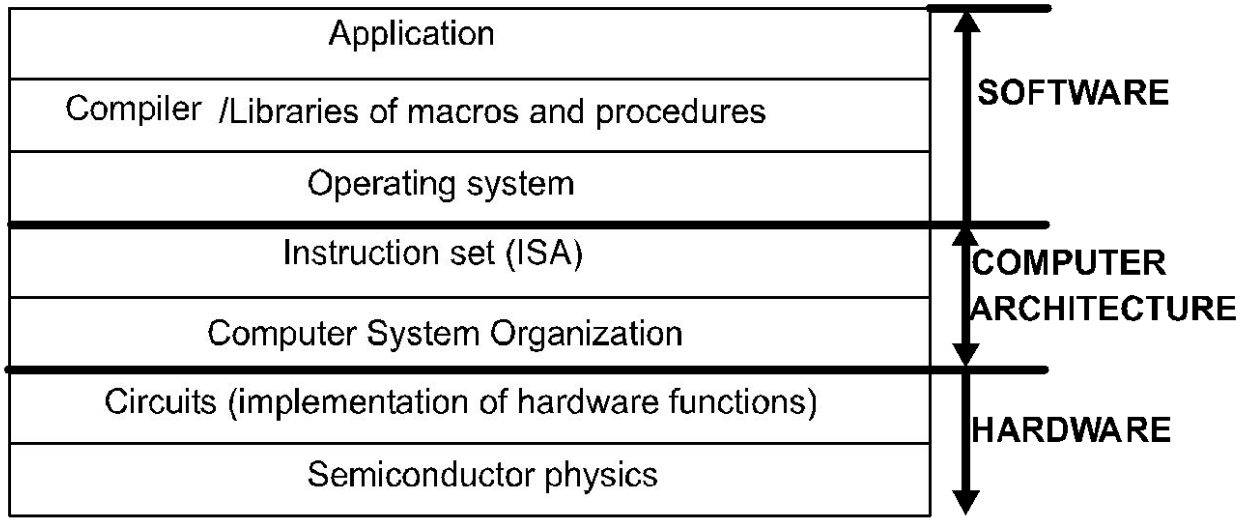
- Computer Architecture: Definition
- System components
- Technological factors and trends
- Parallelism in architectures
- Energy and Power

WHAT IS COMPUTER ARCHITECTURE?

- **Old definition:** Instruction Set Architecture (ISA)

- **Today's definition is much broader:** hardware organization of computers (how to build computer)--includes ISA

- **Layered view of computer systems**



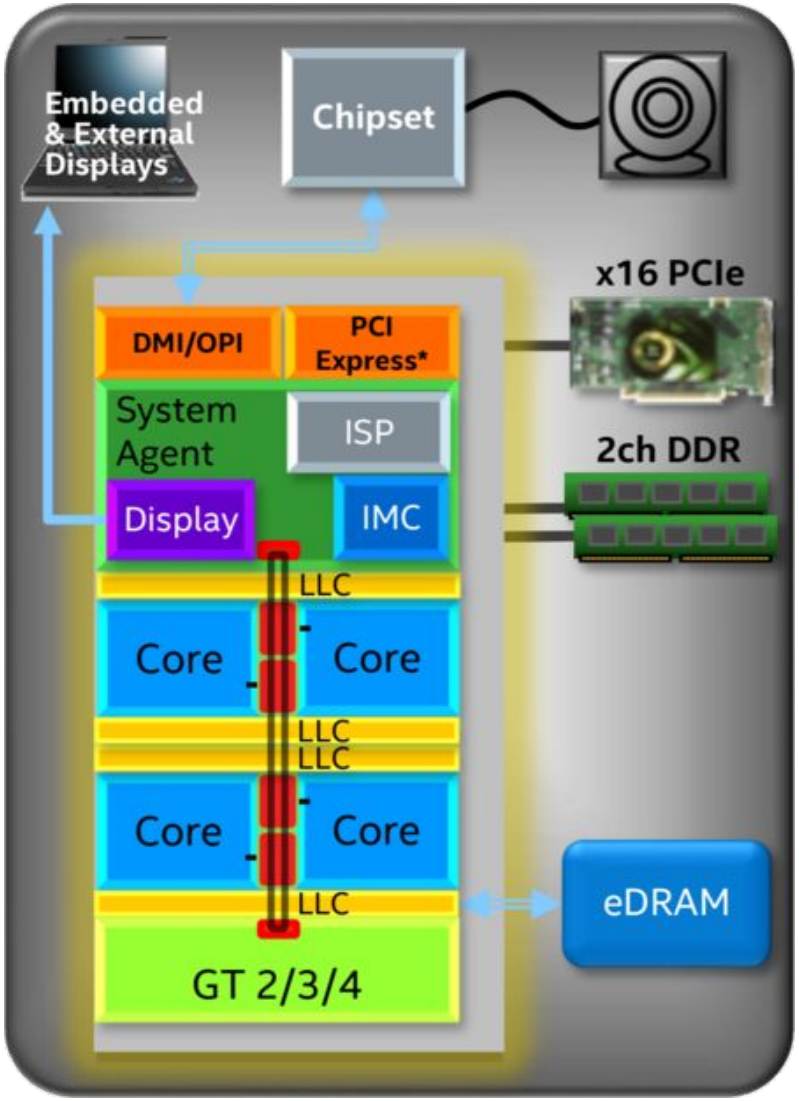
- **Role of the computer architect:**
  - To make design trade-offs across the hw/sw interface to meet functional, performance and cost requirements

SYSTEM COMPONENTS

COMPUTER ORGANIZATION

Intel Skylake System-on-Chip (SoC) (mobile/desktop)

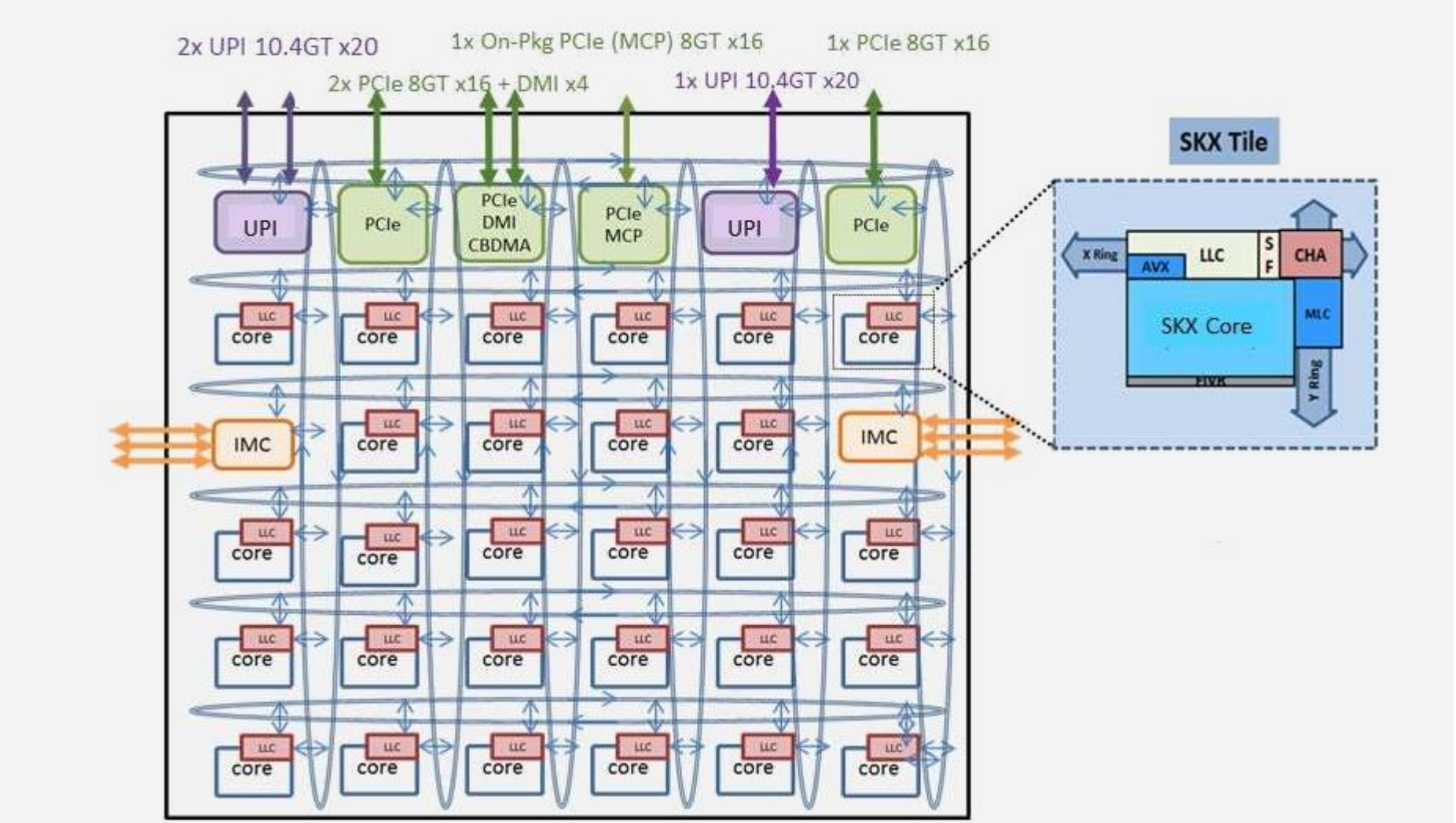
- Features:
  - CPU core
  - LLC
  - Ring interconnect
  - System agent
  - Integrated graphics
- System Agent:
  - integrated memory controller (IMC)
  - Display Controller
- PCIe, DMI (provides access to I/O)



