HPC on a Wafer

Matthias Fouquet-Lapar
Cerebras Systems, Inc.

ICES Foundation Biennial Workshop VI
Geneva , 30 September 2022

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

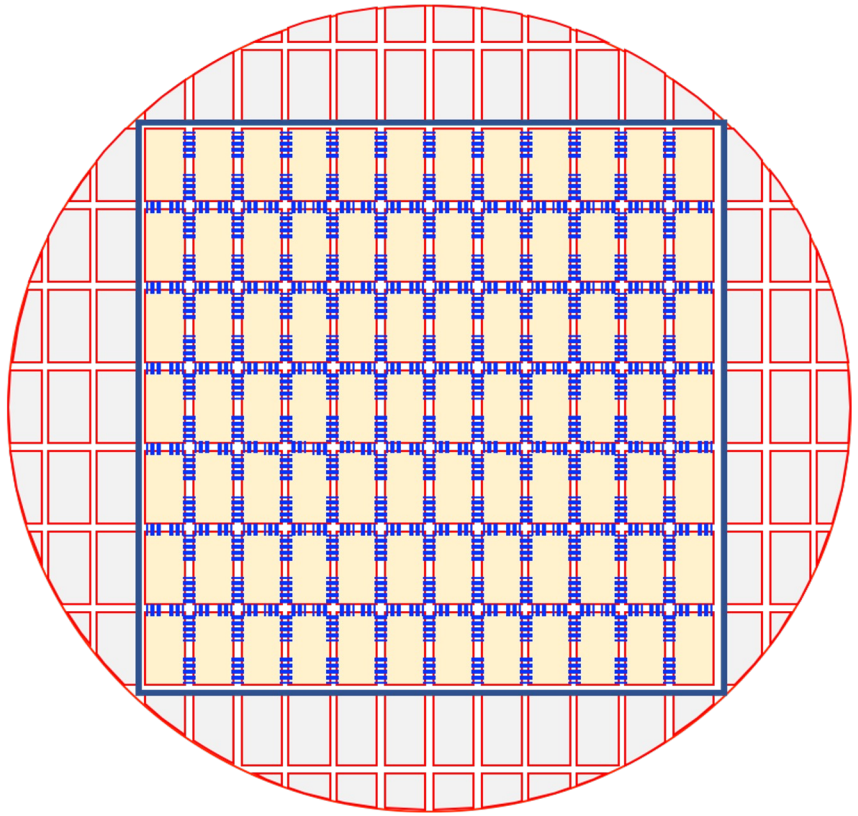
Cluster-scale acceleration on a single chip

cerebras

What's Inside?

- World's first successful wafer-scale chip
- Full 8-inch square in 12-inch wafer
- TSMC 7 nm
- Extra cores and cxns, logical mesh avoiding defects
- 850k cores
- fast floating point, local memory, router
- message triggered computing

- 40 GB fast SRAM
- Only possible on a wafer:
  - Memory bandwidth: vector ops (a x + y) at full speed
  - 1 clock memory latency
  - 1 clock message latency



Cerebras CS-2: Cluster-scale Performance in a Single System

15 RU standard rack-compliant server

1.2 Tbps I/O via 12x100GbE

23 kW power

Customer Adoption Across Verticals

Pharma

Drug discovery acceleration using techniques such as virtual drug screening and new kinds of neural network models

AstraZeneca

GSK increased model complexity while **decreasing training time by 10X** vs. 16-GPU system

gsk

Financial Services

Train Natural Language Processing (NLP) models on financial services specific datasets to increase predictive power by ~2.6x

$BERT_{LARGE}$ **trained 15X faster vs. 8-GPU system** consuming **45% less energy**

Web

Q&A system: retrieve relevant answer to a posted question

Training: ~10X faster time to accuracy vs. 8-GPU system to target accuracy and **Inference: No tradeoffs** in latency vs accuracy – full $BERT_{BASE}$ at low latency target

Energy

Accelerate multi-energy research into batteries, biofuels, wind, gas and CO_2

Model seismic wave propagation with stencil-based finite difference algorithm in seconds, **>100 faster than GPU**

