



重庆大学
CHONGQING UNIVERSITY



智能计算系统实验室
Intelligent Computing Systems Lab

Lecture3

Computer Architecture (Fall 2022)

Introduction

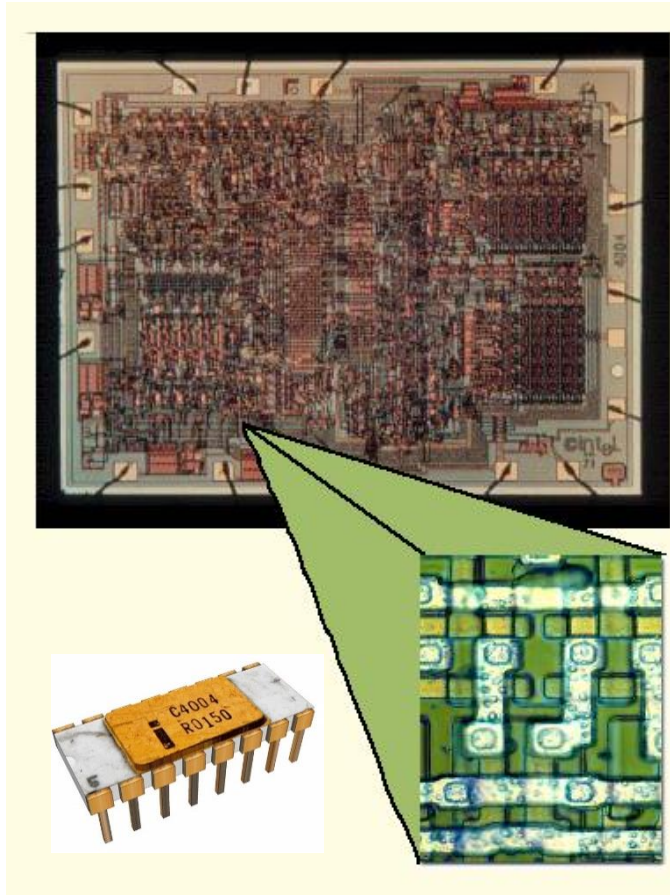
Dr. Duo Liu (刘铎)

Office: Main Building 0626

Email: liuduo@cqu.edu.cn

Intel 4004 (1970)

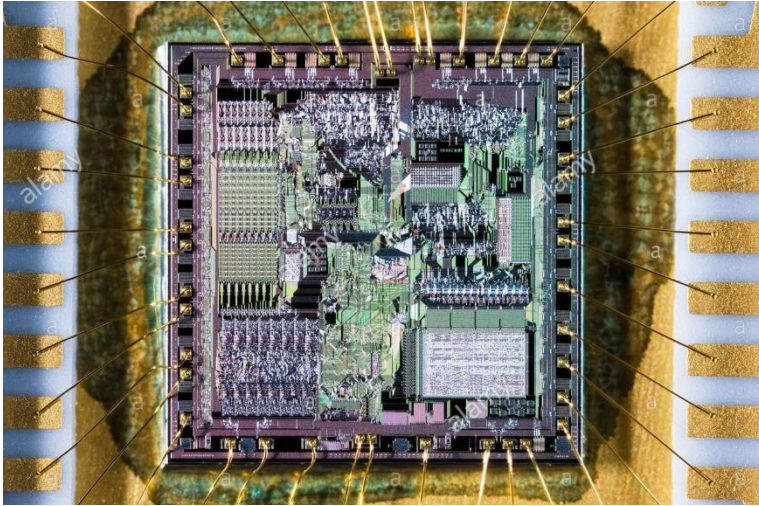
2/51



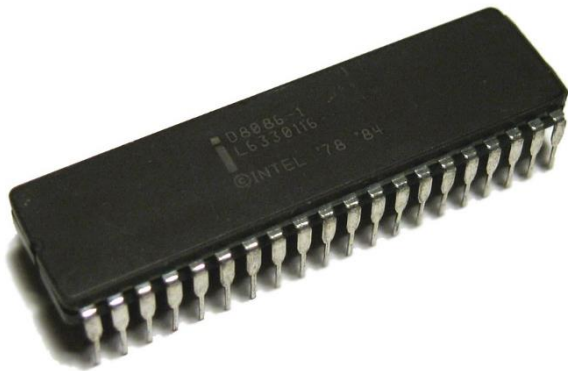
- **The first microprocessor** (all of the components of a CPU on a single chip).
- **2,300 Transistors**
- **108 KHz**
- **Addressable memory: 640 bytes**

Intel 8086 (1978)

3/51

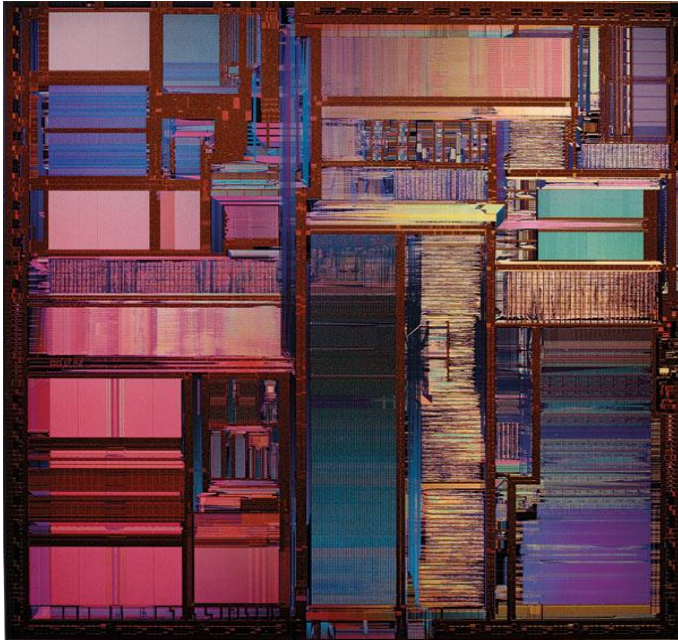


- 16-bit **general-purpose processor**
- 29,000 Transistors
- 5MHz, 8MHz, 10MHz
- 33 square mm



Pentium (1993)

4/51



Wikipedia Appaloosa

- **32-bit** general-purpose processor
- 3.1M Transistors
- 60MHz-166MHz
- 296 Square mm
- **The 1st Superscalar Implementation of IA 32**

Pentium 4 (2000)

5/51



- 42 M Transistors
- 1.3-1.9 GHz
- 146 Square mm

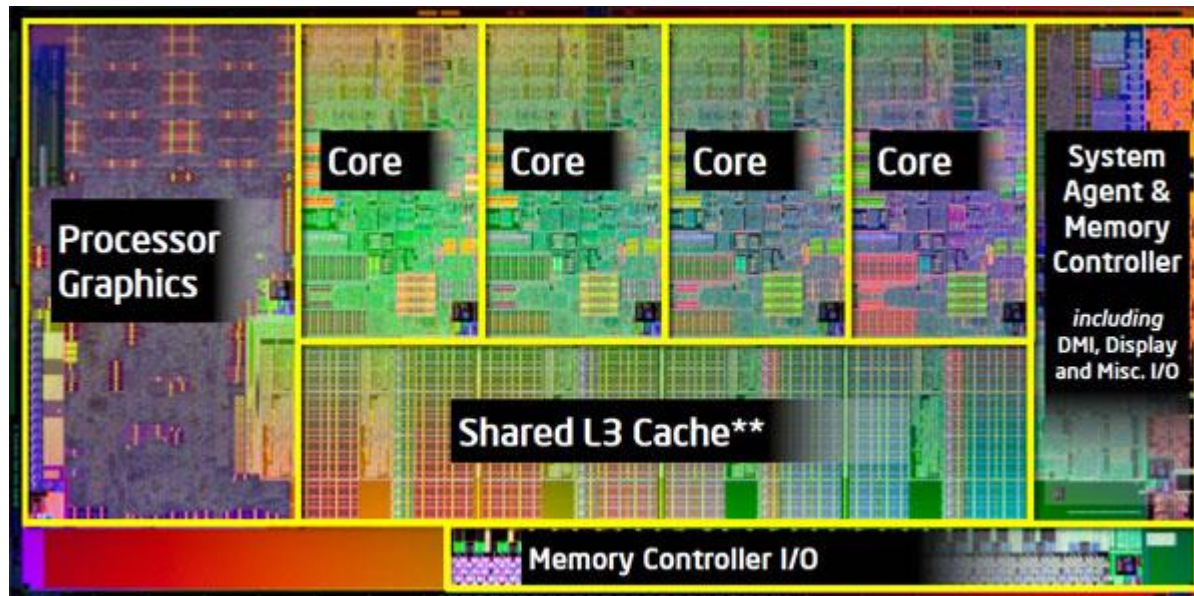


Core i7 (2011)

6/51



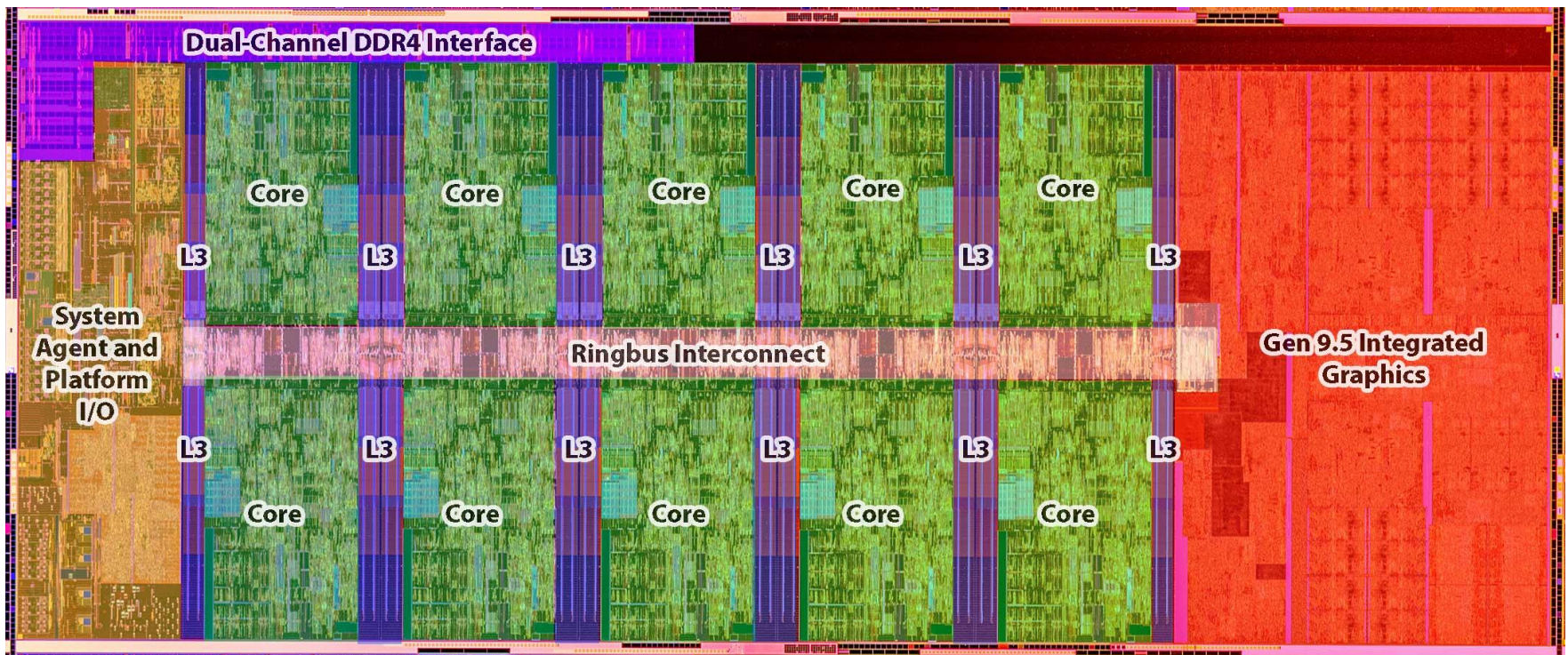
- 731 M Transistors
- 2.66-3.33 GHz
- 263 Square mm