https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(client)

COMPUTER ORGANIZATION

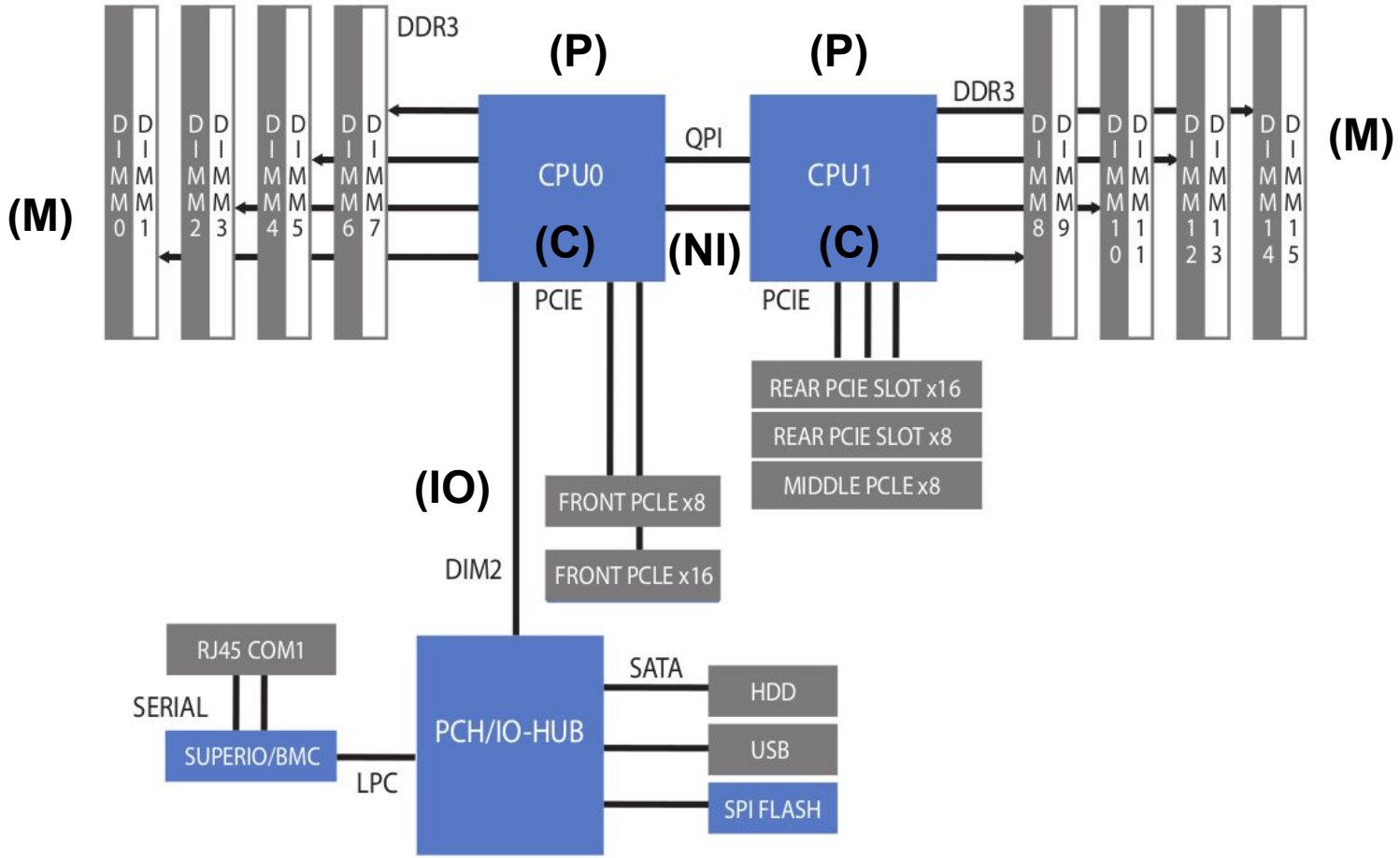
Intel Skylake XCC (Server)



More cores, no client I/Os, no integrated GPU

COMPUTER ORGANIZATION

- **A common parallel server with distributed memory**



1. **Main components:**
   a. **Processor (P)**
   b. **Memory systems (M), Cache hierarchy (C)**
   c. **I/O and Networks (NI + Interconnect)**
2. **Possibly: Accelerators (GPU, FPGA, TPU…)**

Example: IBM SUMMIT (#1 super)

Summit Overview

OpenPOWER

Compute System

10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

Compute Rack

18 Compute Servers
Warm water (70°F direct-cooled components)
RDHX for air-cooled components

39.7 TB Memory/rack
55 KW max power/rack

Compute Node

2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD

25 GB/s EDR IB- (2 ports)
512 GB DRAM- (DDR4)
96 GB HBM- (3D Stacked)
Coherent Shared Memory

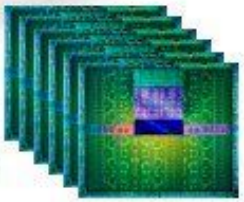
Components

IBM POWER9
- 22 Cores
- 4 Threads/core
- NVLink

GPFS File System

250 PB storage
2.5 TB/s read, 2.5 TB/s write

NVIDIA GV100
- 7 TF
- 16 GB @ 0.9 TB/s
- NVLink

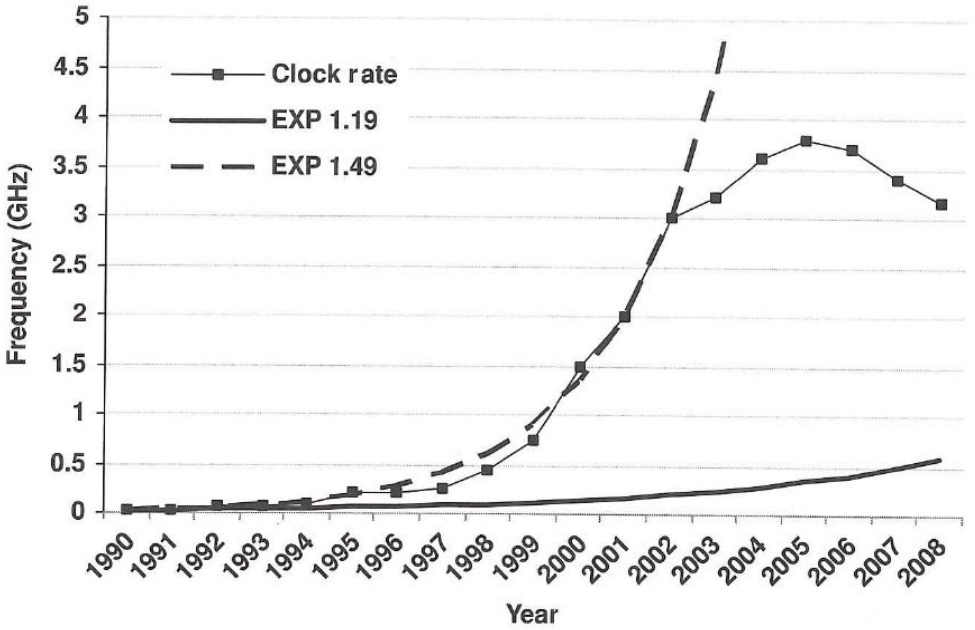


OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

TECHNOLOGICAL TRENDS

PROCESSOR ARCHITECTURE

- **Historically the clock rates of microprocessors have increased exponentially**
  - **Highest clock rate of Intel processors from 1990 to 2008**

