Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

Computers and Intelligent Machines

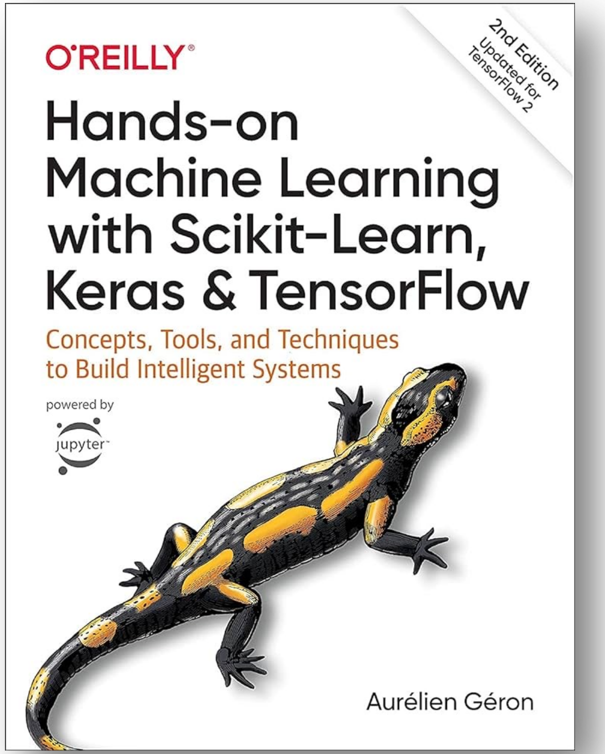
Introduction to Multiuser Cluster Systems at LRZ

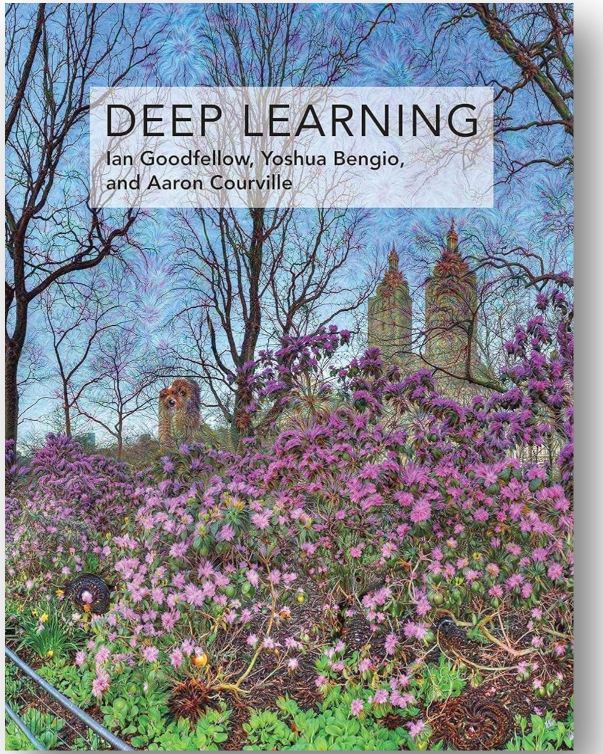
October, 9th 2024

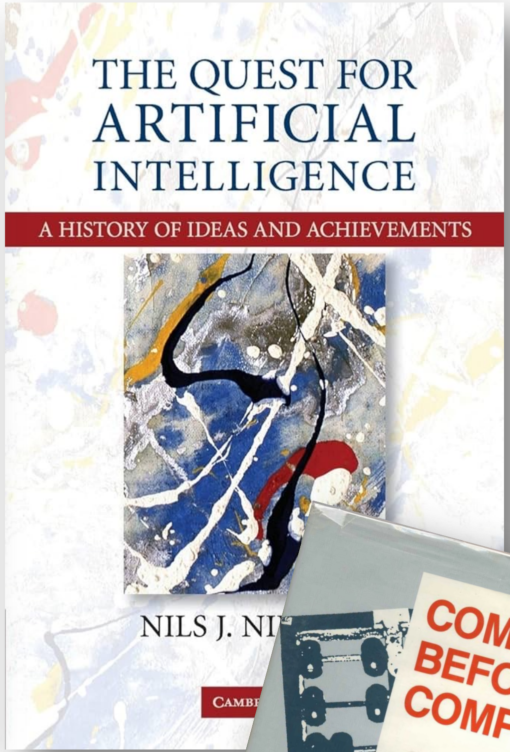
- **Aim**: provide an introduction to multiuser cluster systems in general and to those operated at the Leibniz Supercomputing Centre (LRZ), specifically.
- We will focus on Artificial Intelligence and the machines required to enable Machine Learning and Deep Learning workflows
- We will give enough history that it hopefully explains why multi user clusters are they way they are, even in the 21st century.
- You will probably benefit the most if you're not yet familiar with the LRZ HPC/HPDA/HPAI infrastructure, but plan to work with these systems in the future.
- A majority of systems will be covered in more detail in dedicated sessions.

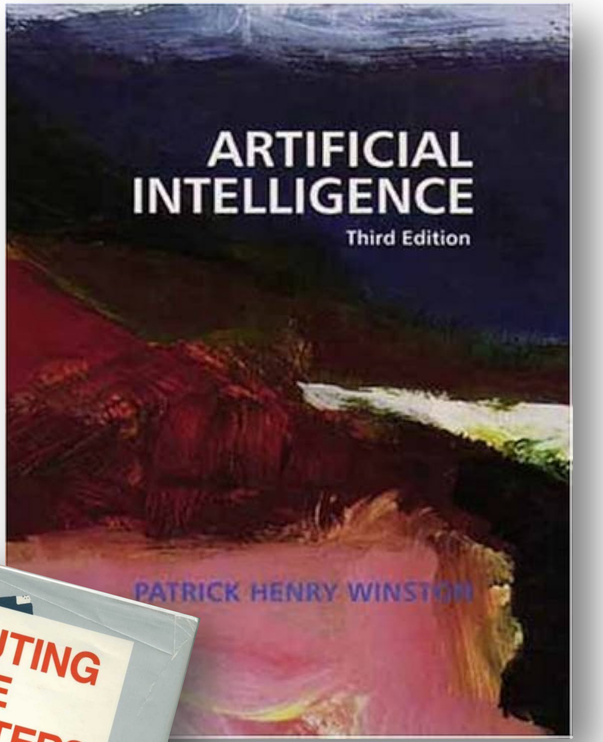
By the end of today's workshop, you should have a general understanding of multiuser HPC/HPDA/HPAI cluster systems and the basic skills to successfully interact remotely with such systems at LRZ.