Core i9-10900x (2019)

7/51

- (unknown, >3.5 B estimated) Transistors
- 3.7-4.5 GHz
- die size: unknown



Generations of Computer

8/51

Gen.	Dates	Technology	Typical Speed	Typical Products
1	1946-1957	Vacuum tube	40,000	ENIAC
2	1958-1964	Transistor	200,000	IBM 7000
3	1965-1971	Small/medium Scale Integrated Circuit	1,000,000	IBM System/360, DEC PDP-11
4	1971-1977	Large Scale Integrated Circuit	10,000,000	Intel 4004
5	1978-	Very-Large-Scale Integrated Circuit	100,000,000	Intel Pentium

- Feature size
 - Minimum size of transistor or wire in x or y dimension
 - 10 microns in 1971 to .032 microns in 2011
 - Transistor performance scales linearly
 - Wire delay does not improve with feature size!
 - Integration density scales quadratically

- **Integrated circuit technology**
 - Transistor density: 35%/year
 - Die size: 10-20%/year
 - Integration overall: 40-55%/year
- **DRAM capacity:** 25-40%/year (slowing)
- **Flash capacity:** 50-60%/year
 - 15-20X cheaper/bit than DRAM
- **Magnetic disk technology:** 40%/year
 - 15-25X cheaper/bit than Flash
 - 300-500X cheaper/bit than DRAM

Technology Scaling and IT Industry Progress

11/51

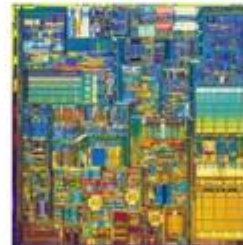
- 1971 (10 μ)
- 4004 μ P
- 5k trans.
- 4mm² die
- 108Khz
- 0.2W
- Busicom calculator



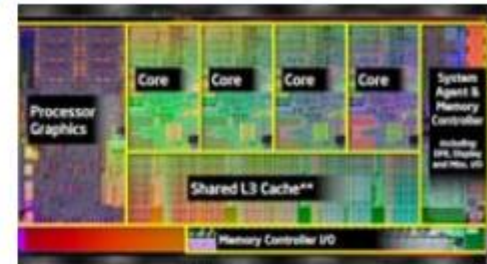
- 1982 (1.5 μ)
- 286 μ P
- 134k trans.
- 47mm² die
- 8Mhz
- 3W
- 15M PCs sold in 6yr



- 1994 (0.6 μ)
- Pentium®
- 3.2M trans.
- 147mm² die
- 100Mhz
- 10W
- sound, images



- 2000 (0.18 μ)
- Pentium® 4
- 42M trans.
- 217mm² die
- 1.5Ghz
- 58W
- internet

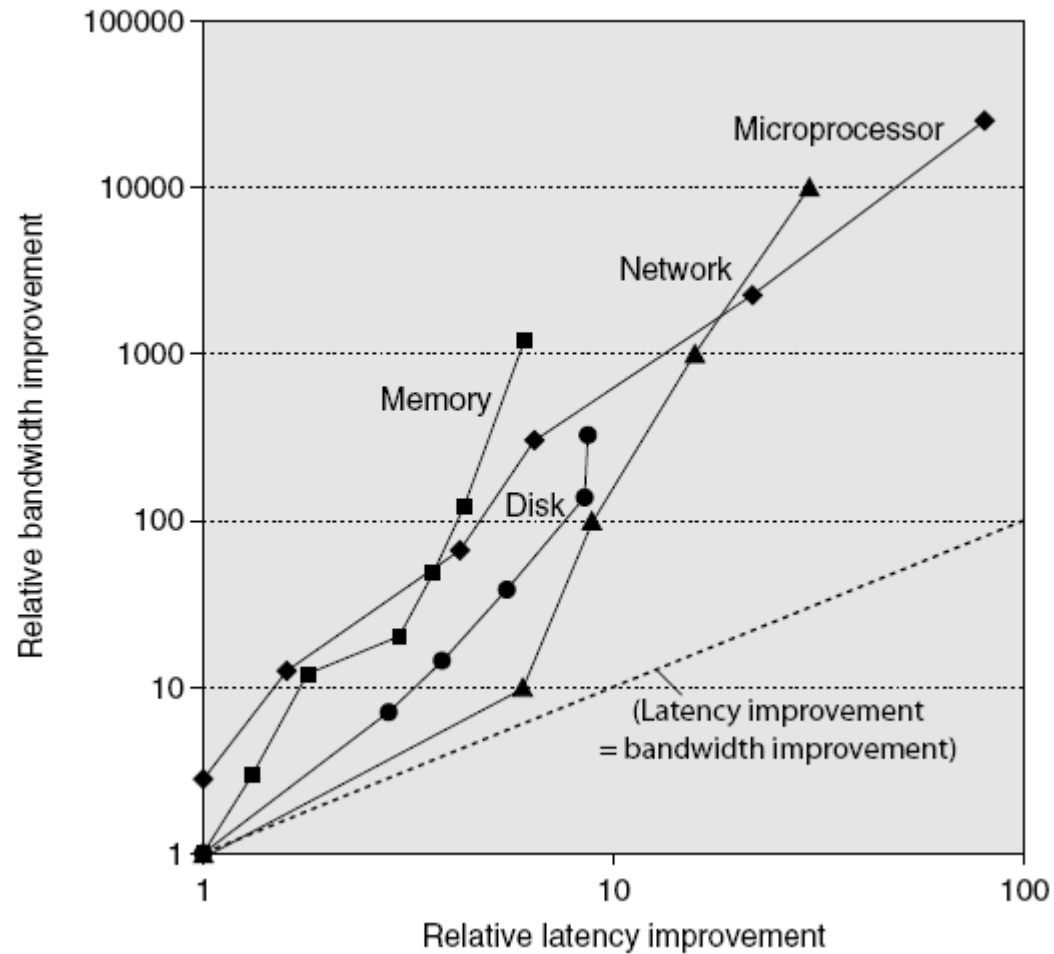


- 2011 (0.035 μ)
- Core® (Sandy Bridge)
- 995M trans.
- 216mm² die
- 3.6Ghz
- 95W
- content creation, immersive gaming, pervasive computing

- Bandwidth or throughput
 - Total work done in a given time
 - 10,000-25,000X improvement for processors
 - 300-1200X improvement for memory and disks
- Latency or response time
 - Time between start and completion of an event
 - 30-80X improvement for processors
 - 6-8X improvement for memory and disks

Bandwidth and Latency

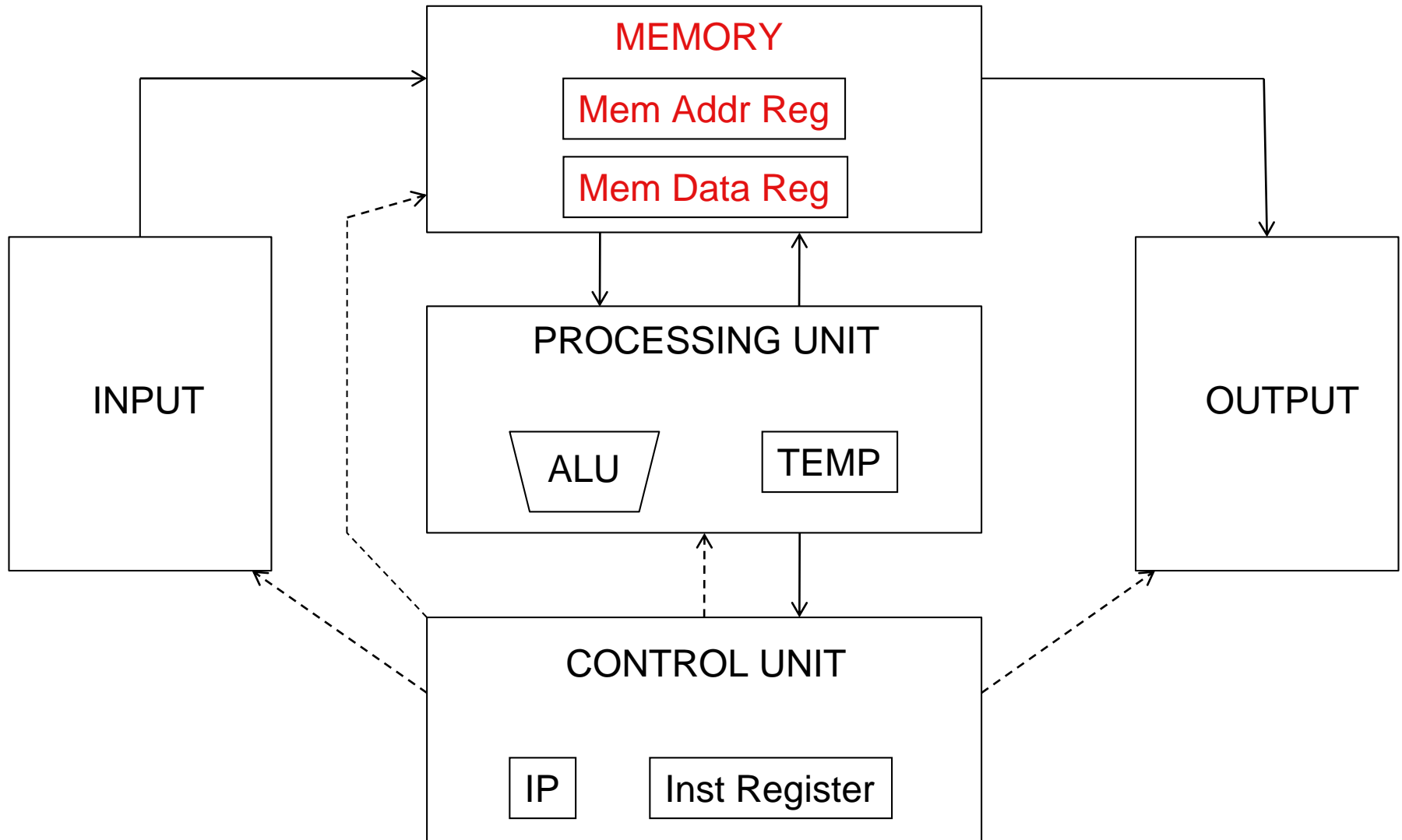
13/51



Log-log plot of bandwidth and latency milestones

Von Neumann Model

14/51



- Also called *stored program computer* (instructions in memory). Two key properties:
- Stored program
 - **Instructions** stored in a linear memory array
 - Memory is **unified** between instructions and data
 - The **interpretation** of a stored value **depends on the control signals**
- Sequential instruction processing
 - One instruction processed (fetched, executed, and completed) **at a time**
 - **Program counter** (instruction pointer) identifies the current instr.
 - Program counter is advanced sequentially **except for control transfer instructions**

