The Parallel Computing Revolution Is Only Half Over

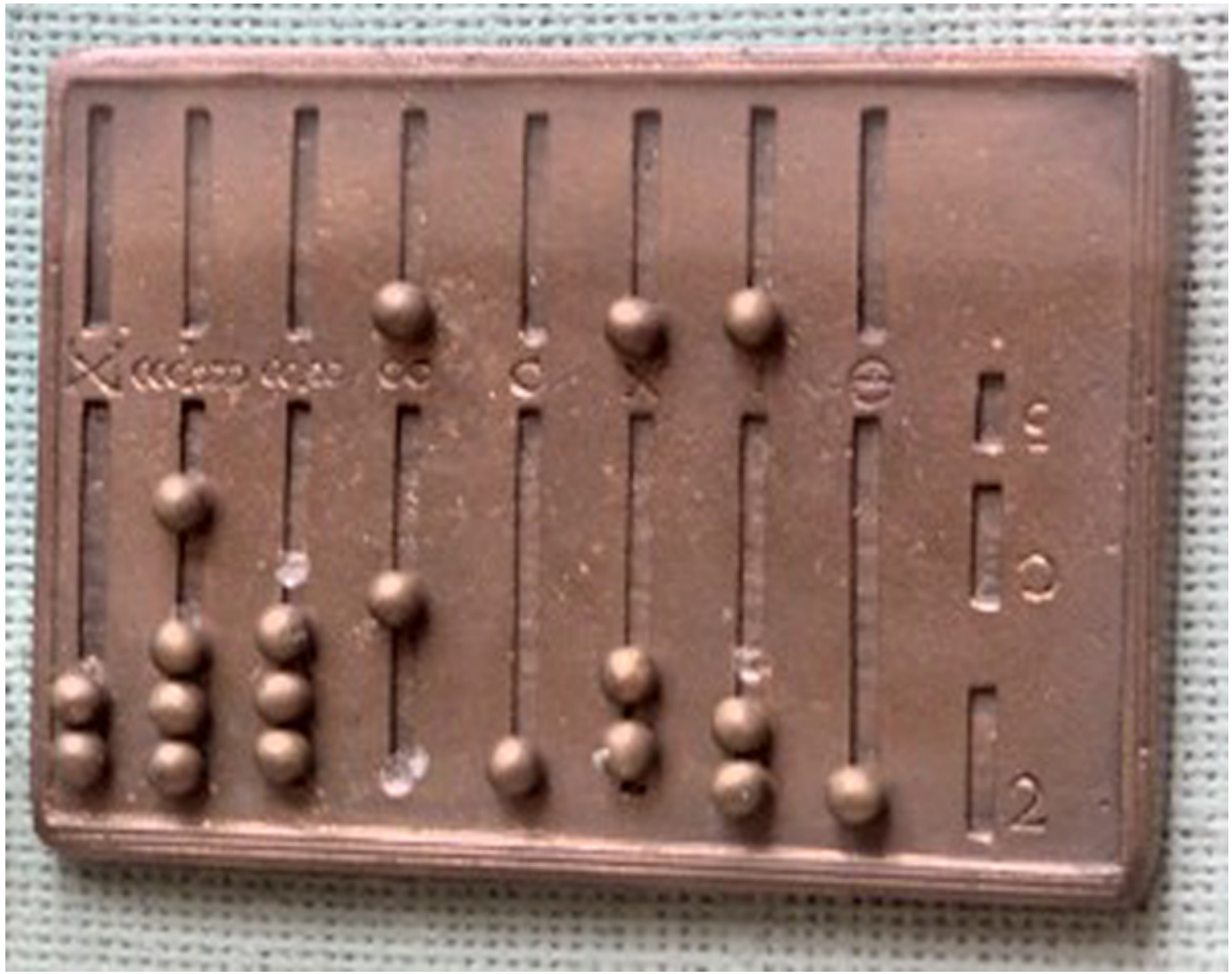
Rob Schreiber, Cerebras Systems, Inc.

ATPESC, August 1, 2019

The Past

What this is?

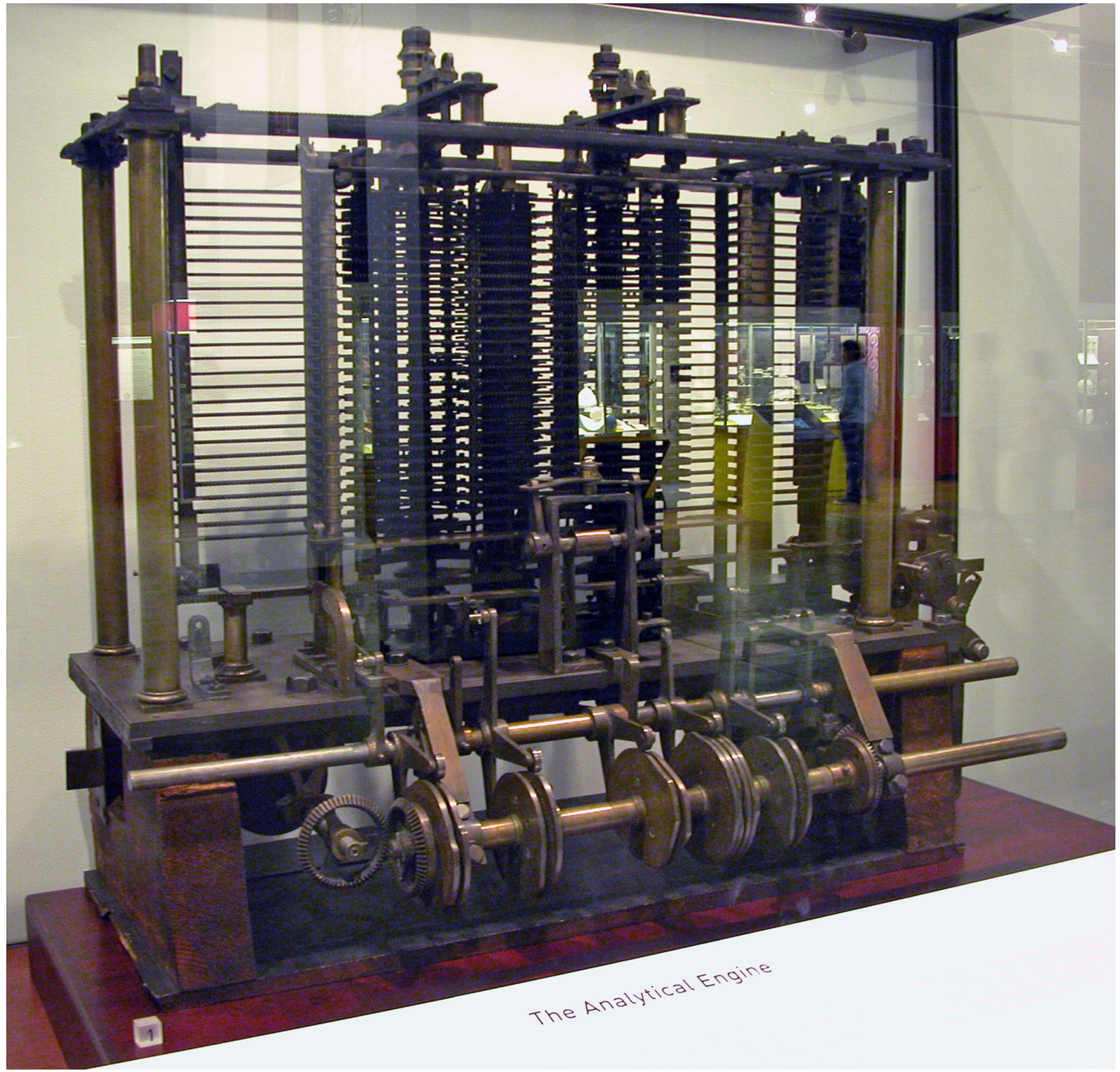
And this?