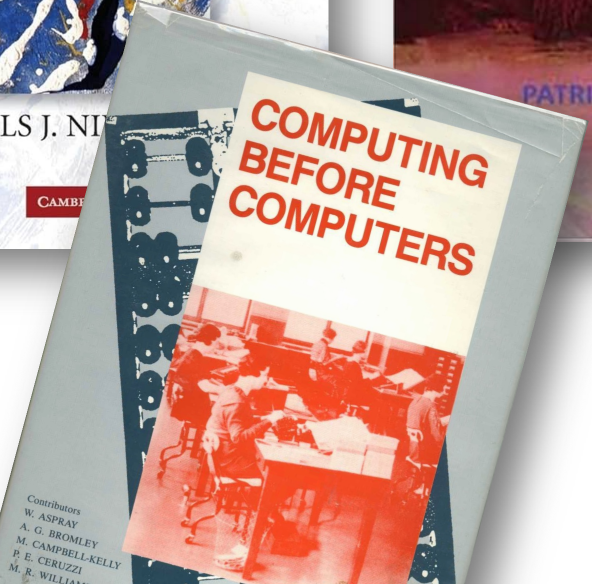
Book Recommendations



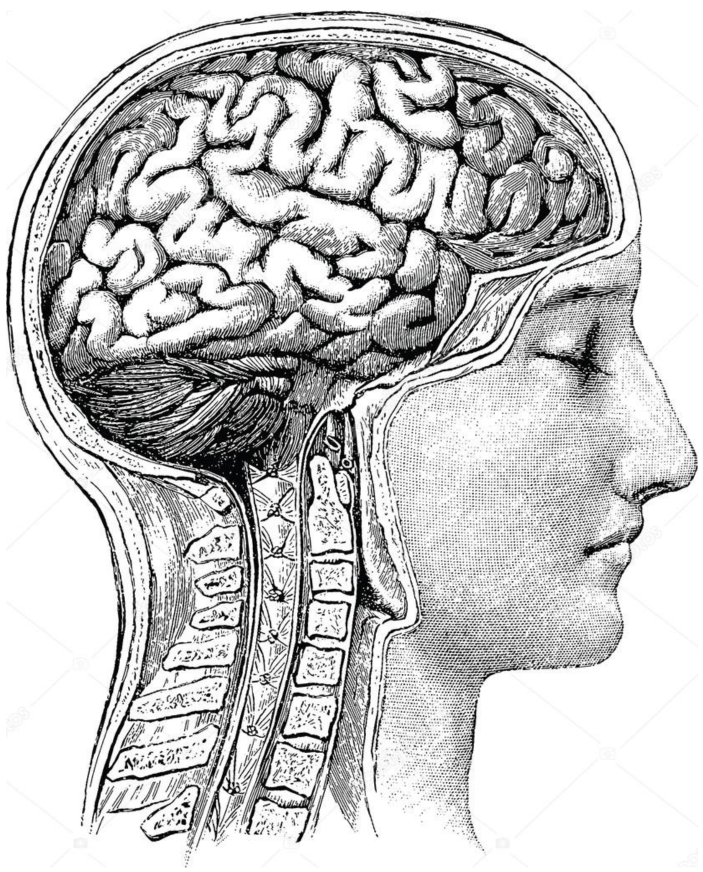








The Hardware of Intelligence

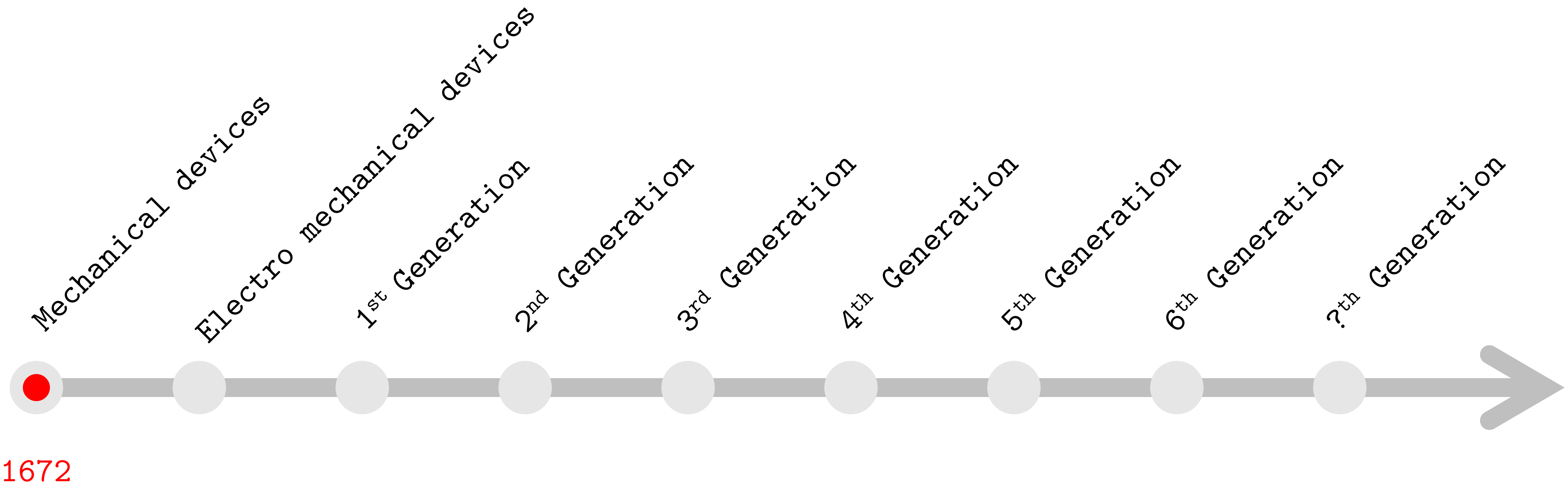


	Adult Human Brain	NVIDIA H100
Neuron Count	86B	20B fp32
FLOPs	~Exaflop Scale? (low precision)	30-3000 teraFLOPs
Memory	PB Scale?	80 GB
Energy (W)	20	300
Mass (g)	1300	1700
Made of	C, H, O, N	Silicon
Bandwidth	~25Mb/s	3TB/s
Price		35k€

Un-Intelligent Machines

History of Computing

The Generations of Computing Devices

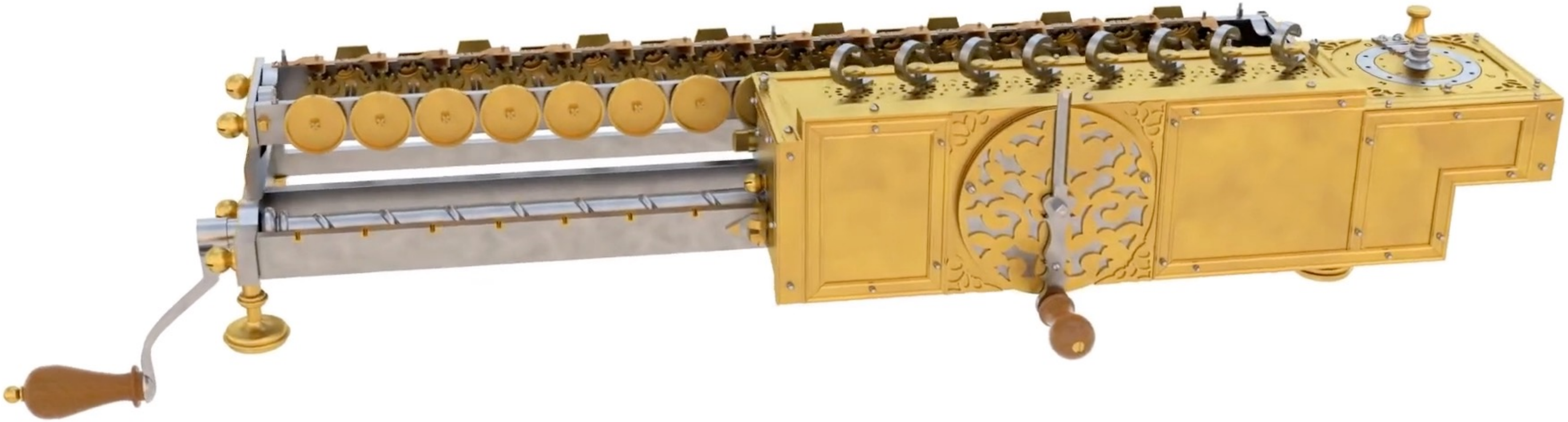


1672: Gottfried Wilhelm von Leibniz

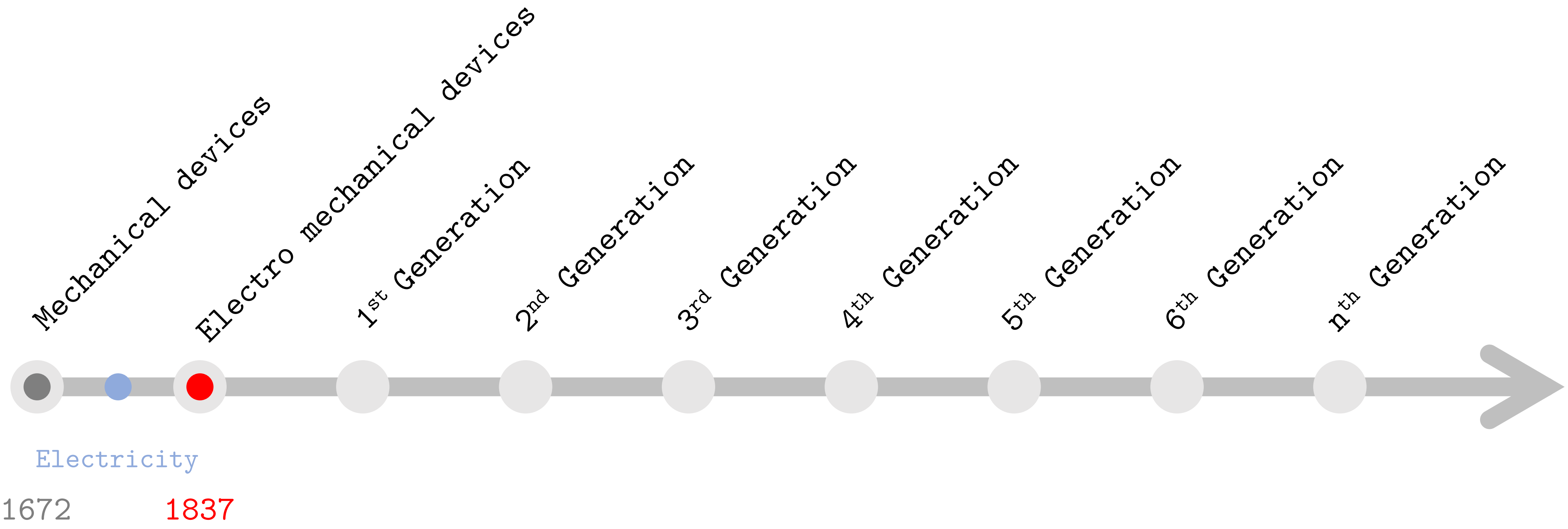
It is beneath people *the dignity of excellent ~~men~~ to waste their time in calculation when any* one *~~peasant~~ could do the work just as accurately with the aid of a ~~machine~~.* ChatGPT

