Wafer-Scale Hardware for ML and Beyond

Rob Schreiber, Cerebras Systems, Inc.

2nd International Workshop On Machine Learning Hardware (IWMLH), July 2021

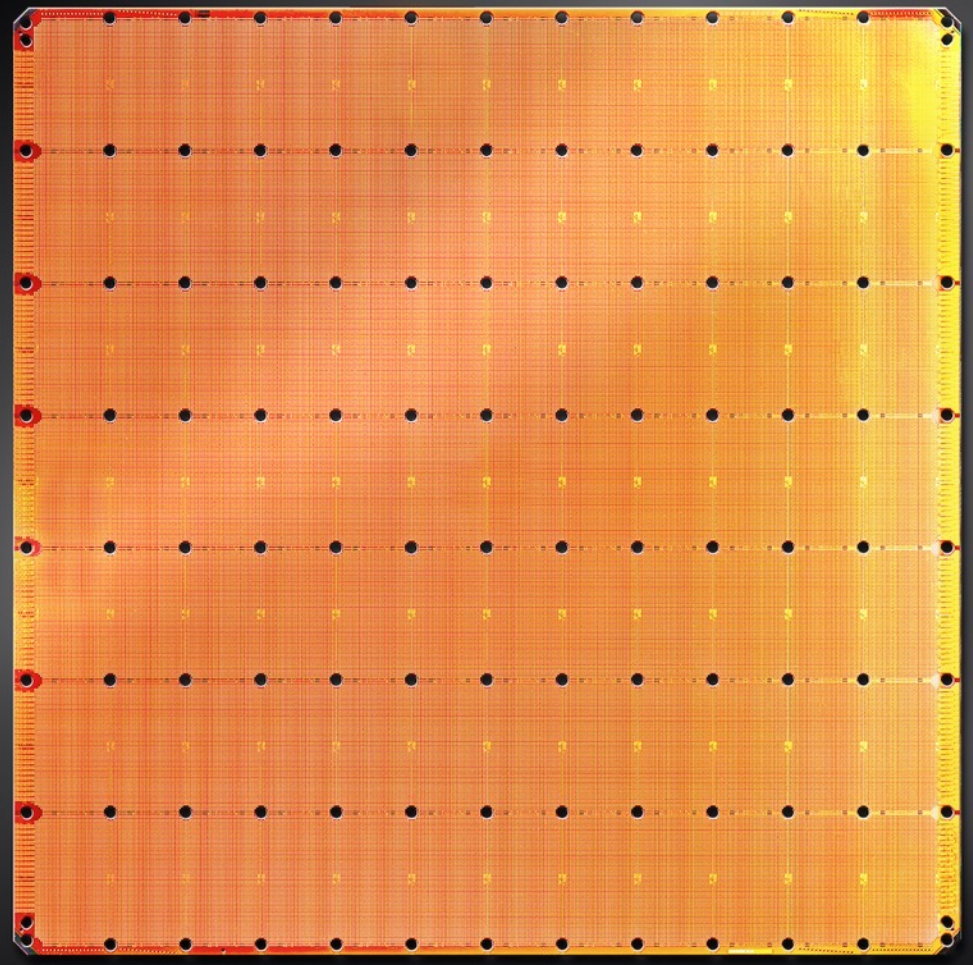
cerebras

Moore's Law

- transistors per mm^2

- mm^2 per chip

Cerebras Wafer Scale Engine

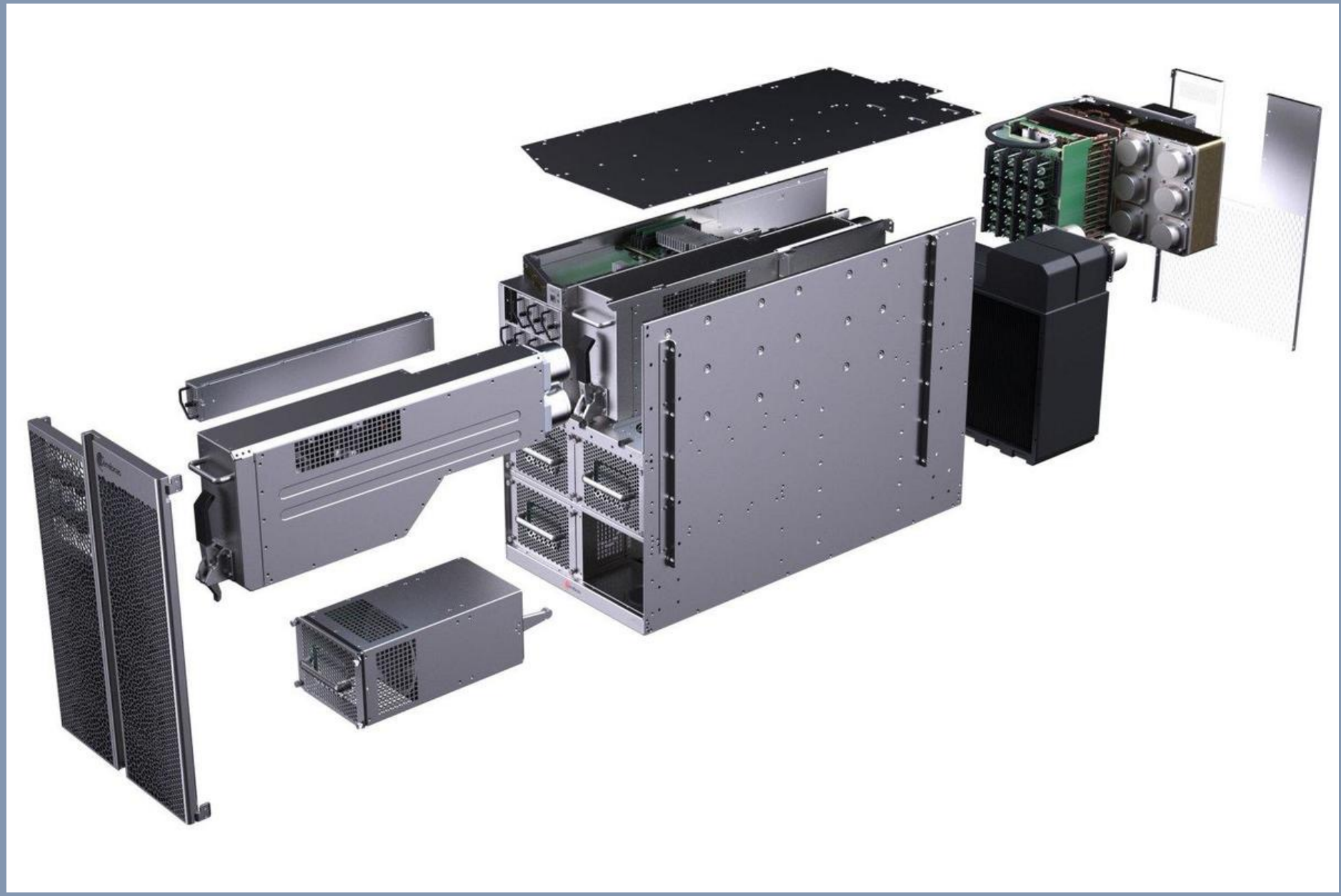


Cerebras WSE-2
2.6 Trillion Transistors
46,225 mm² Silicon

Largest GPU
54.2 Billion Transistors
826 mm² Silicon

	Cerebras WSE	A100	Cerebras Advantage
Chip size	46,225 mm^2	826 mm^2	56 X
Cores	850,000	6912 + 432	123X
On-chip memory	40 Gigabytes	40 Megabytes	1,000 X
Memory bandwidth	20 Petabytes/sec	1555 Gigabytes/sec	12,733 X
Fabric bandwidth	220 Petabits/sec	600 Gigabytes/sec	45,833 X

CS-1 Chassis



How did we do it?