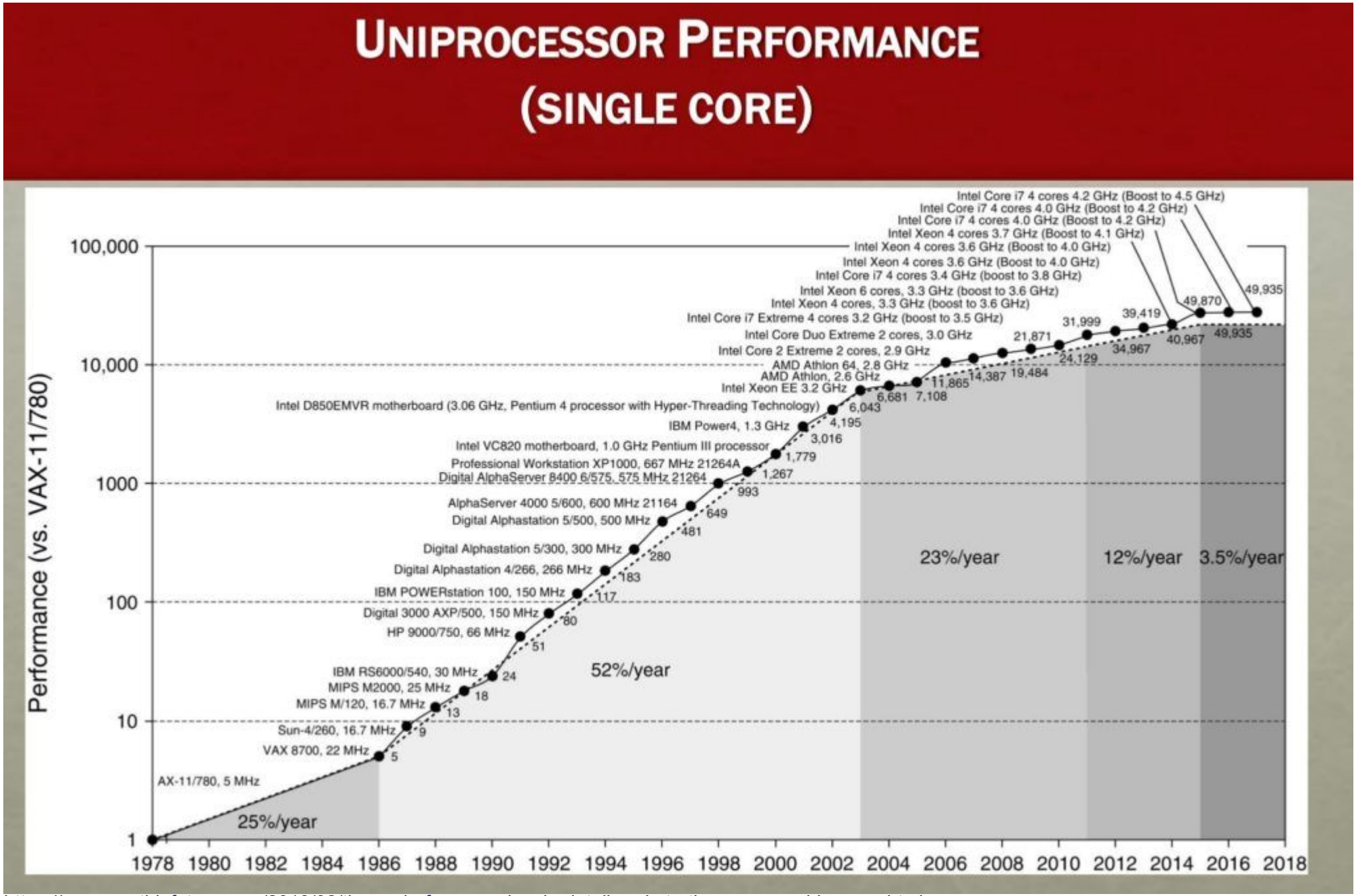


kaby lake@14nm

technology only

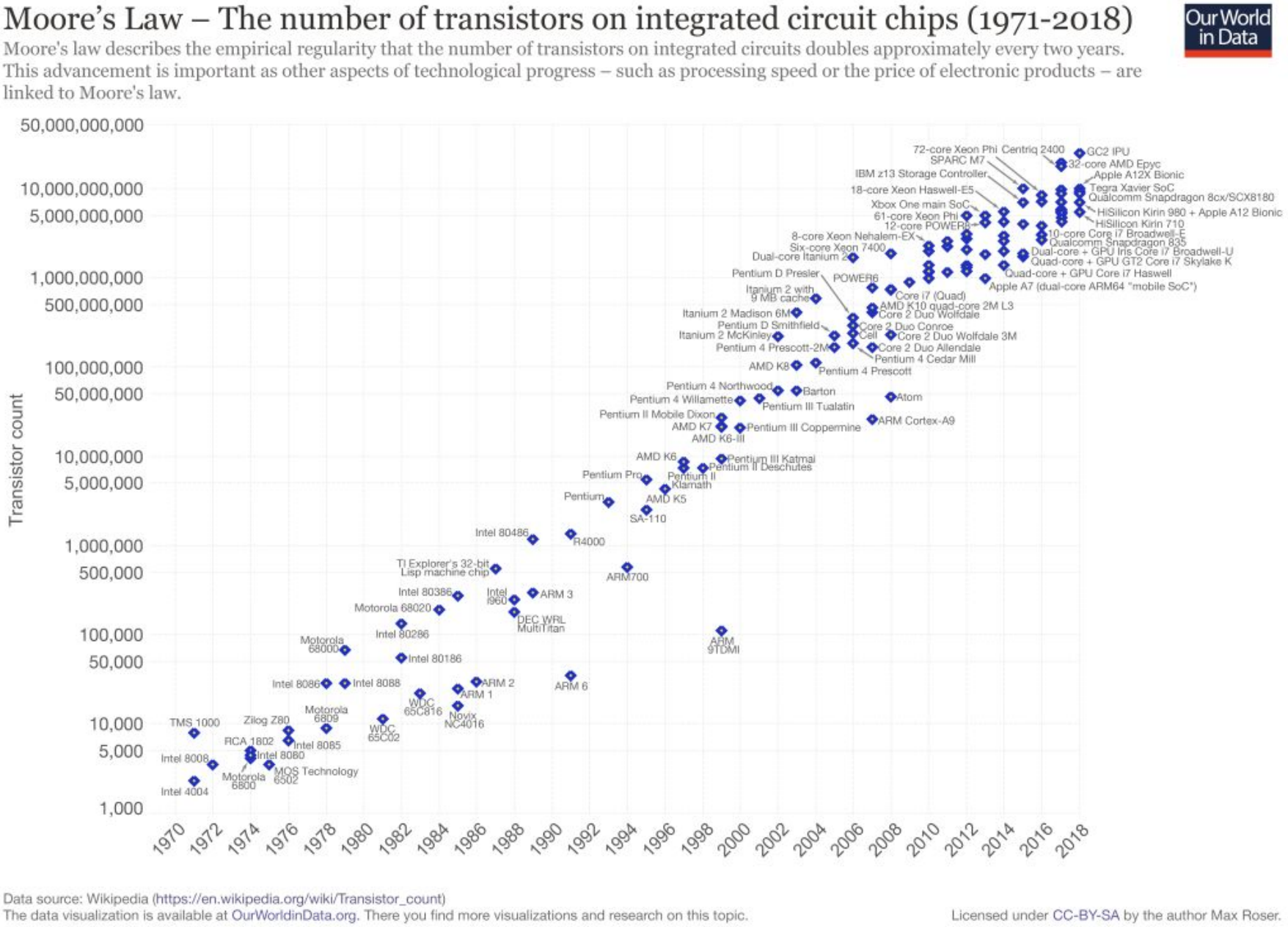
- Exp 1.19 Due To Technology (Process) Improvements
- Rest Due To:
  - Deeper Pipeline
  - Circuit Design Techniques
- **Today's Clock Rates Have Not Changed Much Since ~2005**

