计算物理

Lecture 1
傅子文

Outline of Lecture1

- Fundamentals of computer design
  - von Neumann architecture
    - Memory
    - Process Unit
    - Control Unit
  - Instruction cycle
  - Five generations of computer design
- Moore's Law

Definition of a computer

- Definition has evolved out of Turing's (circa 1930s) discussions:
  - Takes input
  - Produces output
  - Capable of processing input somehow
  - Must have an information store
  - Must have some method of controlling its actions

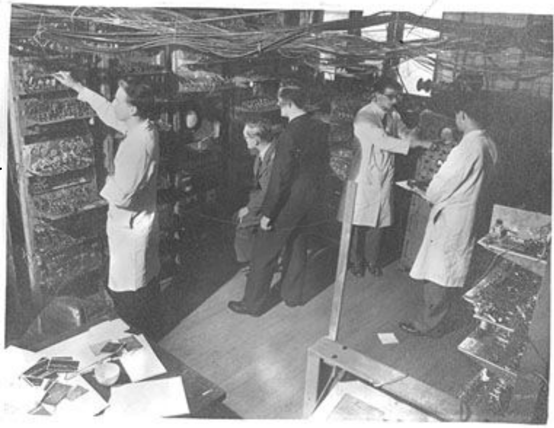
Birth of Electronic Computing

- Prior to 1940 a computer was assumed to be a person
- Turing (1936) definition of the Turing machine (formalized notion of algorithm execution)
- Atanasoff-Berry (1938) & Zuse (Z-series 1941) both developed electronic computing machines with mechanical assistance (e.g. rotating memory drums)
- Colossus (UK) first all electronic configurable computer (1941)
- ENIAC (1945) arguably first fully programmable electronic computer

First generation of computers – used vacuum tubes

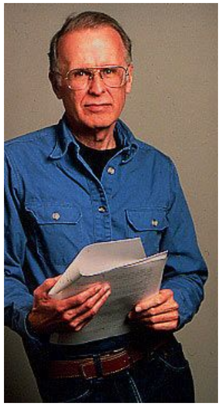
Key events, 1953-1974

- 1953: First transistor based computer (Manchester MkI)

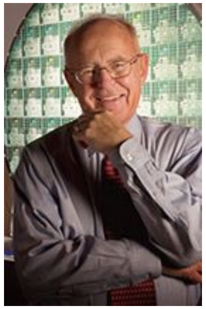


- 1954: FORmula TRANslation language (first high level language) developed at IBM by Backus

2nd Generation (1953-64)



- 1964: IBM System/360, first integrated circuit (IC) computer
- 1965: Moore's Law published by Gordon Moore

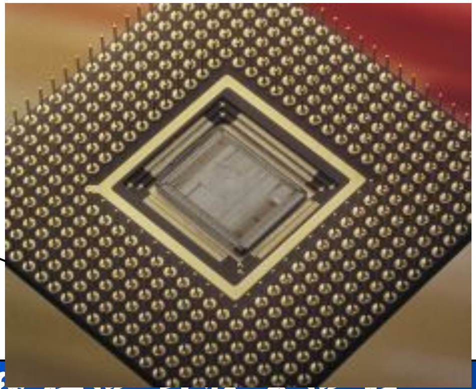


3rd Generation (1964-72)

1972-present

- 1972: 8008 microprocessor released by Intel, uses very large scale integration on IC (3500 transistors, 8 bits)
- C programming language developed by Ritchie this year
- 1985: Intel 386, first 32bit Intel processor
- 1992: First commercially available 64bit processor, DEC Alpha 21064

4th Generation
(1972-now)



Fifth Generation?

- All modern computers are similar in principle to the 8008 – just scaled up in size enormously

- Some historians claim that the fifth generation of machines is constituted by the appearance of parallel computers in the early 1990's
  - -Japan invested heavily in such machines in the 1980s

- Others argue that the appearance of AI based computers will be the next step?

Humour: Famous computer gaffes

- 1943: “I think there is a world market for maybe five computers” Thomas Watson (IBM Chairman)
- 1950: “We’ll have to think up bigger problems if we want to keep them busy” Howard Aiken
- 1957: “I have travelled the length and breadth of this country and talked with the best people, and I can assure you that data processing is a fad that won’t last out the year” Prentice Hall business books Editor
- 1977: “There is no reason anyone would want a computer in their home” Ken Olsen (DEC Chairman)
- 1981: “640k ought to be enough for anybody” Bill Gates

Latency and Bandwidth

- ***Latency***
  - The time taken for a message to get from A to B
  - Typically used when discussing the time taken for a piece of information in memory to get to the central processing unit (CPU)

- ***Bandwidth***
  - The amount of data that can be passed from point A to B in a fixed time
  - Typically used when discussing how much information can be transferred from memory to the CPU in a second