And this?


Centaines de Mille
Dizaines de Mille
Milles
Centaines
Dizaines
Nombres Simples
Sols
Deniers

And this?

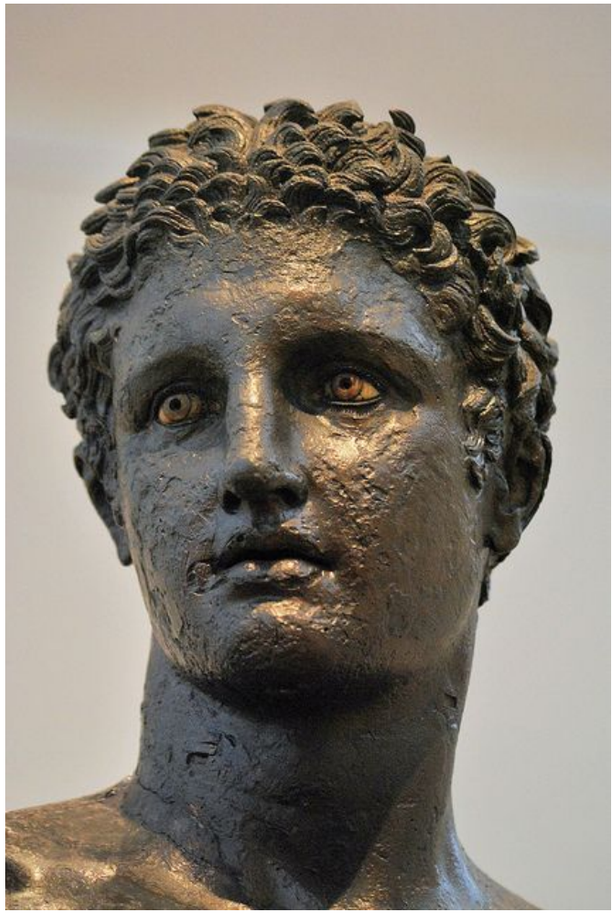


Computer History Museum Timeline: 1933

And this?

2.4 million cores

The pioneers



The parallel computing era

Controversy

Success!

Technology Triumphant

The Crisis

A Pressing Need

A New Era

Parallel computers were once controversial

Amdahl

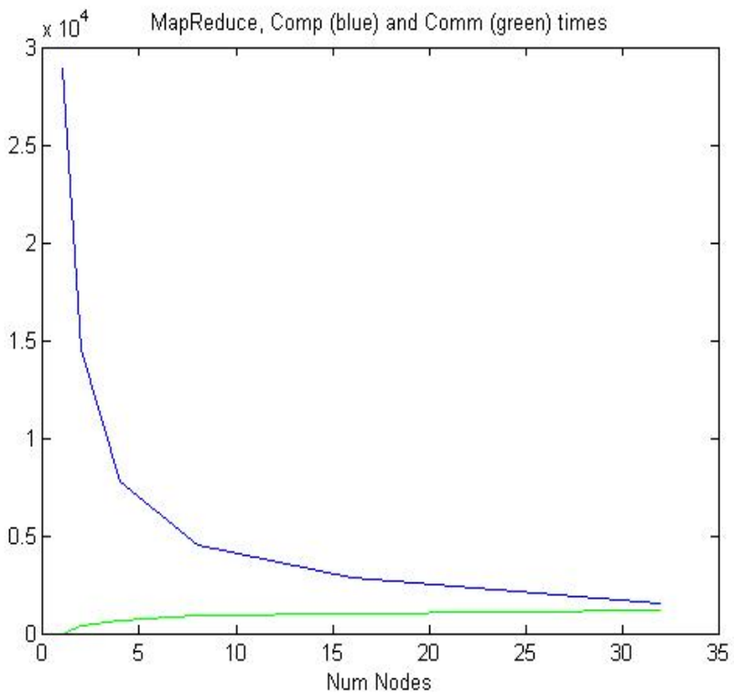
If the parallel fraction is *f* then the maximum possible speedup is
1/ *f*

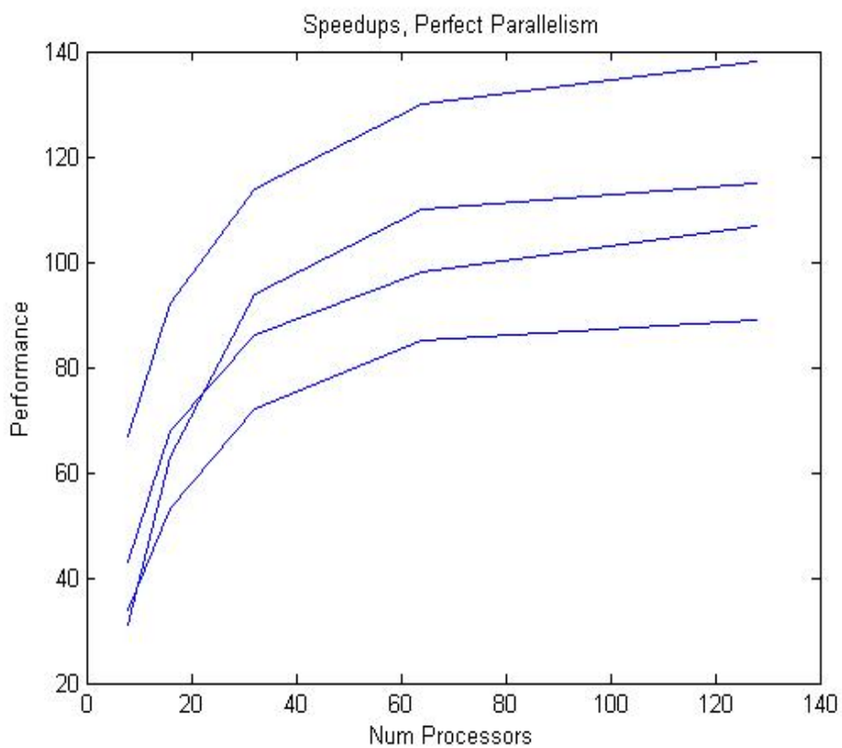
Speedup, efficiency

$$S(p) = T(1) / T(p)$$

$$E(p) = S(p) / p$$

Mapreduce: Parallelism is easy. Performance is hard





It’s the memory and the network

And the algorithm

Dusty Decks

Vendor: “I can build you a machine that is a ***billion*** times more powerful than the one you used earlier”

Customer: “But will I need to rewrite my code?”

