Intro to AI on supercomputers

Huihuo Zheng

Computer Scientist
Argonne Leadership Computing Facility
October 1, 2024

www.alcf.anl.gov

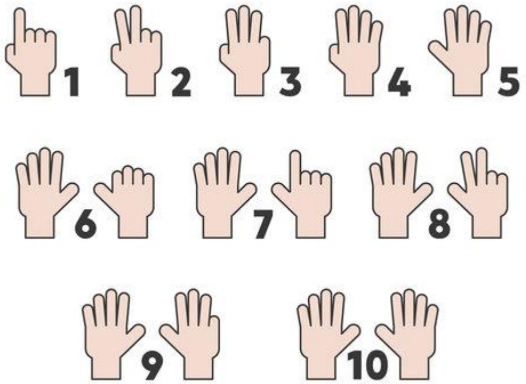
Outline

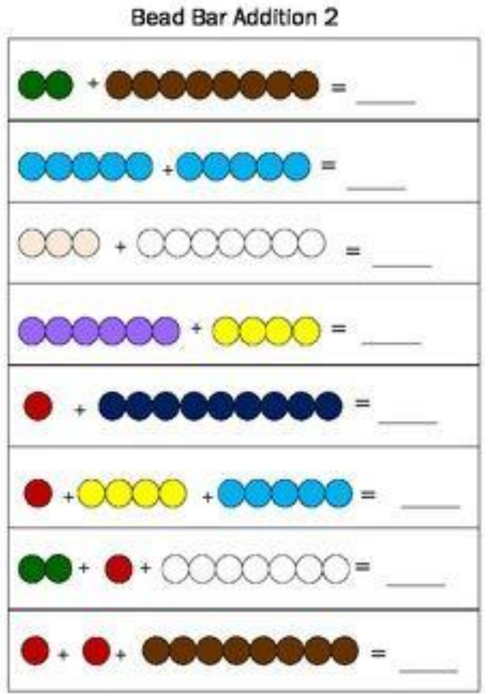
1. Evolution of computing systems
2. Parallel computing
3. AI in a nutshell

Key words: supercomputer, parallel computing, AI

Journey of computing

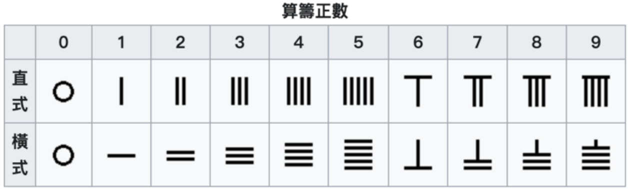
What is the first “computer” in the history?
When did it come out?





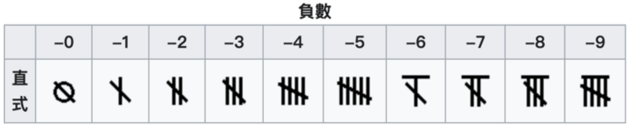
How my daughter calculates addition at school.

Manul “Calculator”