Cross-wafer connectivity
Yield

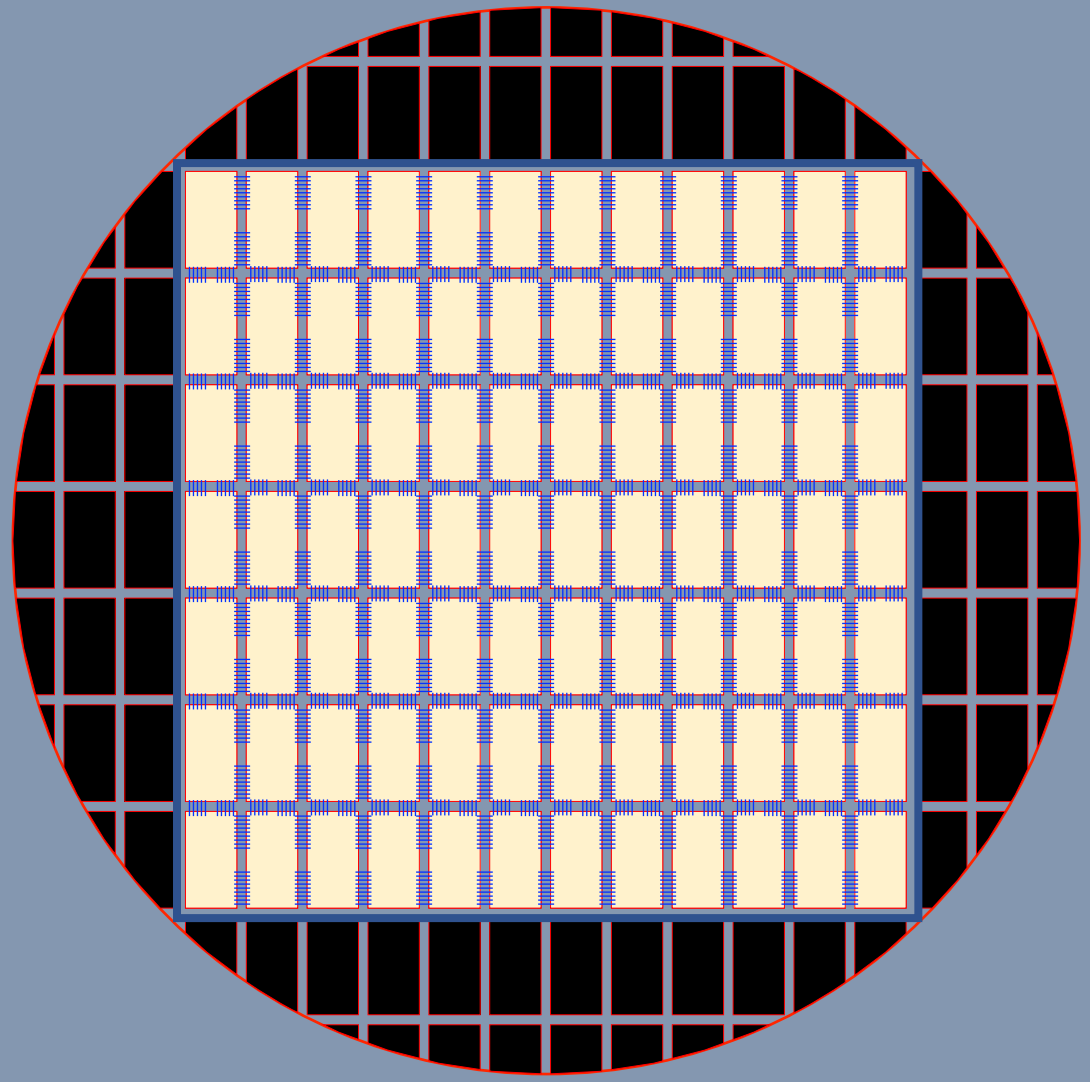
Power and cooling

Thermal expansion

Cross-Die Wires

Developed with TSMC

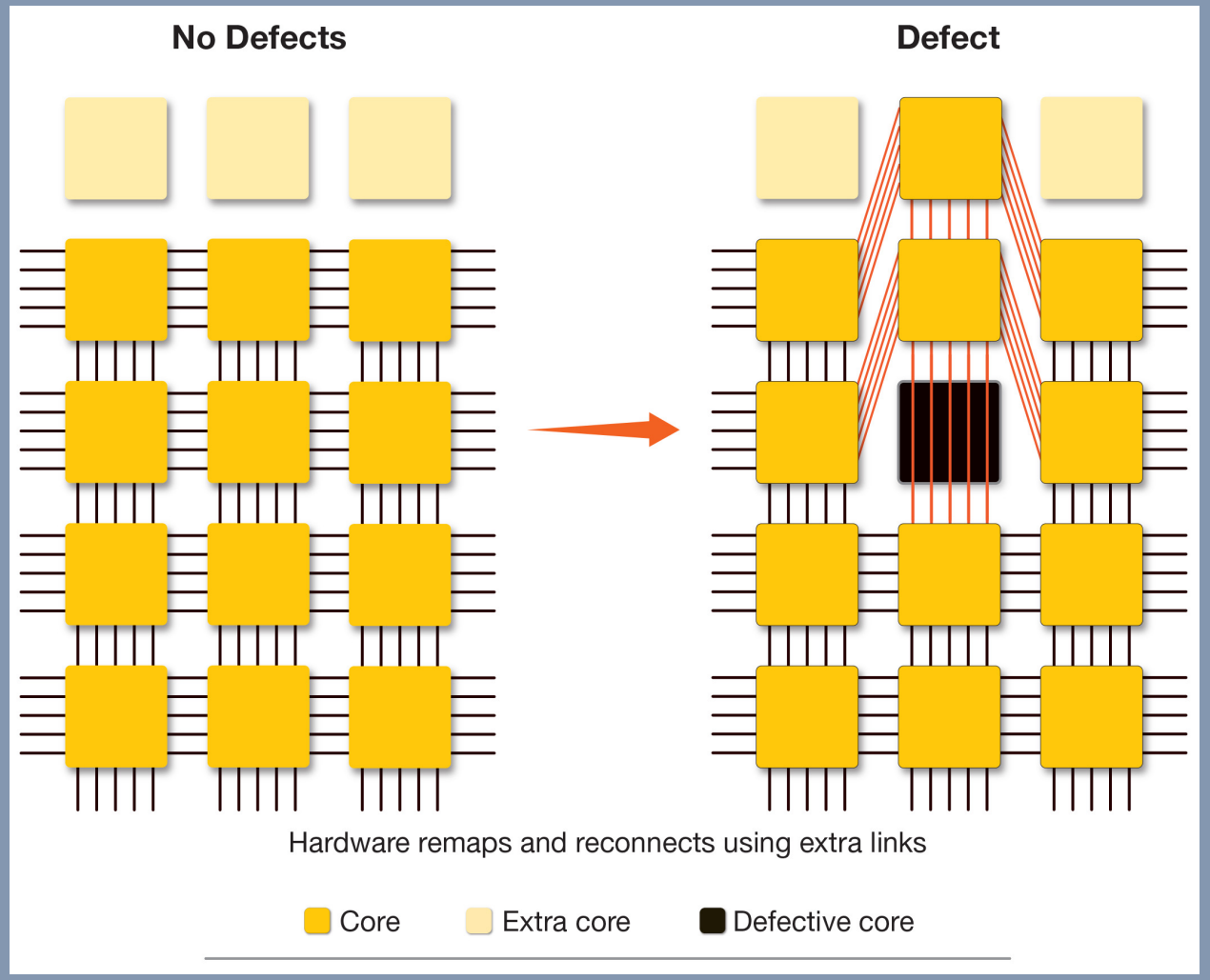