Automatic parallelization

The compiler should create an optimized, parallel implementation of the algorithm in the code

Alan Perlis

Adapting old programs to fit new machines usually means adapting new machines to behave like old ones.

Optimization hinders evolution.

http://www.cs.yale.edu/homes/perlis-alan/quotes.html

Success

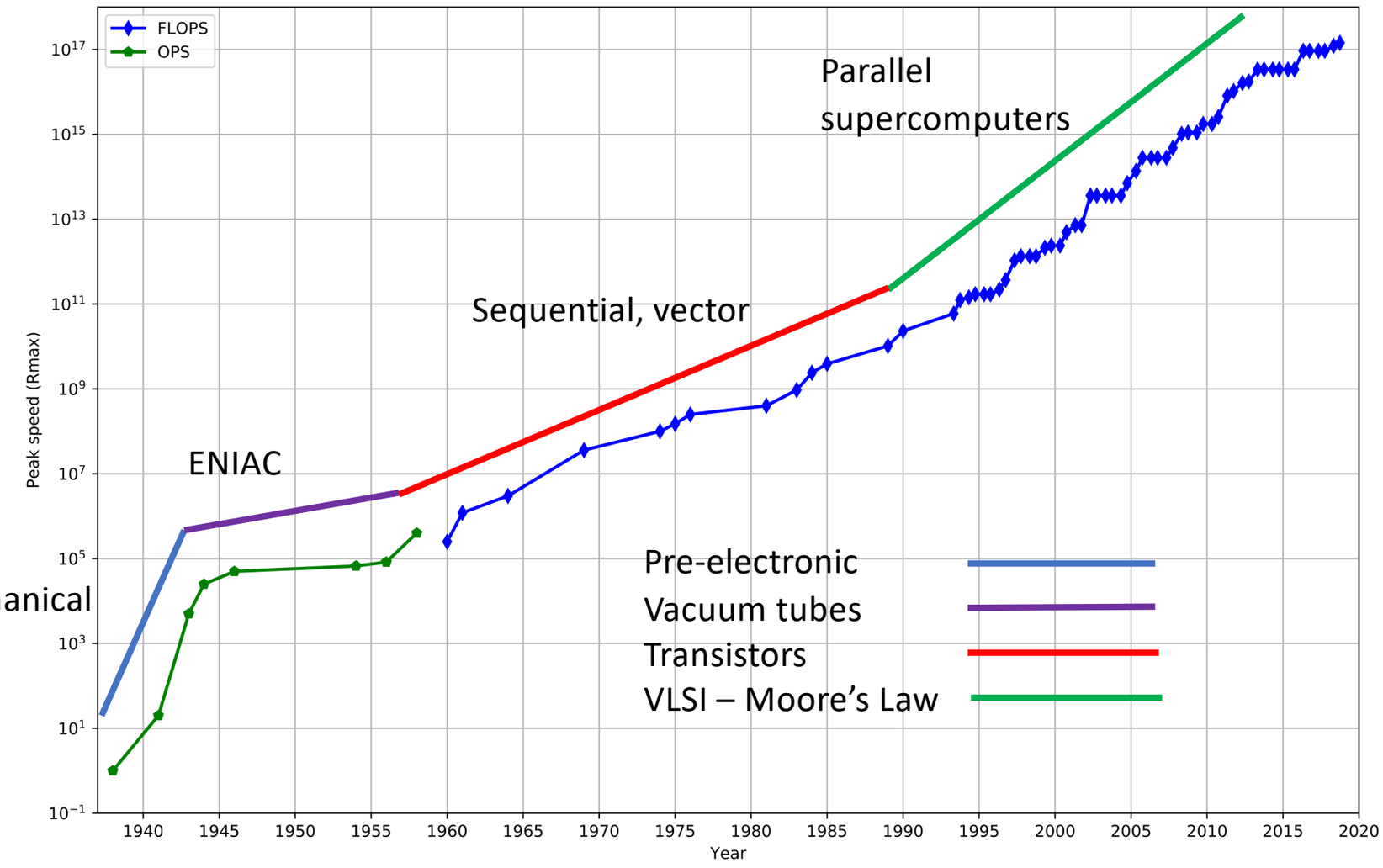
Almost all of HPC *machine* performance is due to parallelism

- 1975 – 2019: 10^8 flops → 10^{17} flops
  - 1 OOM from clock rate
  - 8 OOM from parallelism

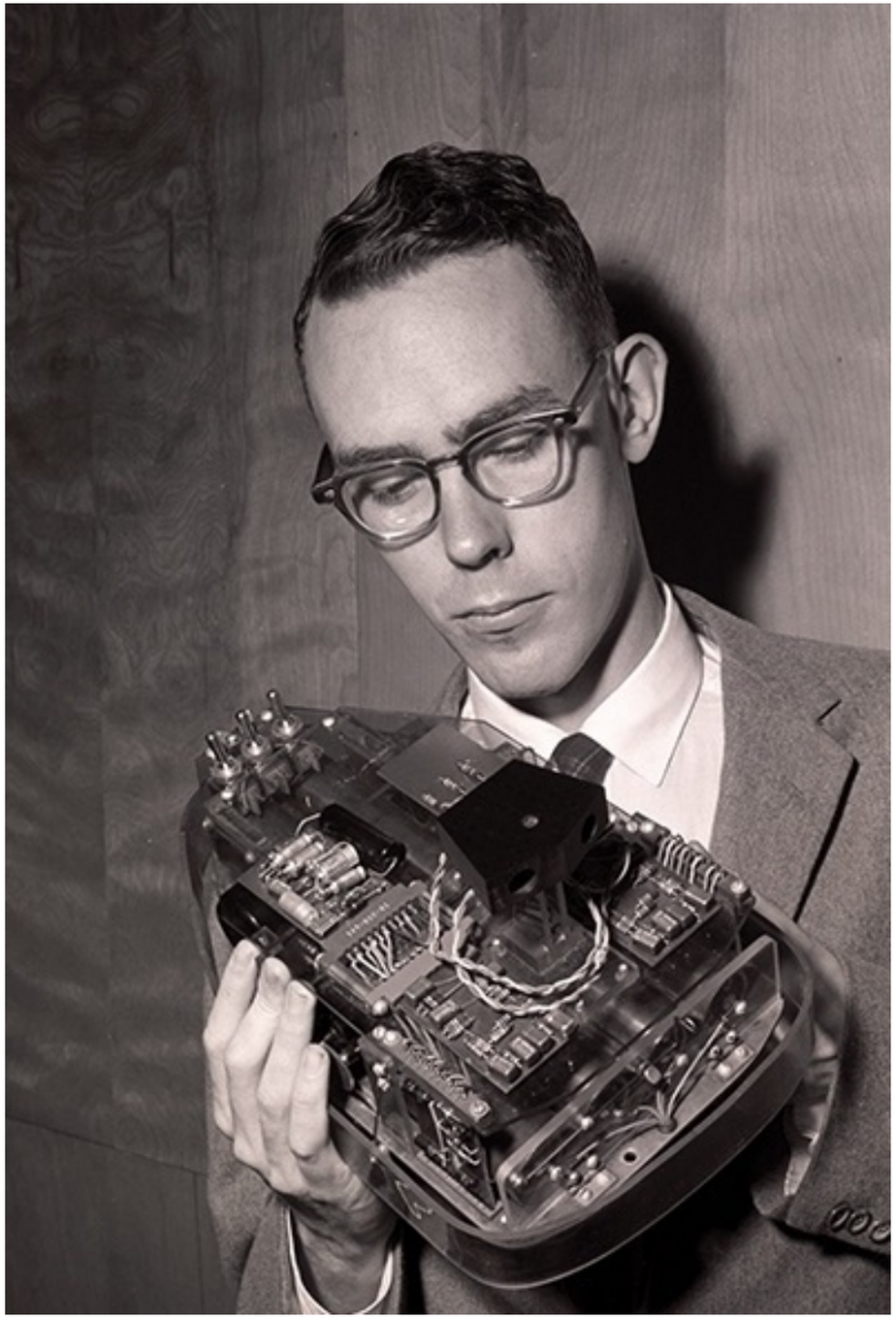
- 2 OOM more power required
- 1 OOM more $ required