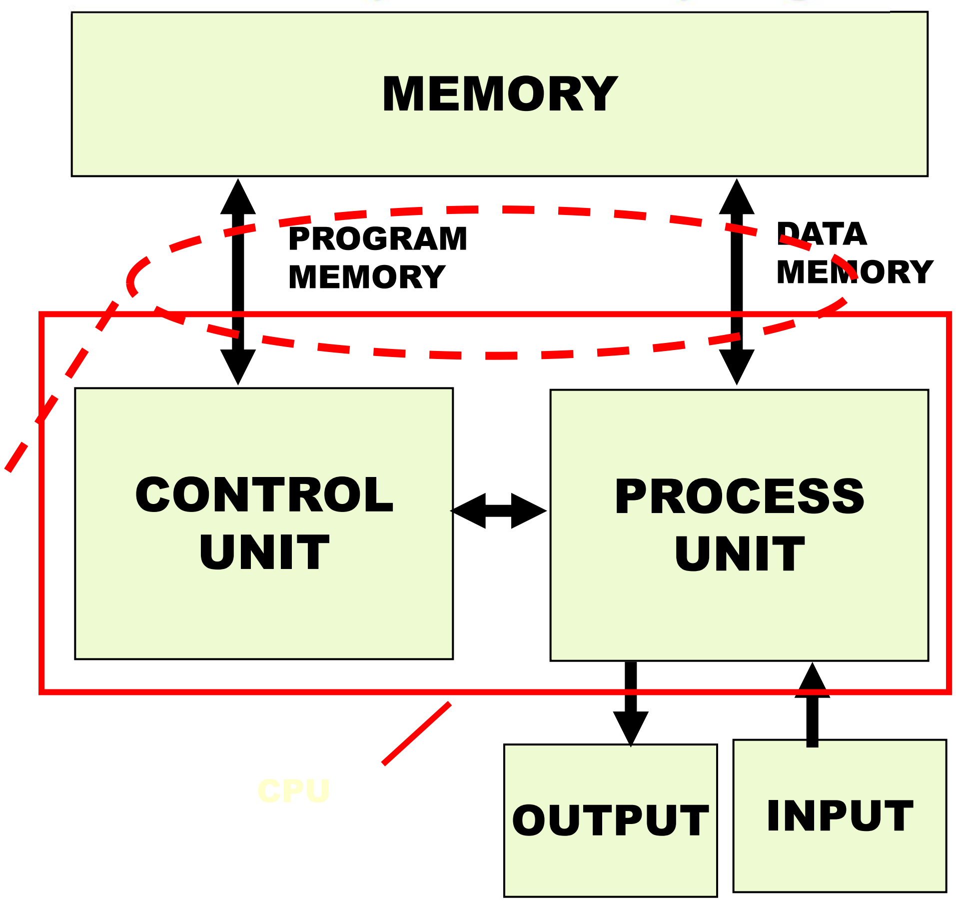
Bits, Bytes & Nybbles

- Just to recap: 1 bit is a zero or one, usually represented (but not always) by a lower case ‘b’
- Byte = 8 bits – usually represented by capital ‘B’
  - Nybble = 4 bits
- Things can be slightly confusing when people talk about bandwith in Mb/s
  - Do they mean megabits per second or megabytes second? Most of the time it is megabytes…

von Neumann architecture

- First practical stored-program architecture
- Still in use today
- Speed is limited by the bandwidth of data between memory and processing unit
  - “von Neumann” bottleneck

Machine instructions are encoded in binary & stored – key insight!



Developed while working on the EDVAC design

Instruction encoding

- Machine instructions are represented by binary codes
- Encoding of instructions is extremely important in CPU design
  - No two instructions can have same value
- Each binary value must be decoded
  - For k-bits in each instruction word require k lines of incoming circuitry
  - However, there are 2^k possible instructions, so one naively requires 2^k output lines
  - Gets expensive very quickly!
- *Instruction Set Architecture* (ISA): A computer's instructions and their formats

von Neumann bottleneck

- This has become a real factor in modern computer design
- In a landmark paper in 1994, Wulf & Mckee pointed out that next generation computers would be seriously limited unless more bandwidth was made available
- Witness the push for higher bandwidth memory in modern computers

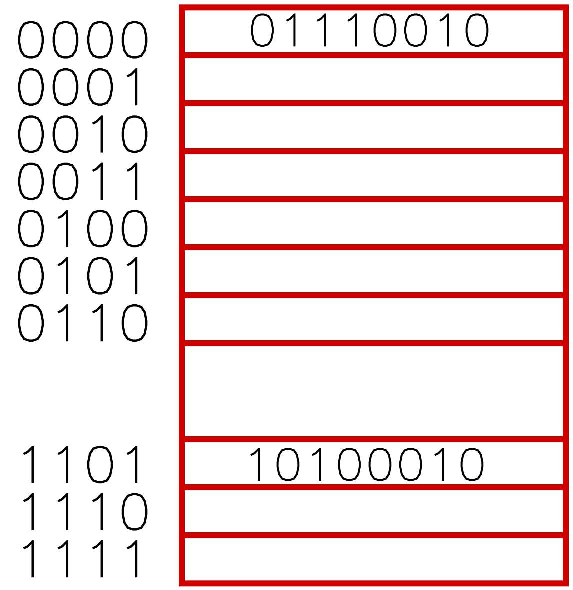
Hitting the Memory Wall: Implications of the Obvious

Wm. A. Wulf
Sally A. McKee
Department of Computer Science
University of Virginia
{wulf | mckee}@virginia.edu

December 1994

Memory

- Memory is divided in cells*, each cell has an *address* and *contents*
  - *Address*: n-bit identifier describing location
    - *Addresses are unique*
  - *Contents*: m-bit value stored at the given address
- k x m array of stored bits (k is usually 2^n)
- Actions:
  - Read a word from memory: LOAD
  - Write a word to memory: STORE



Remember both the program and the data it operates on have to be read from memory

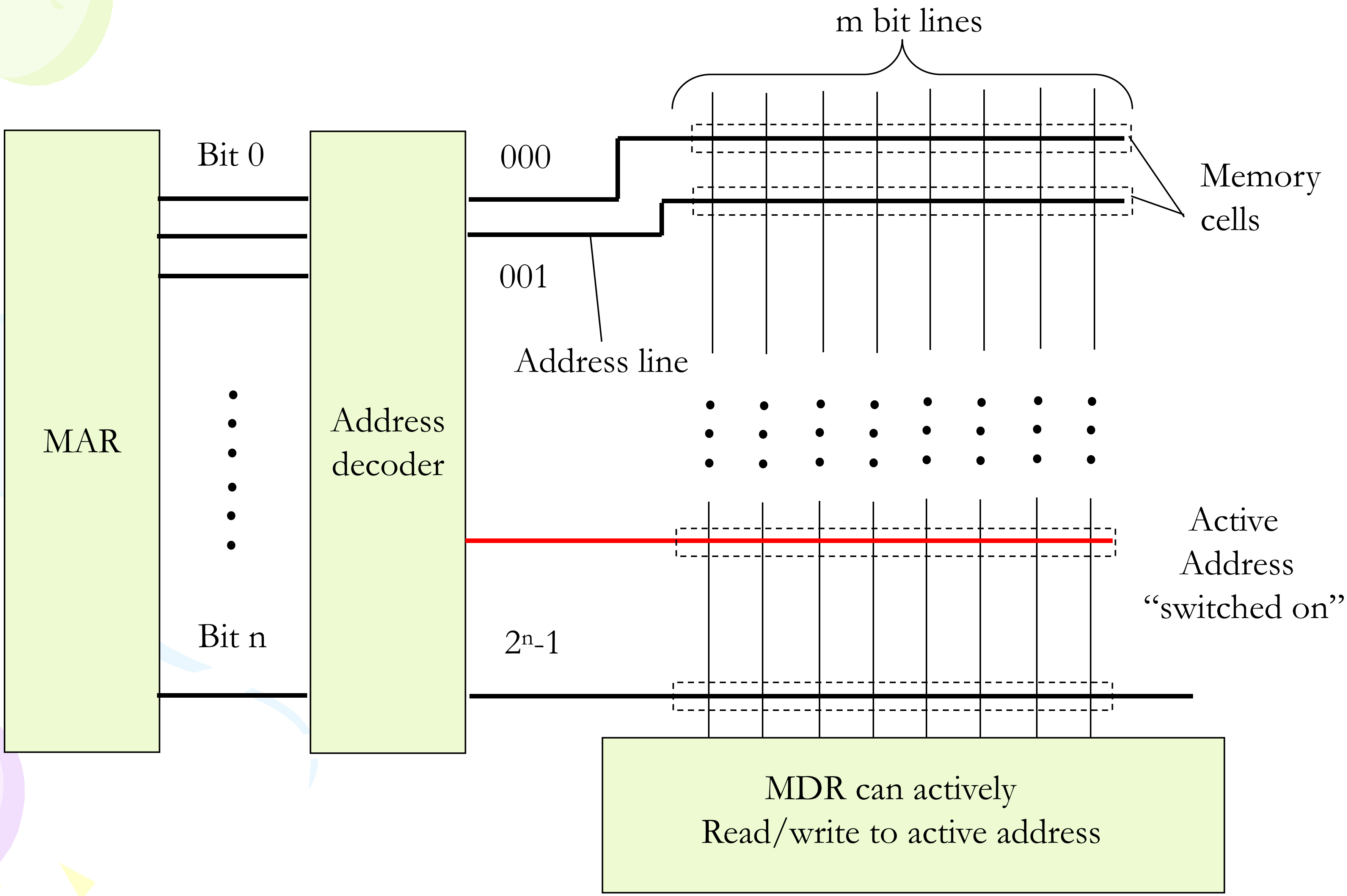
* Technically a cell is a single bit of memory, but we'll use it to mean a single memory unit