Uniform bandwidth across wafer



Redundancy

Extra rows

Logical 2D mesh



Yes, we can build wafer-scale systems

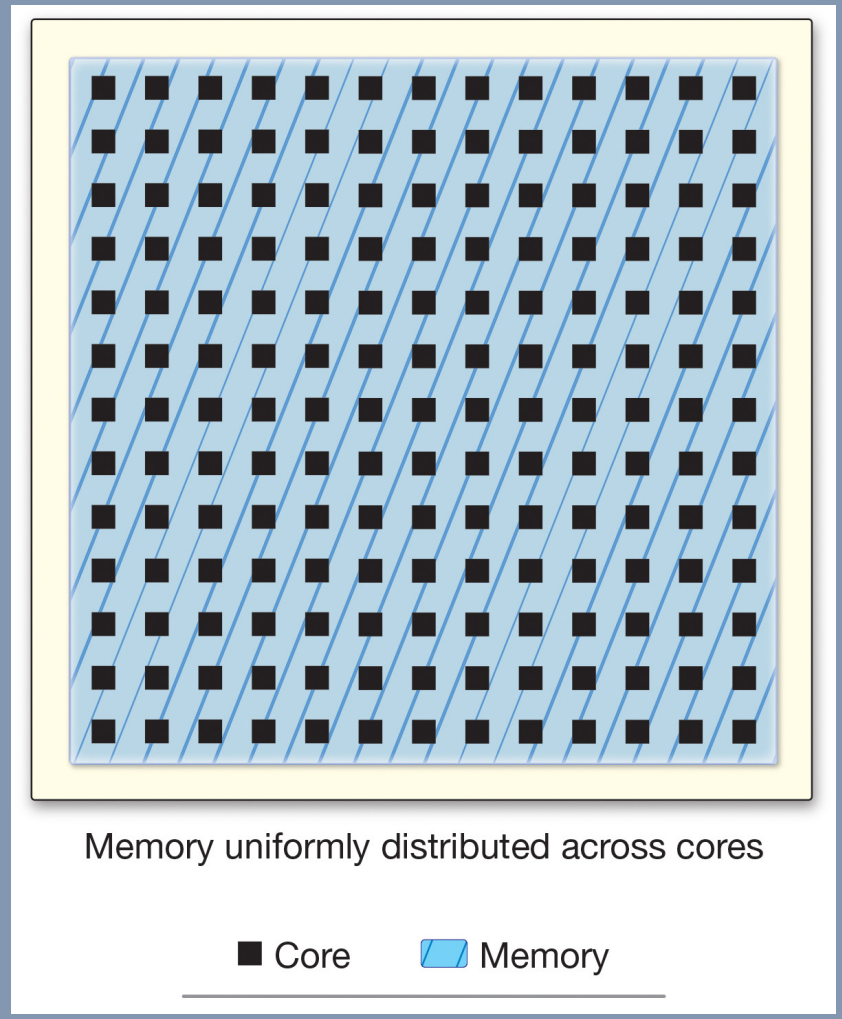
What did we put on the wafer?