Four eras, four exponentials

Electro-mechanical calculators



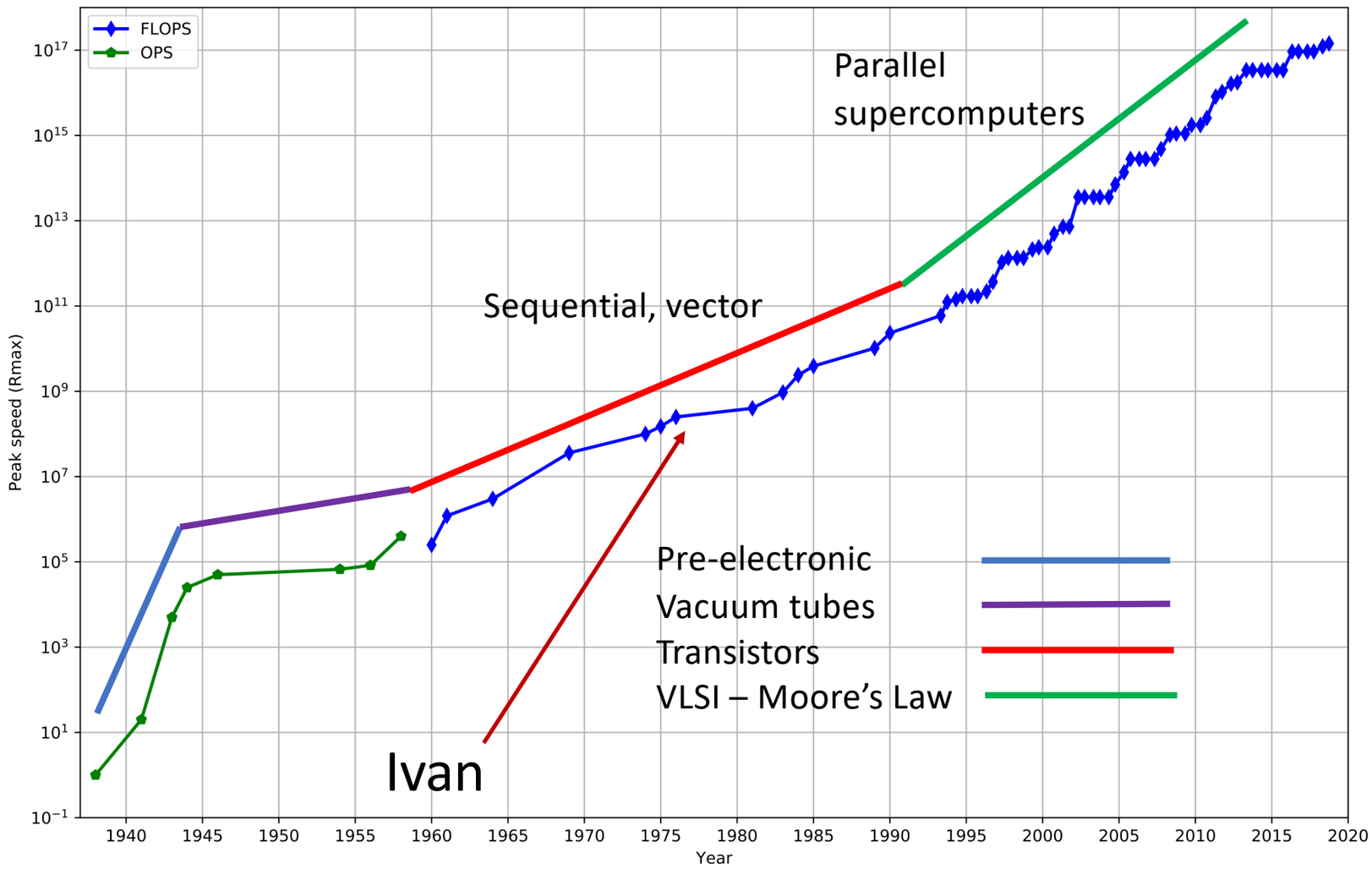
The past is easy to predict. The future?



Ivan Sutherland, at Caltech, 1977

“The VLSI Revolution is Only Half Over”

Was Sutherland right?