- All major *instruction set architectures* today use this model
 - x86, ARM, MIPS, SPARC, Alpha, POWER
- Underneath (at the microarchitecture level), the execution model of almost all *implementations (or, microarchitectures)* is very different
 - Pipelined instruction execution: *Intel 80486 uarch*
 - Multiple instructions at a time: *Intel Pentium uarch*
 - Out-of-order execution: *Intel Pentium Pro uarch*
 - Separate instruction and data caches
- But, what happens underneath that is *not* consistent with the von Neumann model is *not* exposed to software
 - Difference between ISA and microarchitecture

- Recommended reading
 - Burks, Goldstein, von Neumann, “Preliminary discussion of the logical design of an electronic computing instrument,” 1946.
 - Patt and Patel book, Chapter 4, “The von Neumann Model”

- **ISA**

- Agreed upon interface between software and hardware
 - SW/compiler assumes, HW promises
- What the software writer needs to know to write and debug system/user programs

- **Microarchitecture**

- Specific implementation of an ISA
- Not visible to the software

- **Microprocessor**

- **ISA**, **uarch**, circuits
- “Architecture” = ISA + microarchitecture

Problem
Algorithm
Program
Subsystem
ISA
Microarchitecture
Circuits
Electrons

- **Instructions**

- Opcodes, Addressing Modes, Data Types
- Instruction Types and Formats
- Registers, Condition Codes