All the memory

Fine grained parallelism

Shared nothing

Power-efficient, general purpose core



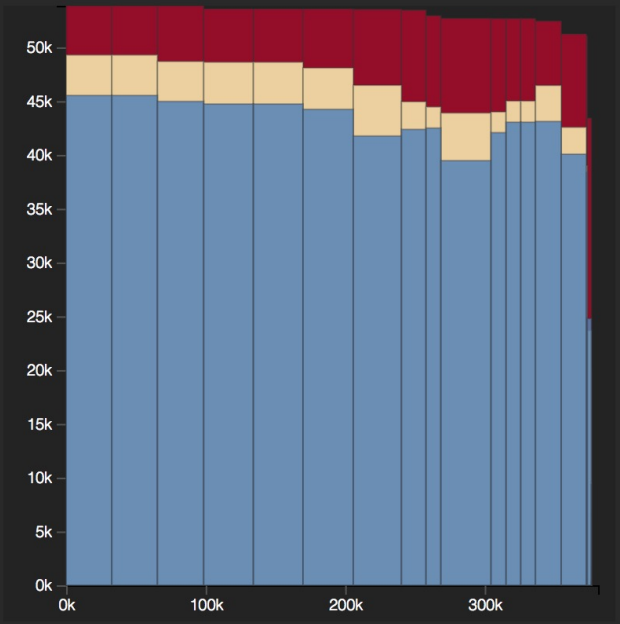
On the Wafer:

- **Huge compute**
- **Huge memory + comm bandwidth**
- **Great flops/watt**
- **40 GB of SRAM memory**

Cerebras Systems: Placement Visualizer

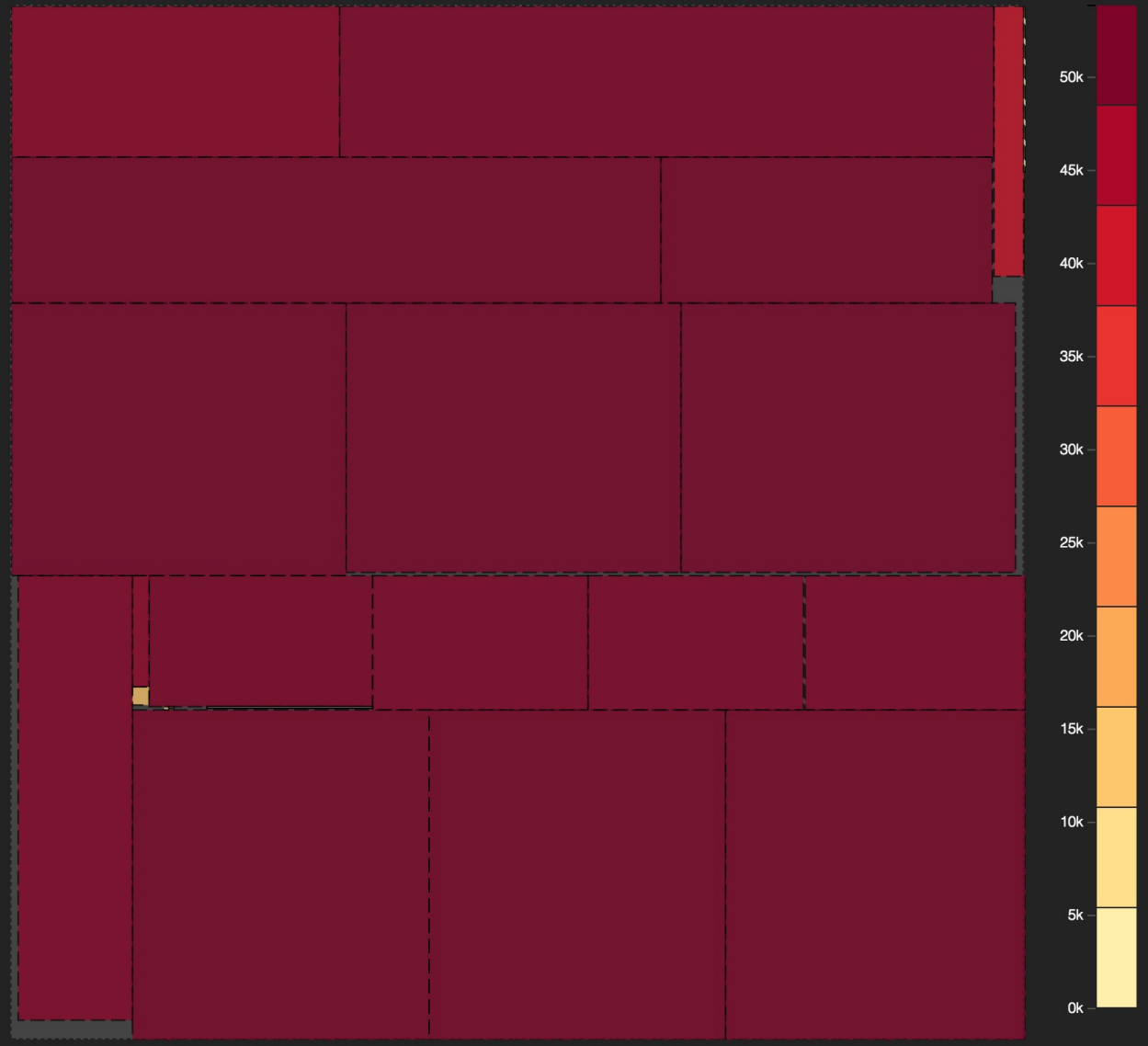
Generated: Sun Oct 15, 2017 22:14 PM [vijay@server1]

vgg_final



612x623	Size
53.9k	Placement Δt
49.9k	Theoretical Best Δt
7.943	Optimality
376,809	Used Area
381,276	Fabric Area
98.8%	Pe Utilization
2.1e+10	Fabric Ops
1.8e+10	Layer Ops
85.8%	Layer Utilization