Memory Registers

- A ***register*** is an extremely fast memory address, close to the circuitry of the control & process units
  - Small number of special purpose cells – they have a *specific function*
  - Can store addresses, data or instructions

- *Example – registers used in memory access:*
  - *Memory Address Register (MAR)*
    - Holds the address of a cell to access
    - Same size in bits as the address length
  - *Memory Data Register* (MDR)
    - Holds content of cell that was fetched or that is to be stored
    - Same size in bits as the cell contents

MAR/MDR & Memory (logical diagram)



Process Unit

- The PU can consist of many specialized units
  - Arithmetic Logic Unit (ALU)
  - floating point arith. or specialized fast square roots
- Needs some kind of small local storage to hold data & output on, so uses *registers*
  - Some modern CPUs have hundreds of registers
- Data is stored in words, with a distinct word size
  - number of bits normally processed by ALU in one instruction

Control Unit

- Oversees execution of the program
  - Responsible for reading instructions from memory
  - Directly interprets the instruction, and informs execution unit what to do
- Two important parts of the CU
  - Instruction Register (IR) holds the current instruction
  - Program Counter (PC) holds the address of the next instruction to be executed
  - The PC is very useful tool for debugging – tells you exactly what is being done at a given time

Execution sequence

Instruction fetch from memory
PC gives address, stored in IR

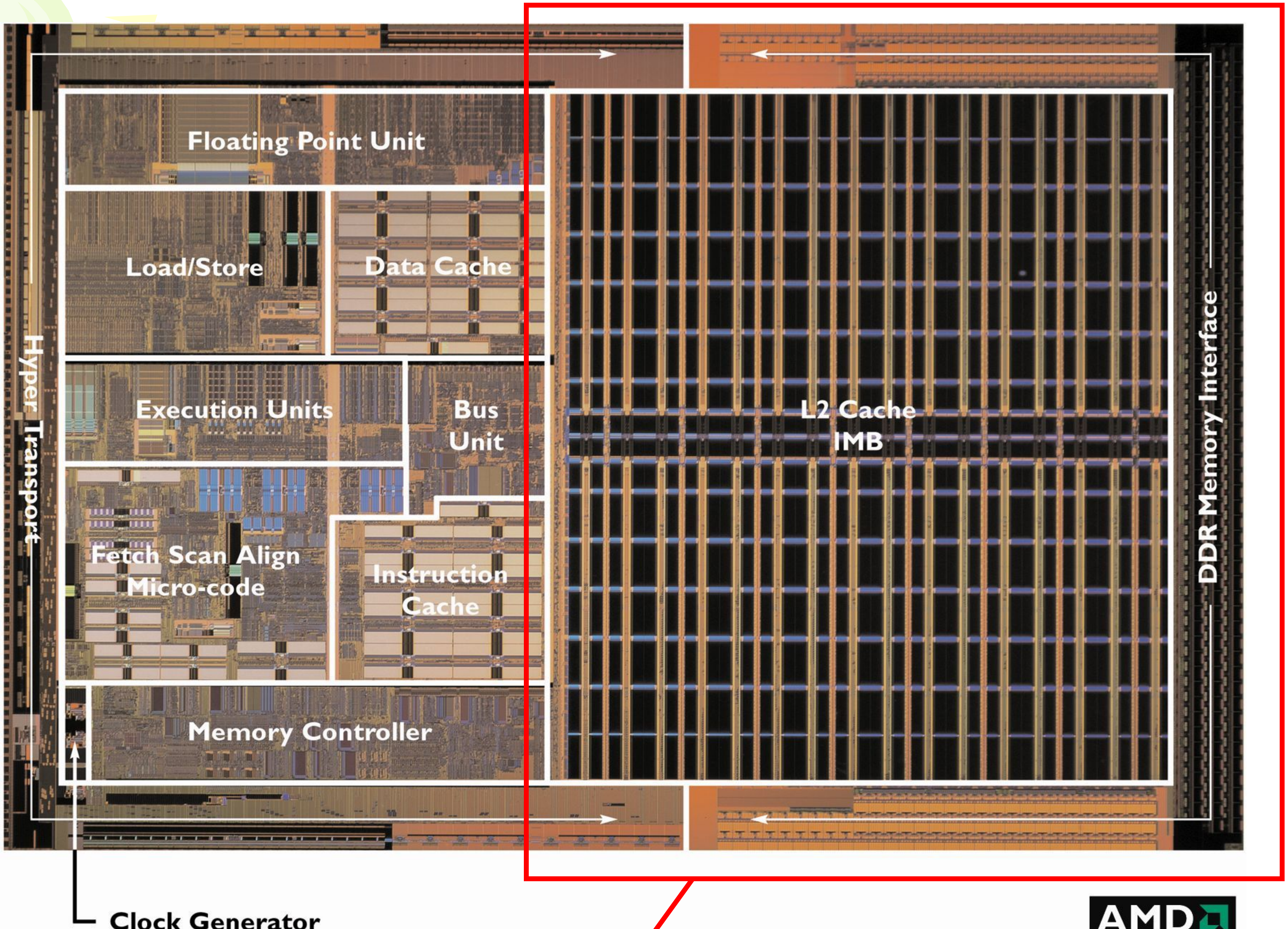
Decode instruction

Evaluate address

Fetch operands from memory

Execute operation on PU

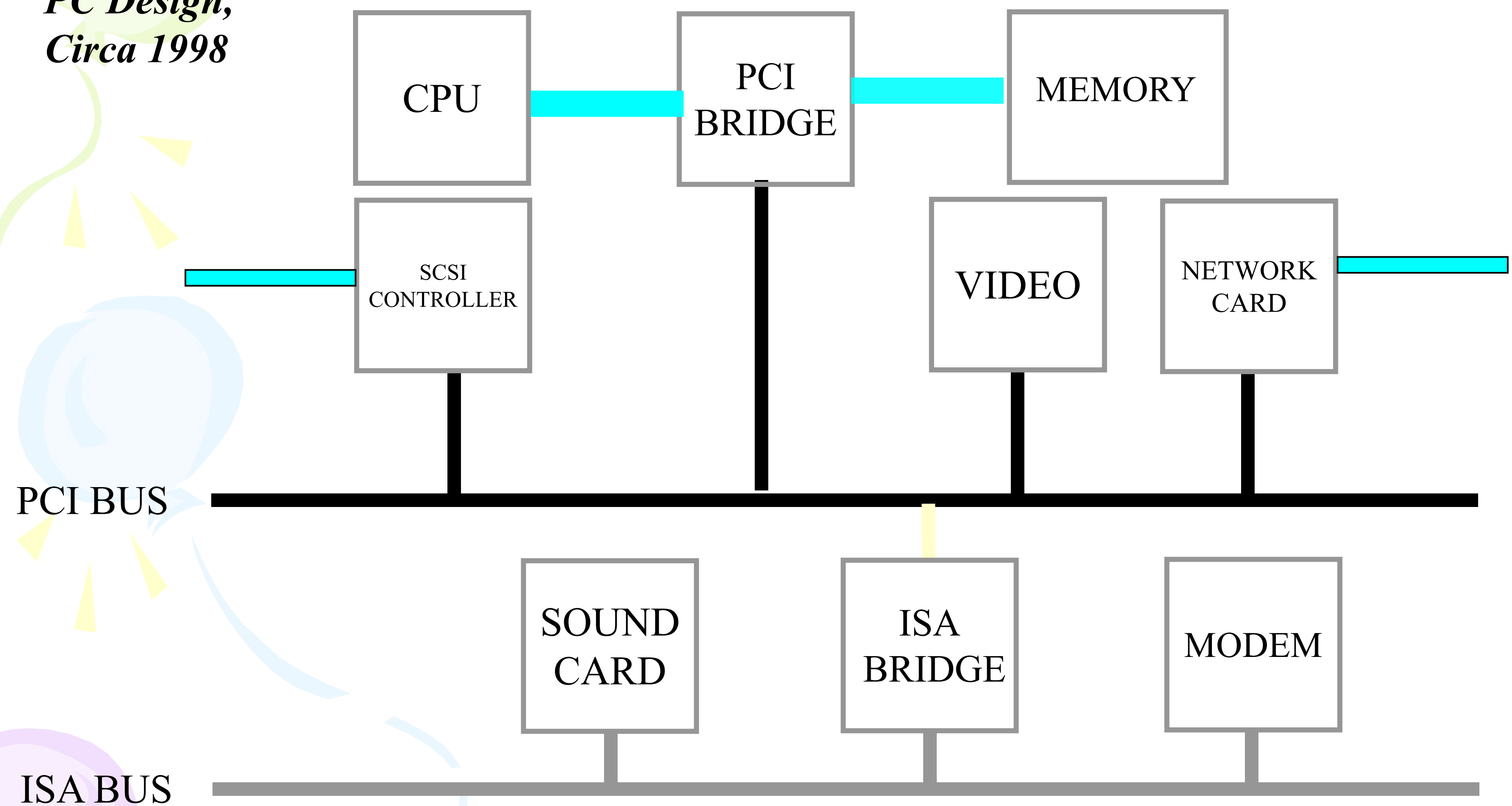
Store result



On die cache is the reason why Moore's Law is still valid

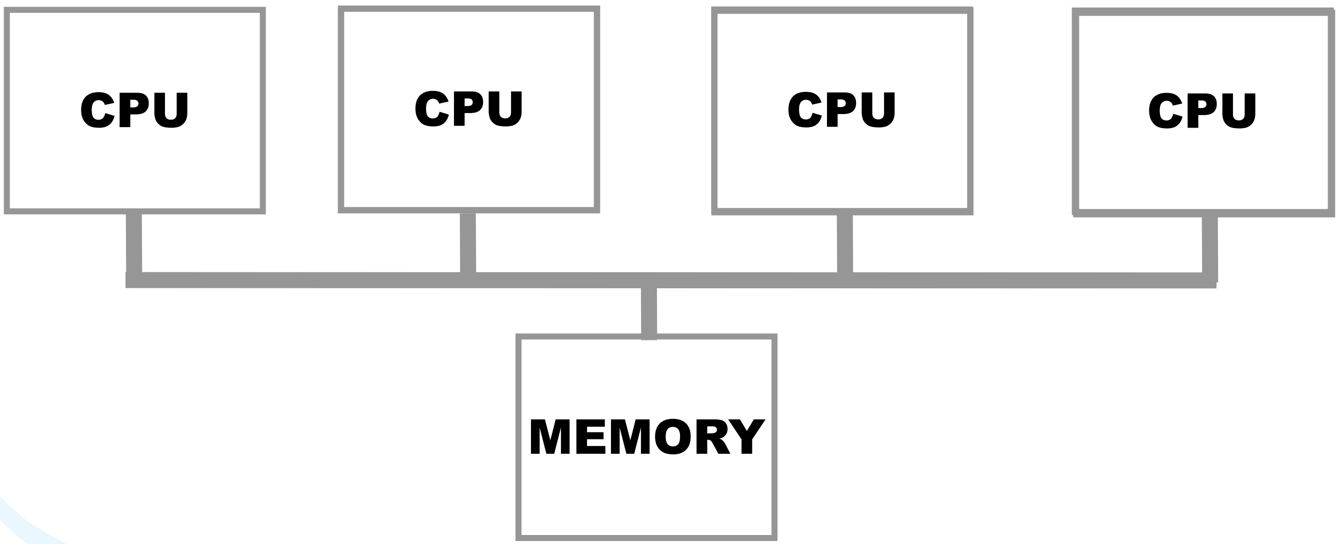
Bus based architectures

PC Design, Circa 1998



A single bus is insufficient for modern machines – multiple buses are used to manage both I/O and data processing