算籌正數

	0	1	2	3	4	5	6	7	8	9
直式	○	𝍠	𝍡	𝍢	𝍣	𝍤	𝍥	𝍦	𝍧	𝍨
橫式	○	𝍩	𝍪	𝍫	𝍬	𝍭	𝍮	𝍯	𝍰	𝍱



負數

	–0	–1	–2	–3	–4	–5	–6	–7	–8	–9
直式	⊘	𝍠̸	𝍡̸	𝍢̸	𝍣̸	𝍤̸	𝍥̸	𝍦̸	𝍧̸	𝍨̸

Counting rods - **1600 BC**

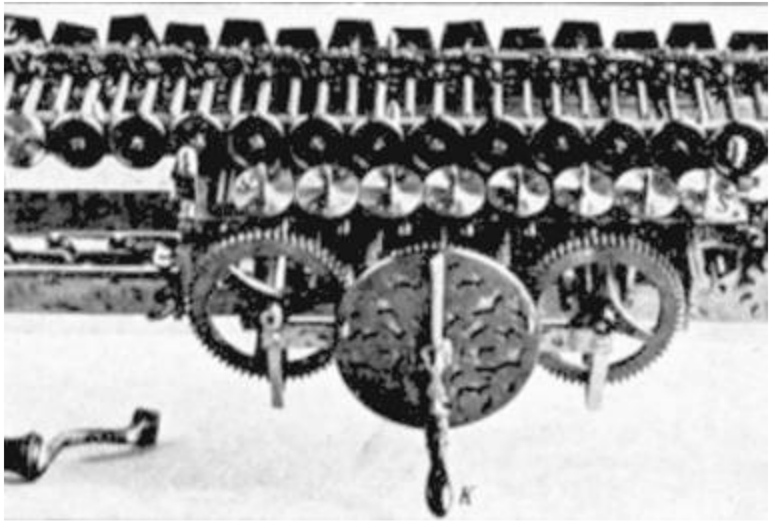
Abacus (算盘) ~ 900 AD

Three, set five remove two (abacus rule)

Slide rule ~ 1600 AD

Mechanical Calculator

Pascal mechanical calculator ~ 1642-1644 AD



Leibnez calculator ~ 1672 AD

Electronic Computers

First computer ENIAC (1946)

Size: 30×50 ft^2,
Weight: 30T
Components: 17000 vacuum cube
Capability: 5000 OP/s.

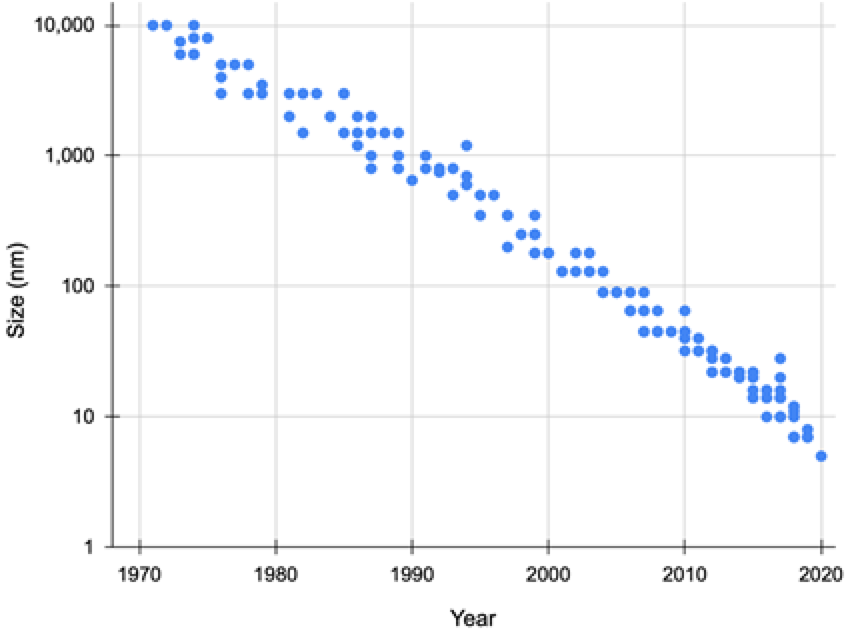
Vacuum cube (1940-1950s)
Transistors generation (1950-1960s)
Integrated Circuits (1960-1070s)