- **Memory**

- Address space, Addressability, Alignment
- Virtual memory management

- **Call, Interrupt/Exception Handling**

- **Access Control, Priority/Privilege**

- **I/O: memory-mapped vs. instr.**

- **Task/thread Management**

- **Power and Thermal Management**

- **Multi-threading support, Multiprocessor support**

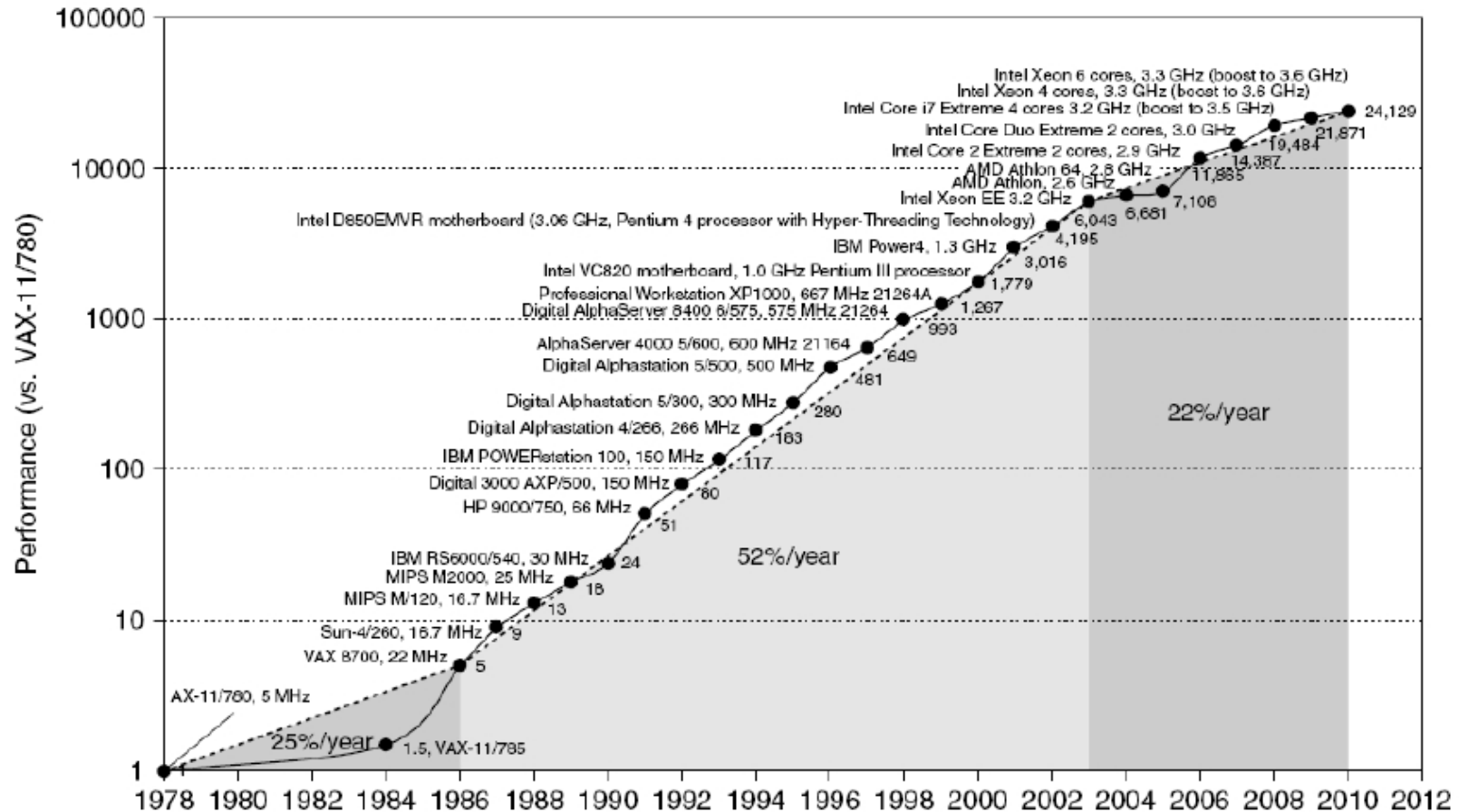


Intel® 64 and IA-32 Architectures
Software Developer's Manual

Volume 1:
Basic Architecture

Sequential Processor Performance

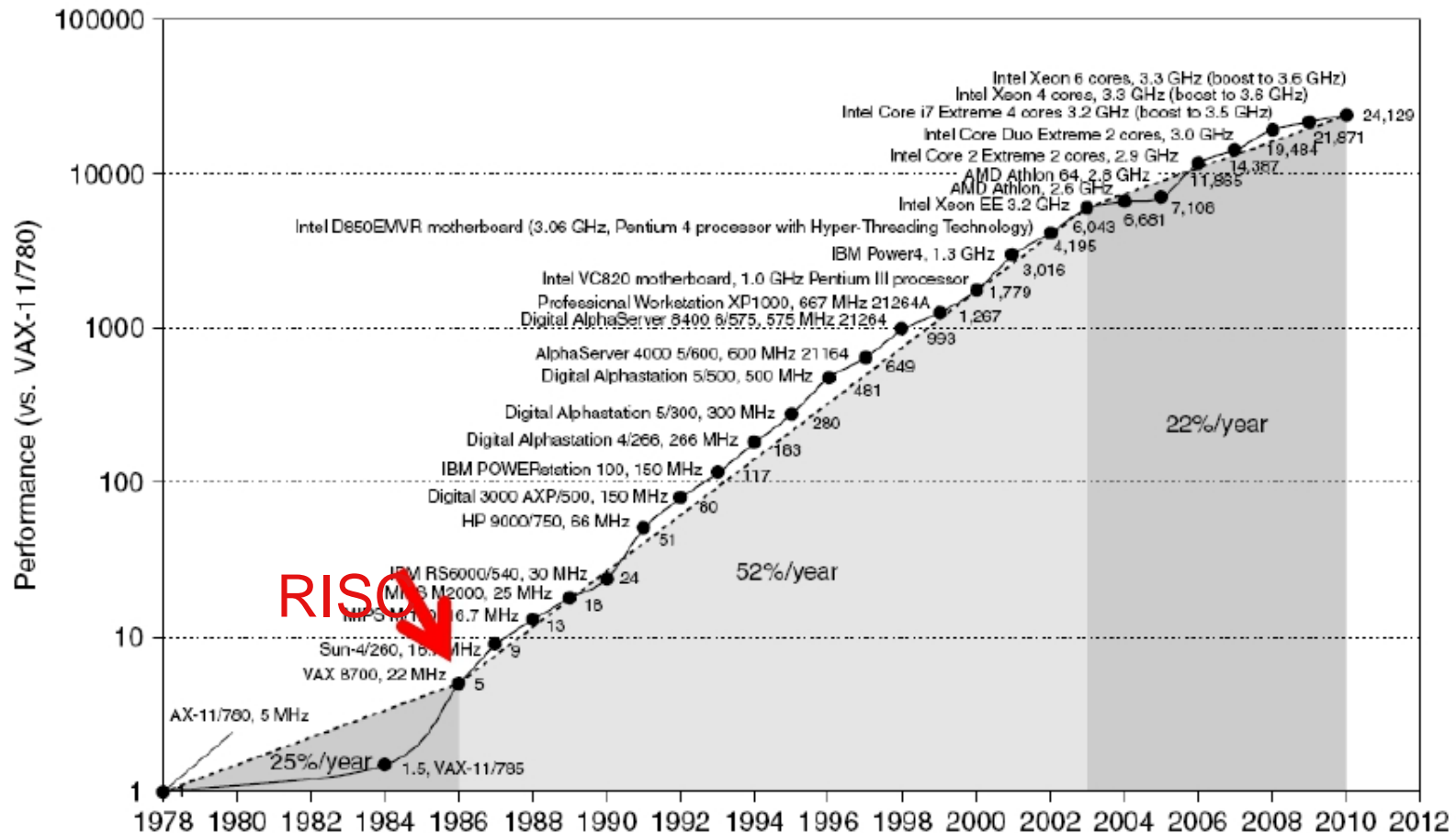
20/51



From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

Sequential Processor Performance

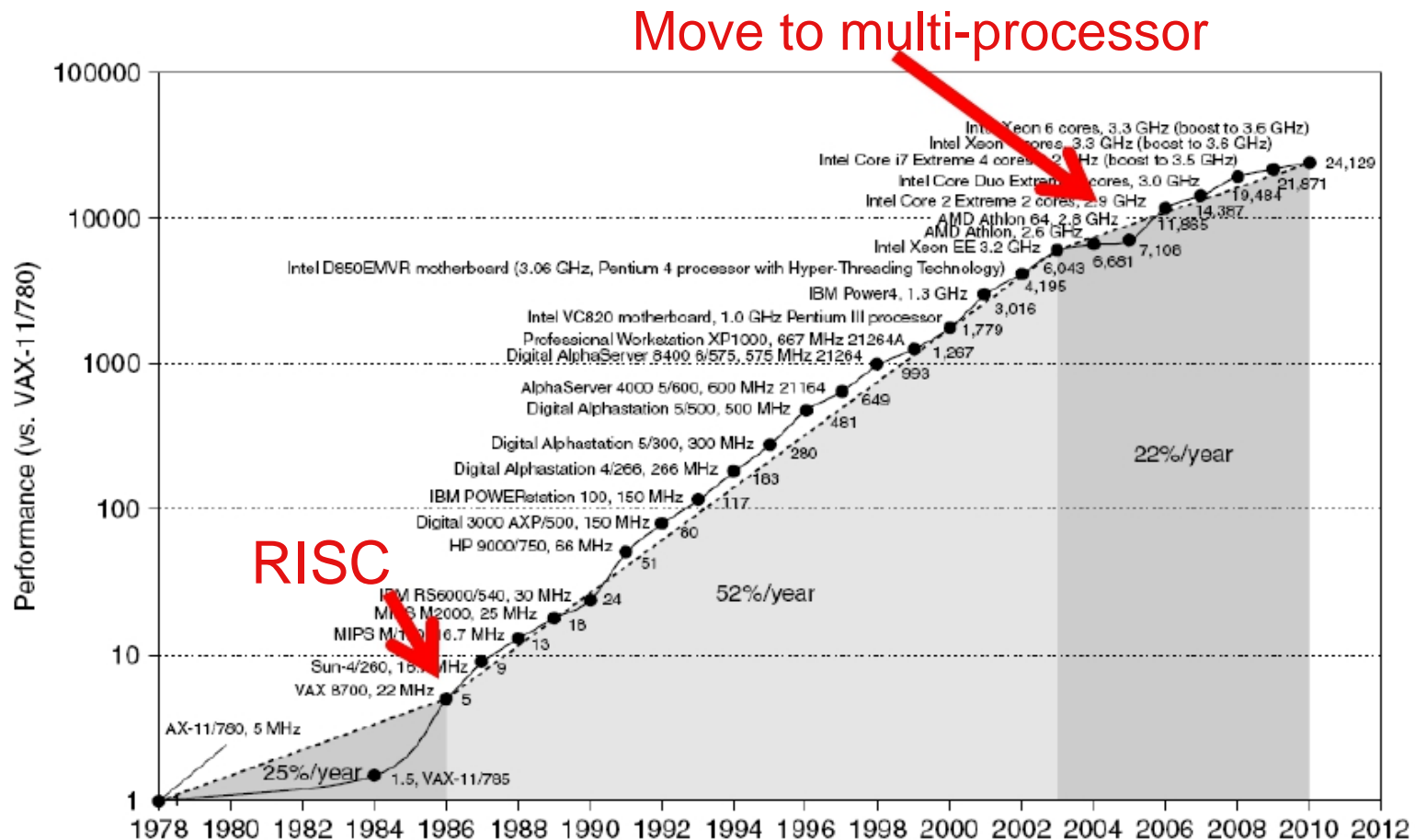
21/51



From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

Sequential Processor Performance

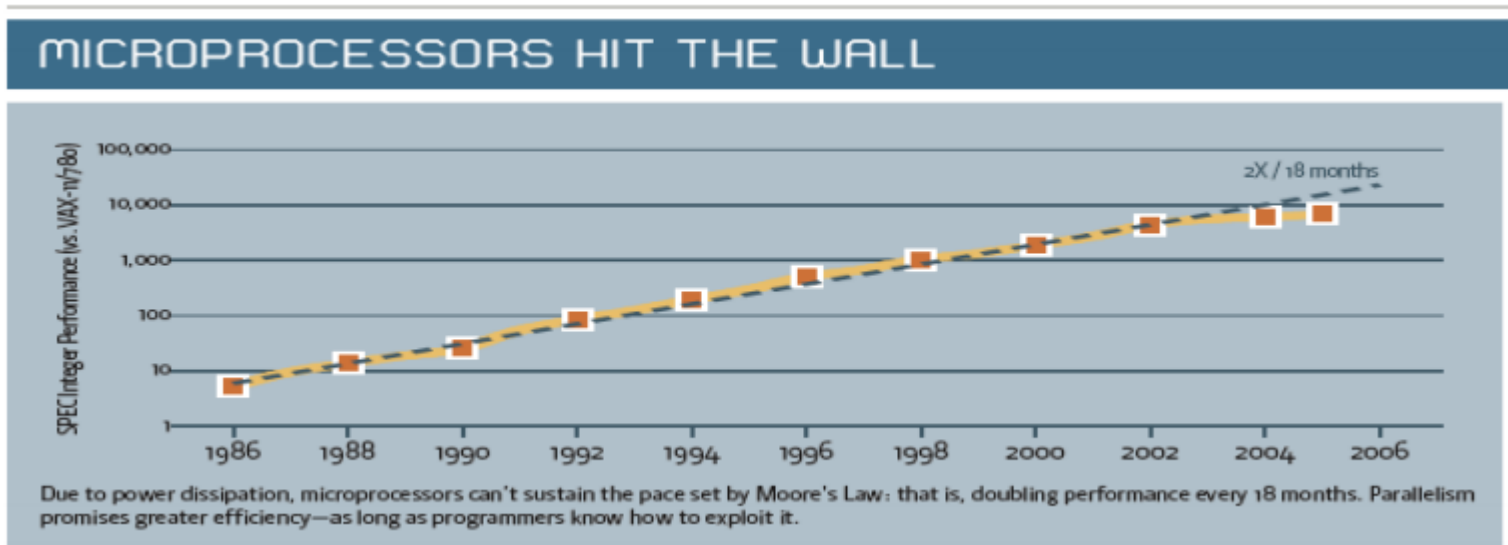
22/51



From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

Microprocessor Hit the Power Wall

23/51



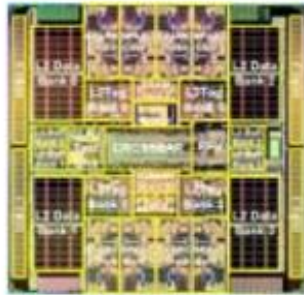
- “The media's mischaracterization of Moore's Law is now evident. Gordon Moore predicted the regular doubling of the number of transistors on a chip. The job of computer architects was to turn twice as many transistors into twice as much performance . Between 1986 and 2002 architects succeeded, and we saw the greatest sustained increase in performance in computing history. The problem was that they kept increasing the power dissipated per chip, and in 2004 it was obvious that the industry had hit a power wall. Today, microprocessors are about a factor of three slower than if we could keep increasing power and doubling performance every 18 months Thus, while Moore's Law continues, power dissipation hit the wall” [D. A. Patterson 2007]

The Processor is the New Transistor

24/51

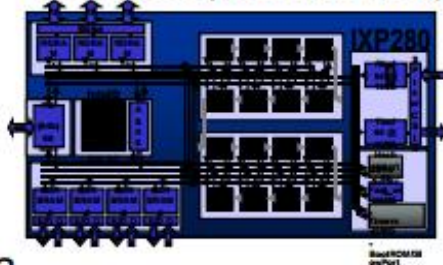
Only way to meet future system feature set, design cost, power, and performance requirements is by programming a processor array

- multiple parallel general-purpose processors (GPPs)
- multiple application-specific processors (ASPs)

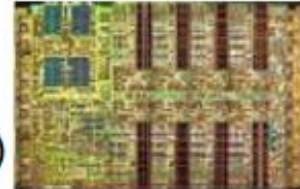
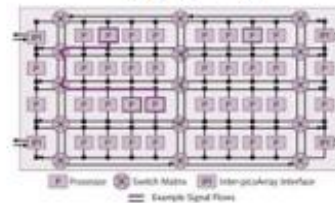


Sun Niagara
8 GPP cores (32 threads)

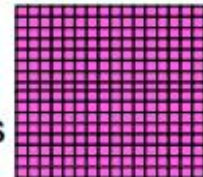
Intel Network Processor
1 GPP Core
16 ASPs (128 threads)




IBM Cell
1 GPP (2 threads)
8 ASPs



Picochip DSP
1 GPP core
248 ASPs



Cisco CSR-1
188 Tensilica GPPs



Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm² chip

- *Many predict that the number of cores will grow exponentially in future years*

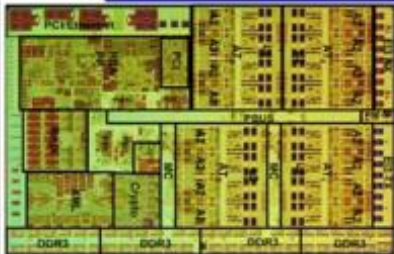
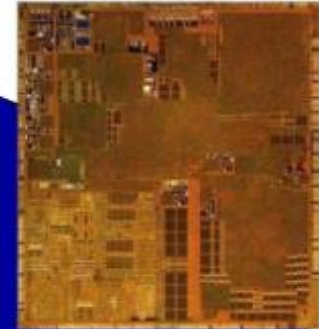
- *"The Processor is the New Transistor"*

Multi-Core Systems-on-Chip (SoC)