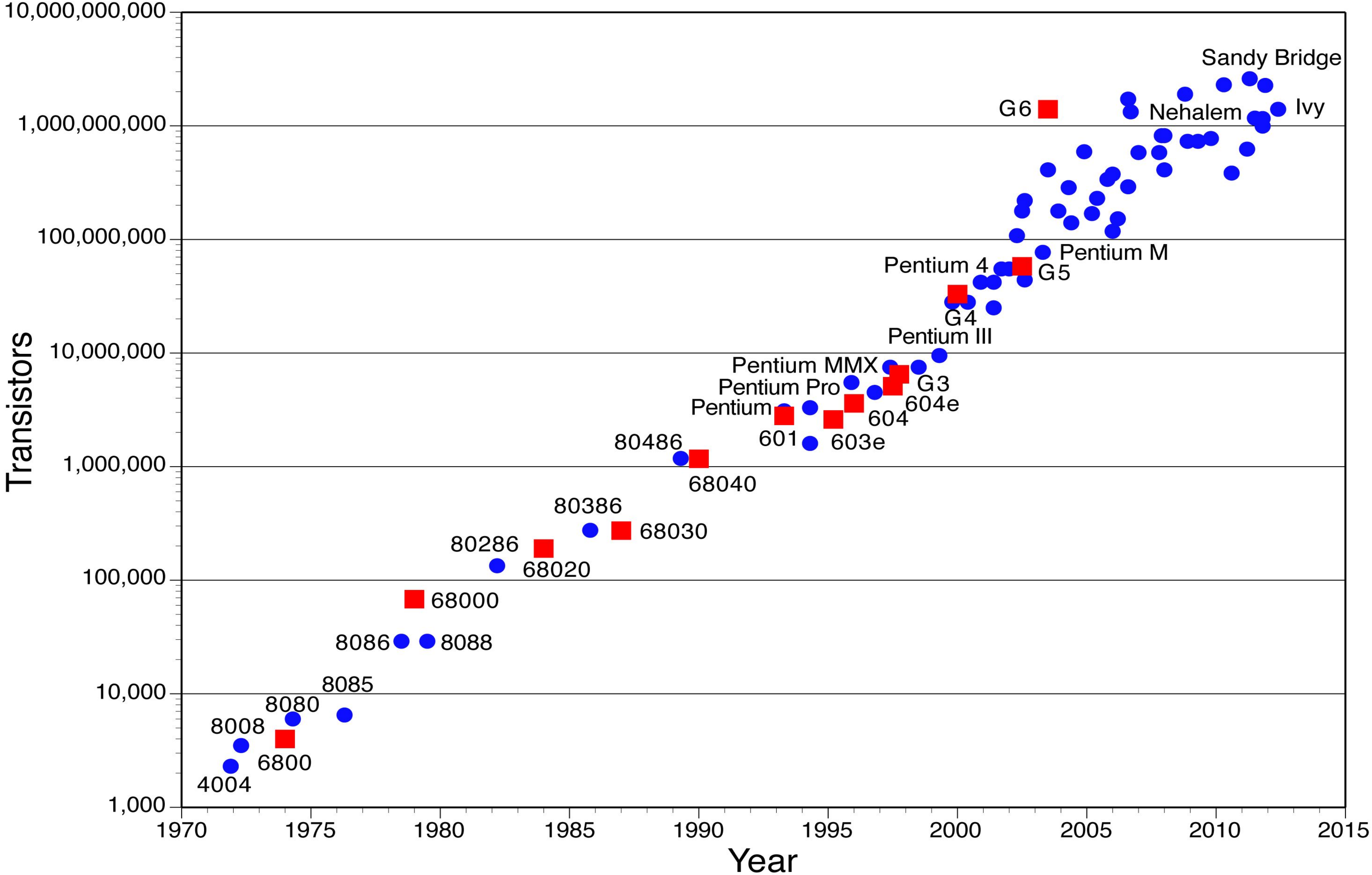
(Symmetric) Multiprocessor machines

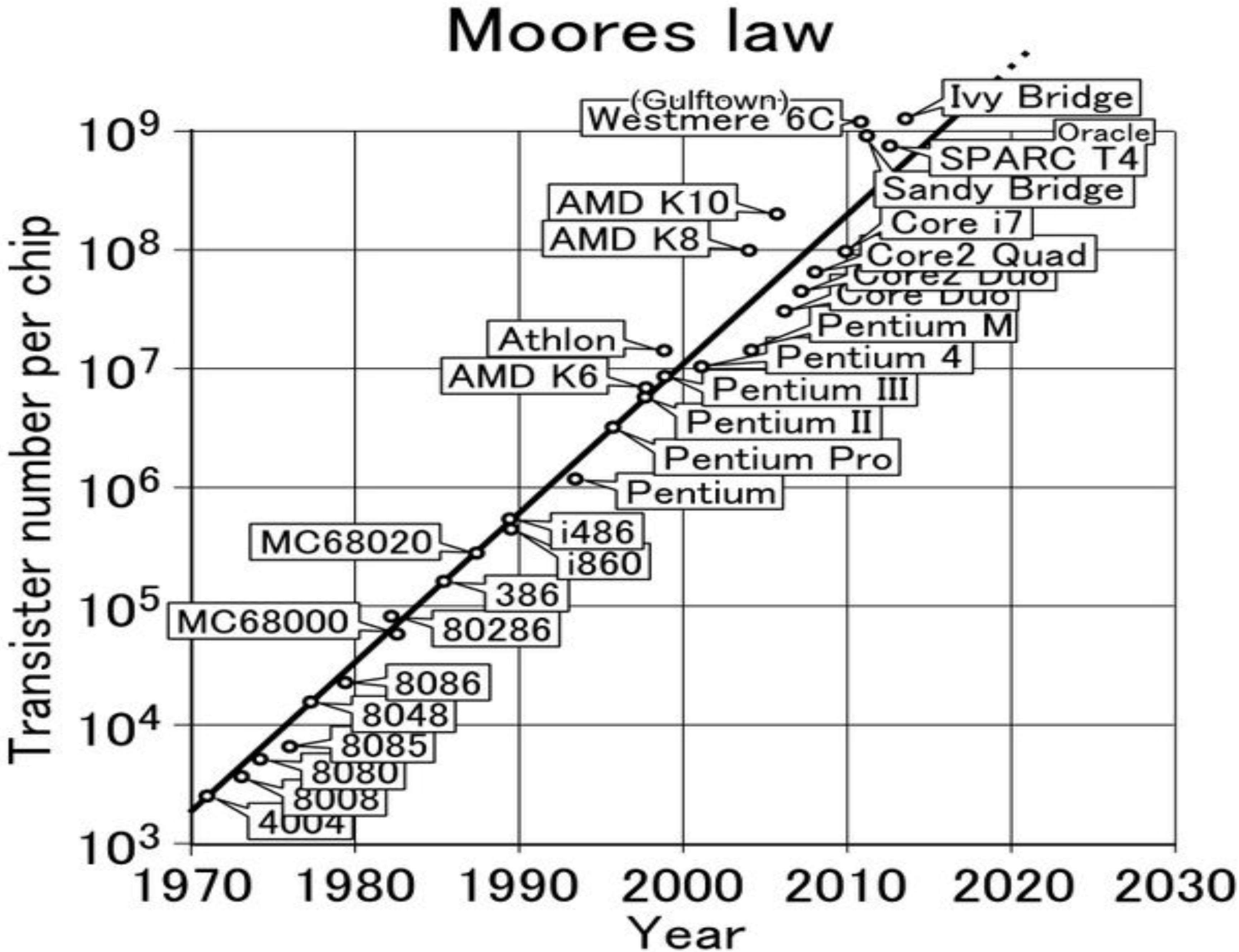


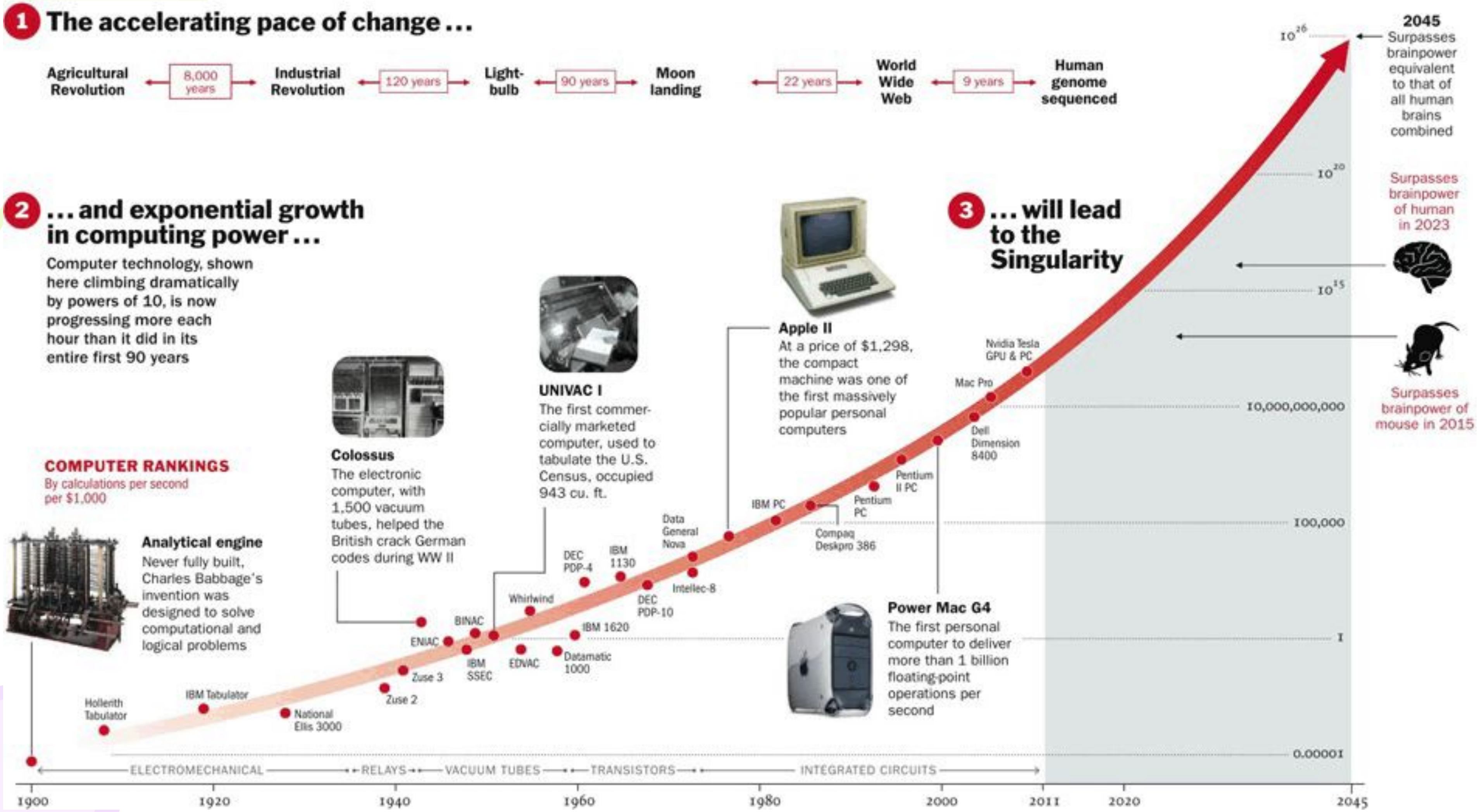
Traditional shared memory design – all processors share a memory bus