PC, Tablet, and
smart phones

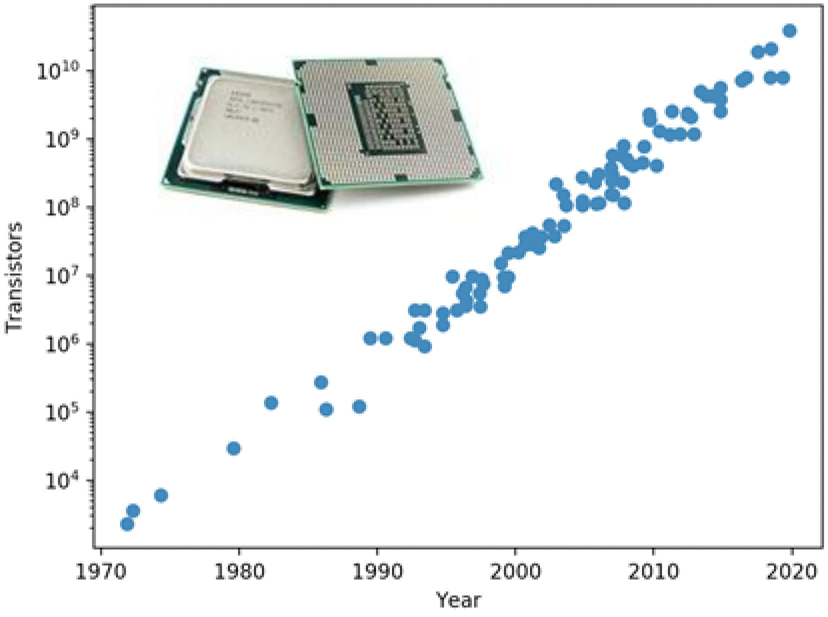
Capability: TFLOPs

The chip development – scale up

Size of transistor

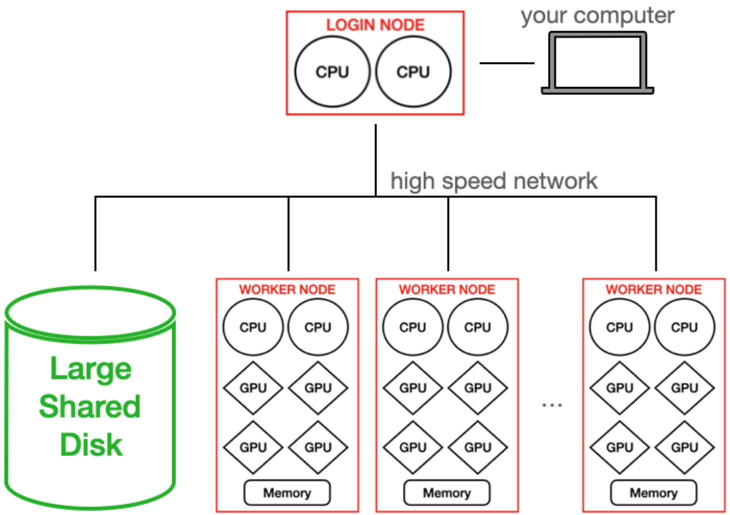


Number of transistor on a chip



https://github.com/karlrupp/microprocessor-trend-data
https://en.wikipedia.org/wiki/Transistor_count