RECENT NEW SAVE delta t



SHOW GRID SHOW NAMES SHOW INPUTS SHOW PORTS

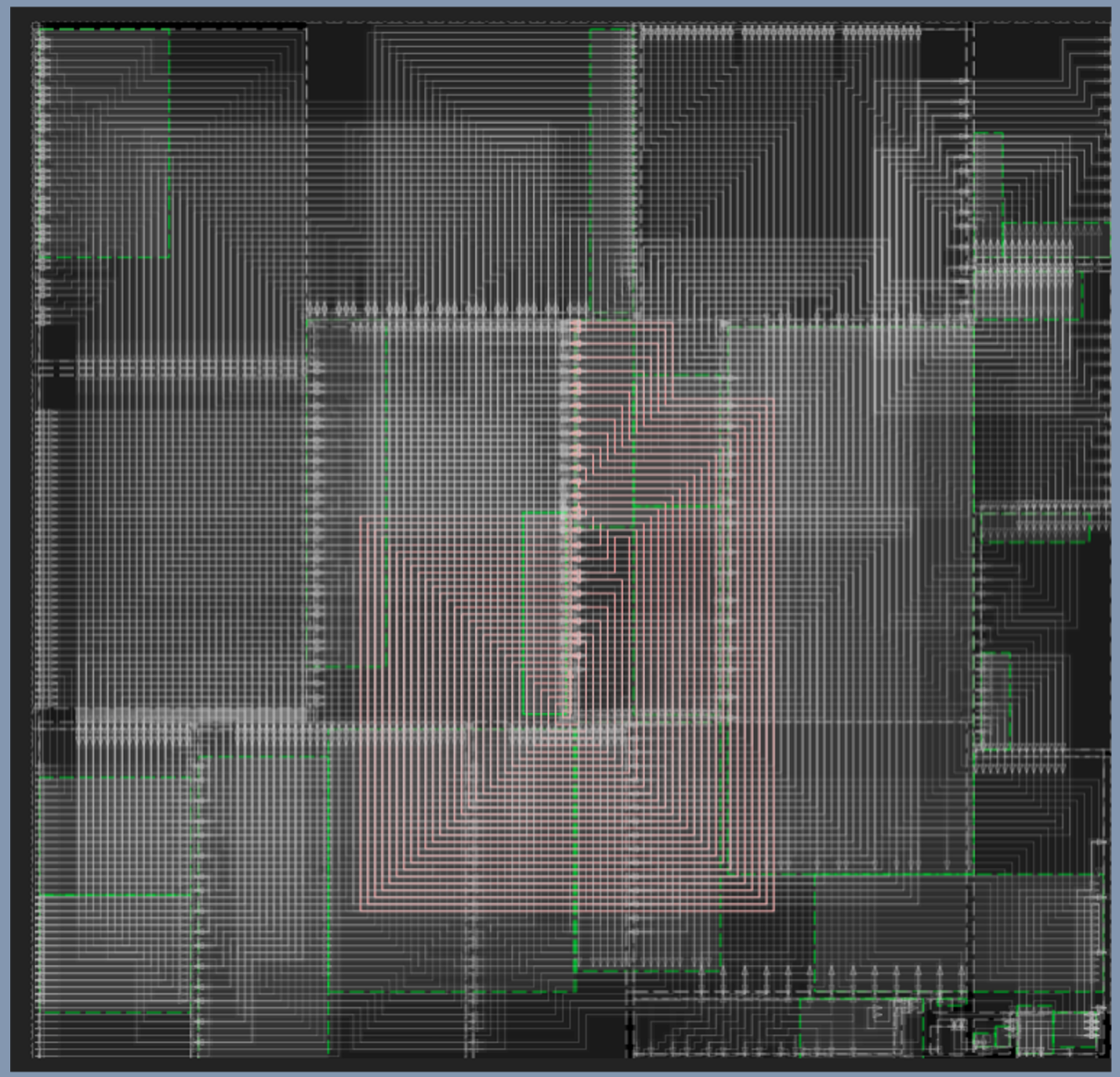
Node Data

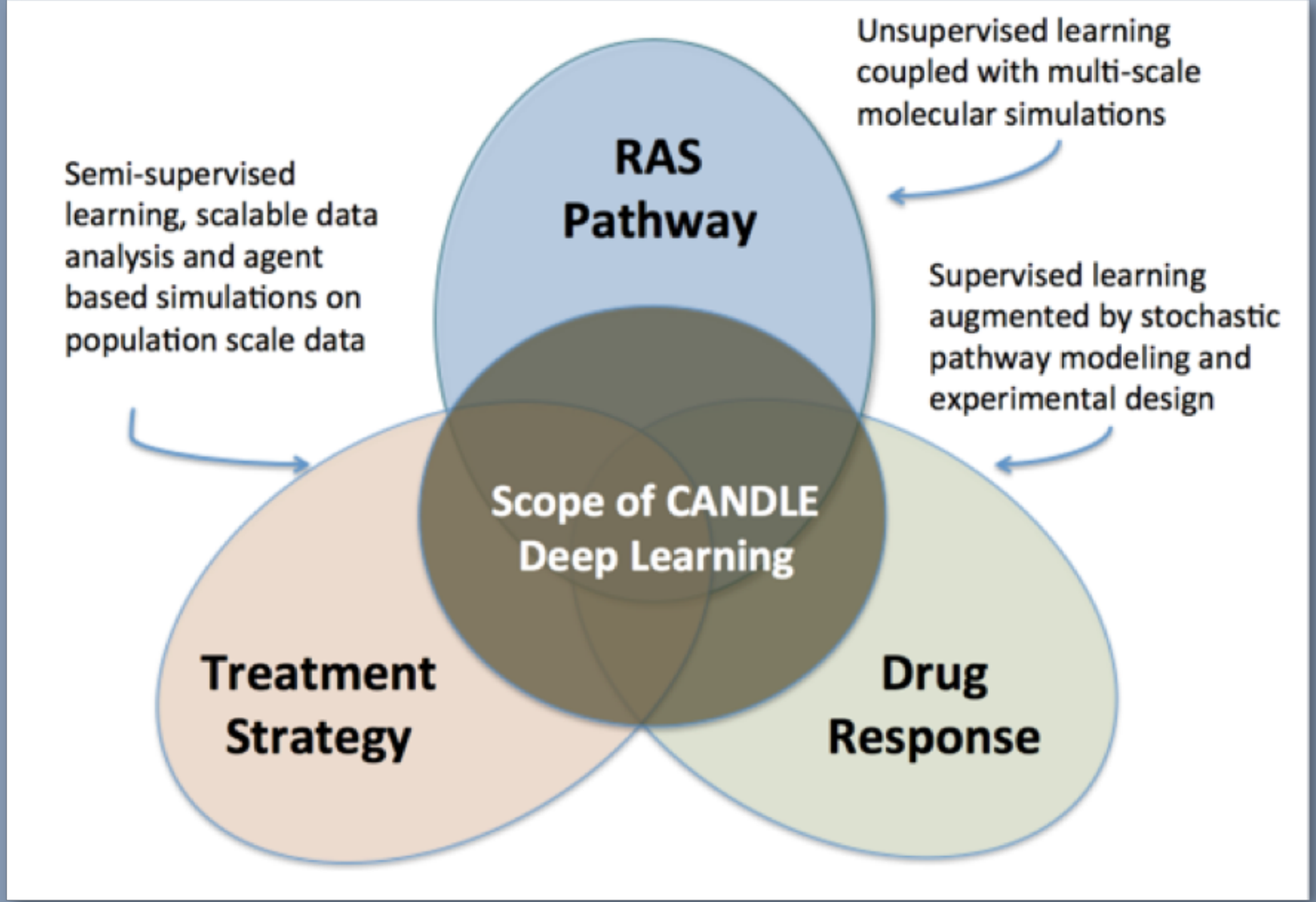

```
{
  "delta_t": 990,
  "depth": 0,
  "in_bw": 1,
  "layer": "GenConvForward",
  "name": "n1.GenForward",
  "ops": 990,
  "out_bw": 1,
  "outputs": {
    "txact": {
      "gate": "out1",
      "index": 0,
      "prop": "fwd"
    }
  },
  "params": {
    "file_name": "input",
    "out1_fan_sz": 100,