— *Gottfried Leibniz (1672)*

1672 – 1673: Mechanical Device

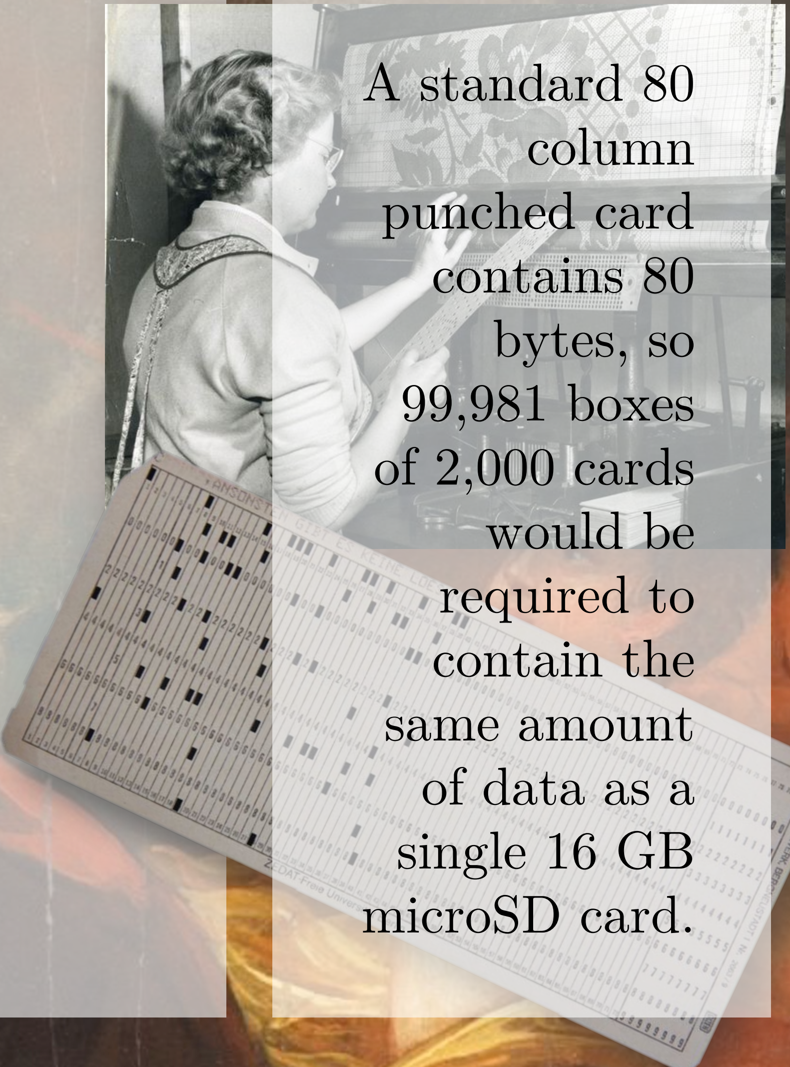


The Generations of Computing Devices



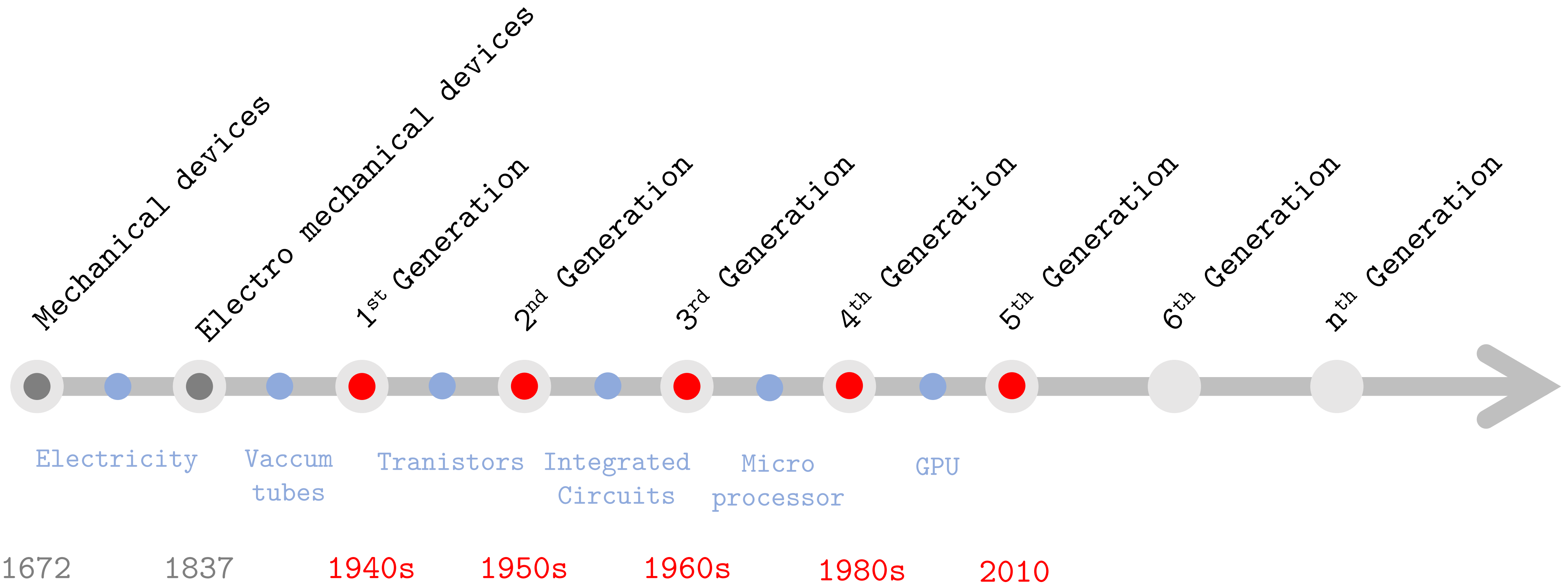
From the Jacquard loom to Electro Mechanical Devices

- **1752: Benjamin Franklin and the kite experiment** 🇺🇸
  - Prove that lightnings are a electrical discharge
  - Contributed to the understanding of electricity
- **1804: The Jacquard Loom** 🇫🇷
  - Was a mechanical loom for cloth weaving
  - First demonstrated by Joseph Marie Jacquard in 1801.
  - Any number of the cards could be chained together into a continuous sequence, with each card corresponding to one row of the design
- **1837: Charles Babbage's Analytical Engine** 🇬🇧
  - First general-purpose mechanical computing device
  - Need for software: Ada Lovelace as the first programmer 🇬🇧
- **Herman Hollerith: Punched card tabulating machine** 🇺🇸
  - Later became the "IBM Punchcards"



A standard 80 column punched card contains 80 bytes, so 99,981 boxes of 2,000 cards would be required to contain the same amount of data as a single 16 GB microSD card.

The Generations of Computing Devices



1950: From Theory to the First Generation of Computers (And AI)

lrz

"Computing Machinery and Intelligence" — 1950

1

Turing test

- Imitation game
- "Can Machines think"
- Simplified version: can a computer fool an interrogator into believing it is human

2

Anticipation

- Arguments against 9 common objectionsthe possibility of achieving intelligent computers: Theological objections, consciousness…

3

Suggestions

- Suggestions on how to produce programs with human level intellectual abilities
- Idea: Start produce a program that simulates a "child" mind

```
------------------------------------------------------------------
      ELIZA -- A Computer Program for the Study of Natural
         Language Communication Between Man and Machine
DOCTOR script by Joseph Weizenbaum, 1966  (CC0 1.0) Public Domain
ELIZA implementation by Ant & Max Hay, 2023  (CC0 1.0) Pub Domain
------------------------------------------------------------------
Type *help and press the Enter key to see a list of commands.


HOW DO YOU DO. PLEASE TELL ME YOUR PROBLEM
```

1945-1955: First Generation (Vaccum tubes and plugboards)

- Used Vacuum tubes made of glass
  - Control flow of electricity between two electrodes: 0 || 1
  - Slow, unreliable, produced a lot of heat: Often would burn and would need to be replaced
- Heavy computers take up a full room
- Used for calculation, storage, and control purposes
- Main memory: Magnetic tapes and magnetic drums (up to KB)
- No OS, no real programming language (machine code)
- Example of machines: EDVAC, UNIVAC 1101
- Applications: numerical computations: tables of sins, cosins, and logarithms

A mainframe is a large computer system that is usually used for multi-user applications. It is so expensive it needs to be share:

- time sharing
- space sharing
- batch processing
- …

… all of which are still relevant.

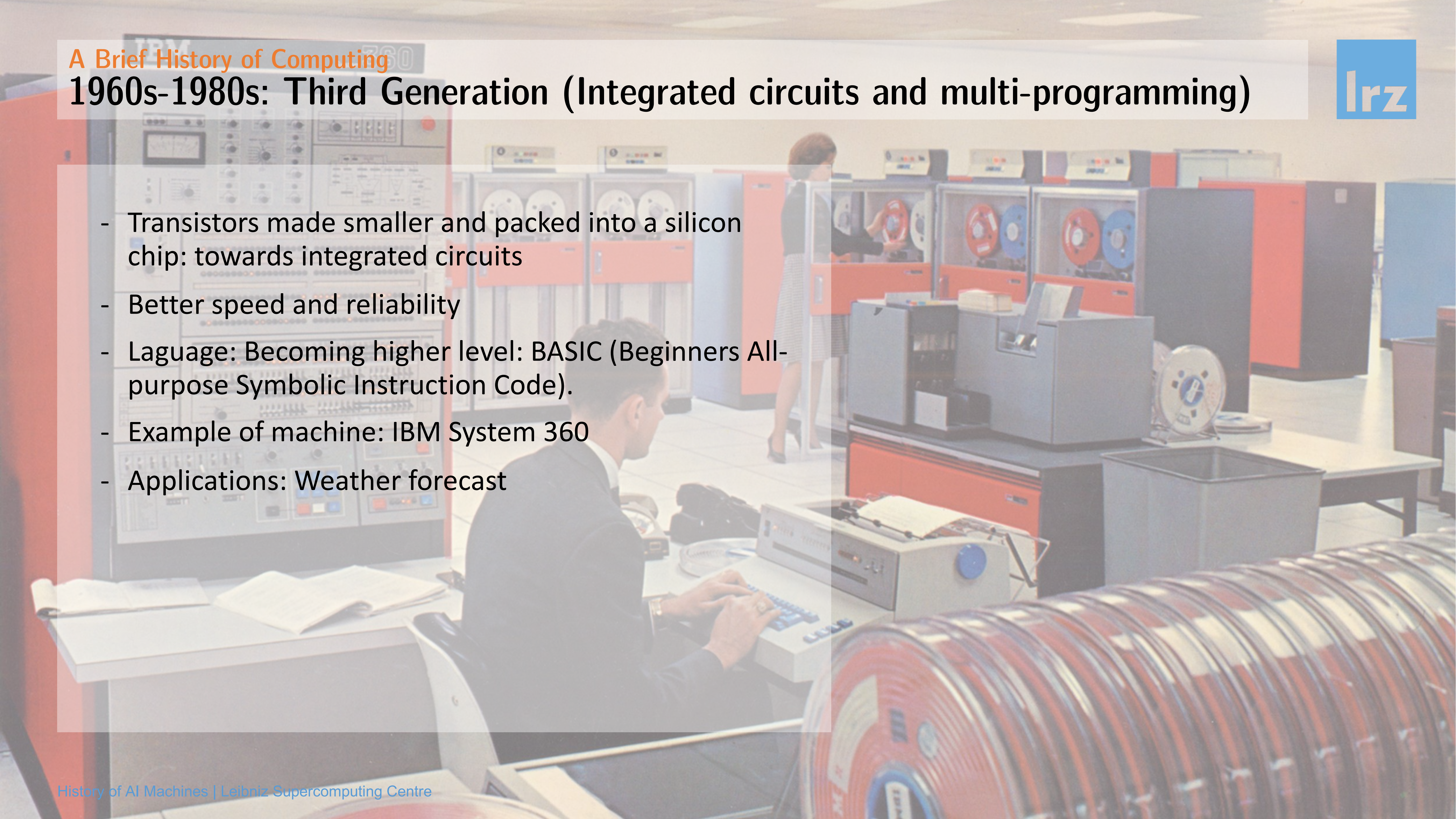
Until the mid-to-late 1950s, the word "computer" referred to people who performed computations, not to machines.