25/51



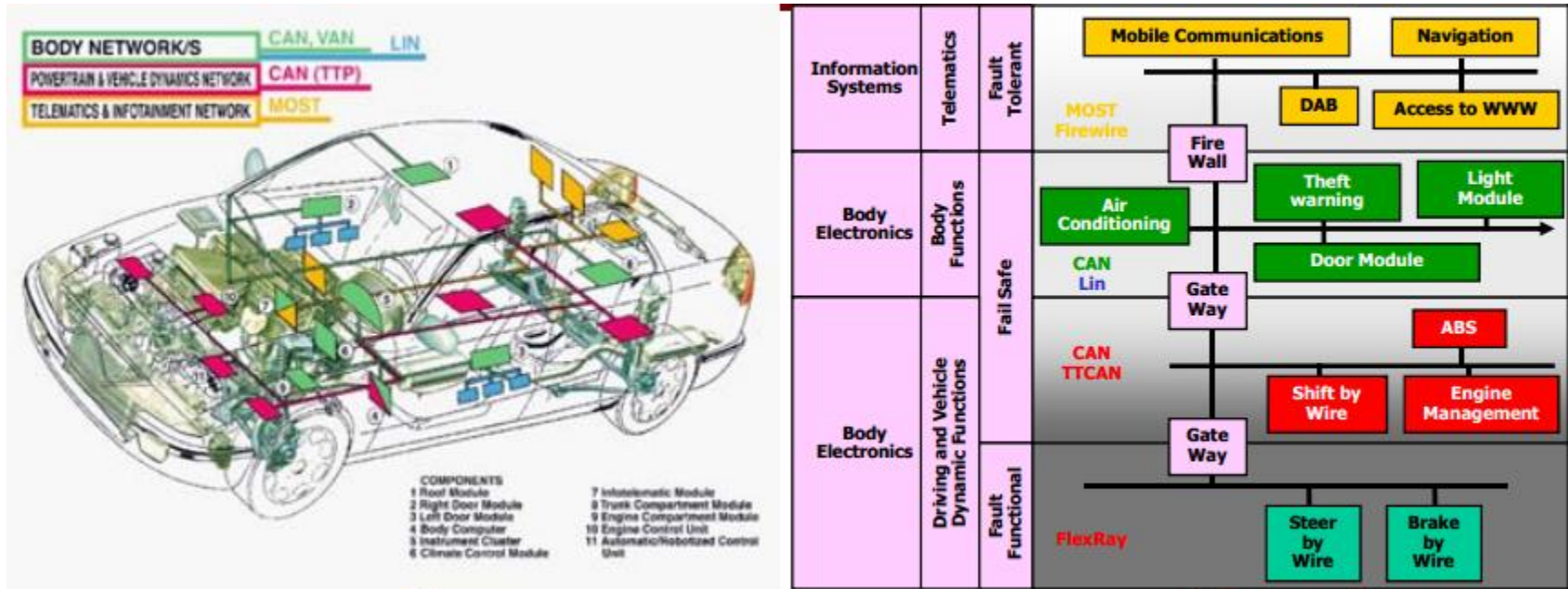
- A rich variety of multi-core architectures, and growing...
 - shift from homogeneous tile-based chip multi-processors (CMP) to heterogeneous multi-core systems-on-chip (SOCs)
- Complexity of design and programming
 - increasing number of heterogeneous cores
 - high-performance design is *power-efficient* design
 - resiliency to parameter variations, component faults,...
 - need for a new HW/SW interface
 - increasing impact of communication



Heterogeneous Embedded Systems

26/51

I Electronics for the Car



- Up to 70 Electronic Computing Units (ECUs) in a modern car like a BMW Series 7
 - Heterogeneous communication networks
 - DSC (dynamic stability control) contains ABS as one of 15 sub-functionalities

- Cannot continue to leverage Instruction-Level parallelism (ILP)
 - Single processor performance improvement ended in 2003
- New models for performance:
 - Data-level parallelism (DLP)
 - Thread-level parallelism (TLP)
 - Request-level parallelism (RLP)
- These require explicit restructuring of the application

Classes of Computers (by 2010)

28/51

- **Personal** Mobile Device (PMD)
 - e.g. smart phones, tablet computers
 - Emphasis on energy efficiency and real-time
- **Desktop** Computing
 - Emphasis on price-performance
- **Servers**
 - Emphasis on availability, scalability, throughput
- **Clusters / Warehouse** Scale Computers
 - Used for “Software as a Service (SaaS)”
 - Emphasis on availability and price-performance
 - Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks
- **Embedded** Computers
 - Emphasis: price, power

I Parallelism

- Classes of **parallelism in applications**:
 - **Data-Level** Parallelism (DLP)
 - **Task-Level** Parallelism (TLP)
- Classes of **architectural parallelism**:
 - **Instruction-Level** Parallelism (ILP)
 - **Vector** architectures/**Graphic** Processor Units (GPUs)
 - **Thread-Level** Parallelism
 - **Request-Level** Parallelism

I Flynn's Taxonomy

- Single instruction stream, single data stream (SISD)
- Single instruction stream, multiple data streams (SIMD)
 - Vector architectures
 - Multimedia extensions
 - Graphics processor units
- Multiple instruction streams, single data stream (MISD)
 - No commercial implementation
- Multiple instruction streams, multiple data streams (MIMD)
 - Tightly-coupled MIMD
 - Loosely-coupled MIMD

