

A large, light gray stylized 'C' logo on the left side of the slide, composed of several concentric, slightly offset arcs.

Wafer-scale: Here Now

Seattle Snowmass
22 July 2022
Rob Schreiber

Dennard scaling ended

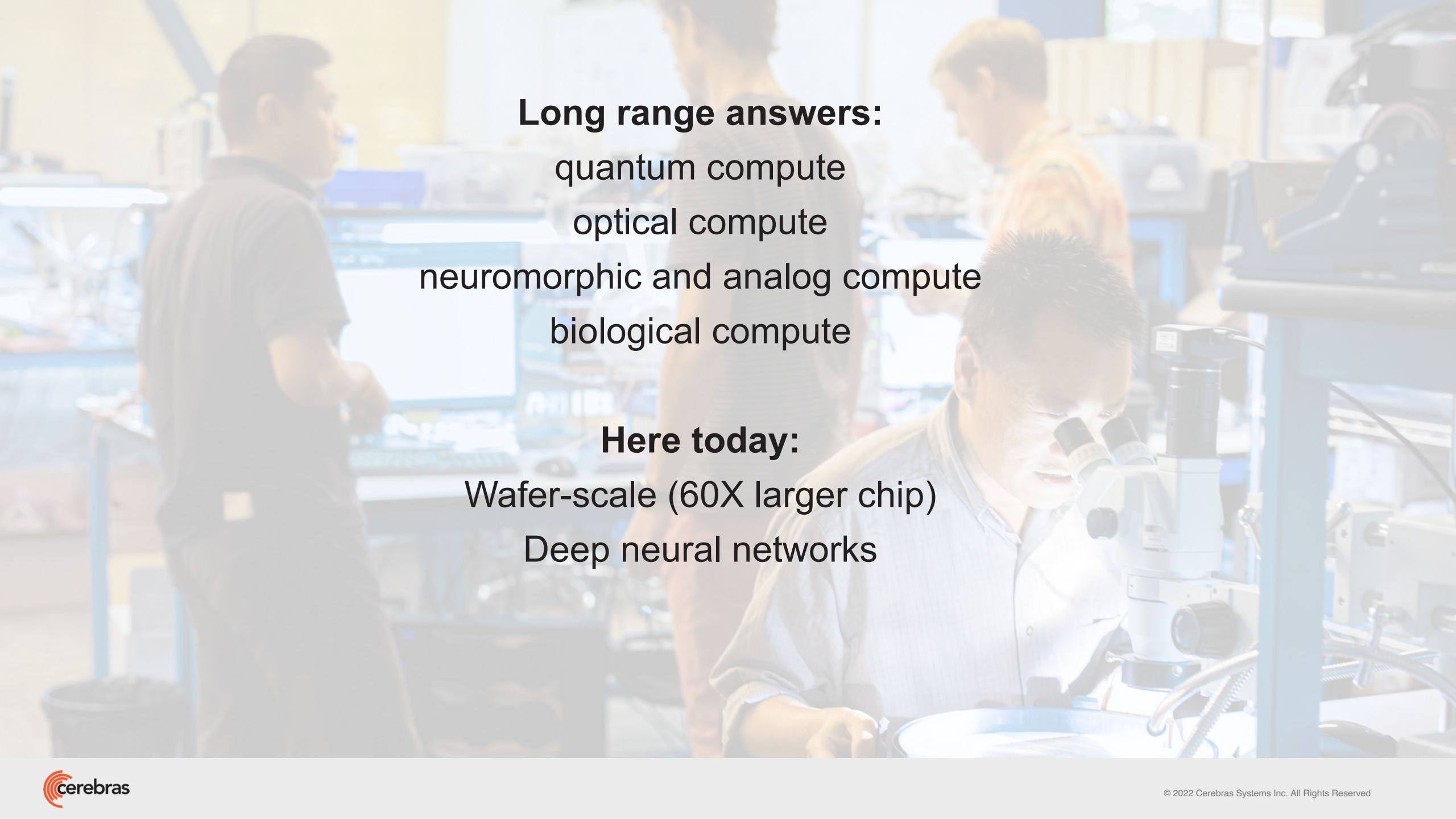
Moore: There's no getting around the fact that we make these things out of atoms

Energy costs dominate

Moving data over macroscopic, high-capacitance wires costs energy

Memory is too slow

The network is a (slow) IO device to the compute node



Long range answers:
quantum compute
optical compute
neuromorphic and analog compute
biological compute

Here today:
Wafer-scale (60X larger chip)
Deep neural networks



Cerebras Wafer-Scale Engine (WSE-2)