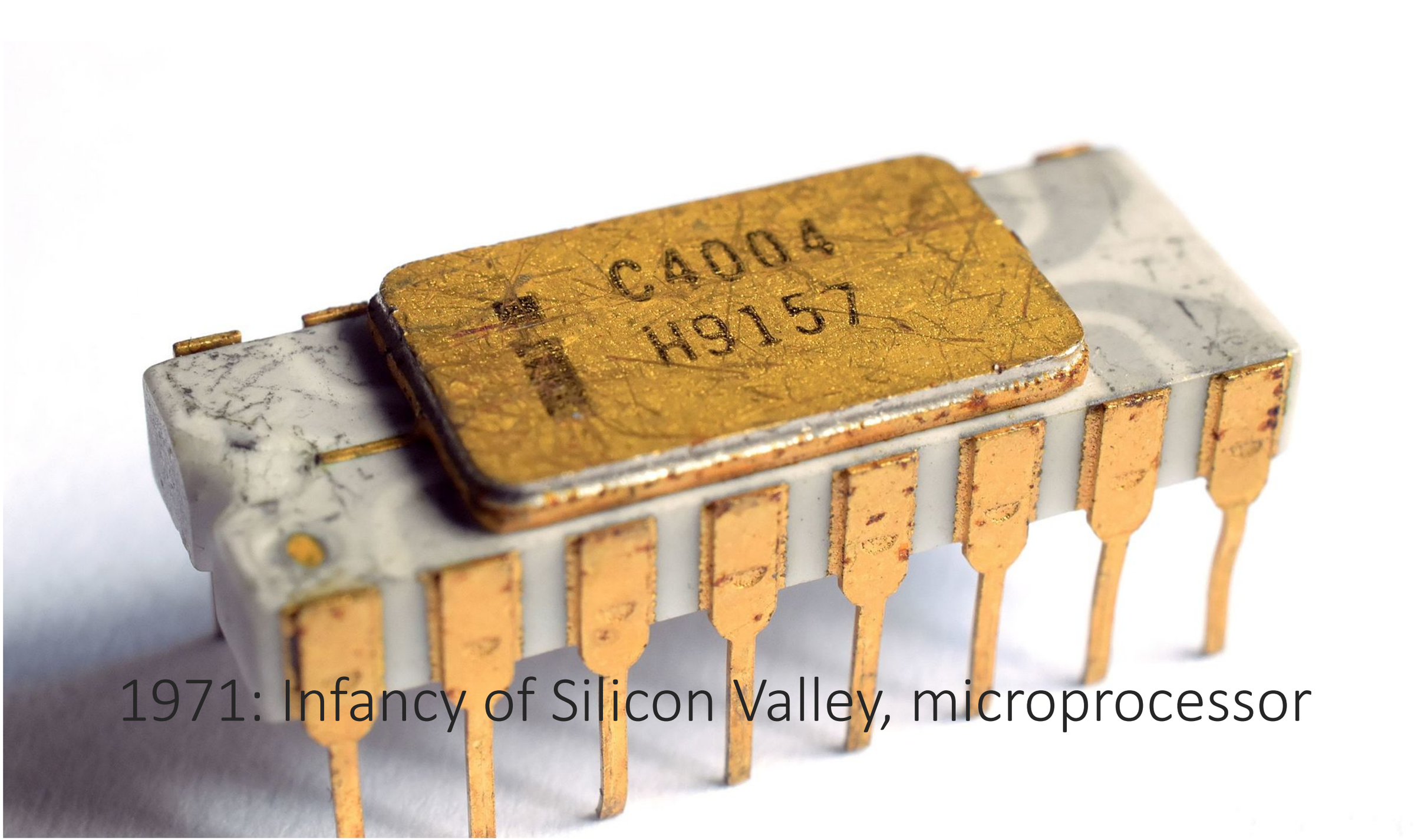


VLSI: Technology Triumphant

After Ivan, What Happened?

1975: HPC is based on high-end technology

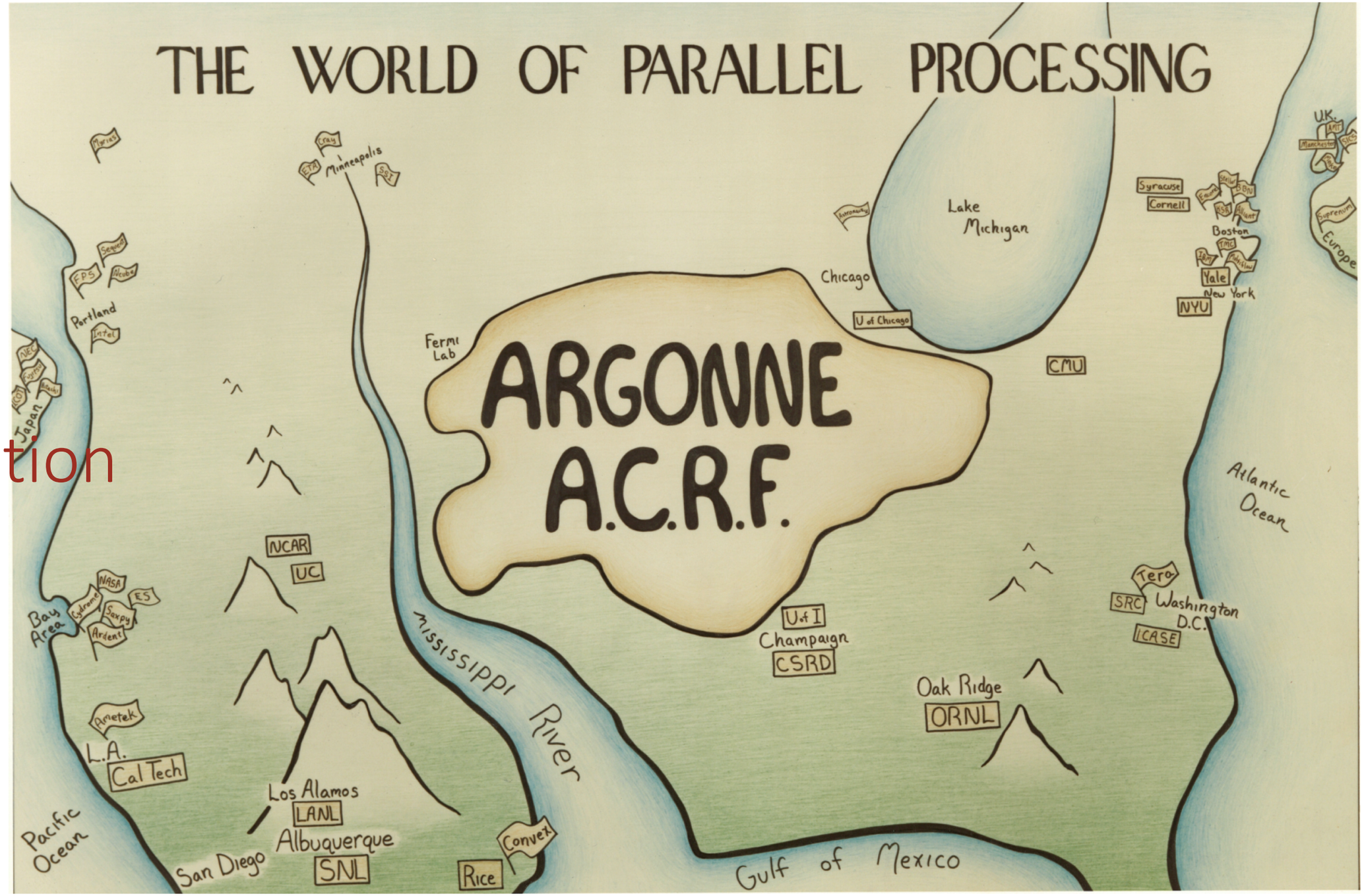
Hardware was customized for HPC



1971: Infancy of Silicon Valley, microprocessor


CHINA
JAPAN
RUSSIA
PACIFIC OCEAN
LOS ANGELES
MEXICO
TEXAS
UTAH
LAS VEGAS
NEBRASKA
KANSAS CITY
CANADA
Chicago
WASHINGTON D.C.
JERSEY
HUDSON RIVER
10TH AVE
9TH AVE
PARK
STEINBERG