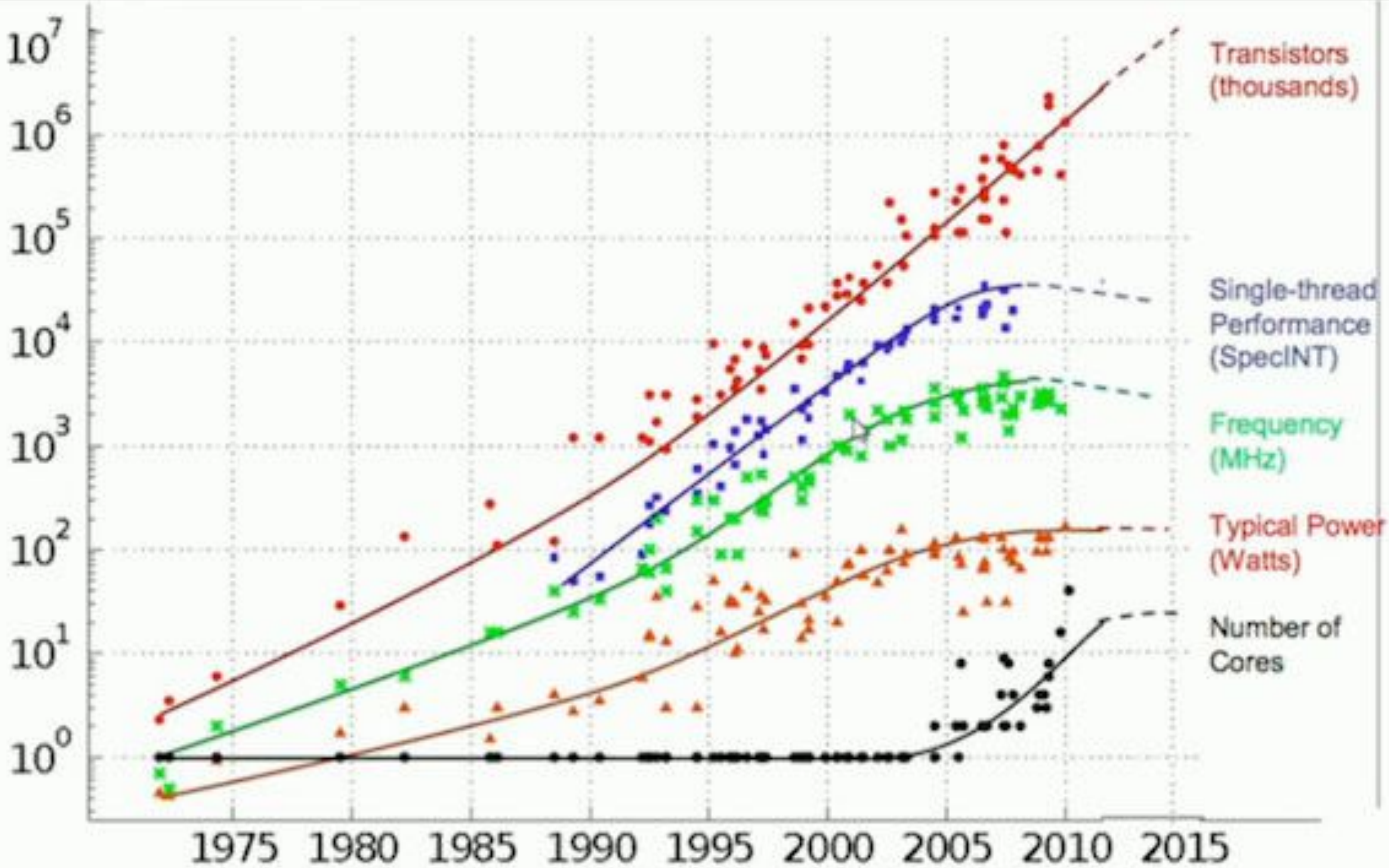
Becoming less popular – extremely hard to design buses with sufficient Bandwidth, although smaller 2-4 processor Intel systems use this design