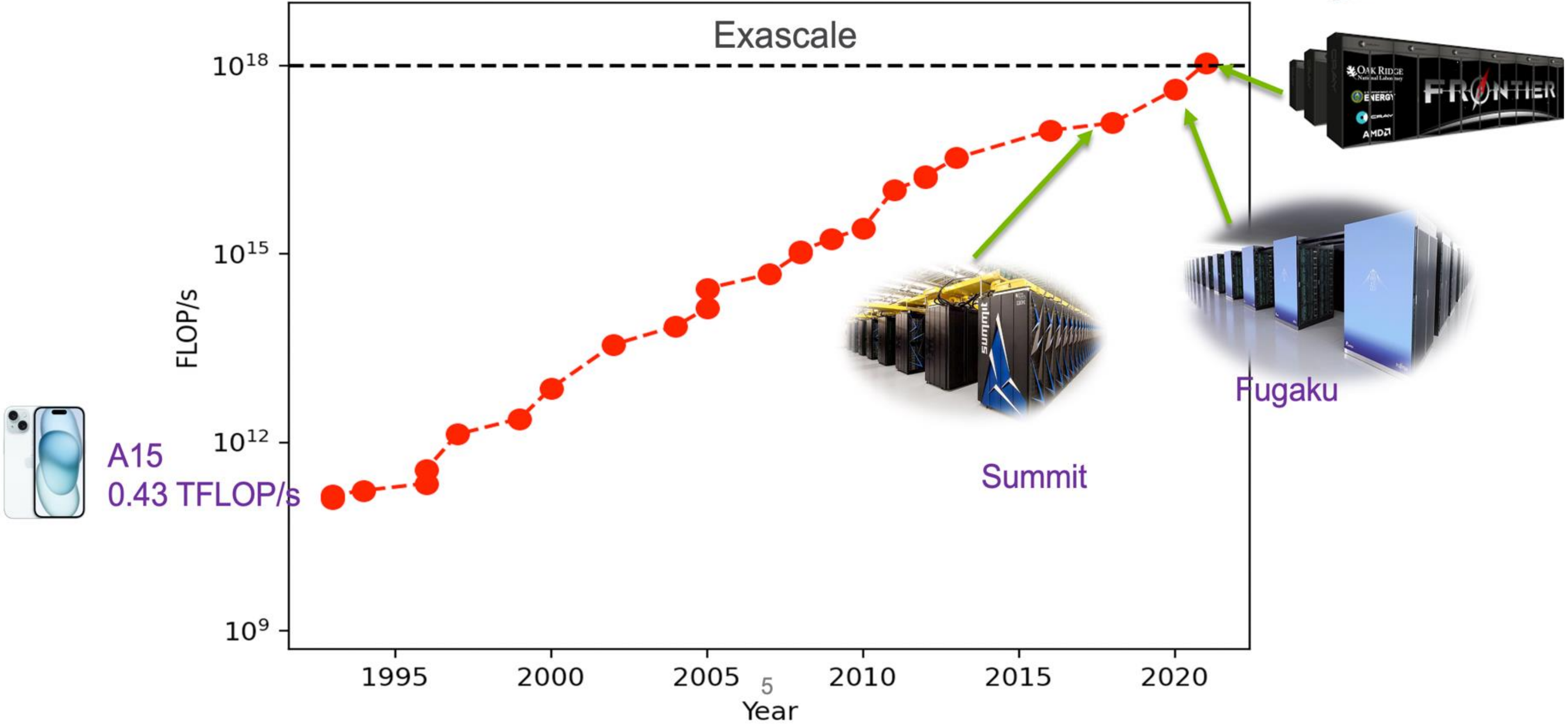
Supercomputers – scale out



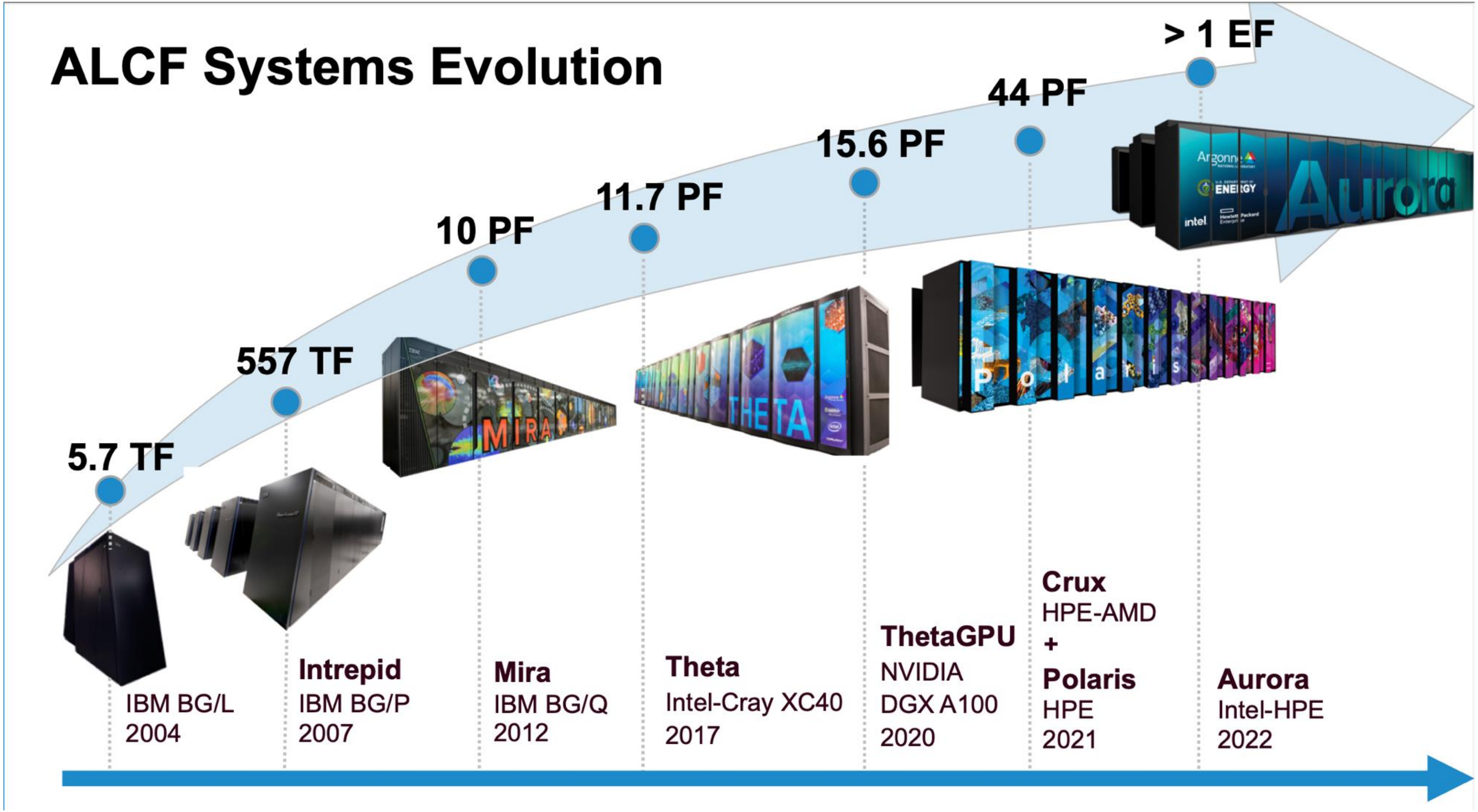
- Multiple CPUs and/or GPUs are combined into a single node.
- All the nodes are connected through high-speed network interconnect that allows it to communicate with other nodes and to a large shared filesystem.