1950s-1960s: Second Generation (Tranistors and batch systems)

- Use transistors instead of vacuum tubes
  - More reliable, smaller, and allow faster clock speeds
  - Transistors shaped the computer revolution and digital age: logical operations are performed by semi conductor devices
- Machine can store up to 2MB of data and run at 1 MHz
- Running Jobs → Batch systems
- Emergence of high-level programming language: FORTRAN (1956), ALGOL (1958), and COBOL (1959).
- Example of machine: IBM1401 / 7094
- Appications: Partial differential equations

The very first transistor — the foundational building block which almost all of modern civilization was built from — was created at AT&T's Bell Labs on December 23 1947.

A transistor is a semi conductor that uses an electric current to control the flow of electrons. It amplifies or switches electrical signals by enabling or preventing the flow of current. Multiple transistors combined form logic gates, arithmetic circuits, memories etc.

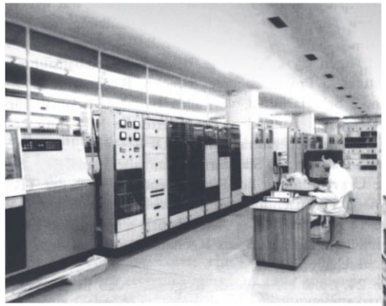


1960s-1980s: Third Generation (Integrated circuits and multi-programming)

- Transistors made smaller and packed into a silicon chip: towards integrated circuits
- Better speed and reliability
- Laguage: Becoming higher level: BASIC (Beginners All-purpose Symbolic Instruction Code).
- Example of machine: IBM System 360
- Applications: Weather forecast

1960s-1980s: Meanwhile in Bavaria

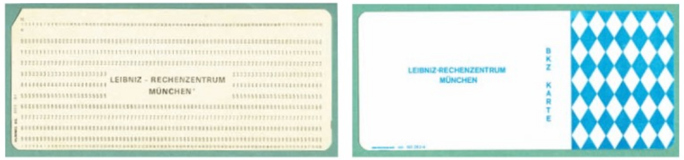
1962 1980 1990 2020


TR4 (1964) in Konstanz vor Auslieferung an das LRZ


LRZ-Gebäude Richard-Wagner-Straße 18

- Telefunken TR 4
- IBM 7090


LRZ-Gebäude I

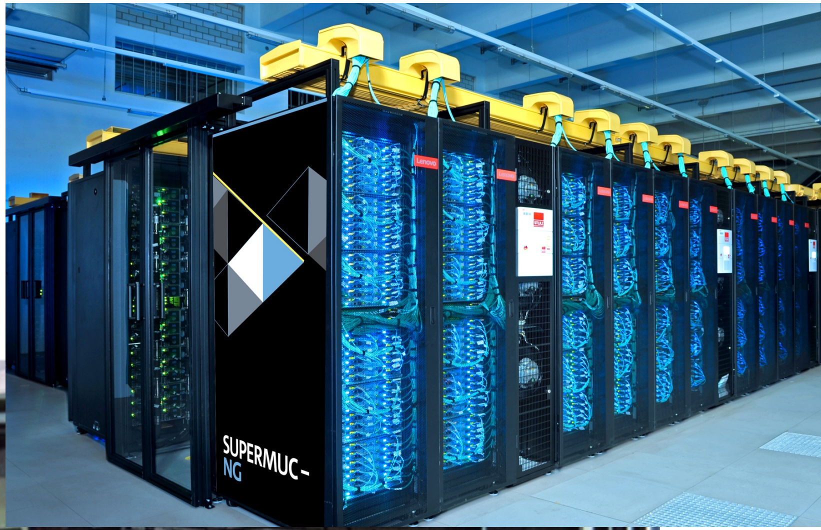


Lochkarten LRZ

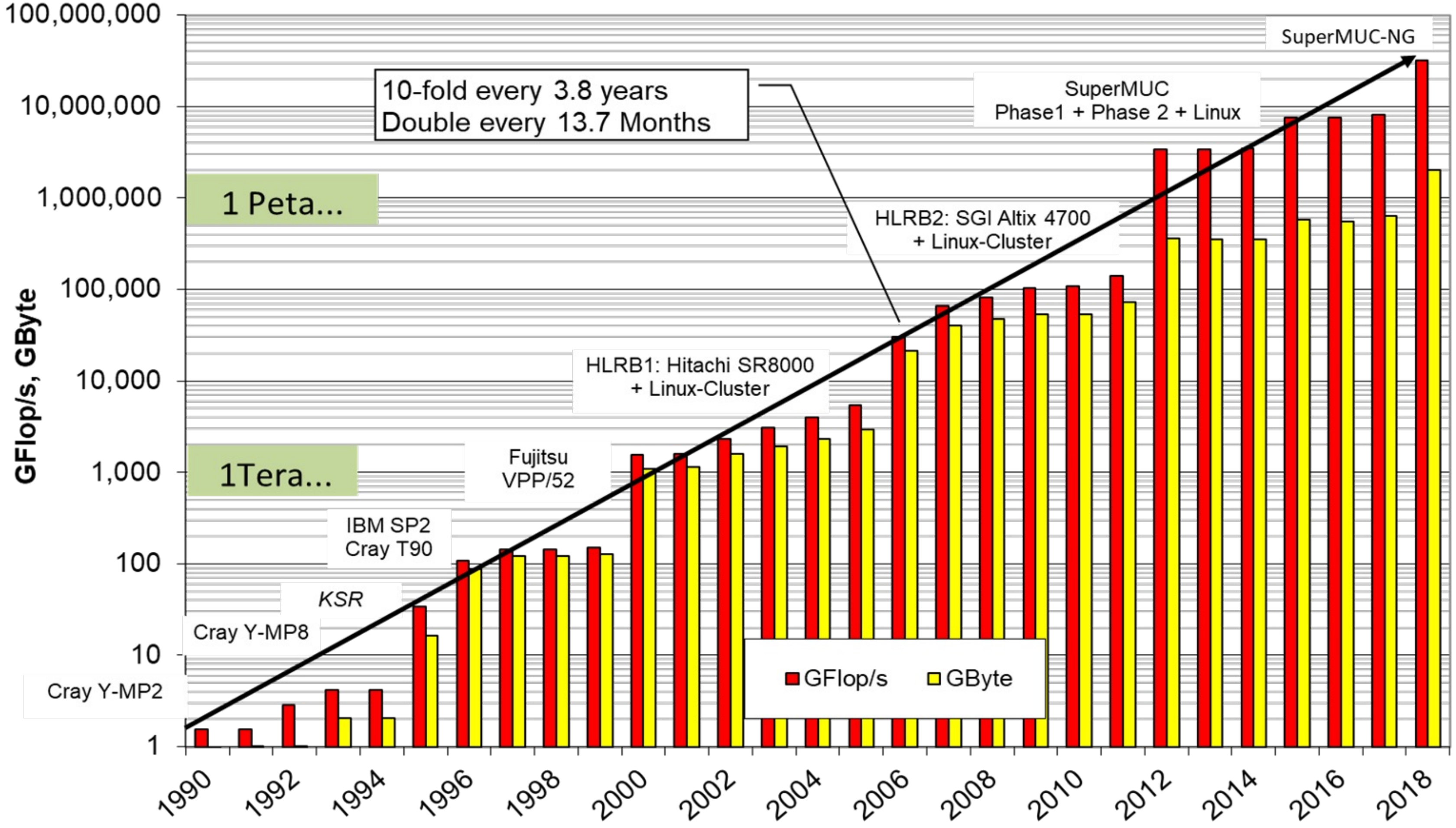

Druckausgabe der LRZ-Datenstation


Lochkarteneingabe in der LRZ-Datenstation


Maschinenraum 1988

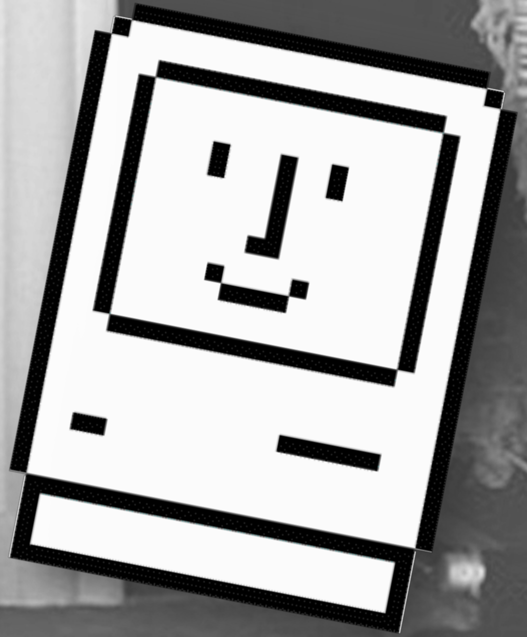


1960s-1980s: Meanwhile in Bavaria



1980s-2010: 4th Generation Computer (personal computer)

- Macintosh 128k released in 1984
- Powered by a microprocessor (8 MHz) / 128 KB RAM
- 400 KB storage space on floppy
- First “real” Personal Computer (PC)
  - vs. IBM PC / Commodore 64
  - Comes with a screen, mouse, keyboard
  - Reaching a new audience: works without a manual
  - User friendly, cute, and adorable
- OS: System I (UN*X family, GUI)
  - Finder, Menu bar
  - Still the current HIG
- Application: MacPaint, MacWrite
- First affordable computer made for personal use ($2,500 ($6,500 in modern dollars))