Moore's Law –transistor *counts* grow exponentially




Moores law
Transister number per chip
Year
10^9
10^8
10^7
10^6
10^5
10^4
10^3
1970
1980
1990
2000
2010
2020
2030
(Gulftown)
Westmere 6C
Ivy Bridge
Oracle
SPARC T4
Sandy Bridge
AMD K10
AMD K8
Core i7
Core2 Quad
Core2 Duo
Core Duo
Pentium M
Athlon
Pentium 4
AMD K6
Pentium III
Pentium II
Pentium Pro
Pentium
i486
MC68020
i860
386
MC68000
80286
8086
8048
8085
8080
8008
4004

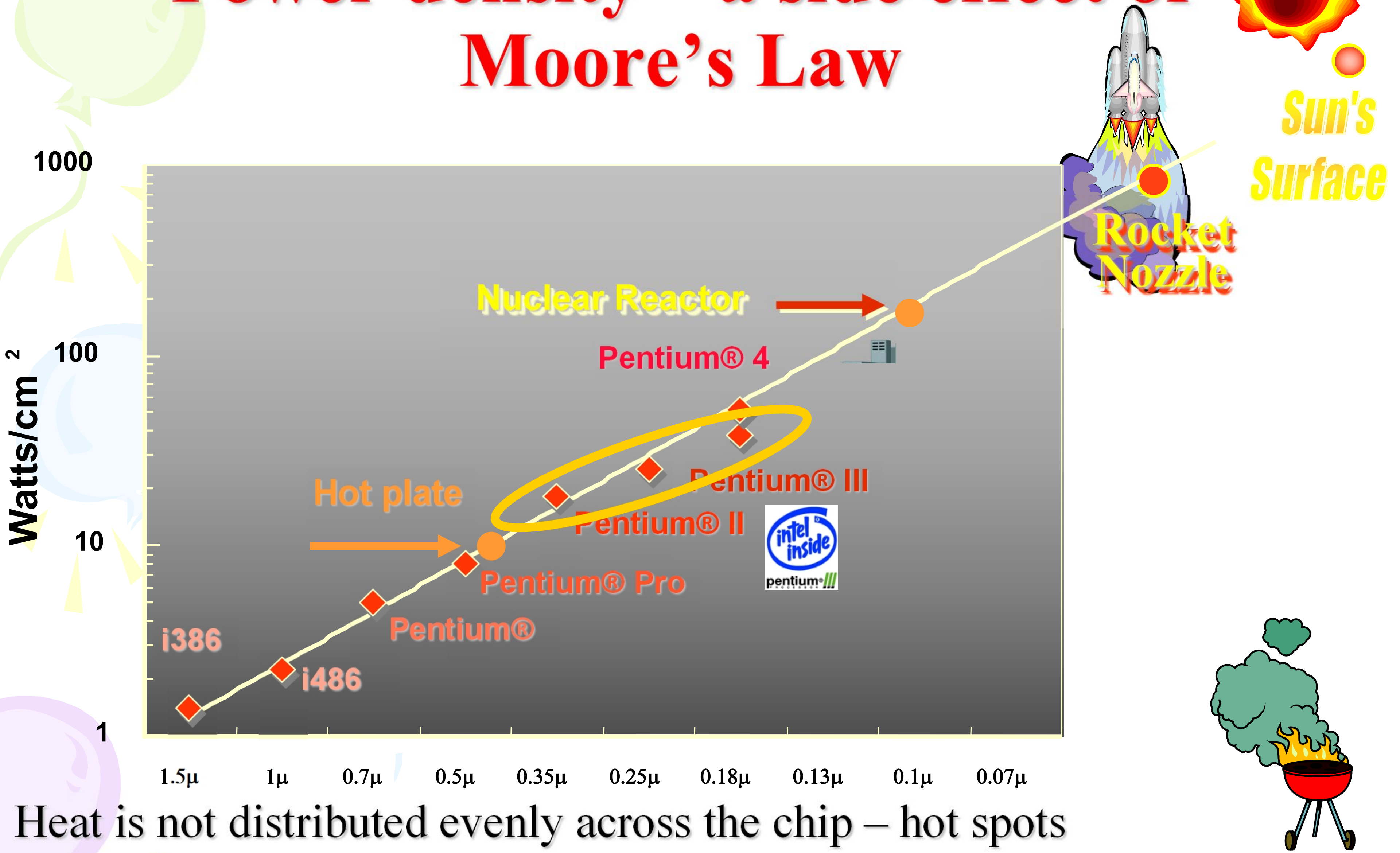

1 The accelerating pace of change …
Agricultural Revolution
8,000 years
Industrial Revolution
120 years
Light-bulb
90 years
Moon landing
22 years
World Wide Web
9 years
Human genome sequenced
2045
Surpasses brainpower equivalent to that of all human brains combined
2 … and exponential growth in computing power …
Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years
3 … will lead to the Singularity
Surpasses brainpower of human in 2023
Surpasses brainpower of mouse in 2015
COMPUTER RANKINGS
By calculations per second per $1,000
Analytical engine
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems
Colossus
The electronic computer, with 1,500 vacuum tubes, helped the British crack German codes during WW II
UNIVAC I
The first commercially marketed computer, used to tabulate the U.S. Census, occupied 943 cu. ft.
Apple II
At a price of $1,298, the compact machine was one of the first massively popular personal computers
Power Mac G4
The first personal computer to deliver more than 1 billion floating-point operations per second
Hollerith Tabulator
IBM Tabulator
National Ellis 3000
Zuse 2
Zuse 3
ENIAC
BINAC
IBM SSEC
EDVAC
Whirlwind
UNIVAC I
IBM 1620
Datamatic 1000
DEC PDP-4
IBM 1130
DEC PDP-10
Data General Nova
Intellec-8
IBM PC
Compaq Deskpro 386
Pentium PC
Pentium II PC
Dell Dimension 8400
Mac Pro
Nvidia Tesla GPU & PC
10^{26}
10^{20}
10^{15}
10,000,000,000
100,000
1
0.00001
ELECTROMECHANICAL
RELAYS
VACUUM TUBES
TRANSISTORS
INTEGRATED CIRCUITS
1900
1920
1940
1960
1980
2000
2011
2020
2045



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Graph by Dr Avi Mendelson, Intel

Power density – a side effect of Moore's Law



Heat is not distributed evenly across the chip – hot spots

Power consumption is already a significant market factor

Leakage?