https://www.alcf.anl.gov/aurora

COMPUTING POWER OF SUPERCOMPUTERS



Plot generated from data https://en.wikipedia.org/wiki/History_of_supercomputing



ALCF Systems Evolution

IBM BG/L
2004

Intrepid
IBM BG/P
2007

Mira
IBM BG/Q
2012

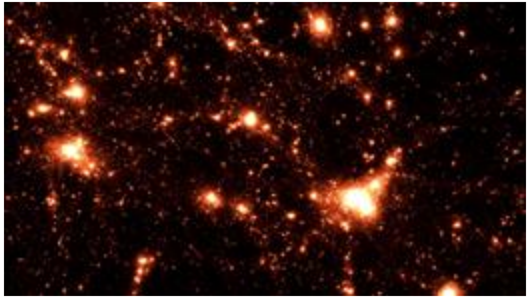
Theta
Intel-Cray XC40
2017

ThetaGPU
NVIDIA
DGX A100
2020

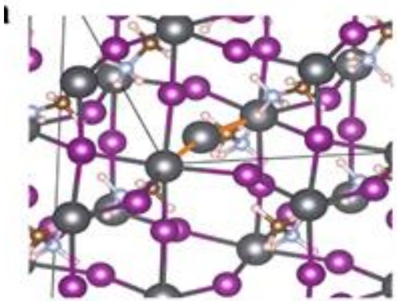
Crux
HPE-AMD
+
Polaris
HPE
2021

Aurora
Intel-HPE
2022

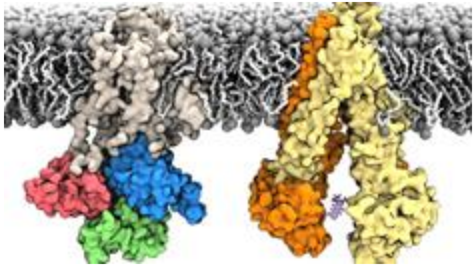
Science on Supercomputer



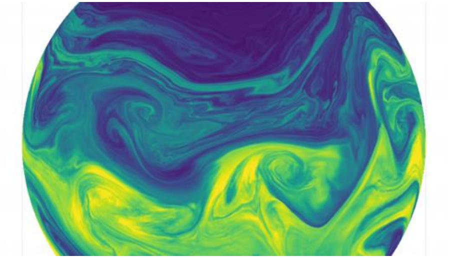
Cosmology



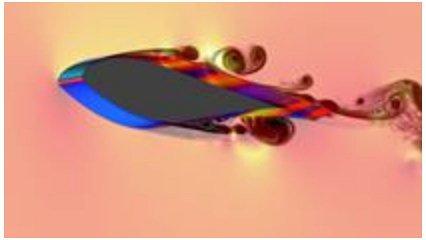
Materials science



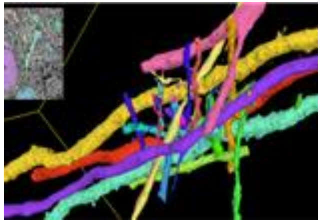
Biology



Climate modeling



Engineering



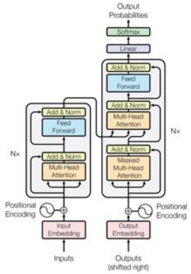
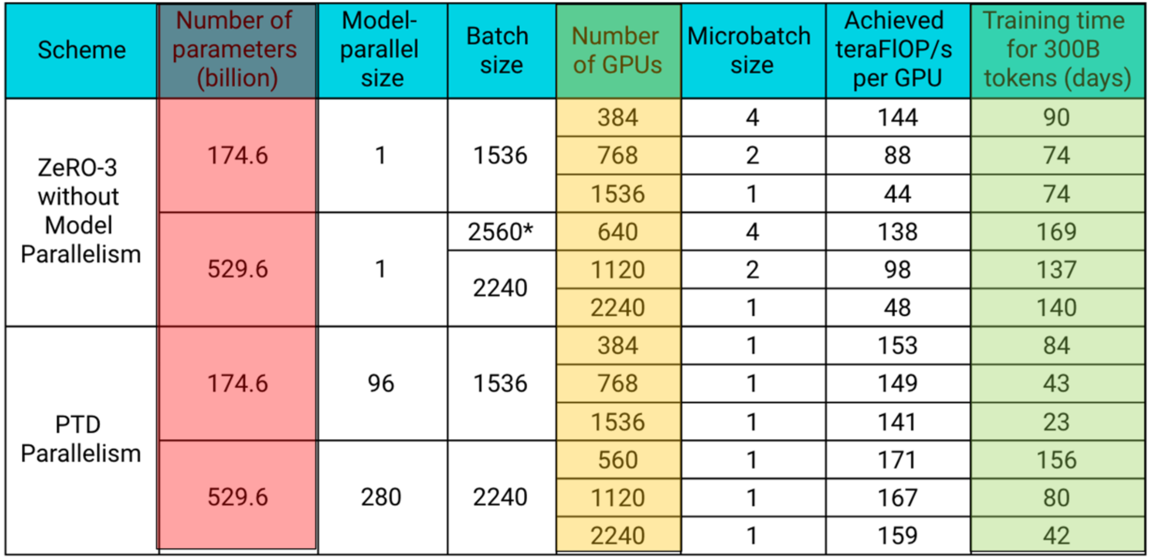


Figure 1: The Transformer - model architecture.