Performance Improvement over Time



https://www.nextbigfuture.com/2019/02/the-end-of-moores-law-in-detail-and-starting-a-new-golden-age.html

Moore's Law



10 µm – 1971
6 µm – 1974
3 µm – 1977
1.5 µm – 1981
1 µm – 1984
800 nm – 1987
600 nm – 1990
350 nm – 1993
250 nm – 1996
180 nm – 1999
130 nm – 2001
90 nm – 2003
65 nm – 2005
45 nm – 2007
32 nm – 2009
22 nm – 2012
14 nm – 2014
10 nm – 2016
7 nm – 2018
5 nm – 2020

https://en.wikipedia.org/wiki/Moore%27s_law

Number of transistors doubles every 2 years (Moore's law). 1B transistors reached in 2008. 100B in 2021??

Cost per Transistor

- The cost per transistor is not going down much with each process node.
- Doubling the number of transistors may also double the cost!
  - E.g. Apple A10 -> A11 -> A12: die size are smaller to keep costs bounded

https://community.cadence.com/cadence_blogs_8/b/breakfast-bytes/posts/linley-on-microprocessor-market

Technology shrinking

As **shrinking** becomes more **complex**, **requiring more capital**, **expertise**, and **resources**, the number of companies capable of providing leading edge fabrication has been steadily dropping. As of 2020, only three companies are now capable of fabricating integrated circuits on the most cutting edge process: **Intel**, **Samsung**, and **TSMC**.

Approaching end of Moore's law more due to economic reasons rather than technological reasons!

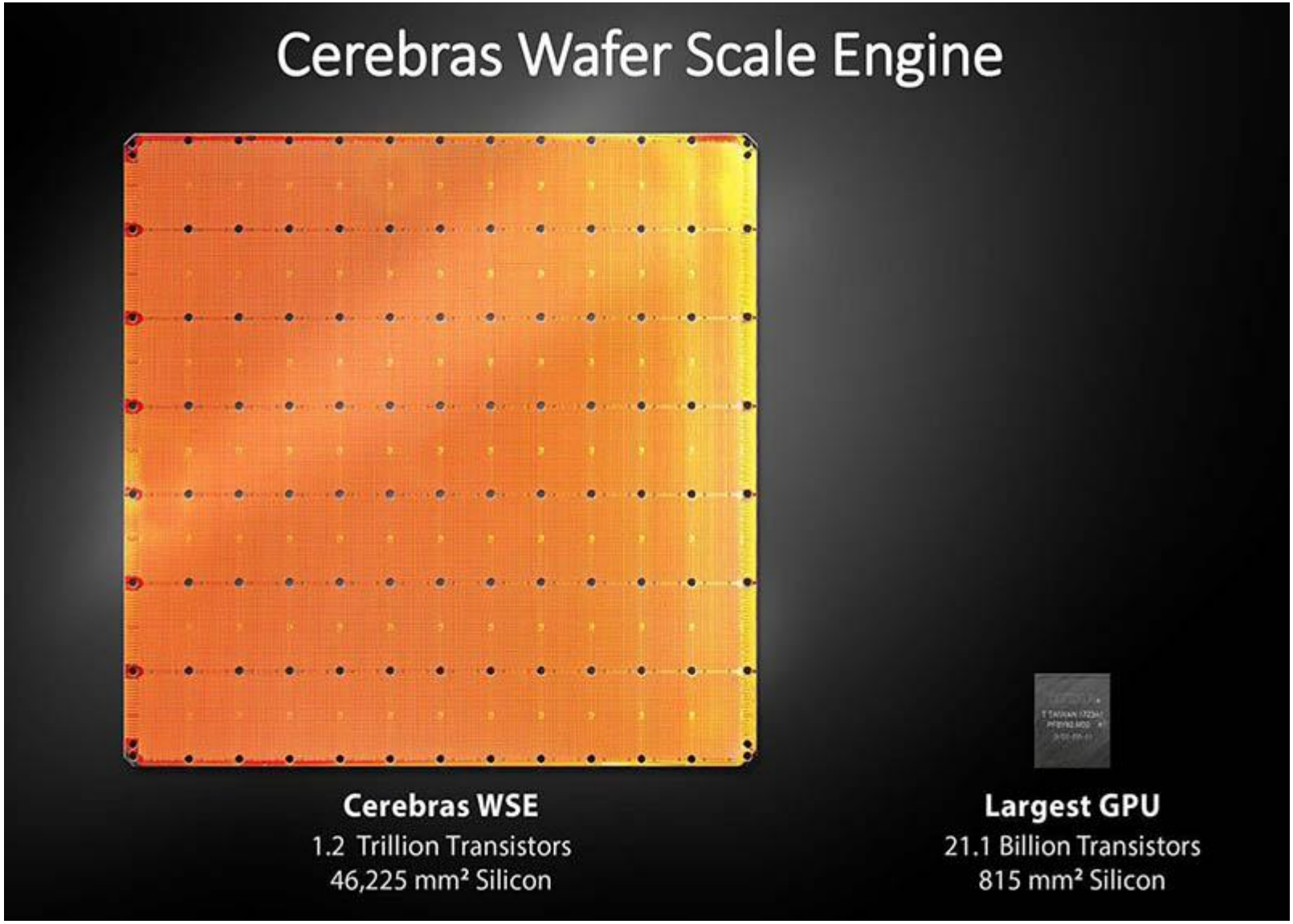
Number of Semiconductor Manufacturers with a Cutting Edge Logic Fab

180 nm	130 nm	90 nm	65 nm	45 nm/40 nm	32 nm/28 nm	22 nm/20 nm	16 nm/14 nm	10 nm	7 nm	5 nm
SilTerra										
X-FAB										
Dongbu HiTek										
ADI	ADI									
Atmel	Atmel									
Rohm	Rohm									
Sanyo	Sanyo									
Mitsubishi	Mitsubishi									
ON	ON									
Hitachi	Hitachi									
Cypress	Cypress	Cypress								
Sony	Sony	Sony								
Infineon	Infineon	Infineon								
Sharp	Sharp	Sharp								
Freescale	Freescale	Freescale								
Renesas (NEC)	Renesas	Renesas	Renesas	Renesas						
Toshiba	Toshiba	Toshiba	Toshiba	Toshiba						
Fujitsu	Fujitsu	Fujitsu	Fujitsu	Fujitsu						
TI	TI	TI	TI	TI						
Panasonic	Panasonic	Panasonic	Panasonic	Panasonic	Panasonic					
STMicroelectronics	STM	STM	STM	STM	STM					
HLMC	HLMC		HLMC	HLMC	HLMC					
UMC	UMC	UMC	UMC	UMC	UMC					
IBM	IBM	IBM	IBM	IBM	IBM	IBM				
SMIC	SMIC	SMIC	SMIC	SMIC	SMIC		SMIC			
AMD	AMD	AMD	GlobalFoundries	GF	GF	GF	GF			
Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung
TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel

https://en.wikichip.org/wiki/technology_node

Motivation for larger chips?