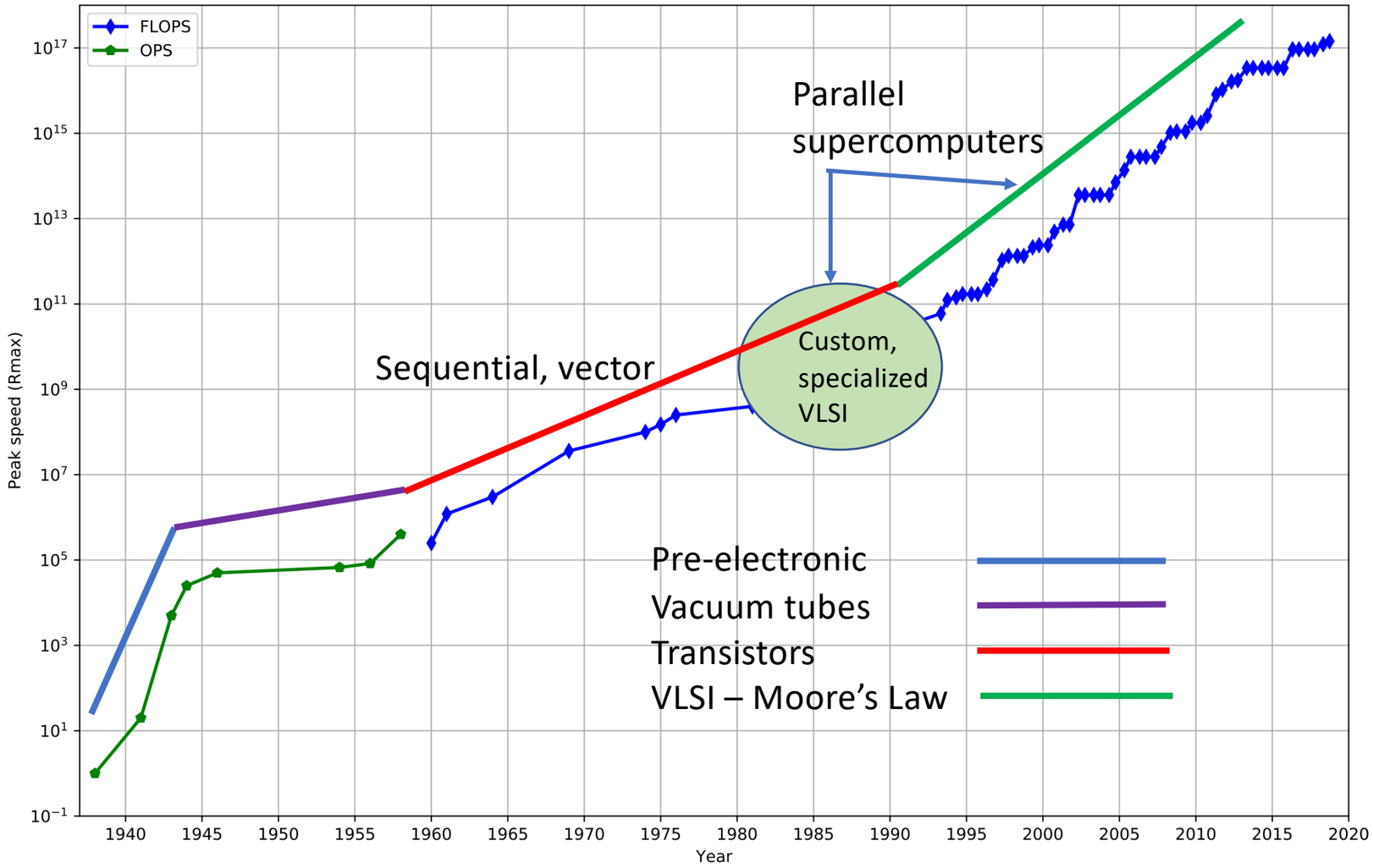
1980s: An era of exploration



Architectures – no one knew what to do

- SIMD / MIMD
- Shared memory / Distributed memory
- COMA
- Hypercubes
- Transputers
- Message-passing dialects
- Latency tolerance
- From workstations to supercomputers

The confused 80s



1971 – 1990: Micros catch up

- Transistor count doubles every two years.
- 1971 – 2000 transistors (i4004)
- 1990 – 1,000,000 transistors (i80486)

- It took 19 years of exponential growth to catch up to old style computers for HPC
- But then…

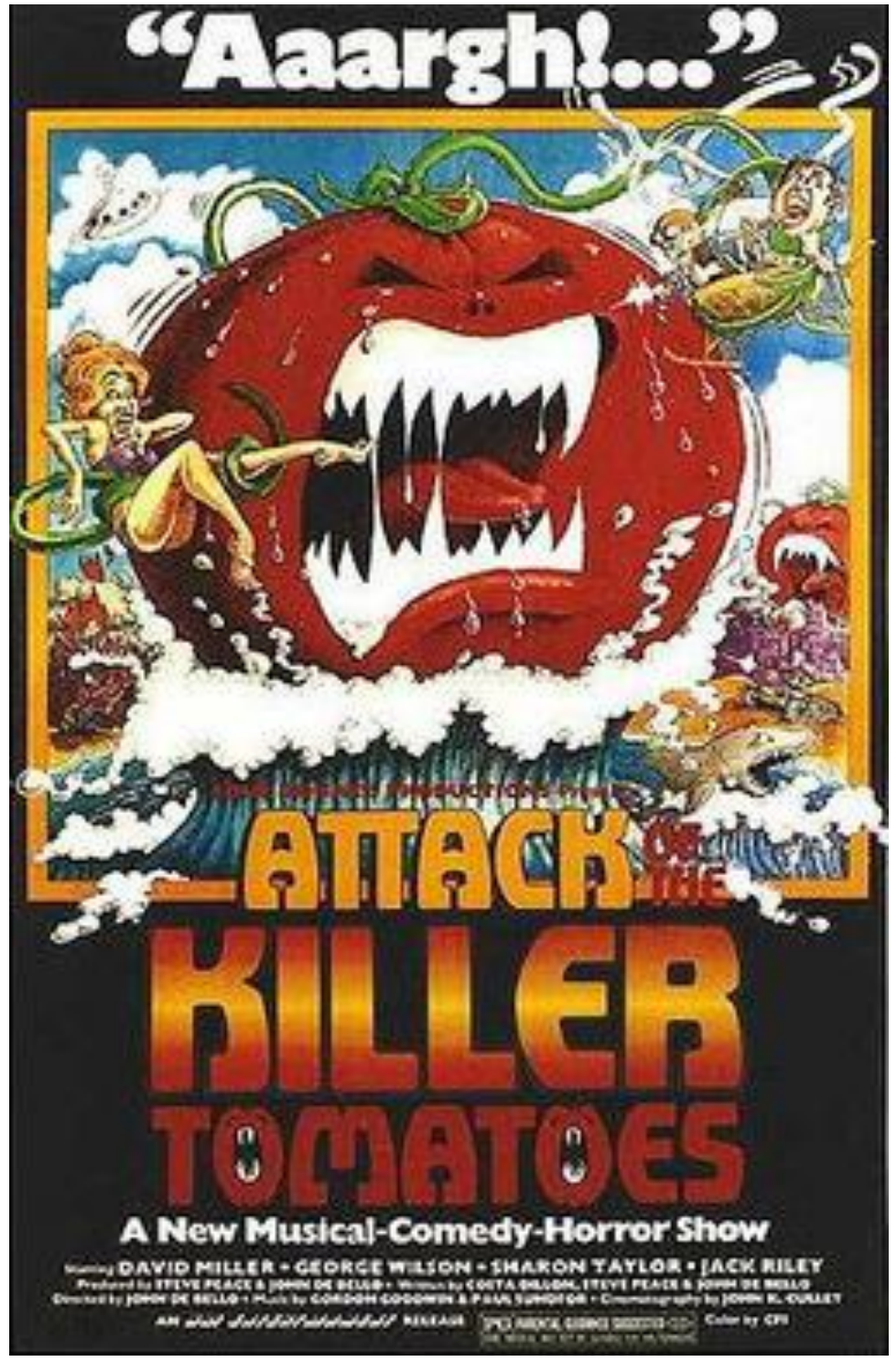
Killer micros

“No one will survive The Attack of the Killer Micros”

Eugene Brooks, Panel talk at Supercomputing 90.