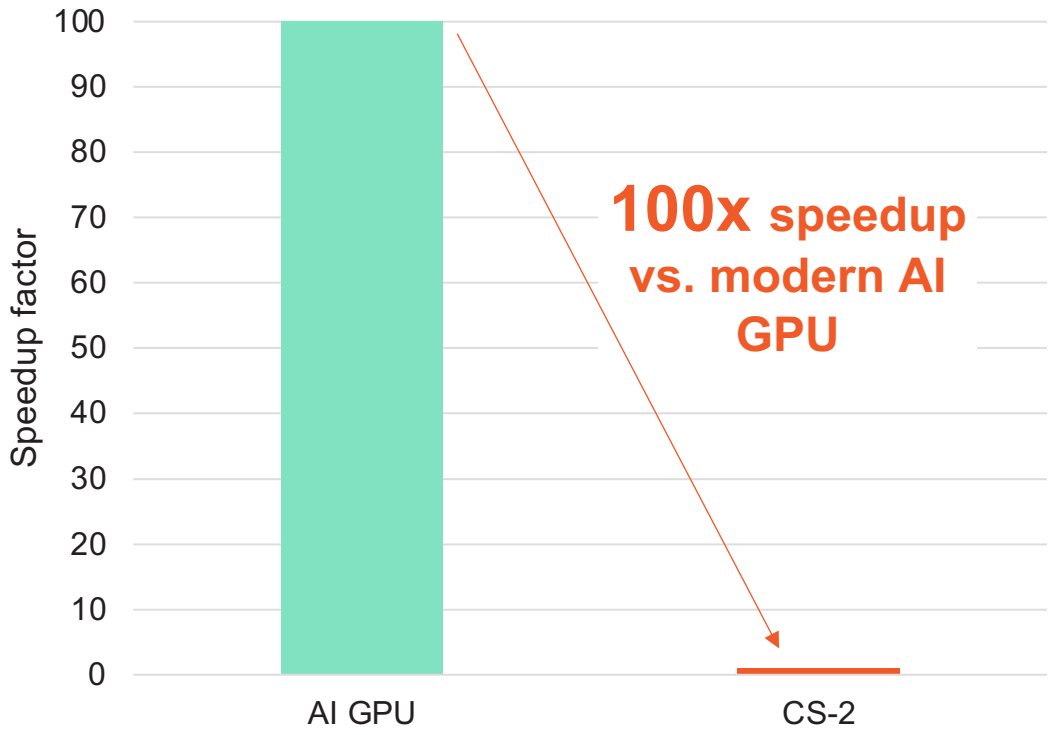
TotalEnergies

TotalEnergies: Accelerate multi-energy research

Objective: enable order-of-magnitude speedups on a wide range of simulations: from batteries to biofuels, to wind flows, drillings, and CO2 storage

Challenge: Participate in their study to evaluate alternative hardware architectures, using finite difference seismic modelling code as a benchmark

Outcome: Cerebras CS-2 system outperformed a modern AI GPU by >100X using code written in the Cerebras Software Language (CSL). System now installed and running at customer research facility in Houston, Texas



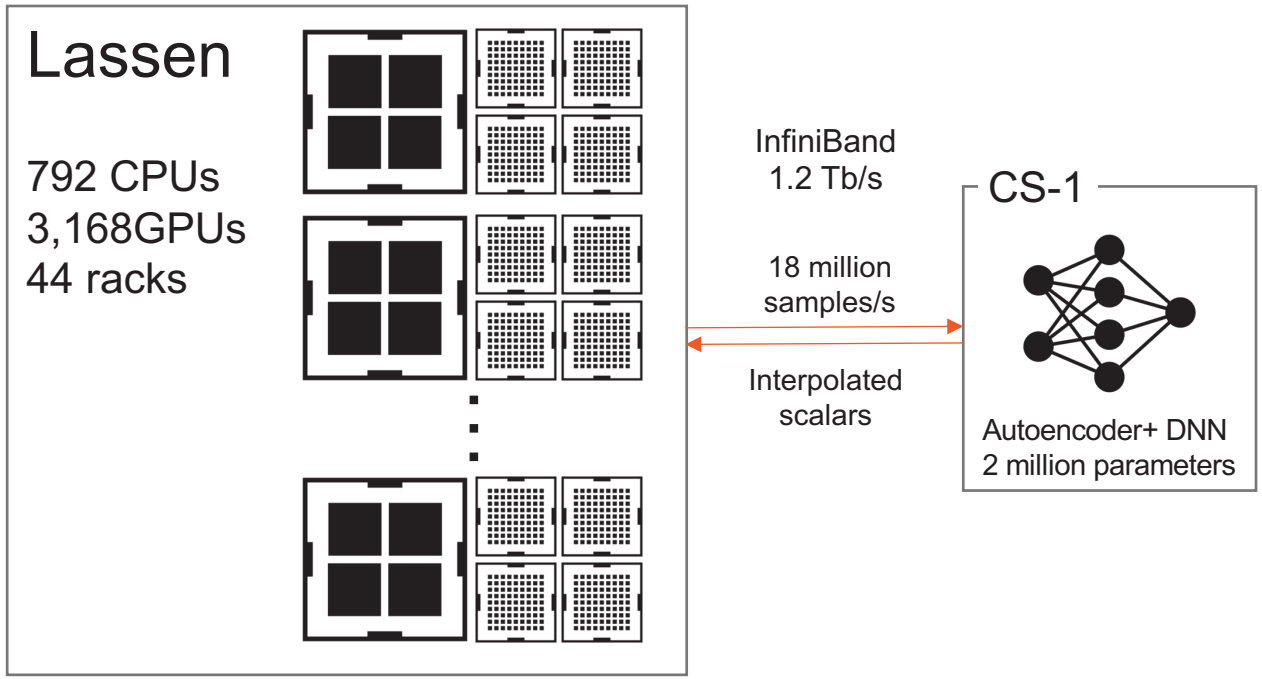
"We count on the CS-2 system to boost our multi-energy research and give our research 'athletes' that extra competitive advantage. "