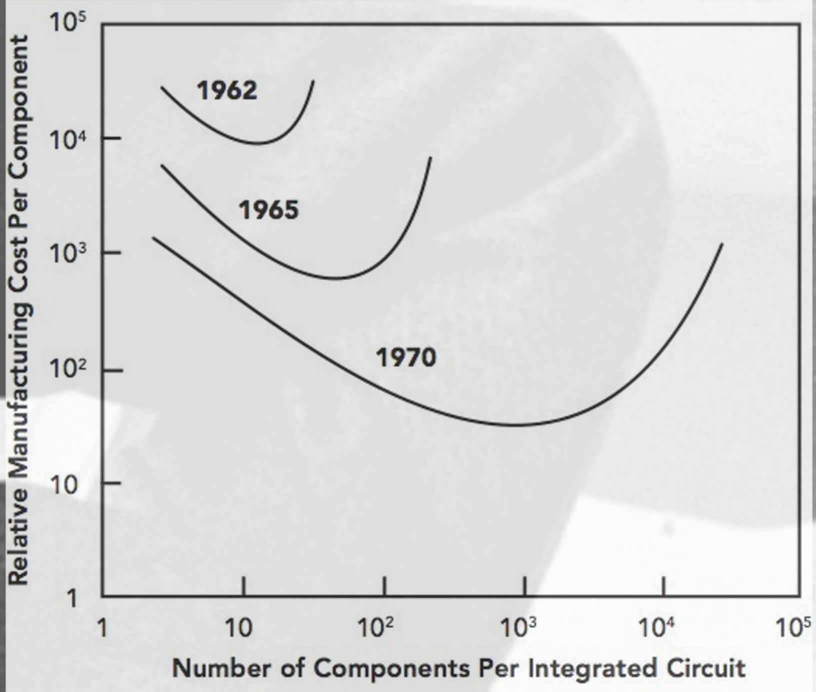
2010s – now: 5th Generation computer

- Key technologies include mobile devices (smartphones, tablets), cloud computing, social networks, high-speed wireless networks, IPv6 networking protocol, touchscreens, solid state storage, virtual/augmented reality, artificial intelligence
- Human like interaction and behaviour
  - Voice recognition
  - Computer vision
- Programming language: Very high level programming, Natural language
- Come in pocket size / wearables / Cloud only
- Digital twins, NVIDIA omniverse, Metaverse

Moore's Law

The number of transistors on integrated circuits is doubling about every ~~18 months~~ 2 years.

— Gordon Moore, 1965

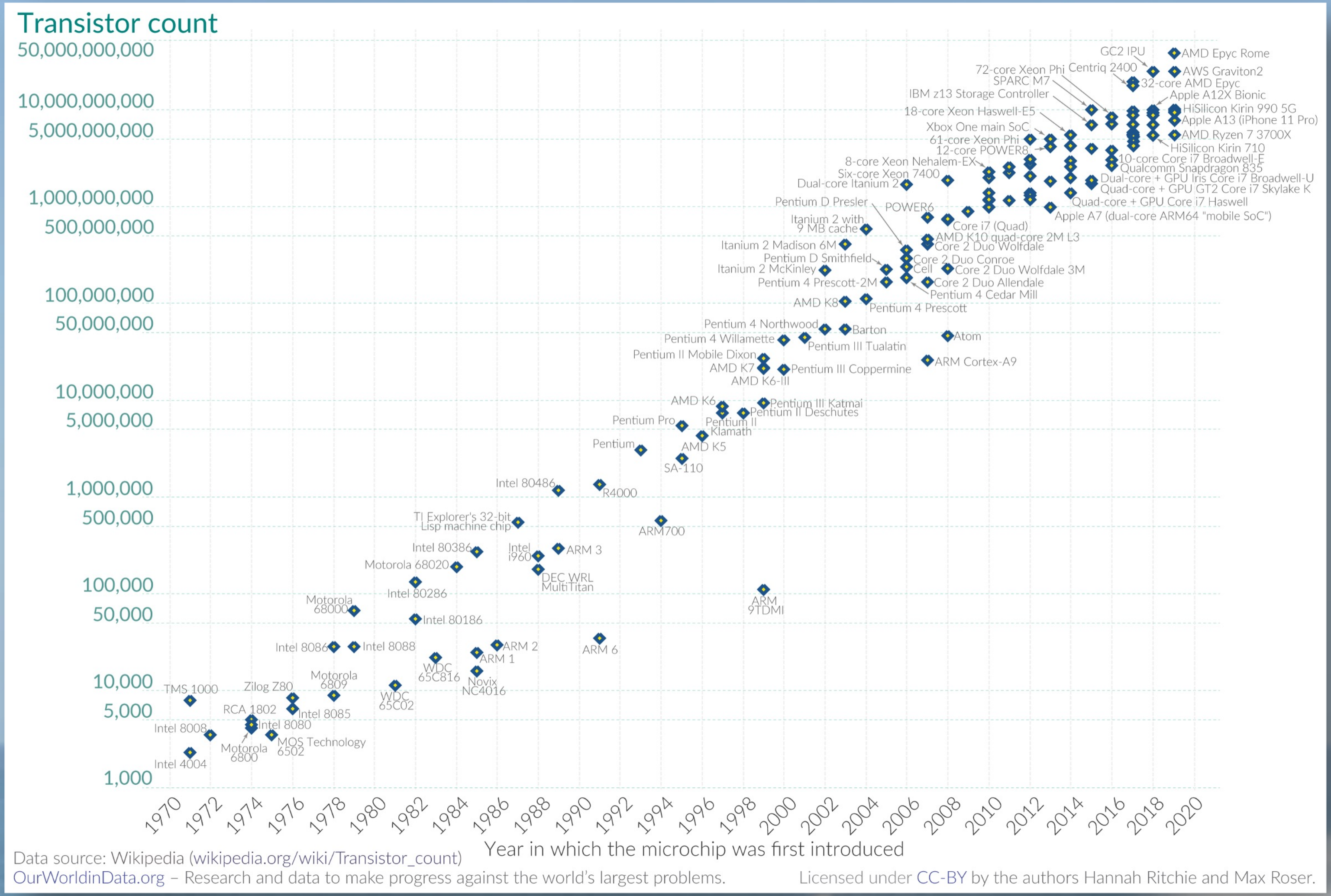


Gordon Moore's curves on this plot show that development of the chemical printing technology makes more complex microchips the cheapest form of electronics.

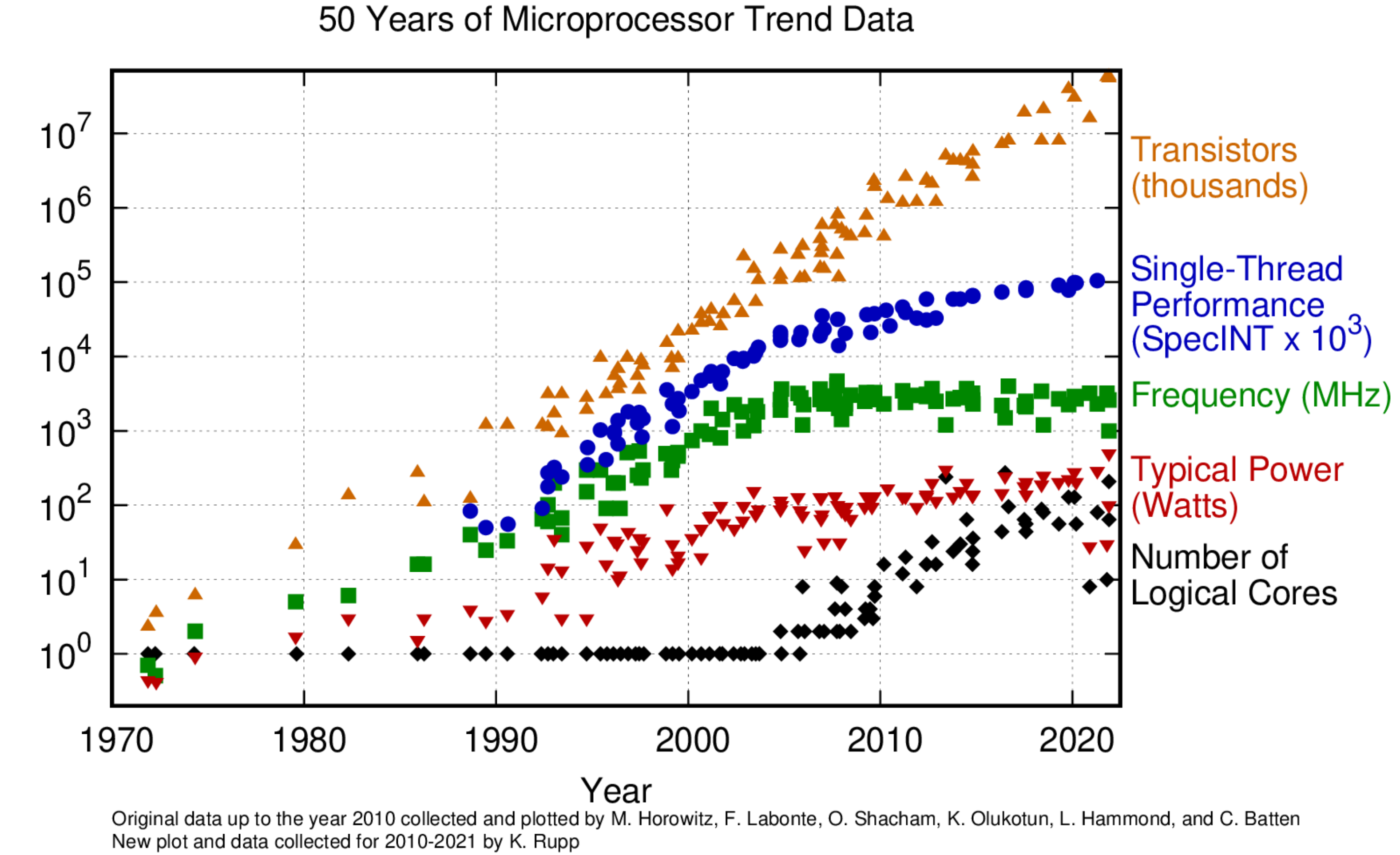
Source: Gordon Moore.

A Brief History of Computing

Is Moore’s Law dead?