You don't know where the electron is anymore.

Pentium chip today has a layer almost down to 20 atoms across, When that layer gets down to about 5 atoms across, it's all over The quantum theory takes over. The Heisenberg Uncertainty Principle says you don't know where that electron is anymore, meaning it could be outside the wire, outside the Pentium chip, or inside the Pentium chip.

So there is an ultimate limit set by the laws of thermal dynamics and set by the laws of quantum mechanics as to how much computing power you can do with silicon.

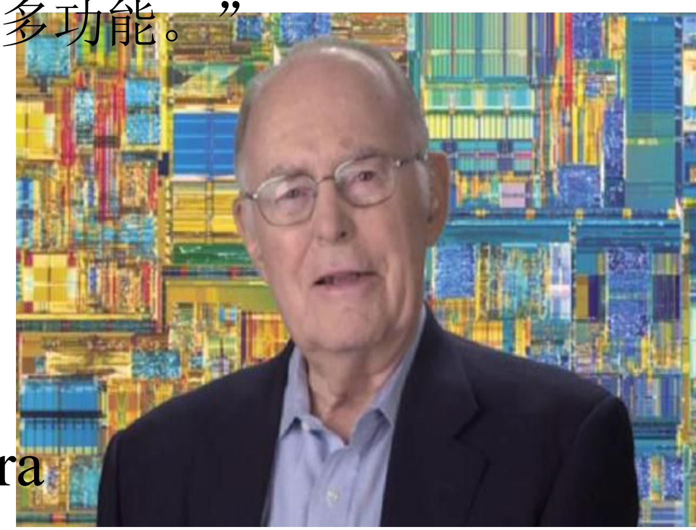
What's beyond silicon? there are a number of proposals:

① protein computers, ②DNA computers, ③optical computers, ④**quantum computers**, ⑤**molecular computers**

End of Moore’s Law in 10 years?

Dr. Michio Kaku says “if I were to put money on the table I would say that in the next ten years as Moore’s Law slows down, we will tweak it. BigThink.com *Computer power simply cannot maintain its rapid exponential rise using standard silicon technology.*

英特尔去年曾表示，与45年前一样，摩尔定律今天依旧有效。特尔高级院士兼制程架构与集成部门总监马博(Mark Bohr)称：“摩尔定律依旧重要，我们将继续遵循该定律。”英特尔在其网站上也称：“英特尔遵循摩尔定律已40多年，为芯片以更低的成本提供了更多功能。”

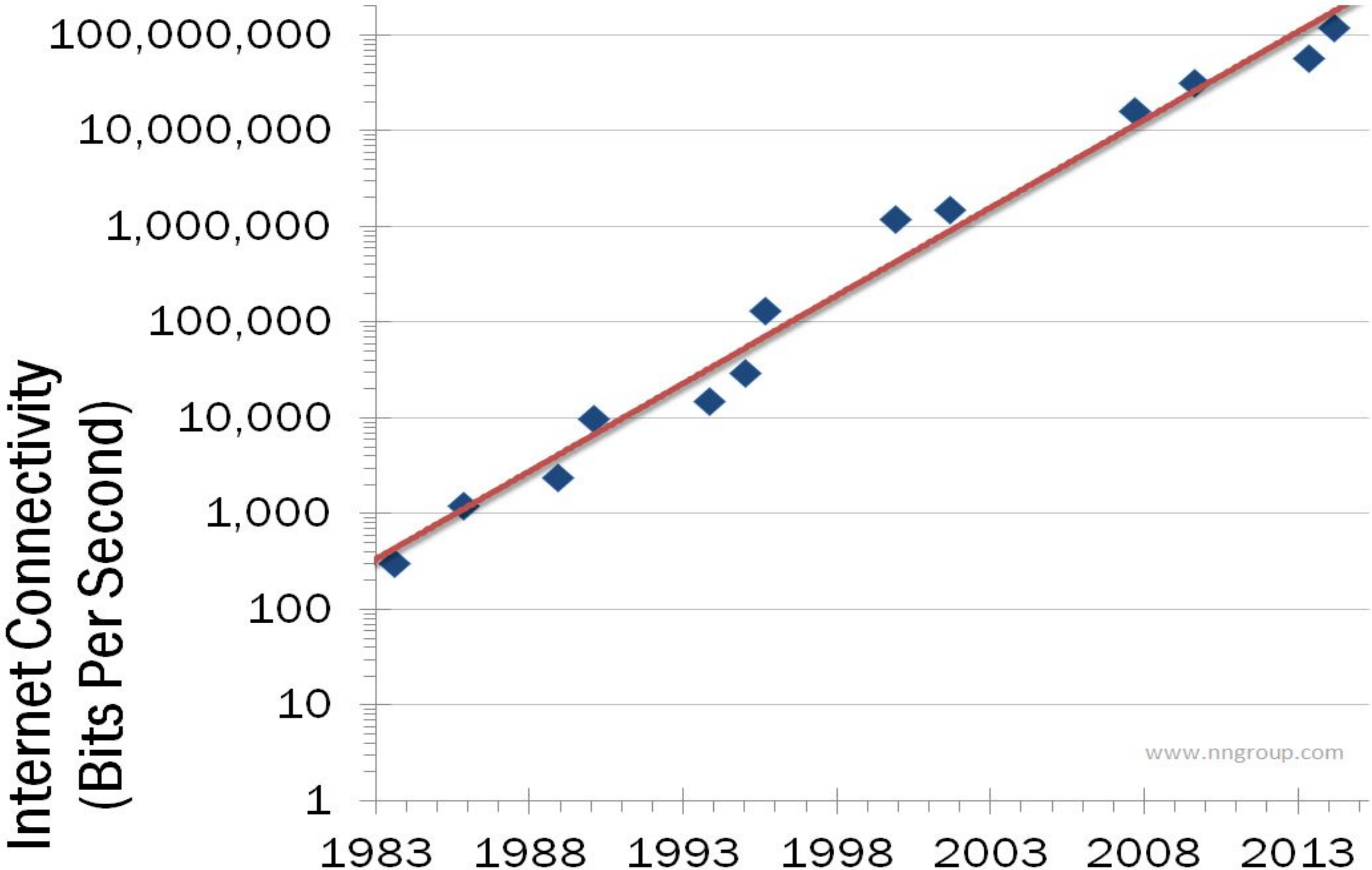


Sooner or later even 3D chips, even parallel processing will be exhausted

Perhaps molecular computers, quantum computers can takeover in post-silicon era

Nielsen's Law of Internet bandwidth states that:

a high-end user's connection speed **grows by 50% per year**



Summary of lecture 1

- Turing developed the fundamental basis on which computation is now discussed
- In the von Neumann architecture the fundamental components are input/output, processing unit, memory
  - Key insight: program can be stored as software
- Modern systems use numerous external buses, connected via bridges