Wafer Scale Engine



If performance requires it and the market justifies it:
E.g. Accelerated Deep Learning

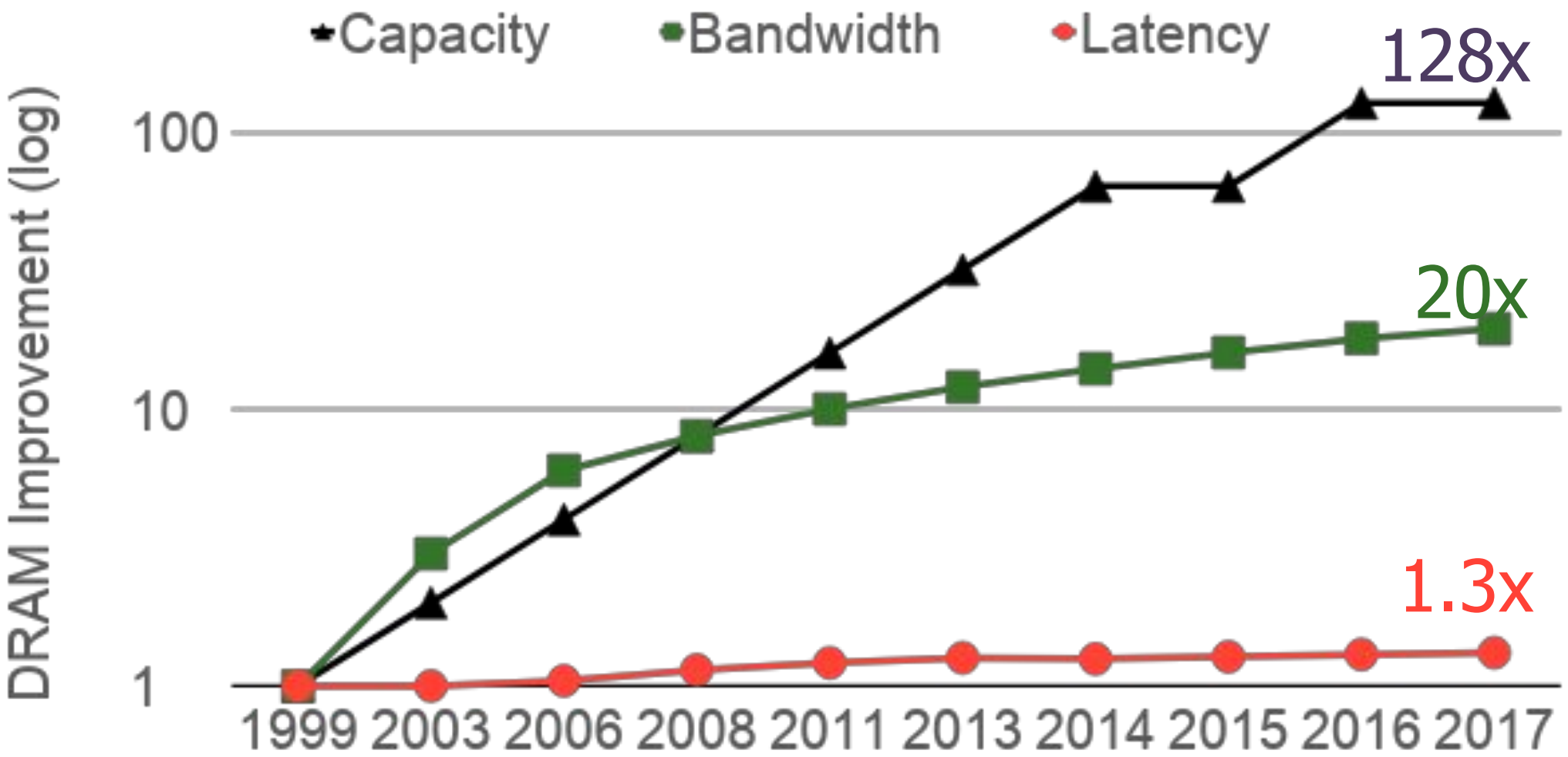
MEMORY SYSTEMS

- **Main Memory Speed Historically Not Growing As Fast As Processors' Speed.**
- Growing Gap Between Processor And Memory Speed (The So-called "**Memory Wall**")
- **We want a Memory System That's Big, Fast and Cheap**
- Use A Multi-level Hierarchy Of Memories
- Memory Hierarchies Rely On Principle Of Locality

Memory	Size	Marginal Cost	Cost per MB	Access Time
L3 Cache (on chip)	12MB	$10/MB	$10	5 nsec
Main Memory	8GB	$10/GB	1c	200 nsec
Disk	4TB	$50/TB	0.005c	5 msec

Approx cost and size of memories in a basic pc (2018)

DRAM Capacity, Bandwidth & Latency



Memory wall = memory_cycle / processor_cycle

In 1990, it was about 4 (25MHz, 150ns). Grew to 200 exponentially until 2002.
Has tapered off since then

Major Trends Affecting Main Memory

- Need for main memory capacity, bandwidth, QoS increasing
  - Multi-core: increasing number of cores/agents
  - Data-intensive applications: increasing demand/hunger for data
  - Consolidation: cloud computing, GPUs, mobile, heterogeneity
- Main memory energy/power is a key system design concern
  - ~40-50% energy spent in off-chip memory hierarchy [Lefurgy, IEEE Computer'03] >40% power in DRAM [Ware, HPCA'10][Paul,ISCA'15]
  - DRAM consumes power even when not used (periodic refresh)
- DRAM technology scaling is ending

WIRE DELAYS

- Wire delays don’t scale like logic delays
- Processor structures must expand to support more instructions
- Thus wire delays dominate the cycle time; slow wires must be local

DESIGN COMPLEXITY

- Processors are becoming so complex that a large fraction of the development of a processor or system is dedicated to verification
- Chip density is increasing much faster than the productivity of verification engineers (new tools, speed of systems)

CMOS ENDPOINT

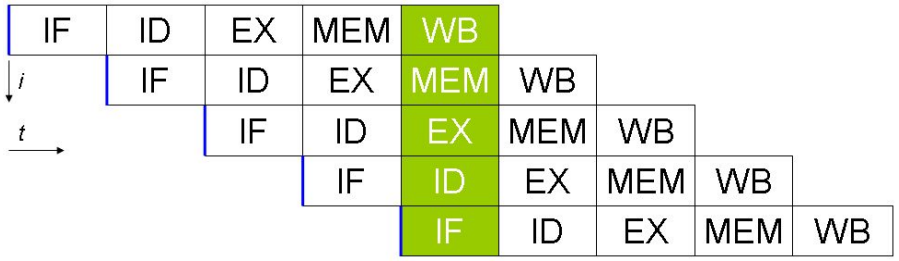
- **CMOS is rapidly reaching the limits of miniaturization**
  - feature sizes will reach atomic dimensions in less than 10 years
  - options????
    - quantum computing
    - nanotechnology
    - analog computing

PERFORMANCE REMAINS A CRITICAL DESIGN FACTOR

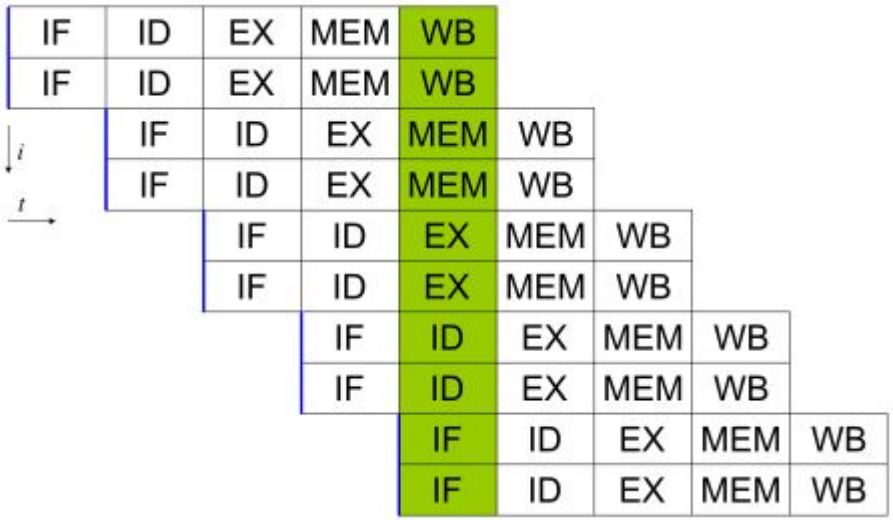
PARALLELISM IN ARCHITECTURES

SCALAR PROCESSOR

- **The Most Successful Microarchitecture Has Been The Scalar Processor**
- A Typical Scalar Instruction Operates On Scalar Operands
- ADD O1,O2,O3 // O2+O3=>O1
  - Execute multiple scalar Instructions at a time
  - Takes Advantage Of **ILP**, I.e., Instruction-level Parallelism, The Parallelism Exposed In Single Thread Or Single Process Execution