The largest chip in the world

850,000 cores optimized for sparse linear algebra

46,225 mm² silicon

2.6 trillion transistors

40 Gigabytes of on-chip memory

20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology

Cluster-scale performance in a single chip

Cerebras CS-2 System

- The world's most powerful AI computer
- Company founded 2016, >400 employees, deep chip-making expertise, AI focus
- Succeeded at wafer-scale integration, a first
- Steady growth in funding, valuation, revenue
- Customers in pharma, energy, government, academia, internet businesses



Glaxo: Epigenomic AI model



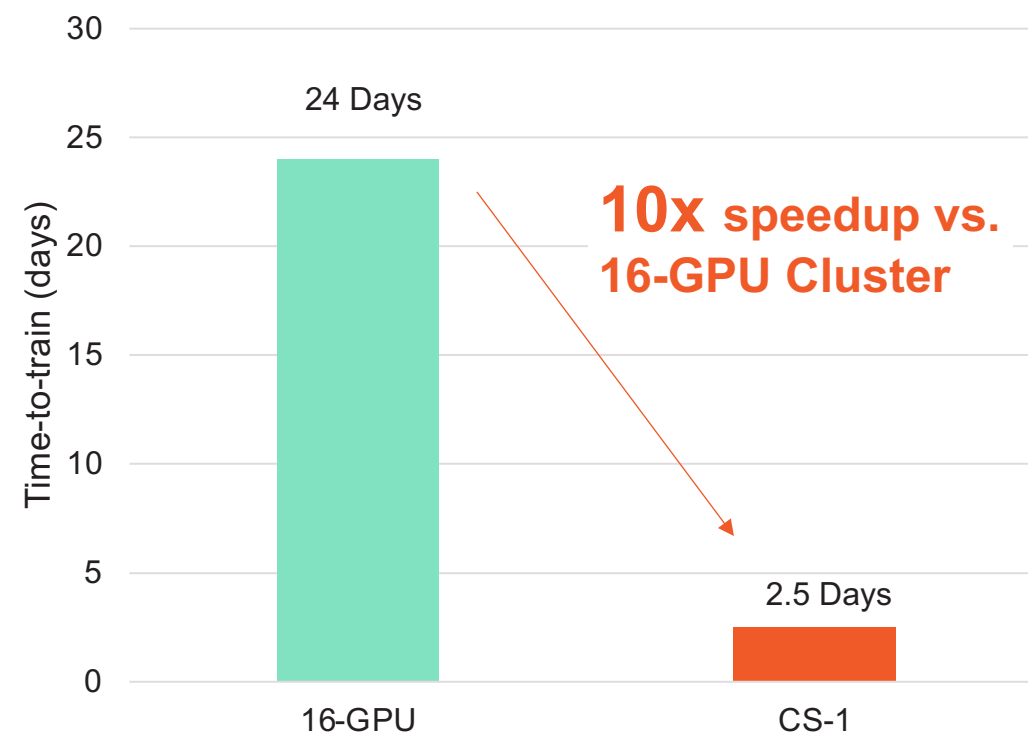
Objective: Genetic validation of drug targets with epigenome + genome in an NLP model



Challenge: Weeks to train on cluster



Outcome: ~10X speedup



“The training speedup afforded by the Cerebras system enabled us to explore architecture variations in a way that would have been prohibitively time and resource intensive on a typical GPU cluster”

“Epigenomic Language Models Powered by Cerebras”, Dec 2021. arxiv.org/abs/2112.07571



Toward real-time computational fluid dynamics: NETL

- CS-1 system solves sparse linear equations 200x faster than Joule 2.0 supercomputer¹
- Faster Spmv due to massive memory bandwidth
- Faster dot product due to low-latency, on-wafer interconnect

Fast Stencil-Code Computation on a Wafer-Scale Processor

Kamil Rocki*, Dirk Van Essendelft†, Ilya Sharapov*, Robert Schreiber*, Michael Morrison*, Vladimir Kibardin*, Andrey Portnoy*, Jean Francois Dietiker†‡, Madhava Syamlal† and Michael James*

* Cerebras Systems Inc., Los Altos, California, USA
Email: {kamil,michael}@cerebras.net

† National Energy Technology Laboratory, Morgantown, West Virginia, USA
Email: dirk.vanessendelft@netl.doe.gov

‡ Leidos Research Support Team, Pittsburgh, Pennsylvania, USA
Email: jean.dietiker@netl.doe.gov

Abstract—The performance of CPU-based and GPU-based systems is often low for PDE codes, where large, sparse, and often structured systems of linear equations must be solved. Iterative solvers are limited by data movement, both between caches and memory and between nodes. Here we describe the solution of such systems of equations on the Cerebras Systems CS-1, a wafer-scale processor that has the memory bandwidth and communication latency to perform well. We achieve 0.86 PFLOPS on a single wafer-scale system for the solution by BiCGStab of a linear system arising from a 7-point finite difference stencil

limited memory bandwidth and high communication latency are primary performance limiters.