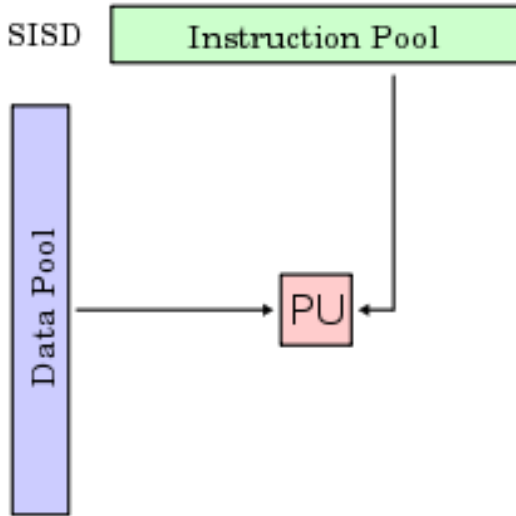
Classes of Computers

31/51

Flynn's Taxonomy

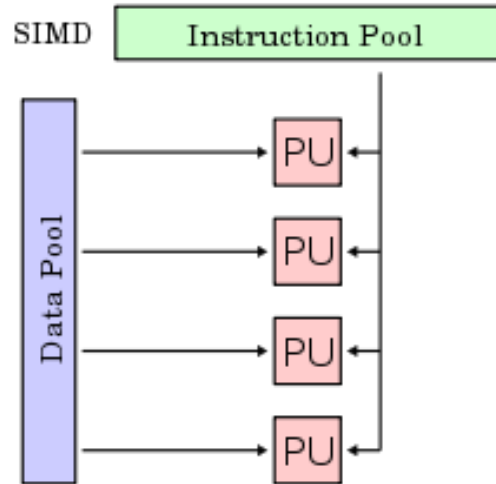
SISD

SISD



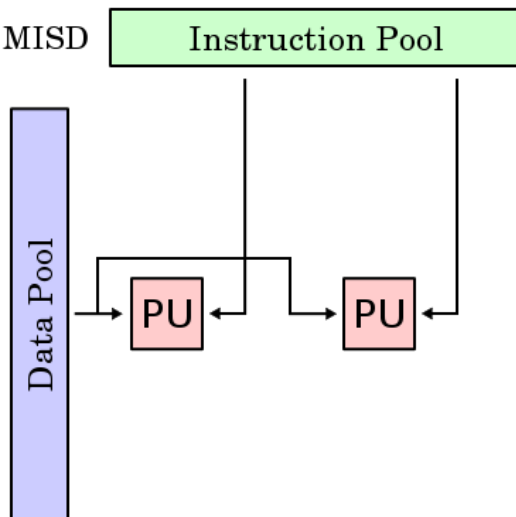
SIMD

SIMD



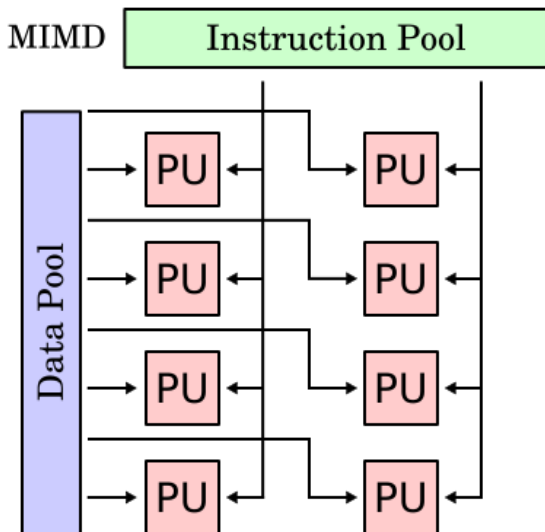
MISD

MISD

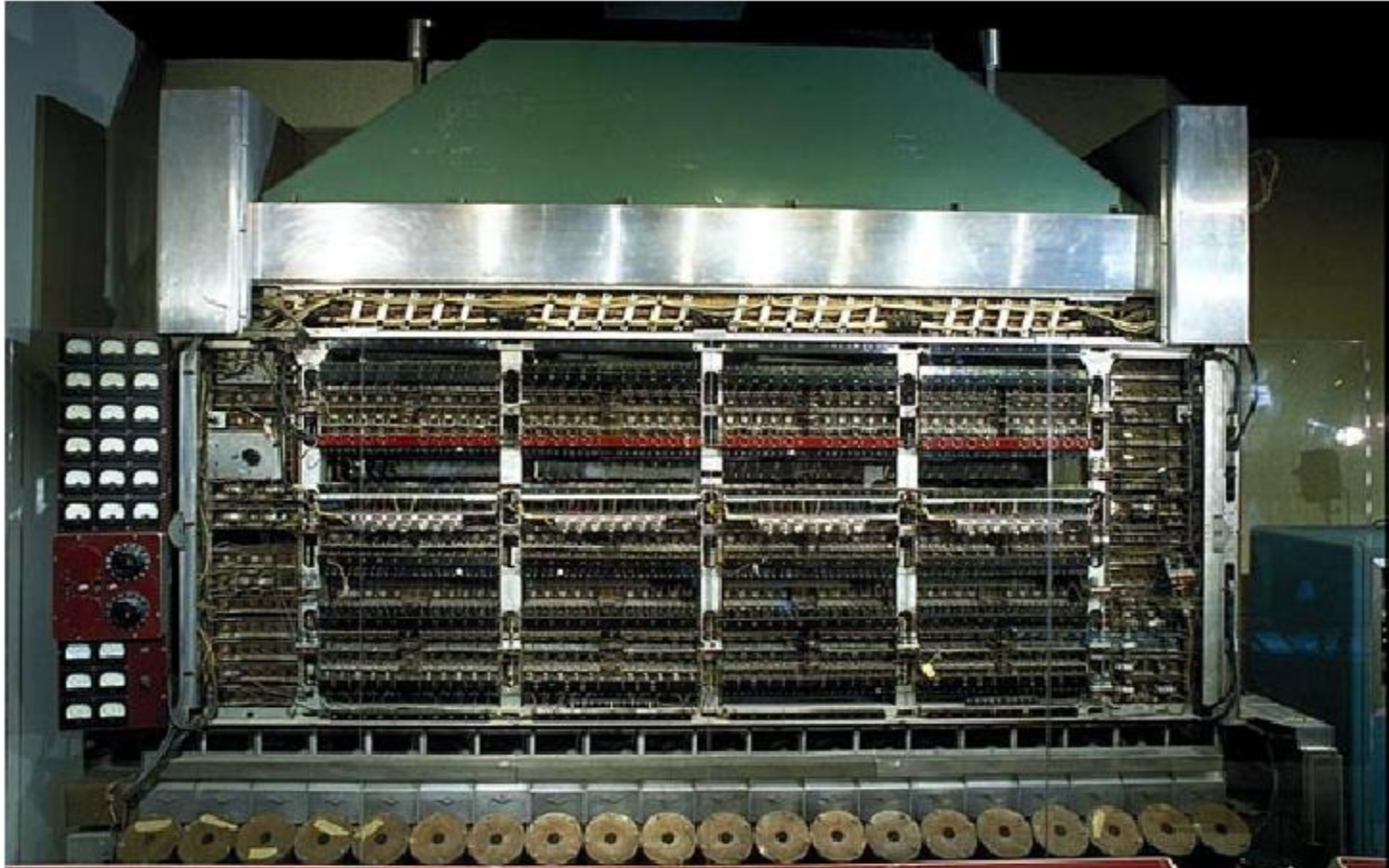


MIMD

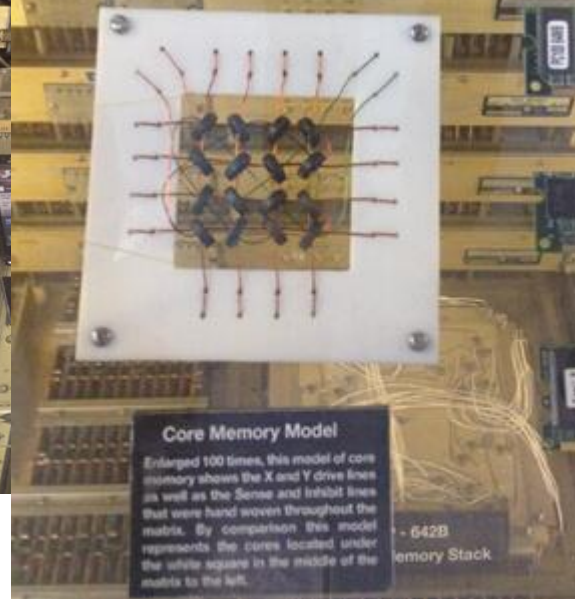
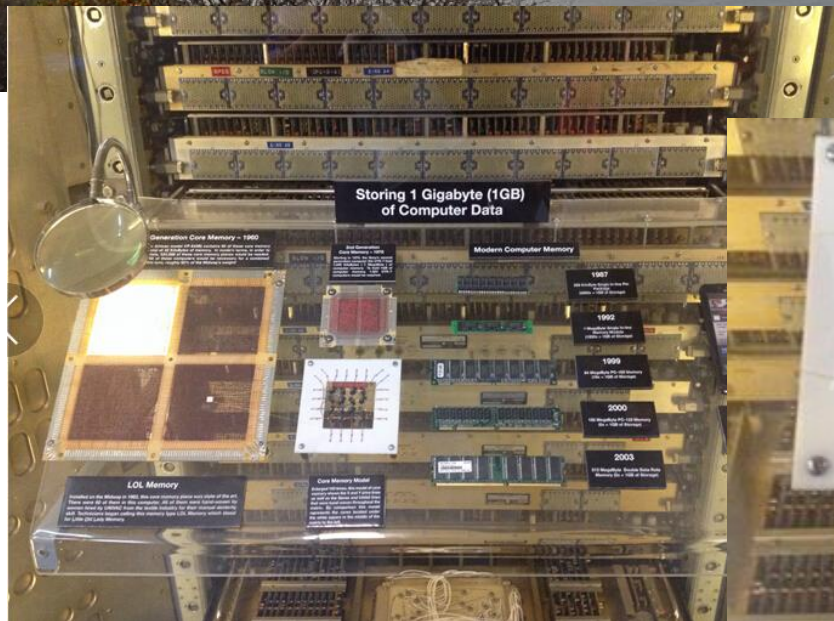
MIMD



I Computers Then



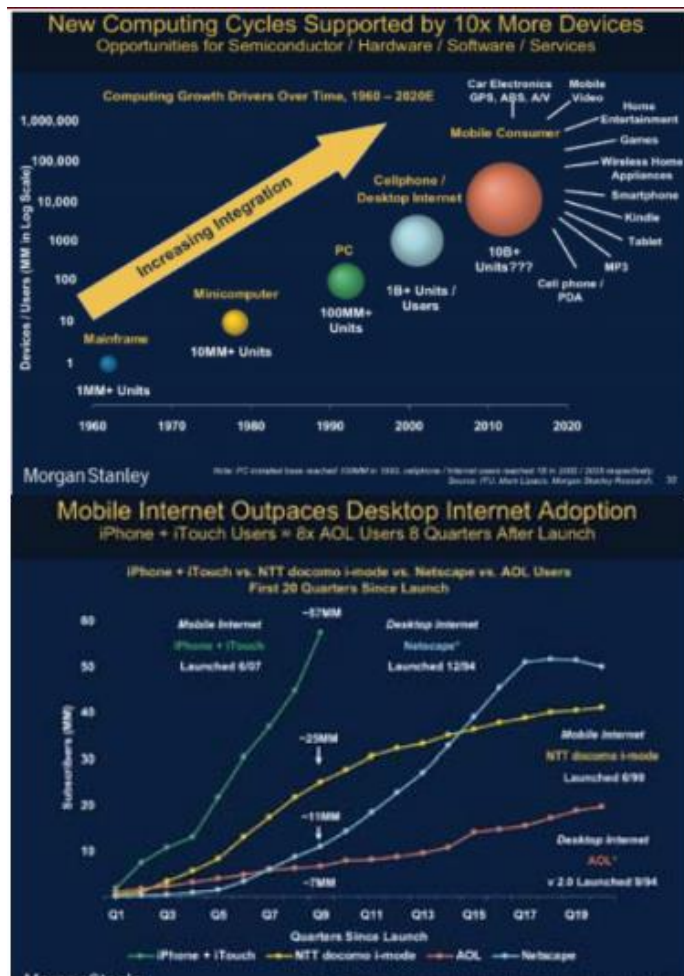
IAS Machine. Design **directed by John Von Nuemann.**
First booted in Princeton NJ in 1952
Smithsonian Institution Archives (Smithsonian Image 95-06151)



I Computers Now

- Sensor Networks
- Cameras
- Smartphones
- Mobile Audio Players
- Laptops
- Autonomous Cars
- Servers
- Game Players
- Routers
- Flying UAVs
- GPS
- eBooks
- Tablets
- Set-top Boxes

Computers Now



The **Emerging IT Scene** and The Emerging Computing Platform

Better Computer-Architectures are **NEEDED** !

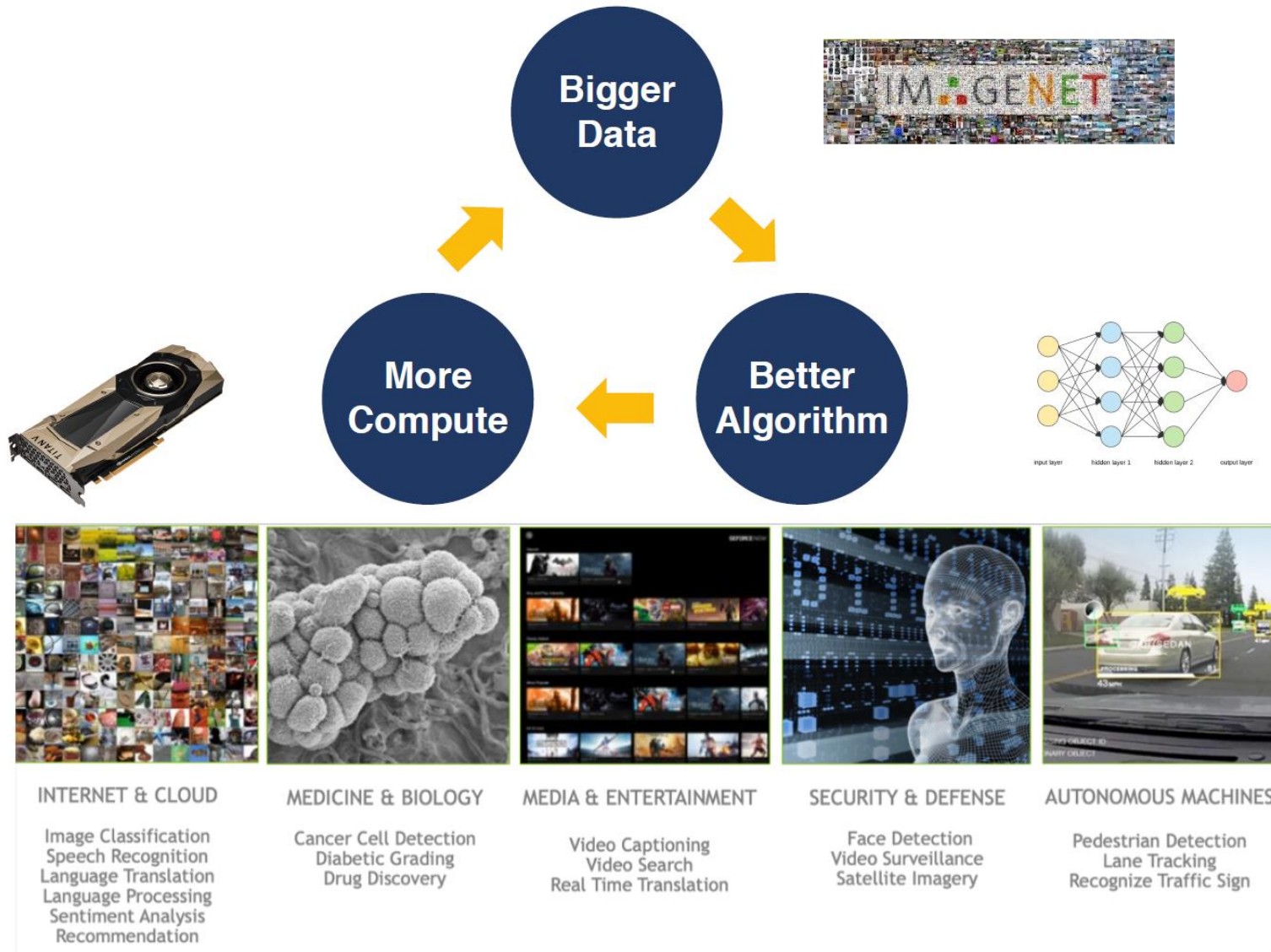
36/51

- We need better computer architectures to fill the gaps
 - **Architecture**
 - von Neumann Architecture & Harvard Architecture
 - VLIW, SuperScalar
 - ?
 - **Instruction Set**
 - CISC, RISC
 - ?
 - **CPU & Memory**
 - Hierarchy of Memory: Cache, Memory, Disk, ...
 - ?
 - **Advanced Peripheral Devices**

- **Pipeline**
 - Separate the execution of an instruction into several stages using separate resources
 - Overlap the execution of multiple instructions
 - e.g., we can have 5 stages: IF → ID → EX → MEM → WB
- **Branch Prediction**
- **Data Flow Analysis**: optimal instruction schedule
- **Speculative Execution**
- **Performance Balance**: Adjust organization and architecture to compensate for the mismatch among various components.