Computer vision & AI

https://www.alcf.anl.gov/science/projects/

Why do we need supercomputer for AI?



Scheme	Number of parameters (billion)	Model-parallel size	Batch size	Number of GPUs	Microbatch size	Achieved teraFlOP/s per GPU	Training time for 300B tokens (days)
ZeRO-3 without Model Parallelism	174.6	1	1536	384	4	144	90
				768	2	88	74
				1536	1	44	74
	529.6	1	2560*	640	4	138	169
			2240	1120	2	98	137
				2240	1	48	140
PTD Parallelism	174.6	96	1536	384	1	153	84
				768	1	149	43
				1536	1	141	23
	529.6	280	2240	560	1	171	156
				1120	1	167	80
				2240	1	159	42

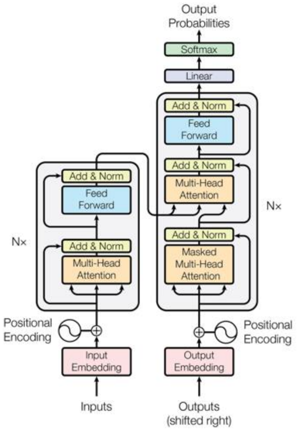
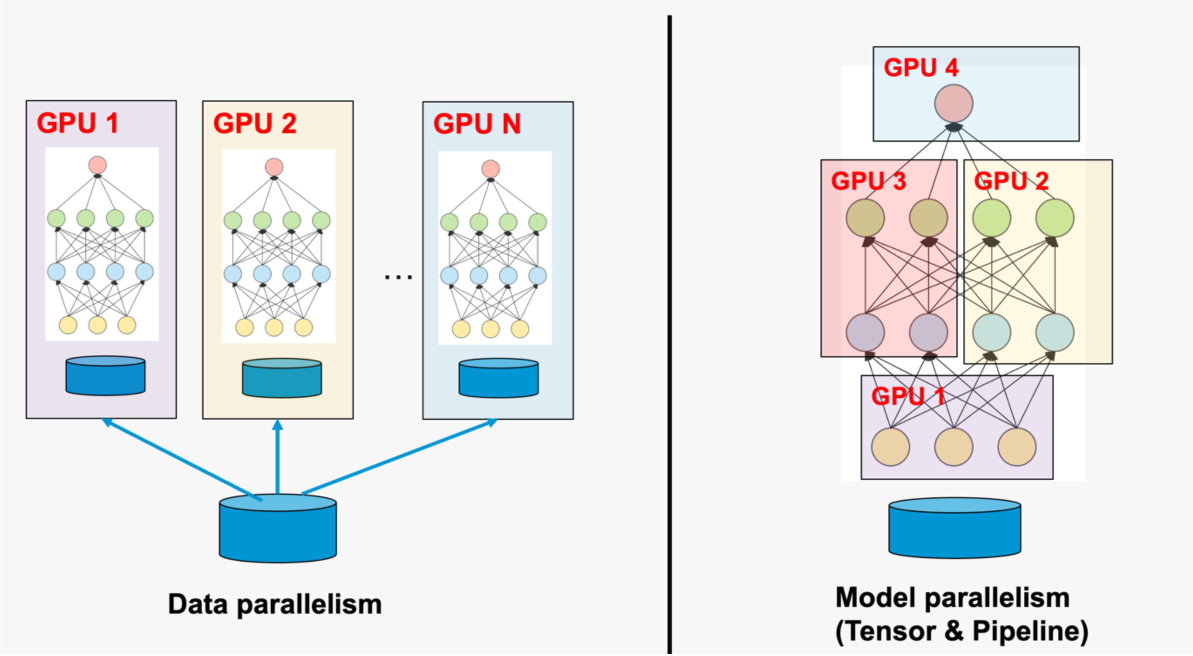


Figure 1: The Transformer - model architecture.