    "splits_c": 1,
    "splits_h": 1,
    "type": "gen",
    "wf_sz": 0
  },
  "placement": {
    "loc": {
      "ht": 1,
```


Graph Data

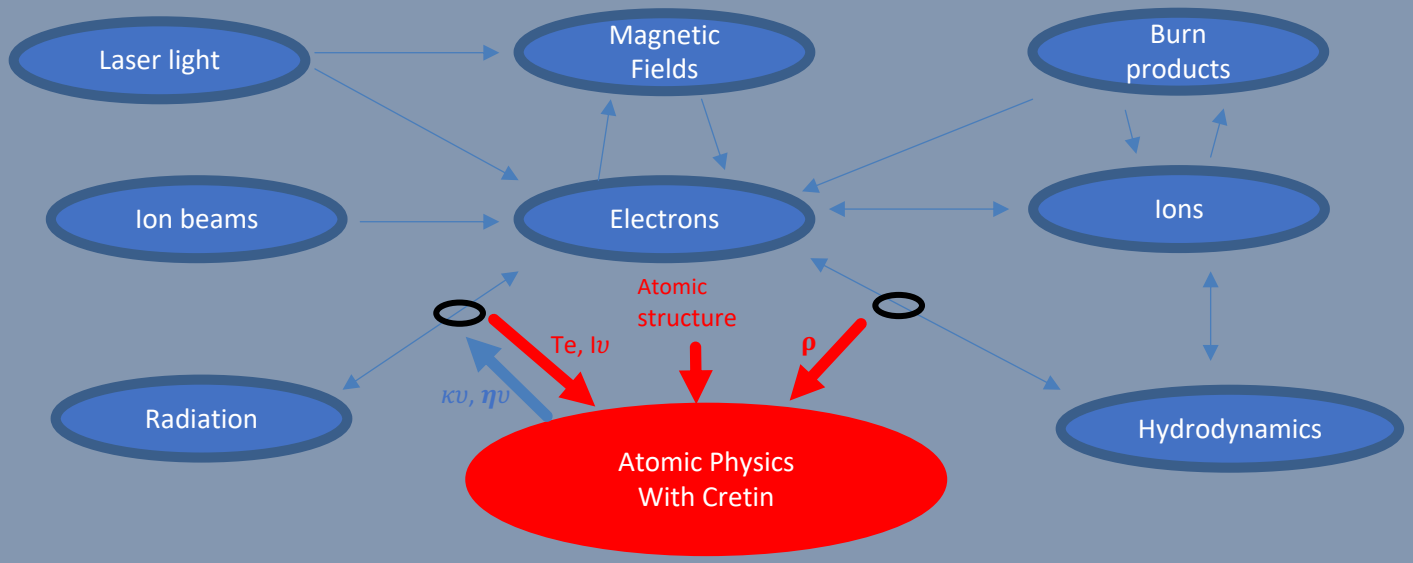

```
{
  "autoroute": 1,
  "connections": 0,
  "directed": 0,
  "fabric_height": 623,
  "fabric_utilization": 0.8,
  "fabric_width": 612,
  "forward_depth": 20,
  "log_level": "normal",
  "max_tilings": 1,
  "meta": {
    "date_created": "2017-10-15T22:14:44".
```