Dr. Vincent Saubestre, CEO and President, TotalEnergies Research & Technology USA

TotalEnergies

Lawrence Livermore National Laboratory

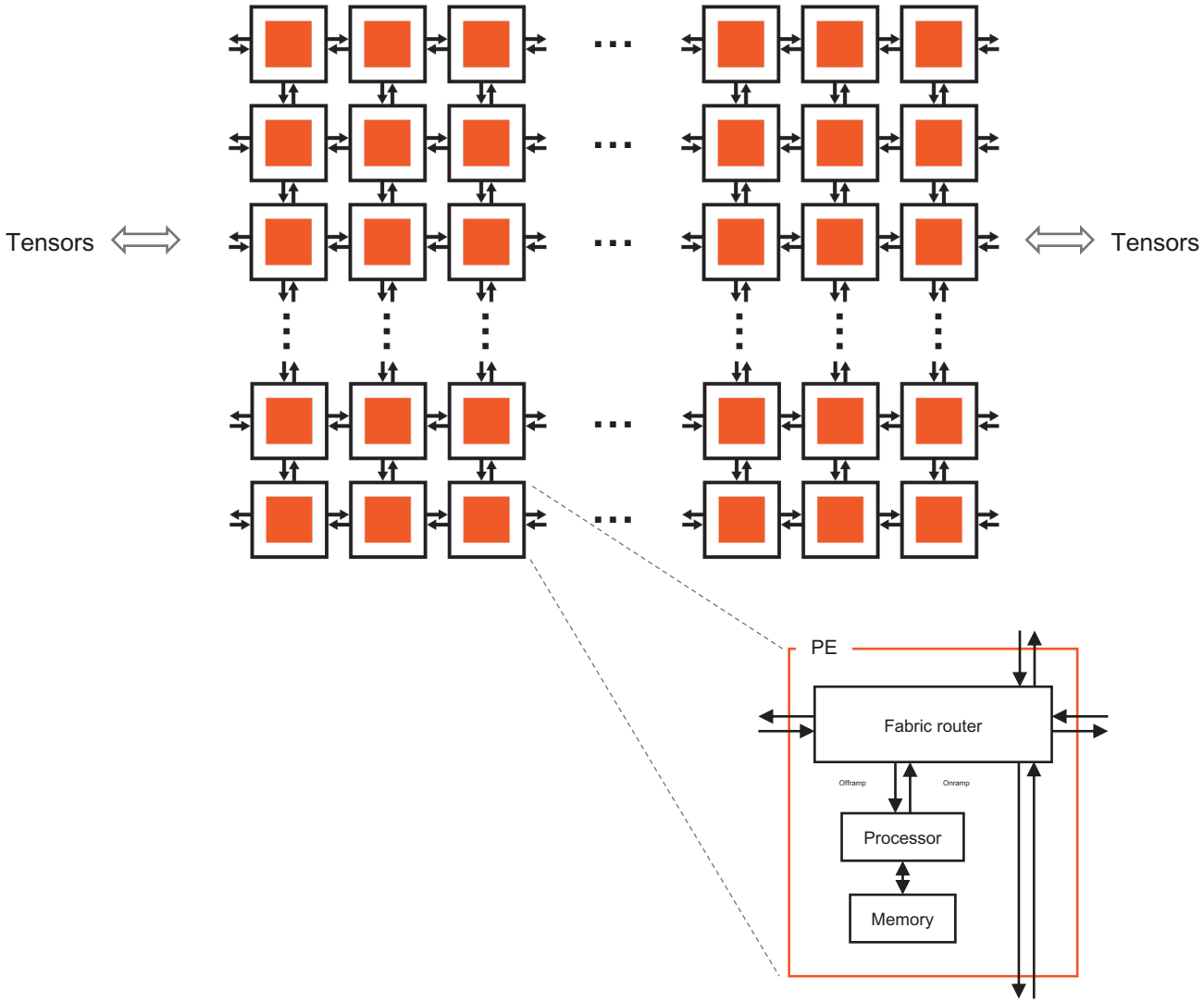
System-Level Heterogeneity



18 million DNN inferences per second
37x performance of Lassen GPU
20 hours from crate to “hello world!”