PIPELINING



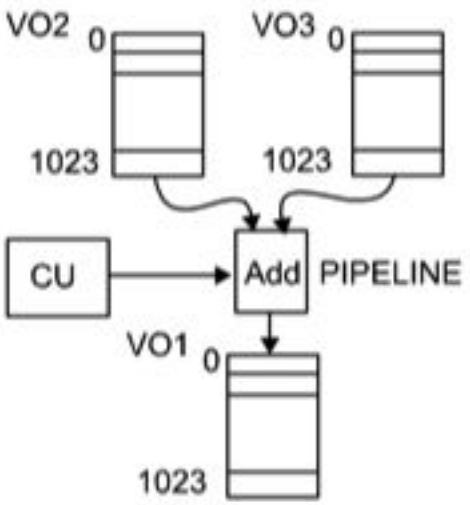
SUPERSCALAR

VECTOR PROCESSOR

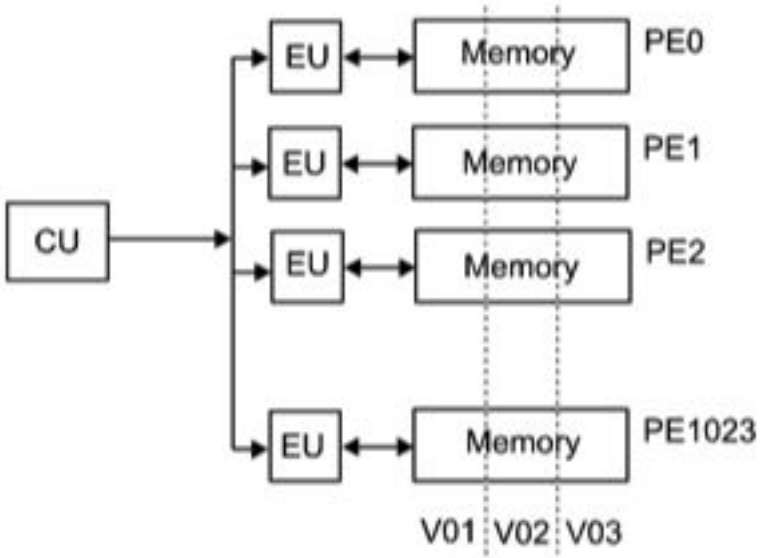
- **vector and array processors**
  - a typical vector instruction executes directly on vector operands

    VADD VO1,VO2,VO3 // VO2+VO3=>VO1
    - VOk is a vector of scalar components
    - Equivalent to computing
      - VO2[i]+VO3[i] => VO1[i], i=0,..,N
- **Vector instructions are executed by vector pipelines or parallel arrays.**

```
for(i=0;i<1024;i++)
  A[i] = B[i] + C[i]
```