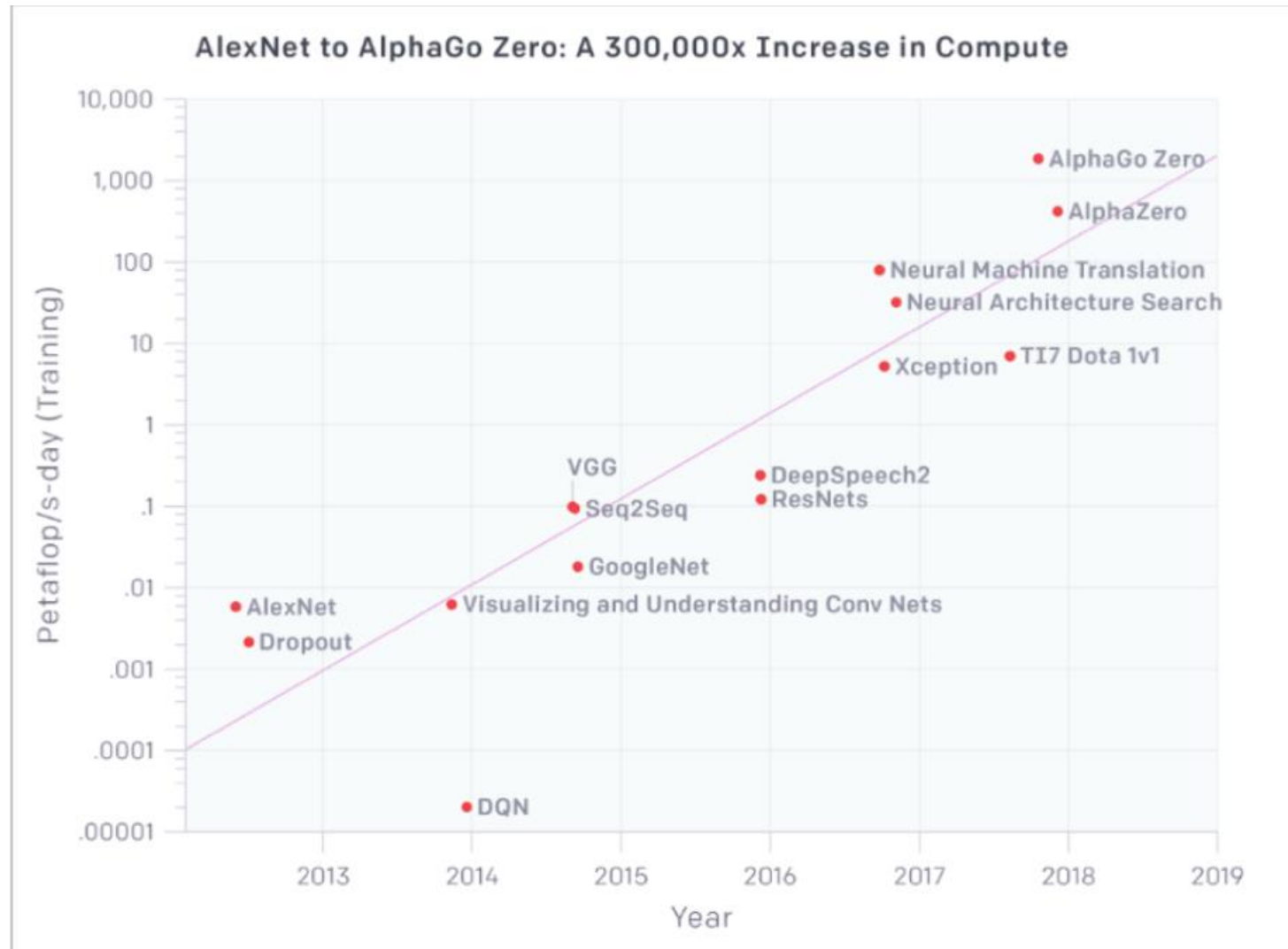
- Use **wide data buses** so we can retrieve more bits at the same time.
- Include a **cache /other buffer scheme** to make memory chip work more efficiently.
- Put **cache into processors**
- Use **high-speed buses** to interconnect processor and memory.
- **Near Data Computing, Processing in Memory.....**

Artificial Intelligence



Why AI Chip

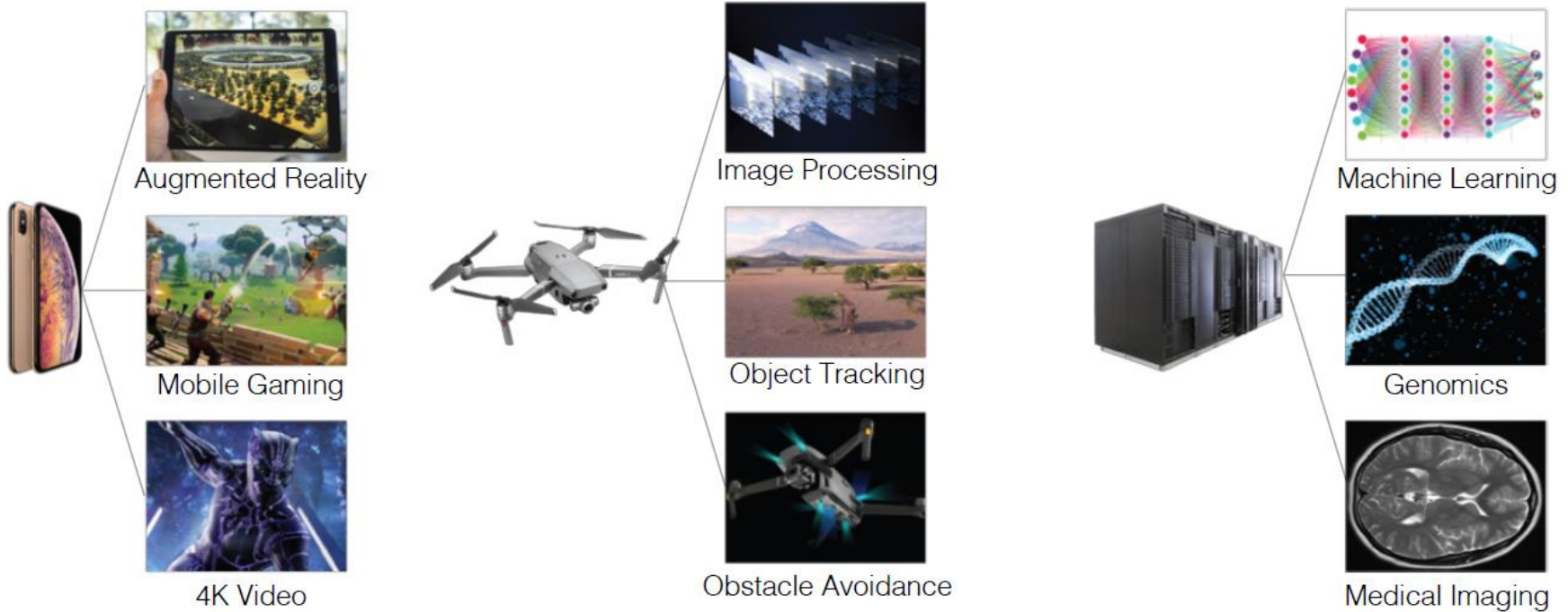
40/51



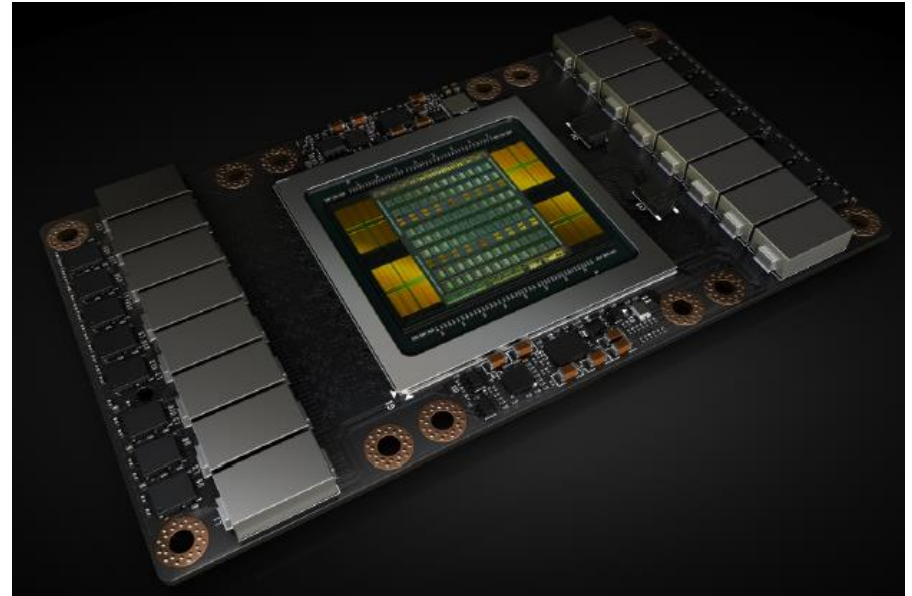
Why AI Chip

41/51

Increasing Demand for Computing



- Volta V100 GPU
 - 21 billion transistors
 - Die size 815 mm²
 - TSMC 12 nm FinFET
 - 15.7 TFLOP/s of single precision (FP32) performance
 - 125 Tensor TFLOP/s of mixed-precision matrix-multiply-and-accumulate
 - TDP 300W