Cray Research says it is worried by "killer micros" – compact, extremely fast work stations that sell for less than $100,000.

John Markoff, in the New York Times, 1991



VLSI and supercomputers after 1990

Then the attack really happens

- A mass market drove investment and innovation
  - Dennard scaling made micros faster, cheaper, same power
  - Commodity pricing killed architectural specialization

- **By 2000, all HPC machines are clusters of commodity designs**

- **AND – we knew how to program them (MPI)**

From 1990 to 2010

7 OOM, from 0.1 teraflops to 0.1 exaflops

Greater than the gains from 1942 (pre-electronic) to 1990

A loss of computer design diversity; the one-size-fits-all processor

The Crisis

Dennard Scaling Ended

… in every technology generation the transistor density doubles, the circuit becomes 40% faster, and power consumption (with twice the number of transistors) stays the same.

- 2005-2007: leaky transistors. V_{dd} stops dropping. Clock rate hits a wall.

Moore’s Law limit

“There’s no getting around the fact that we build these things out of atoms.”

-- Gordon Moore

The transition, to the post-Moore era

A pressing need

AI --- for science too

- Compute demand for DNN training has grown 300,000x since 2012
- Doubling time 3.5mo vs 18 mo (Moore)
- Training a DNN can take weeks

- see, ref https://blog.openai.com/ai-and-compute/

cerebras

Processor Specialization

	Area (mm^2)	Transistors (billion)	Teraflops
CPU	800	20	1
GPU	500	10	10

First GPGPU. Now, a new class of AI-optimized accelerators

- **Purpose-built** compute engine
- **More parallel compute on each chip**

AI-optimized accelerators are here

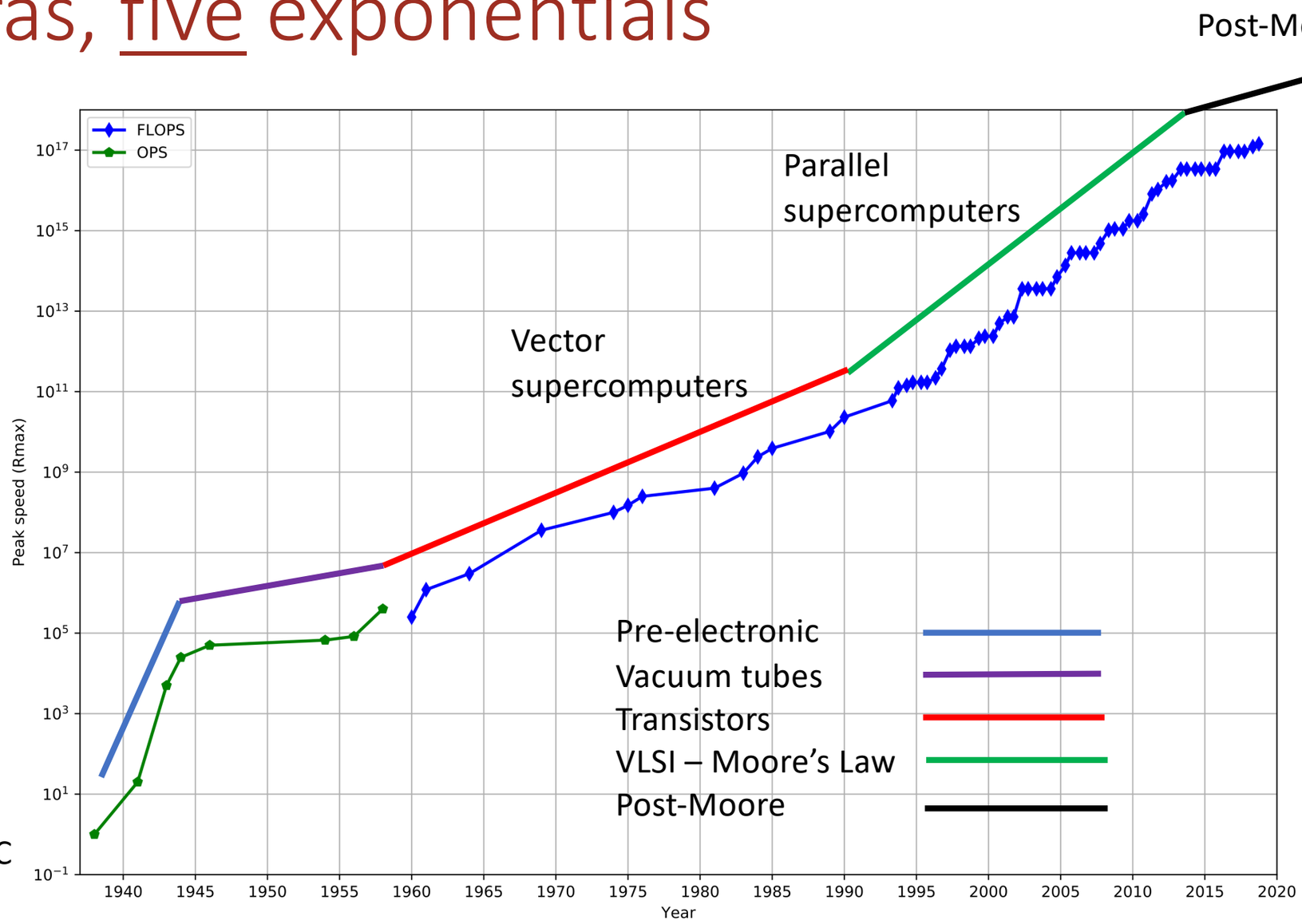
Compare chips

	Area (mm^2)	Transistors (billion)	Teraflops
CPU	**800**	**20**	**1**
GPU	**500**	**10**	**10**
AI-optimized (Volta, GC, TPU)	**800**	**20**	**100**

Still not enough! **Days to train Resnet-50 on Volta**

A New Era

Six eras, five exponentials



Mechanical devices, from ~100 BC

Scaling everything else

- **Specialized architectures**
- **A heterogeneous Top 500 list**
- **Heterogeneous clusters of heterogeneous nodes**
- **Better algorithms**
- **The AI revolution**
  - **In science as well as all other applications of computing**
- **New memory technology**
- **Photonics**
- **2.5D and 3D interconnects**
- **Wafer scale**
- **Quantum**
- **Other stuff too weird to mention, yet**