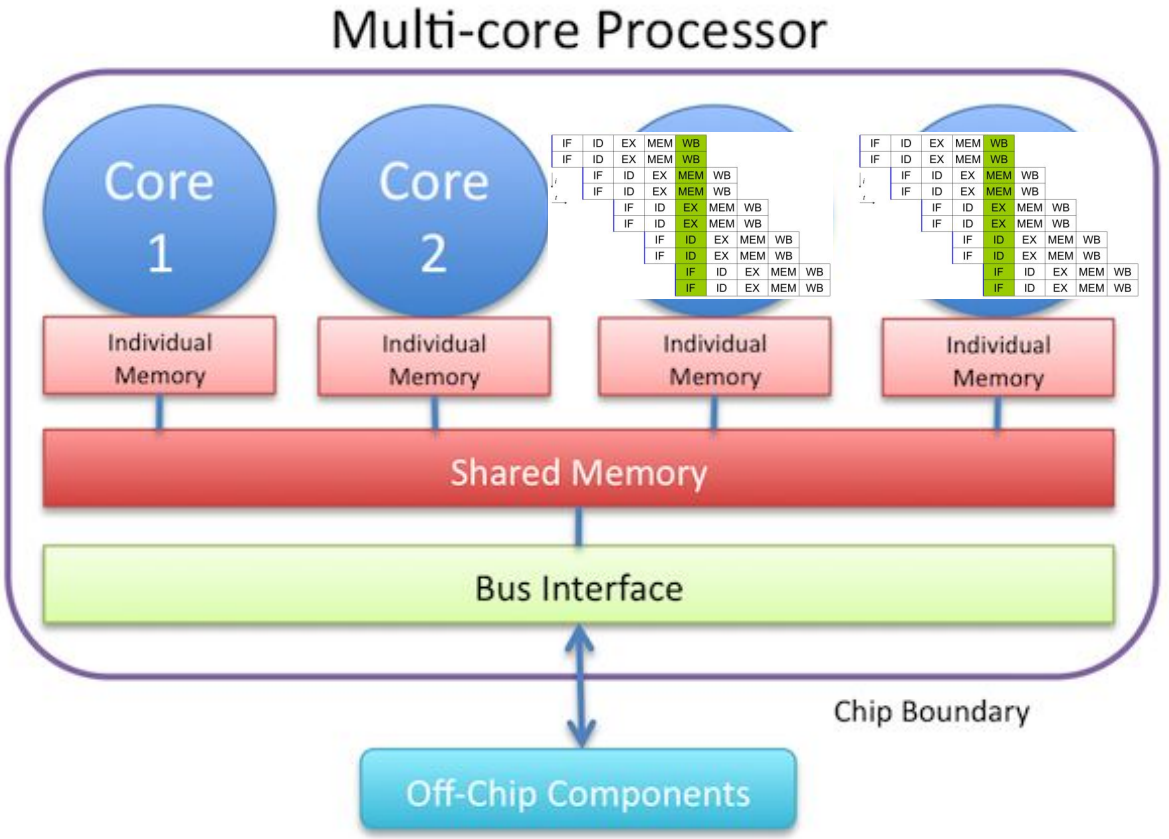
(a) Vector



(b) Array

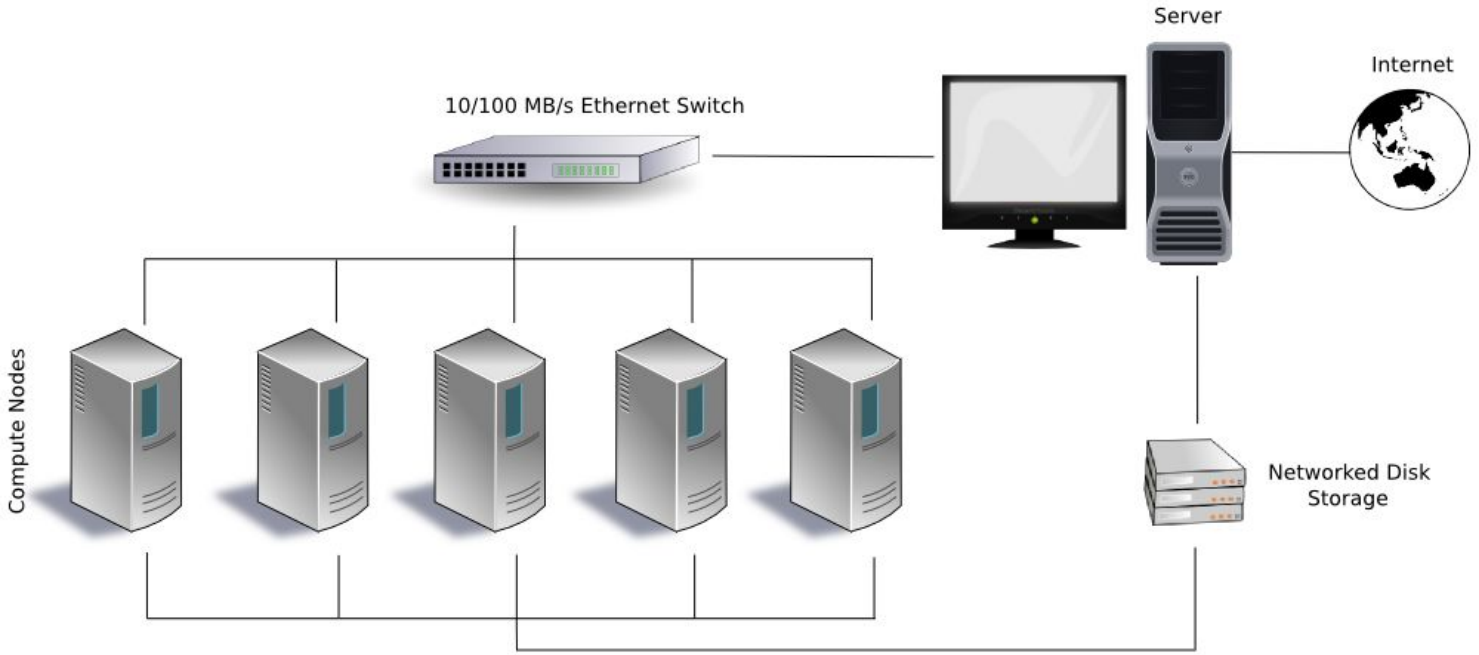
CHIP MULTIPROCESSOR (CMP)

- **CMPs (chip multiprocessors) exploit parallelism exposed by different threads running in parallel**
  - Thread Level Parallelism or **TLP**
  - Can be seen as multiple scalar processors running in parallel



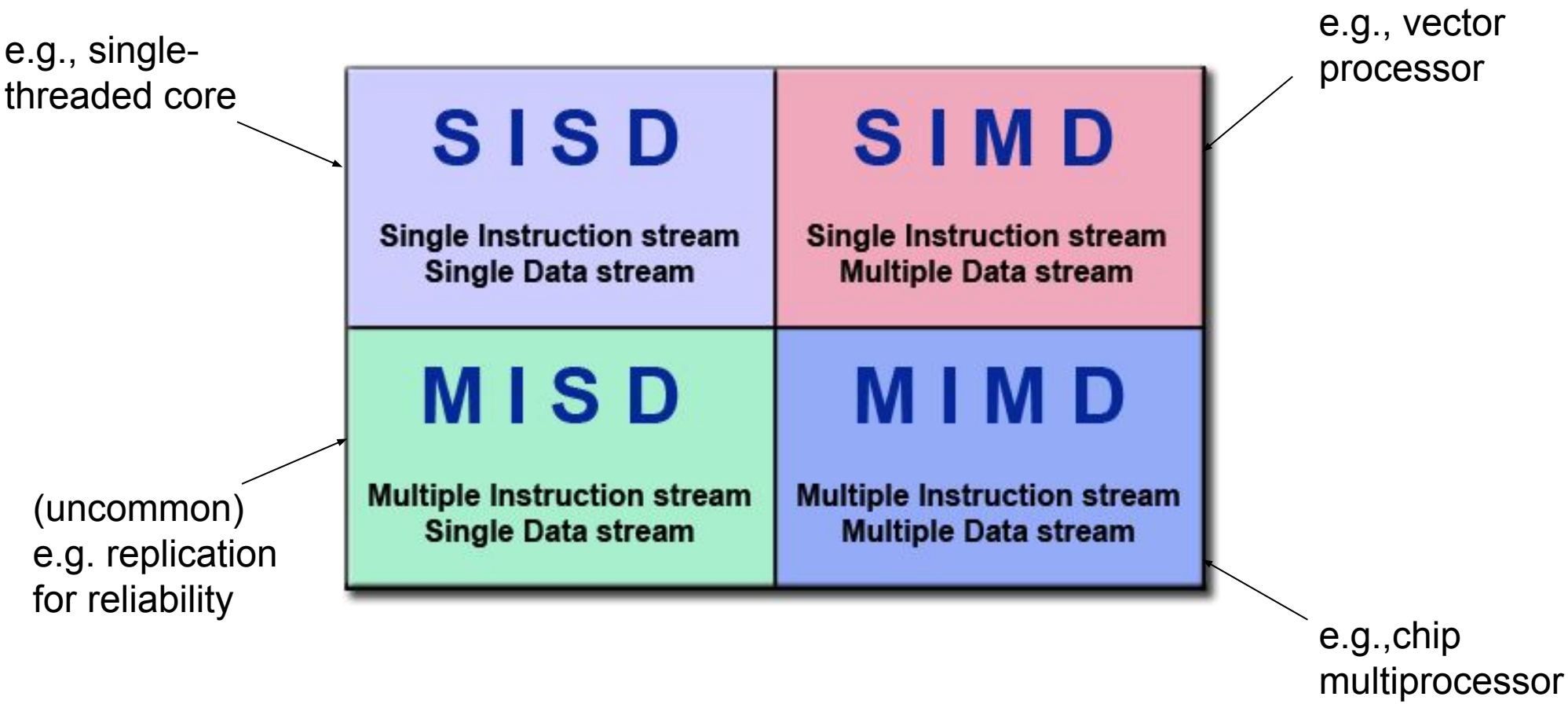
COMPUTER CLUSTER

- **Exploits parallelism exposed by different threads running in parallel on compute nodes**
  - Commonly: Single Program Multiple Data (SPMD) or Multiple Instruction Multiple Data (MIMD)
  - Nodes typically do not share memory
  - Each node includes CMPs or Vector Processors