cerebras

CS-2 Dataflow Programming



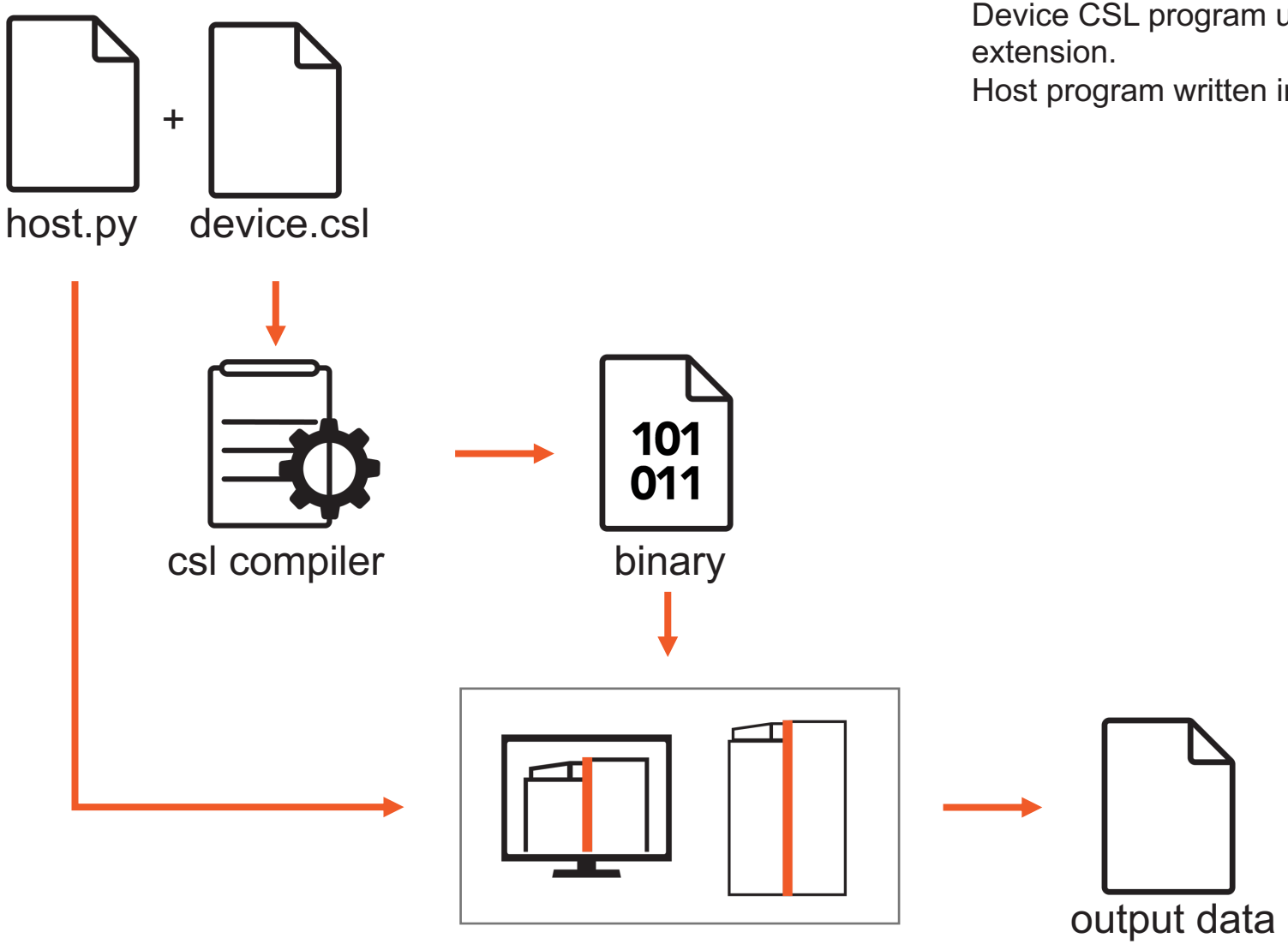
- The CS-2 is a 2D array of ~800K Processing Elements (PEs) offering:
- Flexible Compute:
  - General purpose CPU
  - Types: FP32, FP16, INT16
  - Direct CPU-Network cxn.
  - Data packets trigger compute
- Flexible Communication:
  - Programmable, static or dynamic routes (virtual channels)
  - 32-bit + 6 packets
  - 1 cycle single hop latency
- Fast Memory:
  - 40GB on-chip distributed SRAM
  - 1 cycle read/write

Writing a Host and Device Program

Step 1:
Write host and device code

Step 2:
Compile device code

Step 3:
Run on simulator or CS-2



Device CSL program uses .csl file extension.
Host program written in Python