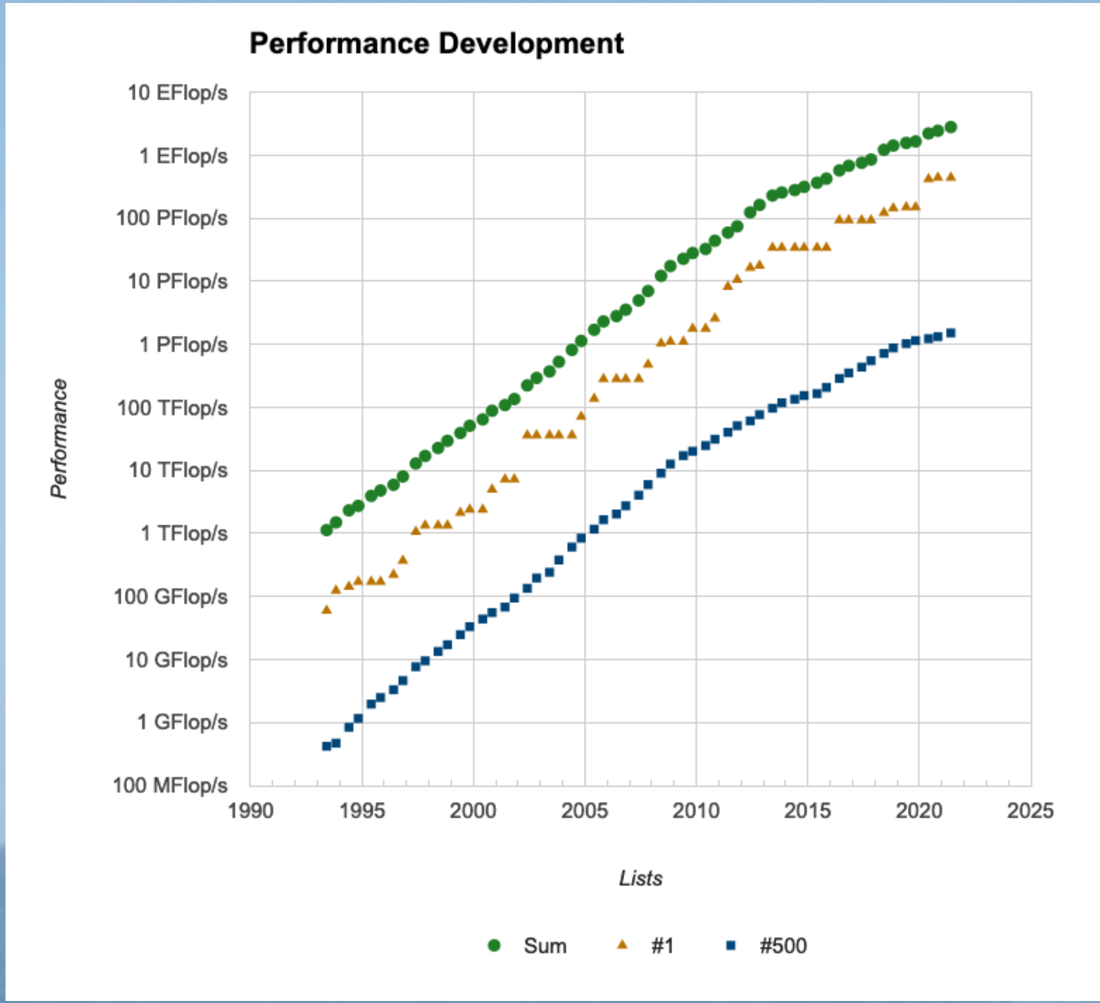


Is Moore's Law dead?

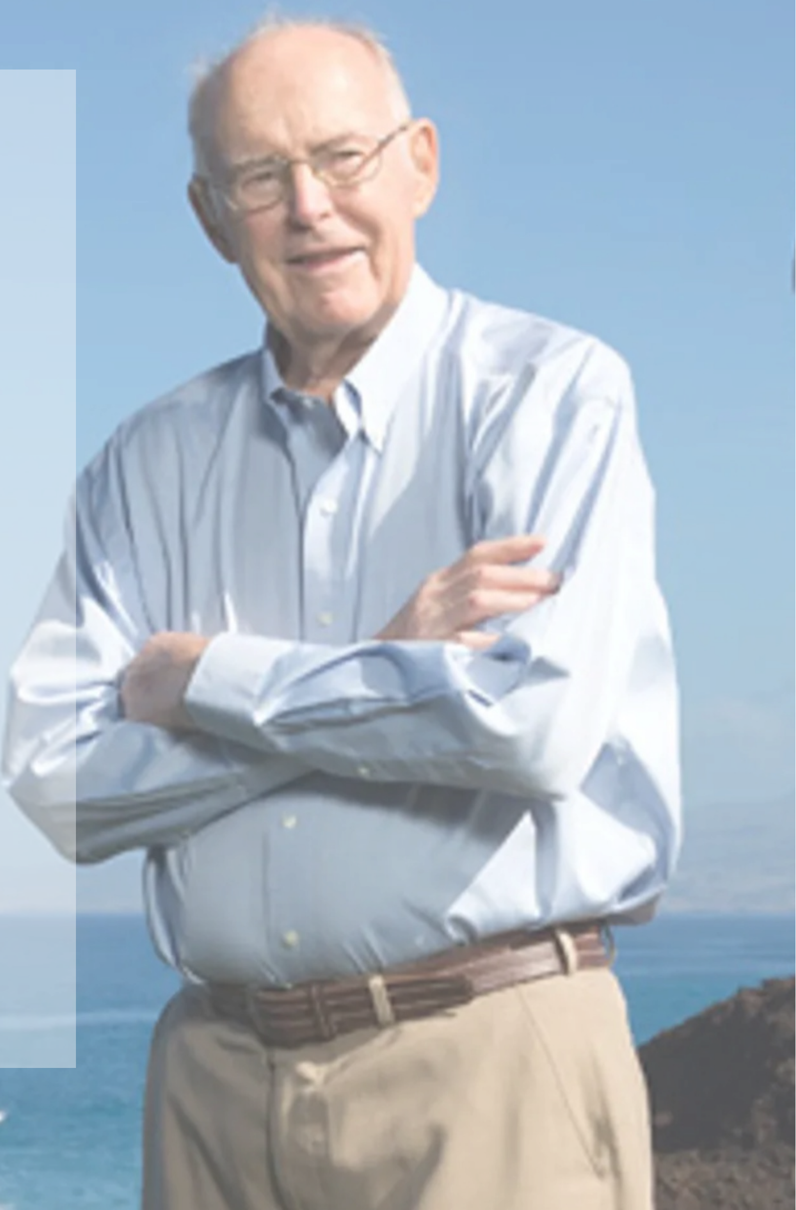


- Mid 2000s: "heat death"
- No more faster processors, only more of them.
- But: 2x3 GHz != 6 GHz

A Brief History of Computing
Is Moore’s Law dead?



- **From #1 to #500:** 6-8 years
- **From #500 to Notebook:** 8-10 years



Multiuser clusters…

… Haven't always been multiuser

… They are big and expensive: They need to be shared to cost-effectively serve a large number of concurrent users.

… They are meant to be used remotely

… System administrators take care of the system for the users

… can achieve much higher performance than individual systems by aggregating resources

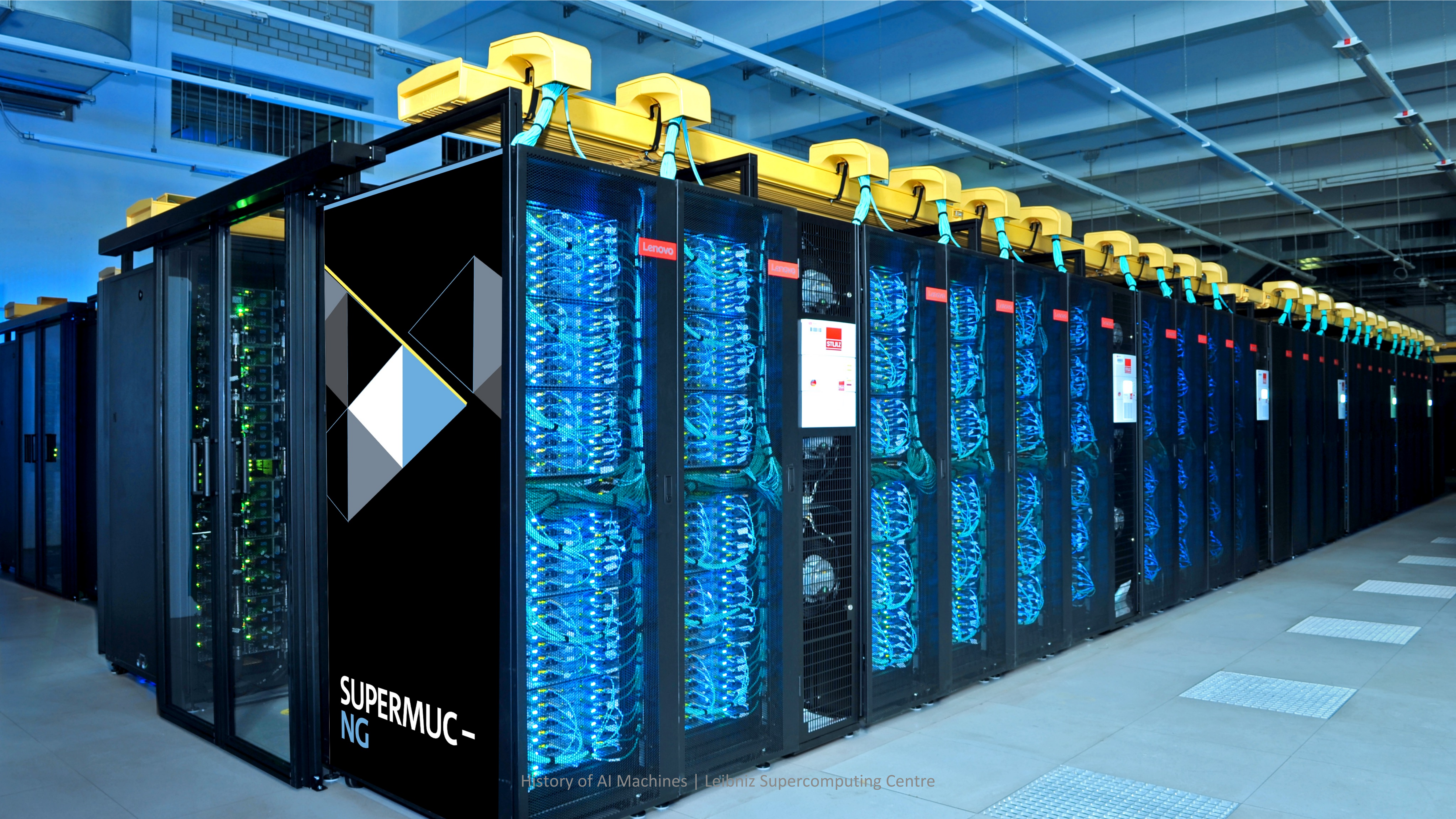
… They provide fault tolerance through redundancy. If any part of the cluster fails, the rest of the cluster can continue operating