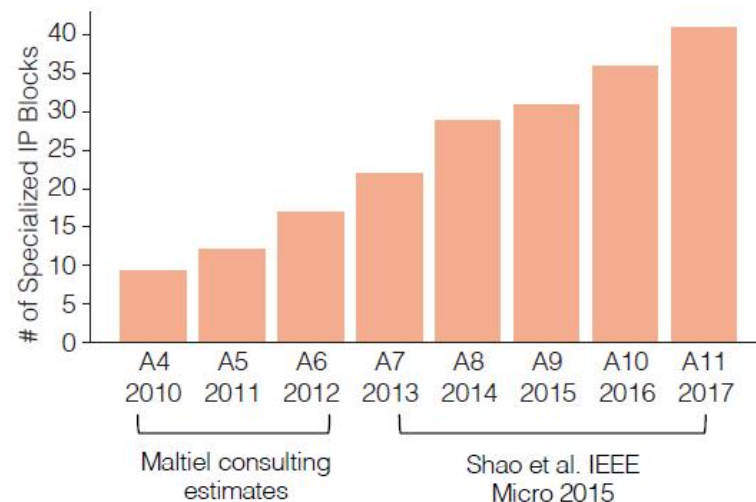
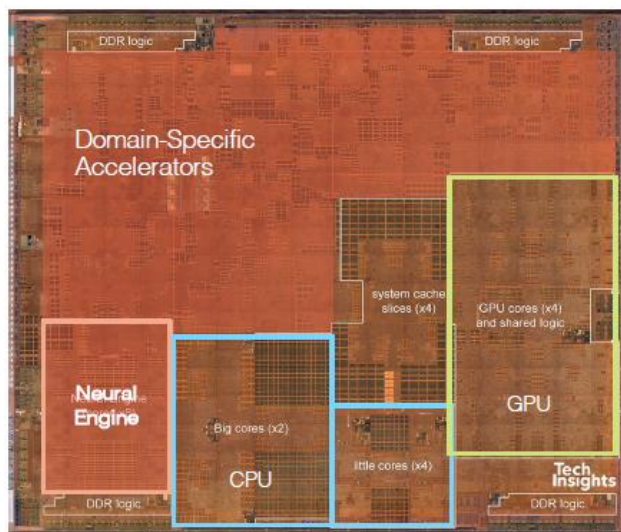
Domain-Specific Accelerators

43/51

Customized hardware designed for a domain of applications



iPhone XS,
2018

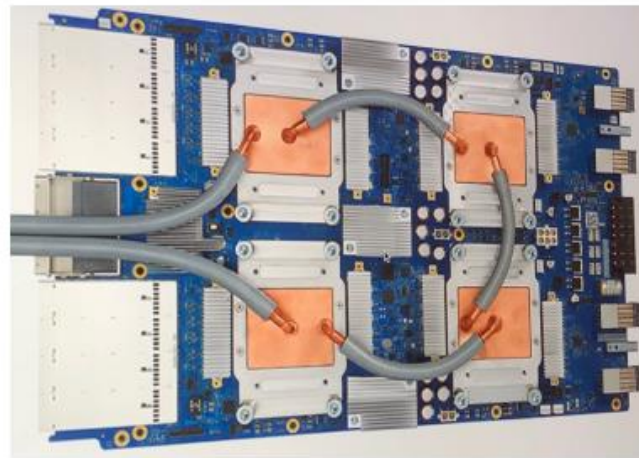


Domain-Specific Accelerators

44/51

Google TPU

- Systolic-array-based architecture
 - V1: Inference only
 - V2: Training with bfloat
 - V3: 2x powerful than v2
- Edge TPU
 - Coral Dev Board
 - 4 TOPS
 - 2 TOPS/Watt
 - Supports TensorFlow Lite

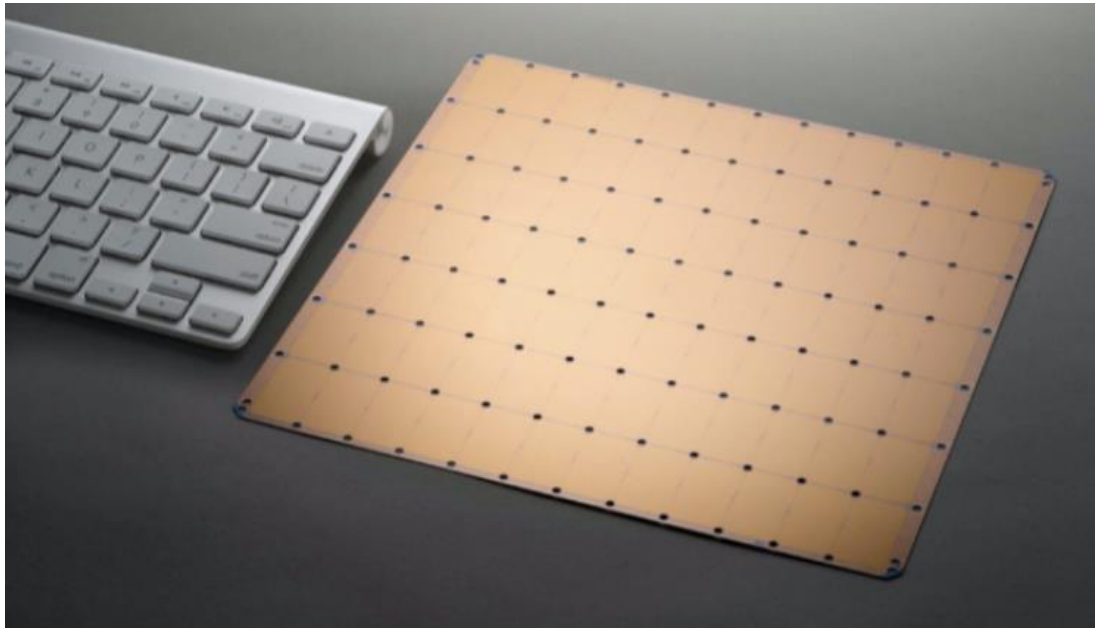


Domain-Specific Accelerators

45/51

Cerebras: Wafer-Scale Deep Learning

- Largest Chip Ever Built!
- 46,225 mm² silicon
- 1.2 trillion transistors
- 400,000 optimized AI cores
- 18 GB of on-chip memory
- TSMC 16nm process



Industry is Booming

46/51



Four Types of Hardware Platforms

47/51



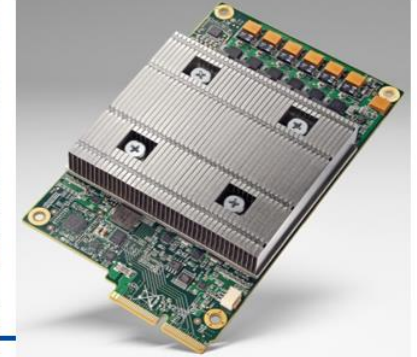
CPU



GPU



FPGA



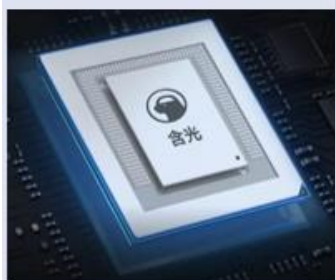
ASIC

Flexibility

Efficiency

Five Types of Accelerator Chips

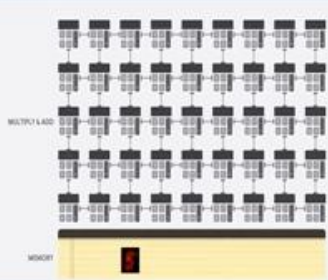
48/51



Instruction-based

High Generality

寒武纪
阿里含光
华为昇腾



Dataflow

High Utilization

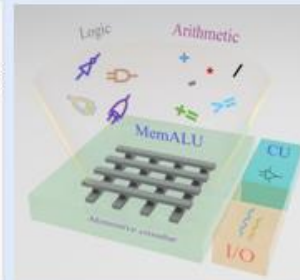


Reconfigurable

**Balance Flexibility
and Efficiency**

Usually Combined

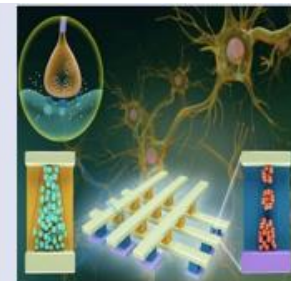
鲲鹏CAISA
Sambanova



In-Memory
Computing

High Bandwidth

IBM相变存内计算
台积电RRAM
英特尔NOR Flash



Brain-Inspired

**AI Chip for
Next Generation**

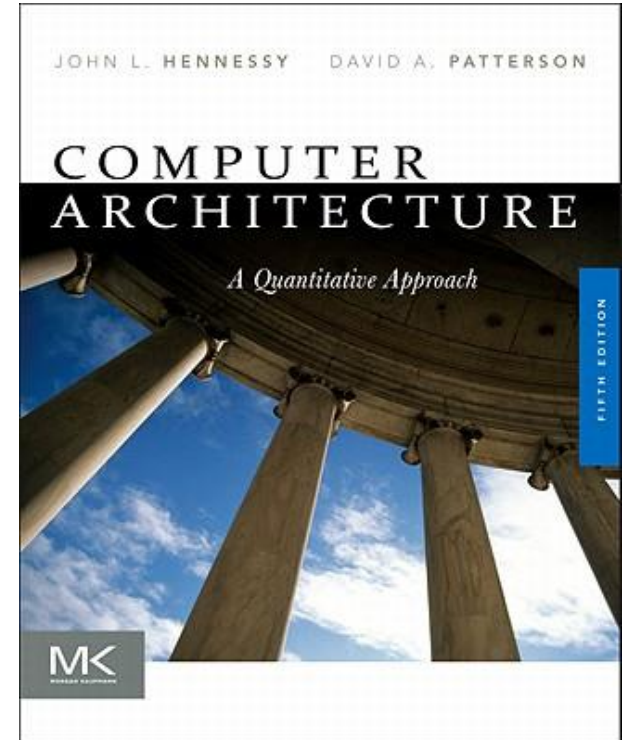
清华大学天机
IBM TrueNorth
英特尔 Loihi

- Technique advance made **dramatic performance improvement** in the past 40 years
 - **Processor**
 - logic capacity: increases about **30%** per year
 - performance: **2x** every **1.5** years
 - **Memory**
 - DRAM capacity: **4x** every **3** years, now **2x** every **2** years
 - memory speed: **1.5x** every **10** years
 - cost per bit: decreases about **25%** per year
 - **Disk**
 - capacity: increases about **60%** per year
- **Moore's Law will continue work in the near future?**
- **Better architectures are needed for better usage of IC capacity**

Assigned Readings

50/51

- **Computer Architecture: A Quantitative Approach, 5th ed.,** John L. Hennessy and David A. Patterson, Morgan Kaufman, 2011
- Sections: 1.1-1.6, Appendix L



Further Readings

51/51

- Martin Davis. “Engine of Logic”. Norton, 2000
- Hermann H. Goldstine. “The Computer: From Pascal to Von Neumann”. Princeton University Press, 1972
- Andrew Hodges. “Alan Turing: The Enigma”. Walker & Company, 2000
- Eloina Pelaez. “The Stored-Program Computer: Two Conceptions”. In Social Studies of Science 29(3), June 1999.
- Computer History Museum. <http://computerhistory.org>