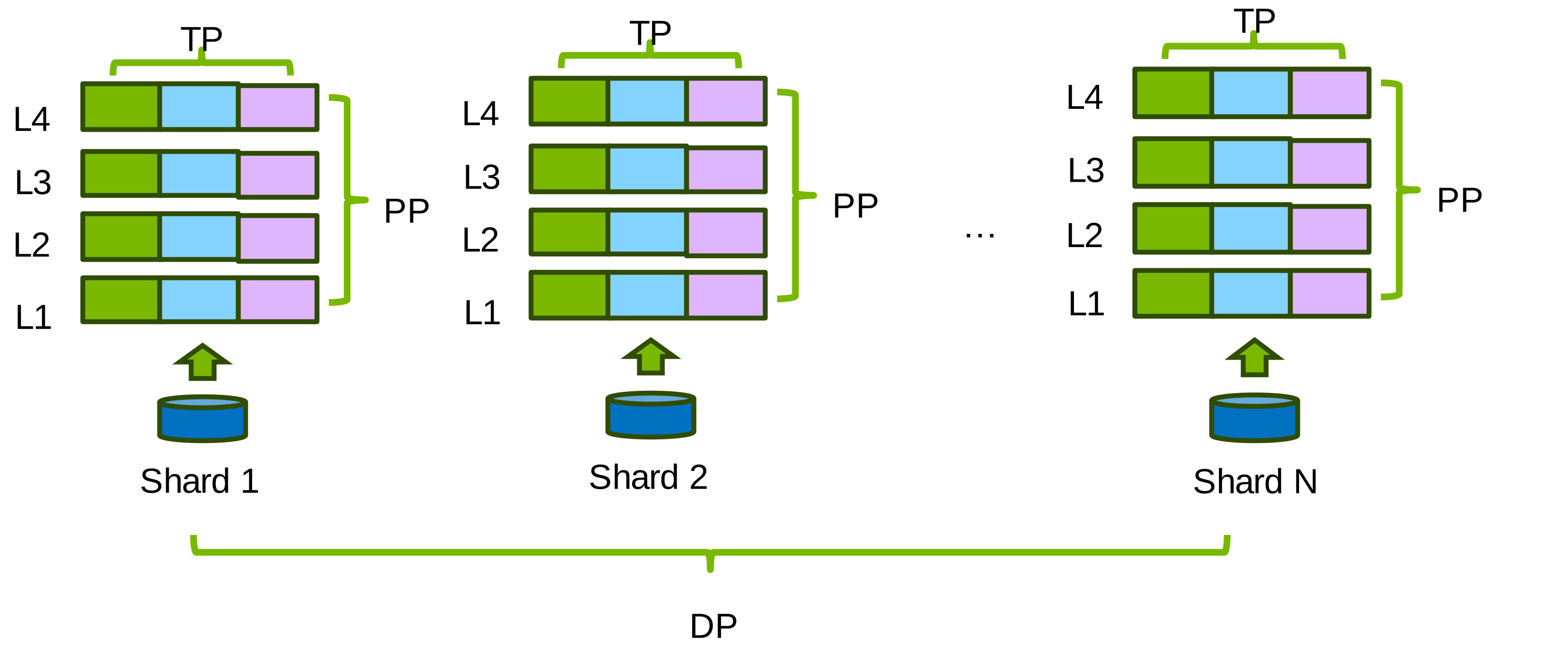
Time for training LLM models

Parallelization for AI – distributed training



3D parallelism for LLM

- Tensor (TP): Split each layer.
- Pipeline (PP): Distribute different layers.
- Data (DP): sharding dataset.



Outline

1. Evolution of computing systems
2. Parallel computing
3. AI in a nutshell

Key words: supercomputer, parallel computing, AI