… Common architectures include…

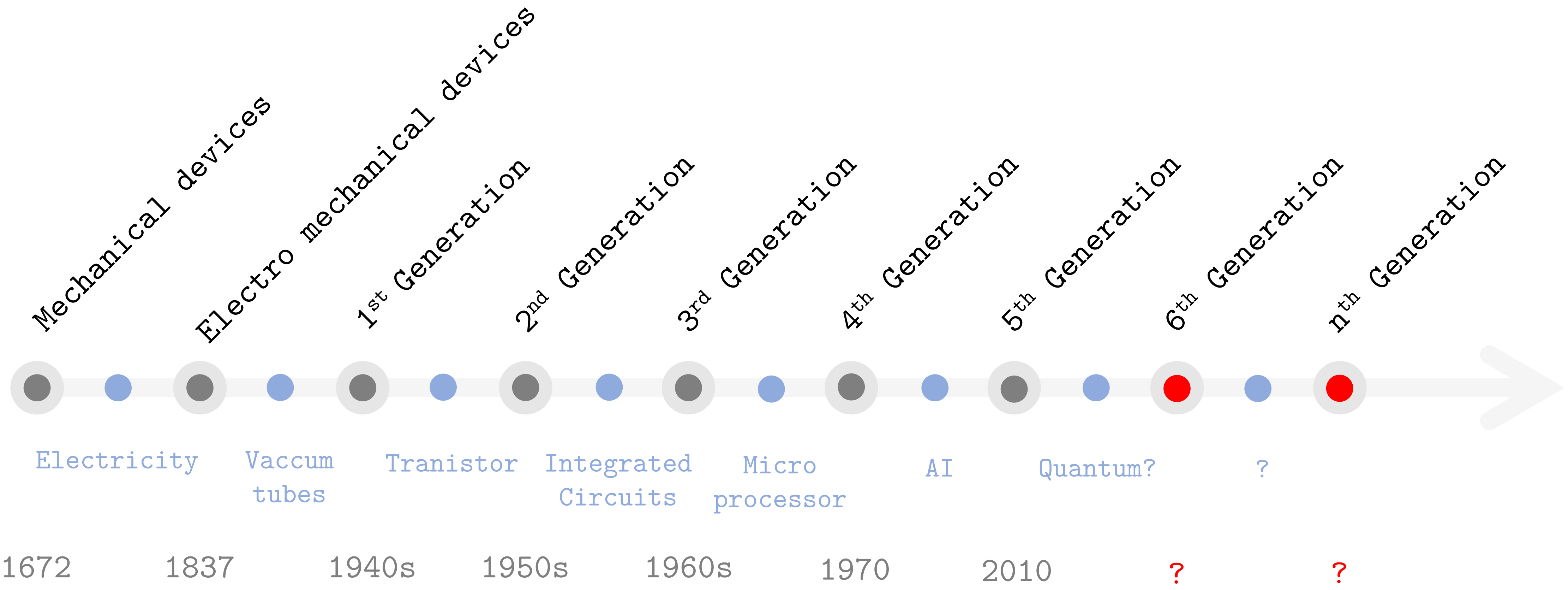
- … Server clusters: Multiple interconnected servers with shared storage and networking. Used for high availability and scalability.
- … High-performance computing clusters: Powerful servers with fast interconnects, used for running highly parallel workloads.
- … Cloud computing clusters: Massive clusters that run cloud platforms and services, accessed by many users over wide-area networks.
- … Edge clusters: Clusters deployed at the edge (near users/devices) to support localized computing, storage and networking needs.

What is a Supercomputer or High-Performance Cluster… (Not)?

It runs Microsoft Windows?	😌 No, no worries
It will run my Excel spreadsheet?	😏 No!
It has overclocked high-speed processors?	🤨 No
The CPU runs faster than a desktop PC?	🤔 Not even
It has a large internal memory (RAM)?	😱 Usually not (except exceptions)
It will run my old tried and tested executable?	😔 Probably not
It will run my software without changes?	😵‍💫 Probably not
It will run my program with millions of threads?	🥺 Probably not
It can be used interactively?	🤯 Probably not
It has shiny RGB lights?	…

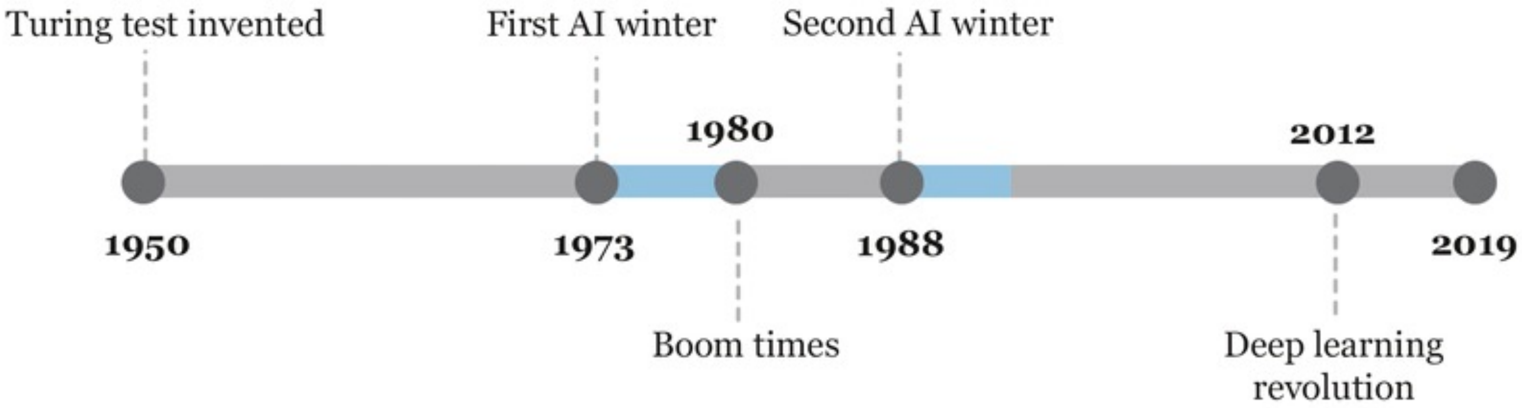

Lenovo
Lenovo
STULZ
SUPERMUC-
NG

The Generations of Computing Devices



Intelligent Machines

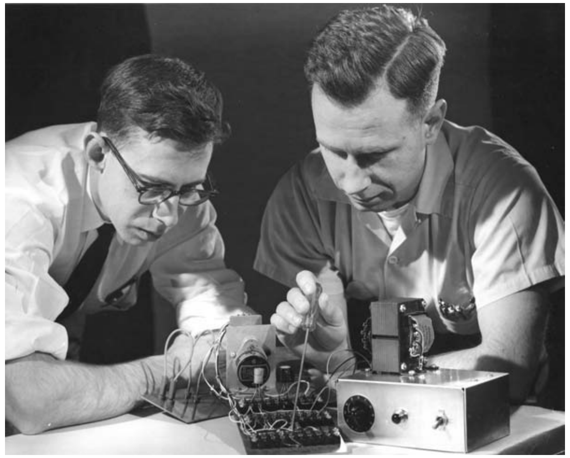
The Four Seasons of AI

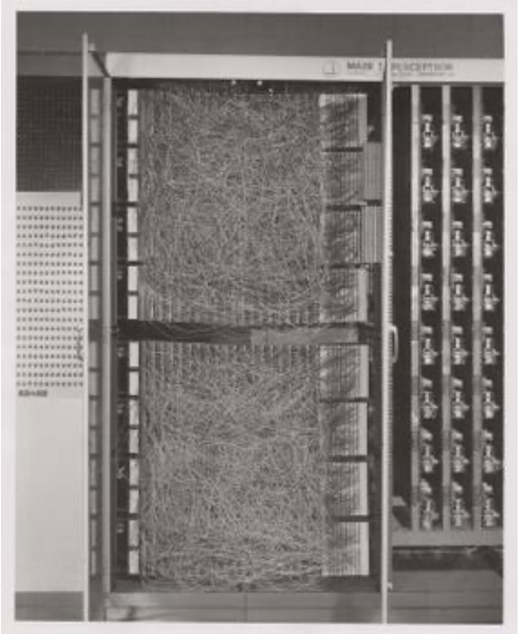


S. Schuchmann, “Analyzing the Prospect of an Approaching AI Winter,” 2019

~~General Purpose~~ Specialized Computers for AI

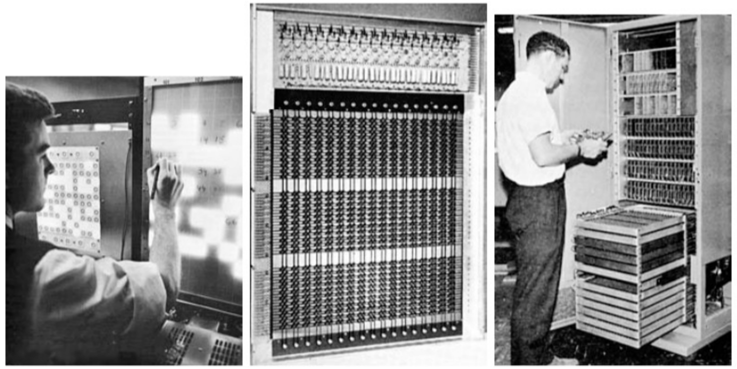
1957
Frank Rosenblatt (left) working (with Charles Wrightman) on a prototype A-unit