HPC memory and communication systems struggle to keep up with processing performance. In 2016 the flops to words ratios for both memory and interconnect bandwidth were in the hundreds, and the flops needed to cover the memory or network latencies were in the 10,000 to 100,000 range, with the trend going higher; see Figure 1.



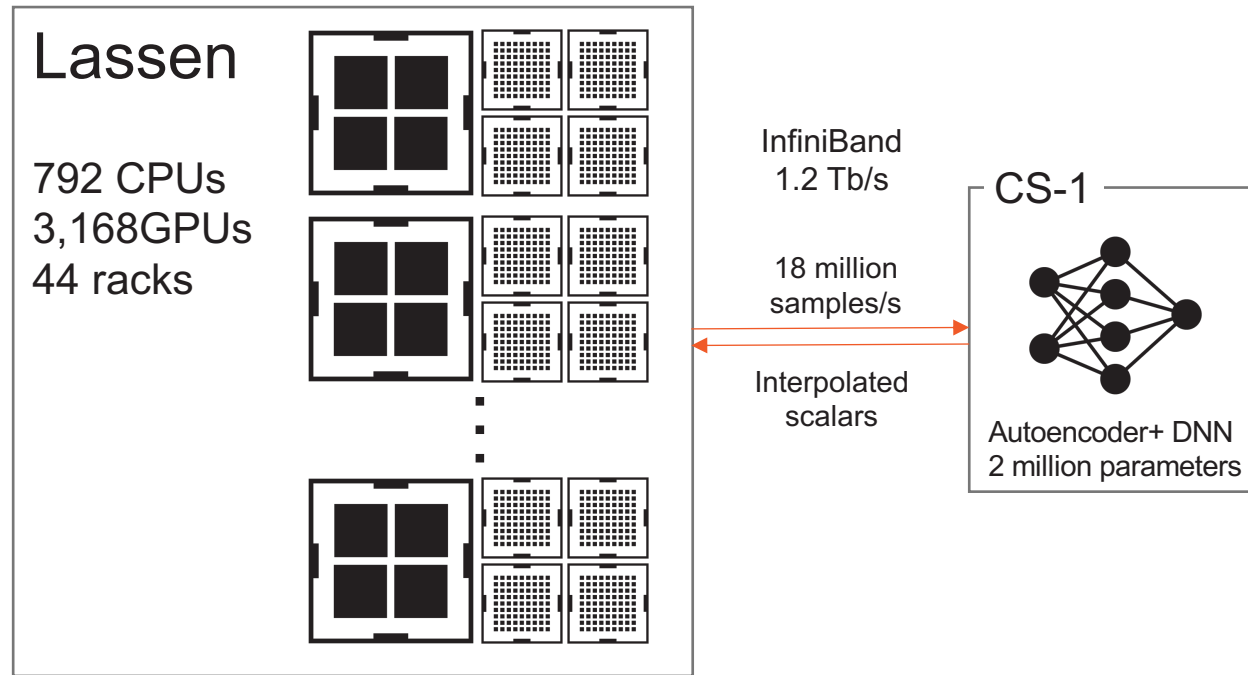
60v1 [cs.DC] 7 Oct 2020

1. Rocki et al., “Fast Stencil-Code Computation on a Wafer-Scale Processor” SC20. arxiv.org/abs/2010.03660



Photo: Christian Kühn, CC BY 3.0

AI surrogate models accelerating HPC at LLNL



Ready to run in under 20 hours
CS-1 64x performance of Lassen GPU



You had questions, I may have answers

Moore's Law and the Industry

We will be fine

Programming models

Productivity and high level abstractions vs Performance

Silicon photonics

On the chip

On the wafer

In networks

Integration technologies

If one wafer is good...

Engagement

- Our product, marketing, sales team are eager to engage and hear about new opportunities
- We are hiring. Many on our team have physics, math, computational science backgrounds
- We have lots of internship positions and very strong intern mentors

Thank you

Questions, Answers: Find us online at www.cerebras.net