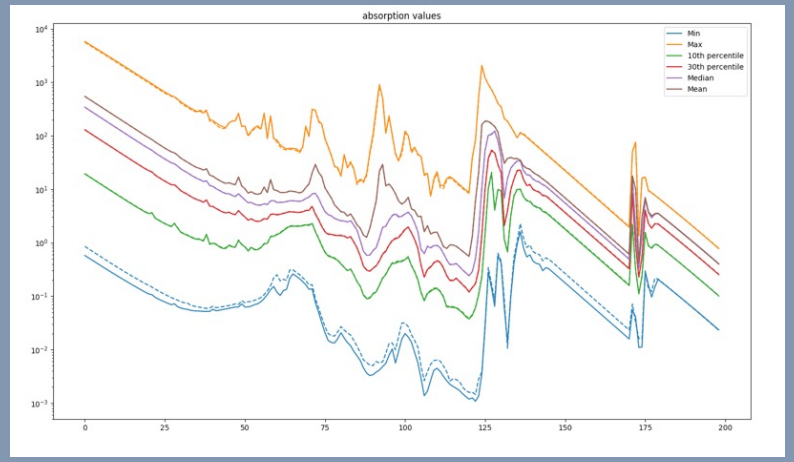
Detailed Routing




Unsupervised learning coupled with multi-scale molecular simulations
RAS Pathway
Semi-supervised learning, scalable data analysis and agent based simulations on population scale data
Supervised learning augmented by stochastic pathway modeling and experimental design
Scope of CANDLE Deep Learning
Treatment Strategy
Drug Response