Chip scaling and feature scaling

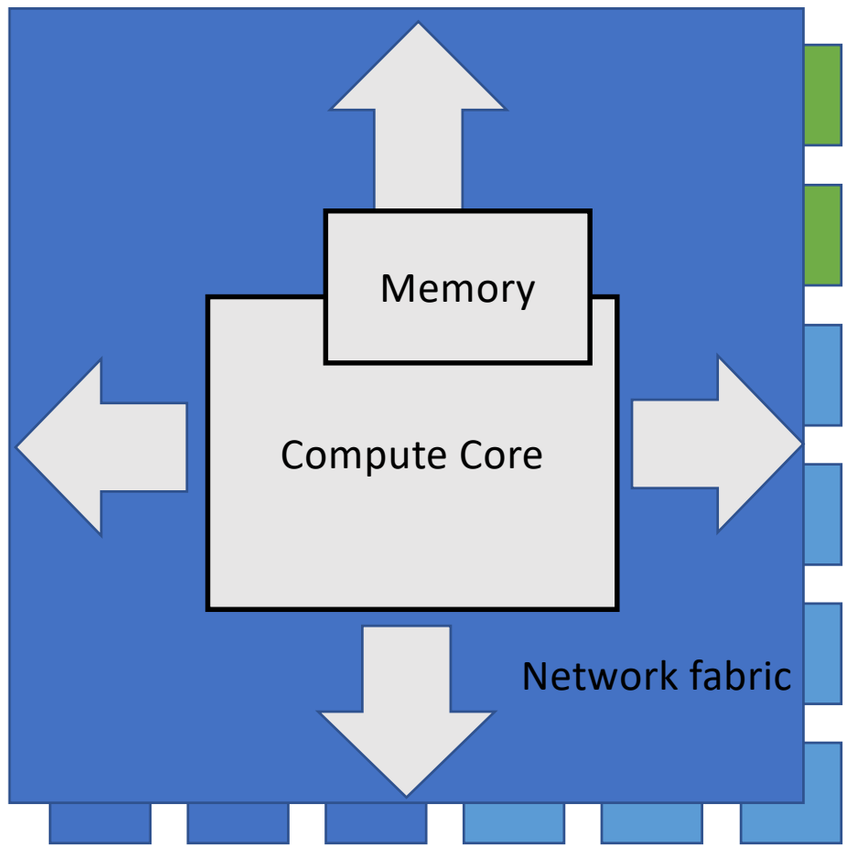
When feature size scaling stops, use bigger chips

EE Times, May 24, 2019

Startups Cerebras, Habana, and UpMem will unveil new deep-learning processors. Cerebras will describe a much-anticipated device using wafer-scale integration.

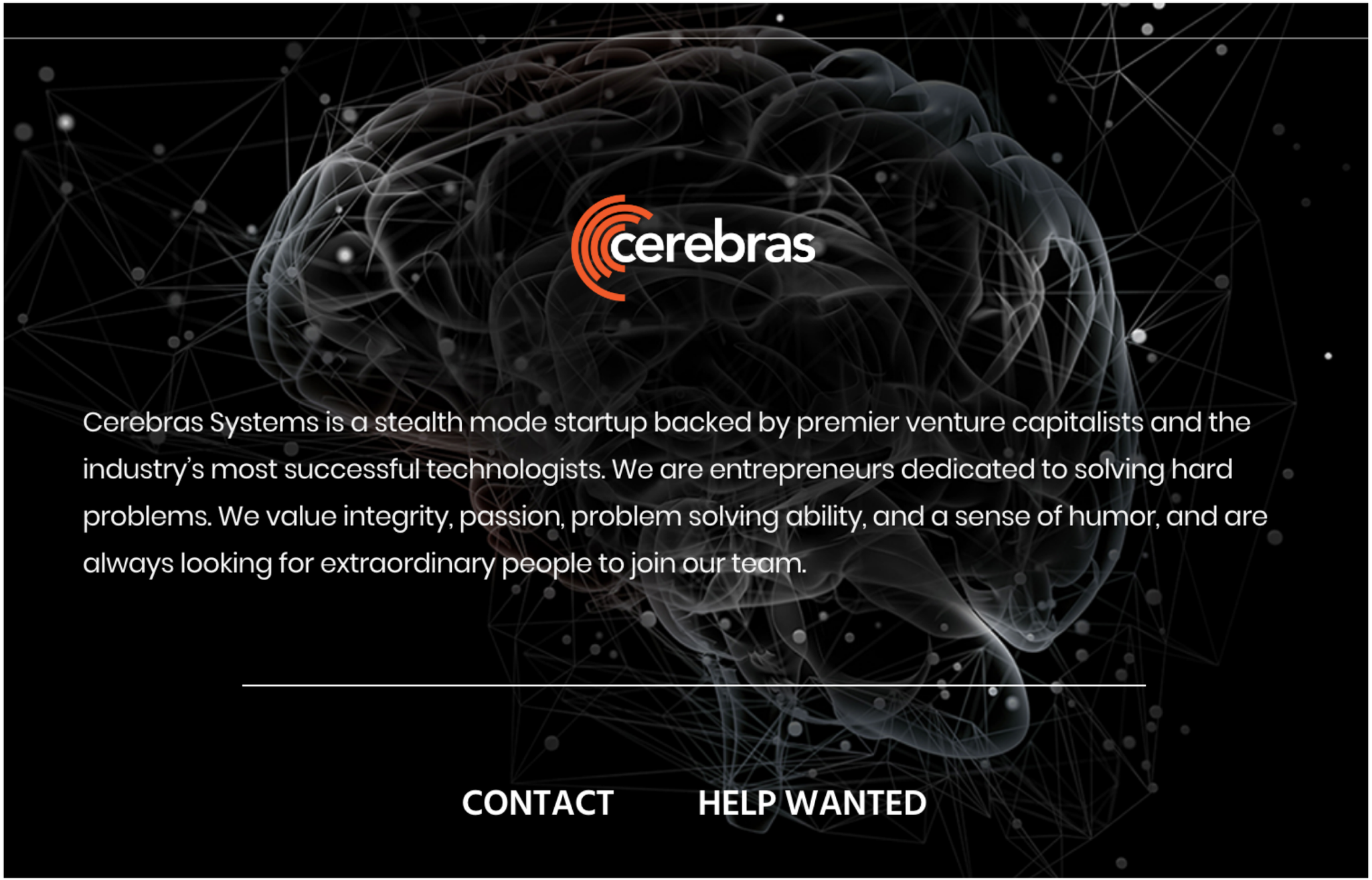
Accelerating AI

- Distributed memory
- Tightly integrated, fine-grain, active messages
- One-clock per hop network latency



Scaling the area too

	Area (mm^2)	Transistors (billion)	Teraflops
CPU	**800**	**20**	**1**
GPU	**500**	**10**	**10**
AI-optimized	**800**	**20**	**100**
Cerebras	**> 50X**	**to be announced**	



Cerebras Systems is a stealth mode startup backed by premier venture capitalists and the industry's most successful technologists. We are entrepreneurs dedicated to solving hard problems. We value integrity, passion, problem solving ability, and a sense of humor, and are always looking for extraordinary people to join our team.

CONTACT **HELP WANTED**