Mark I Perceptron

1960’s
MINOS I, II, III



Intelligent Machines

GPU, AlexNet

2012: Not an AI company

SYSTEM SPECIFICATIONS

GPUs	8X NVIDIA® Tesla® V100
Performance (Mixed Precision)	1 petaFLOPS
GPU Memory	256 GB total system
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz
NVIDIA CUDA Cores	40,960
NVIDIA Tensor Cores (on Tesla V100 based systems)	5,120
Power Requirements	3,500 W
System Memory	512 GB 2,133 MHz DDR4 RDIMM
Storage	4X 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Operating System	Canonical Ubuntu, Red Hat Enterprise Linux
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)

2017: AI Supercomputer in a Box

Intelligent Machines

Bitter Lesson, Transformer, Mamba

The Bitter Lesson

Rich Sutton

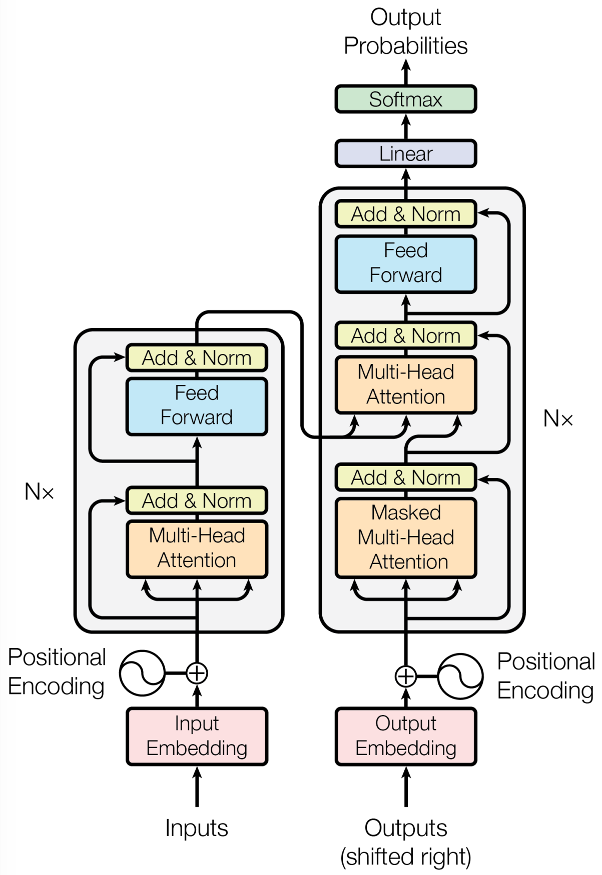
March 13, 2019

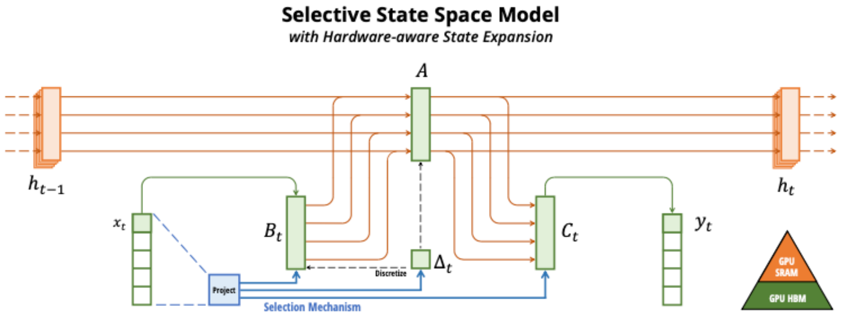
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

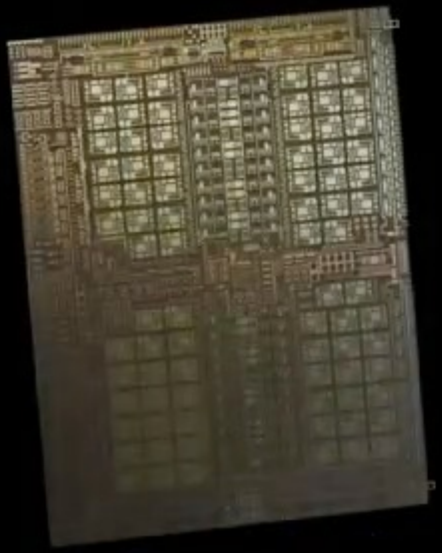
In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that ``brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

In speech recognition, there was an early competition, sponsored by DARPA, in the 1970s. Entrants included a host of special methods that took advantage of human knowledge---knowledge of words, of phonemes, of the human vocal tract, etc. On the other side were newer methods that were more statistical in nature and did much more computation, based on hidden Markov models (HMMs). Again, the statistical methods won out over the human-knowledge-based methods. This led to a major change in all of natural language processing, gradually over decades, where statistics and computation came to dominate the field. The recent rise of deep learning in speech recognition is the most recent step in this consistent direction. Deep learning methods rely even less on human knowledge, and use even more computation, together with learning on huge training sets, to produce dramatically better speech recognition systems. As in the games, researchers always tried to make systems that worked the way the researchers thought their own minds worked---they tried to put that knowledge in their systems---but it proved ultimately counterproductive, and a colossal waste of







BLACKWELL

THE LARGEST CHIP PHYSICALLY POSSIBLE

104 billion transistors

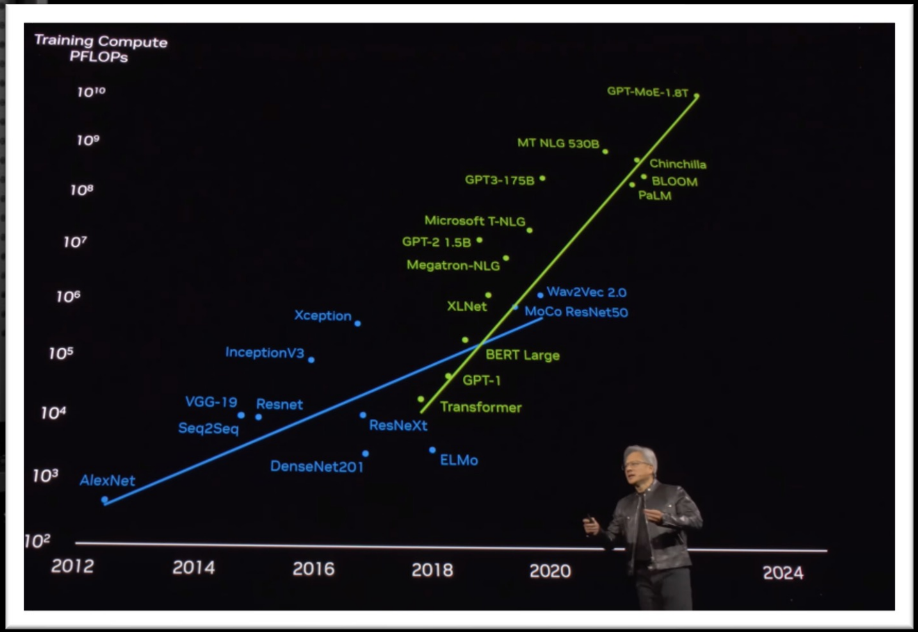
TSMC 4NP process

10TB/s NVIDIA High-Bandwidth Interface

Intelligent Machines
GTC 2024

- "People think we make GPUs — and and we do — but GPUs don't look the way they used to"
- Second Industrial revolution: Foundries of Intelligence

→ Take away message: the commoditization of intelligence is happening.



Current AI workflow for practitioners

lrz

- **Local Machine for experimentation**: PyTorch, TF, Notebook
  - Workstation with Apple MPS
  - Discrete NVIDIA GPU: *e.g.,* a6000 with cuda
- **Training at larger scale on a shared system with** SLURM
  - *e.g.,* LRZ AI Systems
  - *e.g.*, foundation model
  - Share model with community (*e.g.,* Hugging face)
- **Deploy model for inference on the cloud**
  - Specific type of GPU (T4)
  - Inferentia instances AWS
  - …

TensorFlow

PyTorch

Hugging Face

Trends

- Exotic hardware: Cerebras / FPGA / TPU…
- Larger models and multi modalities: Need for more compute and storage (text, tables, videos)
  - Develop common sense: text and video
  - Synthetic data generation (like humane simulation)
- On device, inference: LLMs on iPhone etc., online training
  - Size and Energy Efficiency consideration
  - Framework
  - Also better for privacy
- Trustworthy AI: Security and Privacy
  - Confidential Computing
  - Federated Learning
  - Explainable AI
- Energy efficiency

Example of recent article pre-print:

Environmental impact

Lastly, we would like to give a rough estimate of the climate impact of this study. We assume the average German power mix which as of 2021 according to the German Federal Environment Agency corresponds to 420g CO_2/kWh. Only the final RadImagenet trainings (no hyperparameter optimisation) ran on 8 NVIDIA A40, where we assume a power consumption of 250W on average, each for almost 4 days, 5 privacy levels, and 5 repetitions. Hence, this amounts to around 960kWh and thus more than 400kg of CO_2 equivalents. This almost equals a return flight from Munich to London. Hence, we tried to limit our hyperparameter searches to the necessary. In total, we assume that this study produced at least 2 tons of CO_2 equivalents.

References

[1] Kristina Lång, Viktoria Josefsson, Anna-Maria Larsson, Stefan Larsson, Charlotte Högberg, Hanna Sartor, Solveig Hofvind, Ingvar Andersson, and Aldana Rosso. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (masai): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8):936–944, 2023.

14