FLYNN'S TAXONOMY

Classification of computer architectures, proposed by Michael J. Flynn in 1966



POWER

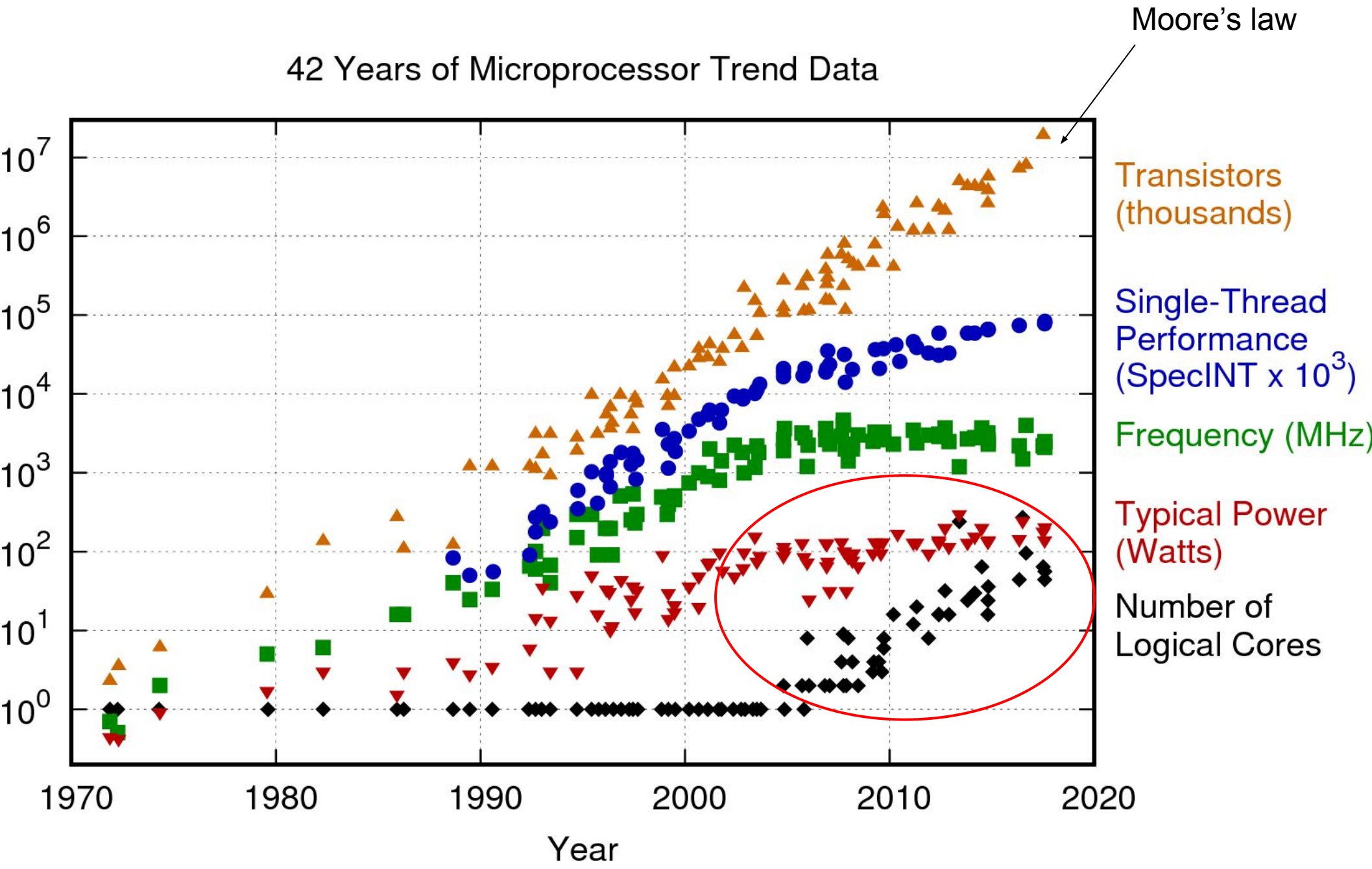
POWER

- **Total power: Dynamic + Static (leakage)**

$$P_{dynamic} = \alpha C V^2 f$$

$$P_{static} = V I_{sub} \propto V e^{-KVt/T}$$

- **Dynamic power favors parallel processing over higher clock rate**
  - dynamic power roughly proportional to f^3
  - take a U.P. and replicate it 4 times: 4x speedup & 4x power
  - take a U.P. and clock it 4 times faster: 4x speedup, but 64x dynamic power!



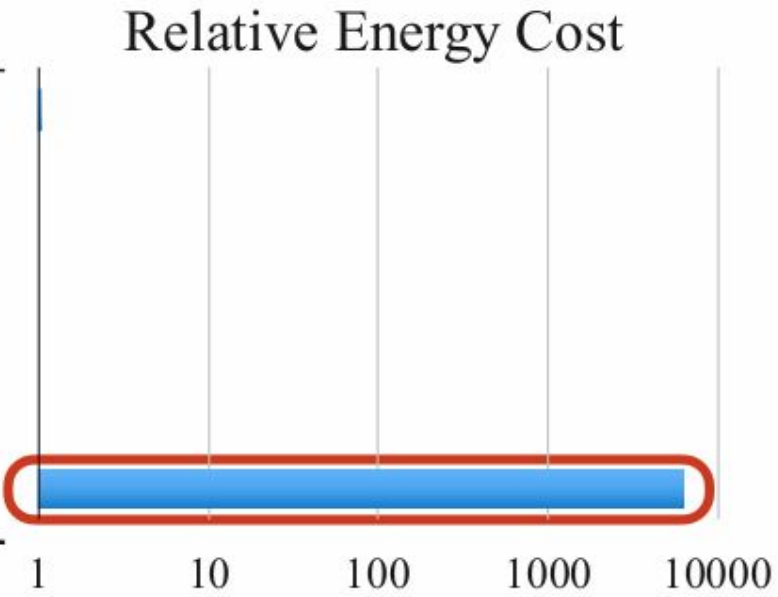
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

POWER

- **Power/energy are critical problems**
  - Power (immediate energy dissipation) must be dissipated
    - Otherwise temperature goes up (affects performance, correctness and may possibly destroy the circuit)
- Energy E=P*T (depends on power and speed)
  - Costly; global problem
  - Battery operated devices

Where is the energy consumed

Operation	Energy [pJ]
32 bit int ADD	0.1
32 bit float ADD	0.9
32 bit Register File	1
32 bit int MULT	3.1
32 bit float MULT	3.7
32 bit SRAM Cache	5
32 bit DRAM Memory	**640**



1

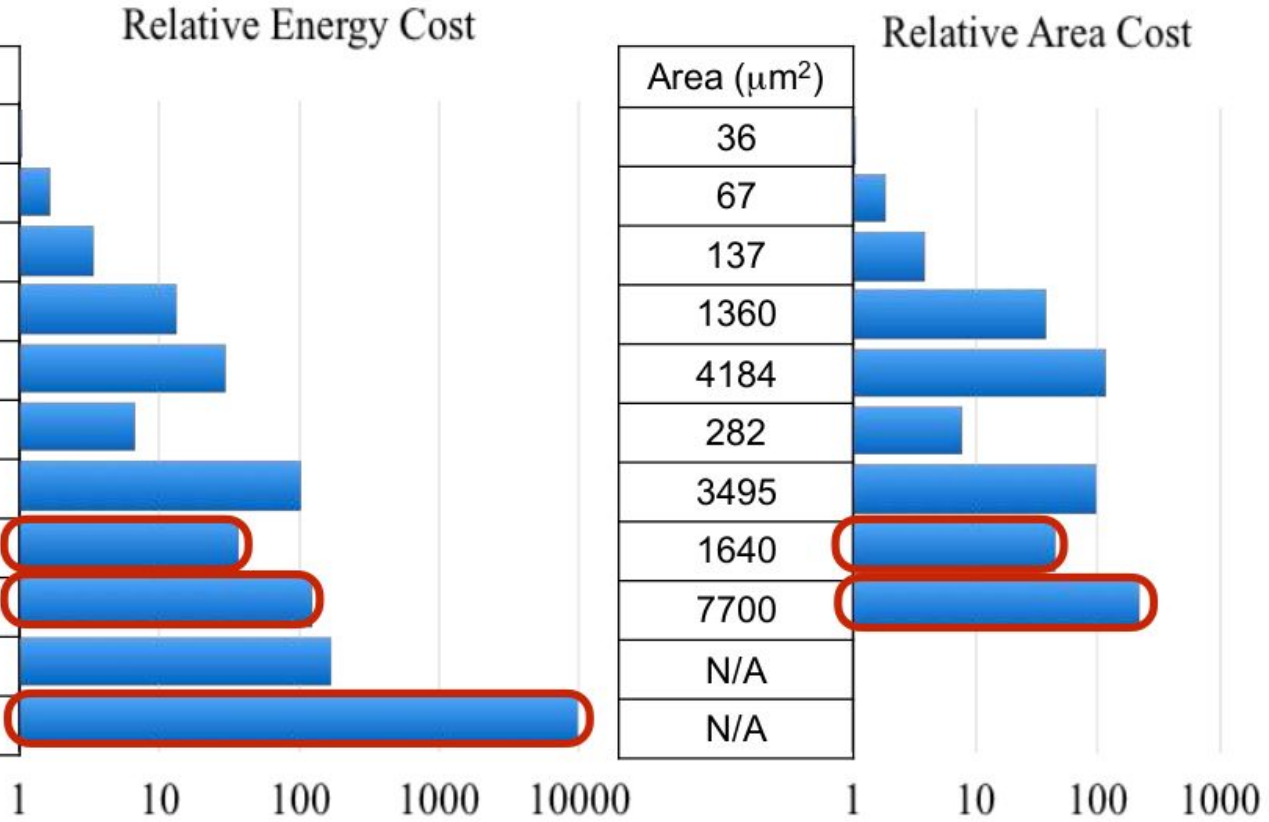
This image is in the public domain

= 1000

Exascale target is 2pJ / FLOP

Energy cost and Accuracy

Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



Energy numbers are from Mark Horowitz "Computing's Energy Problem (and what we can do about it)", ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

SUMMARY

- Computer architecture definition
- System Components
- Technology trends
- Parallelism in architectures
- Power

Next Lecture (Tue, Jan 28)

- Performance Metrics
- Vector Processors
- Discussion of Project