@LLNL: Inertial confinement fusion model





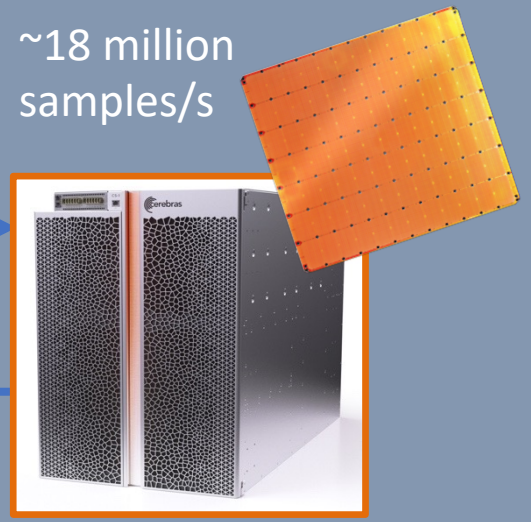
DNN output matches Cretin

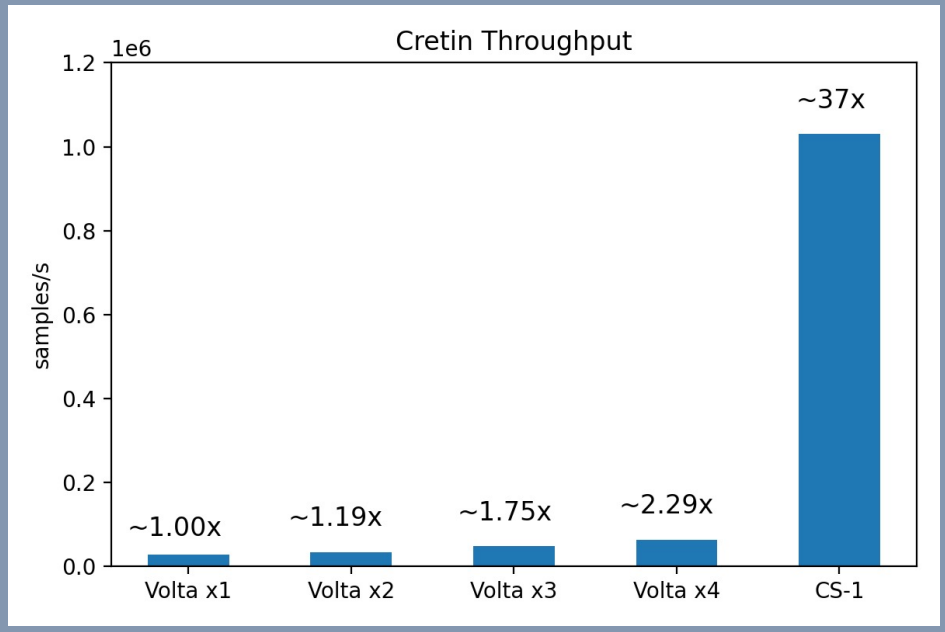
Lassen

42 singles/zone
Requires 3.2 GB/s

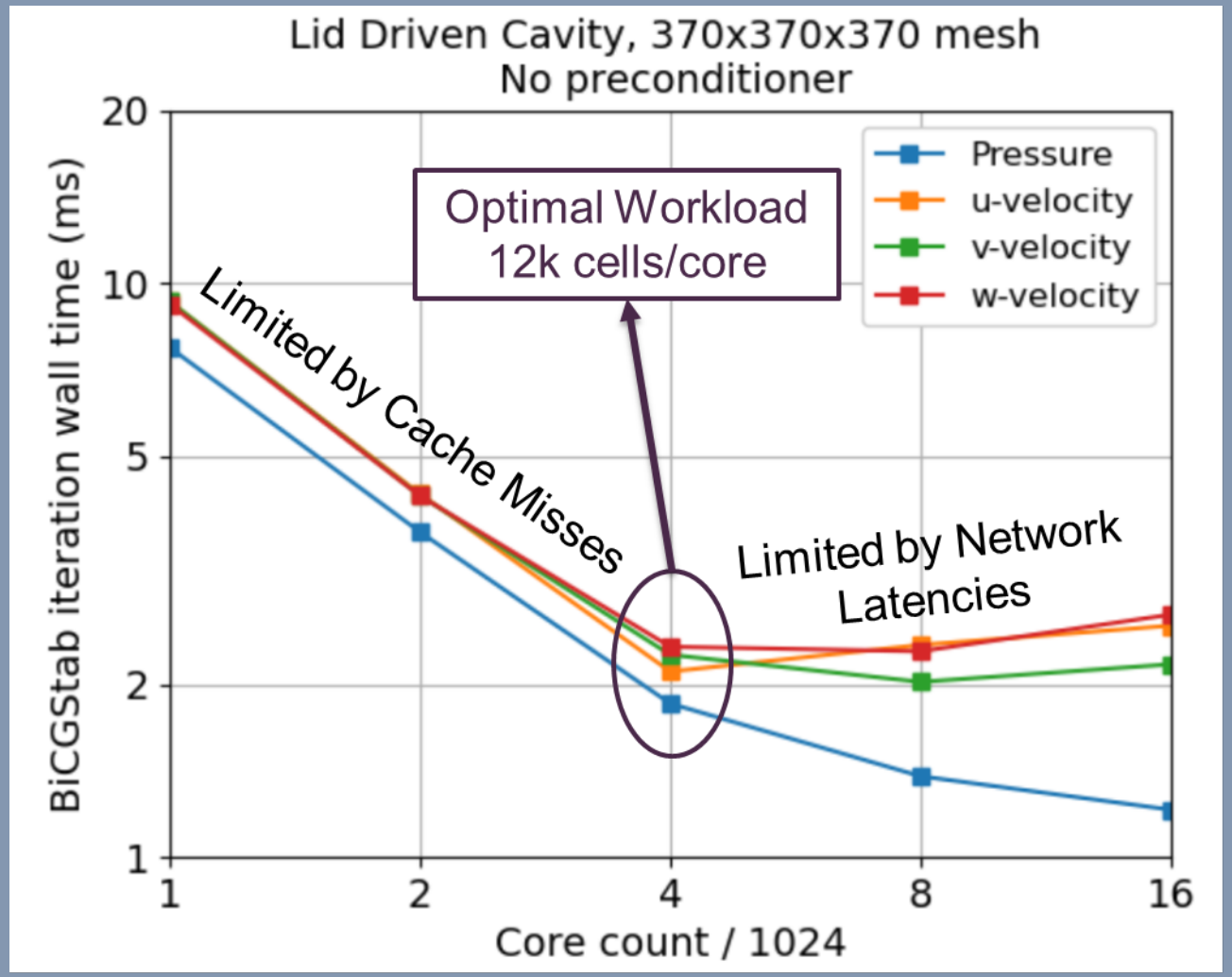
3 interpolated scalars/zone

~18 million samples/s





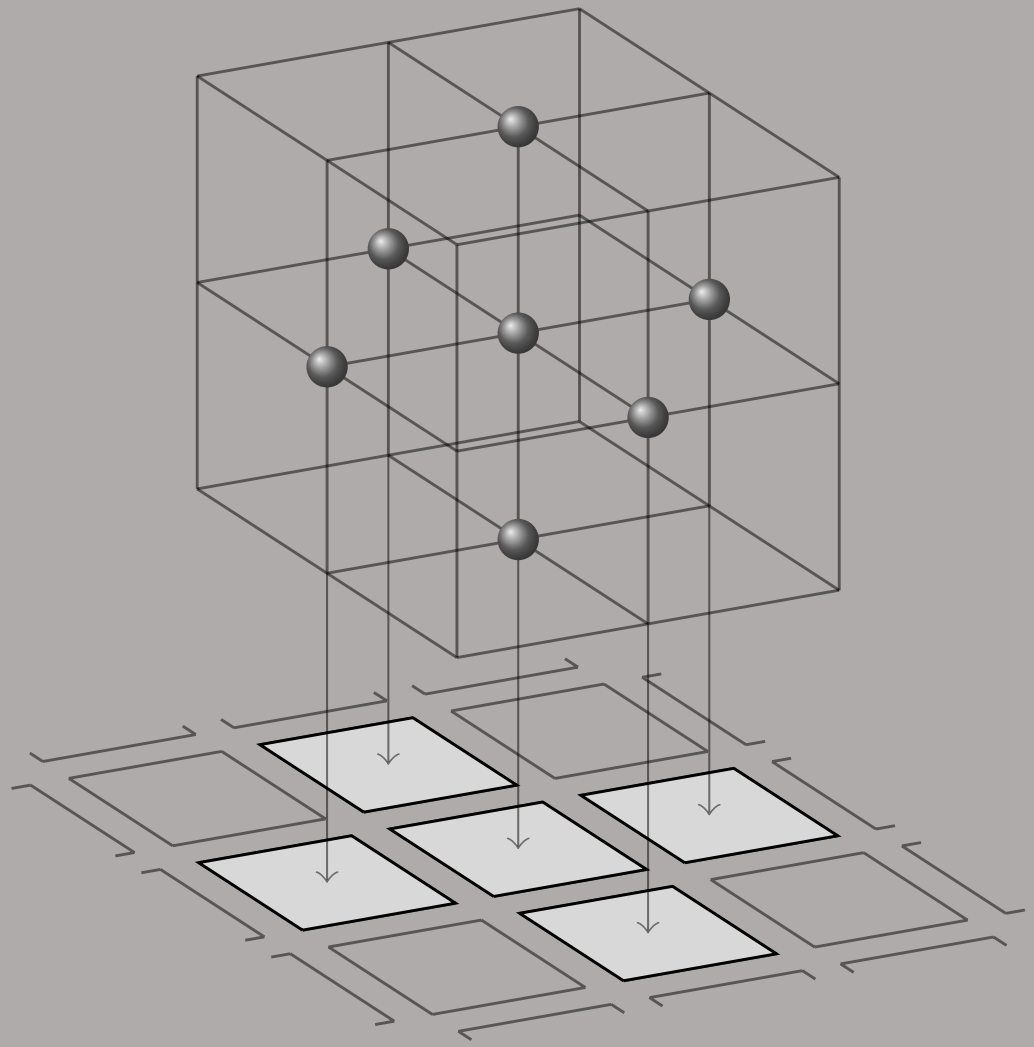
CPU and GPU performance

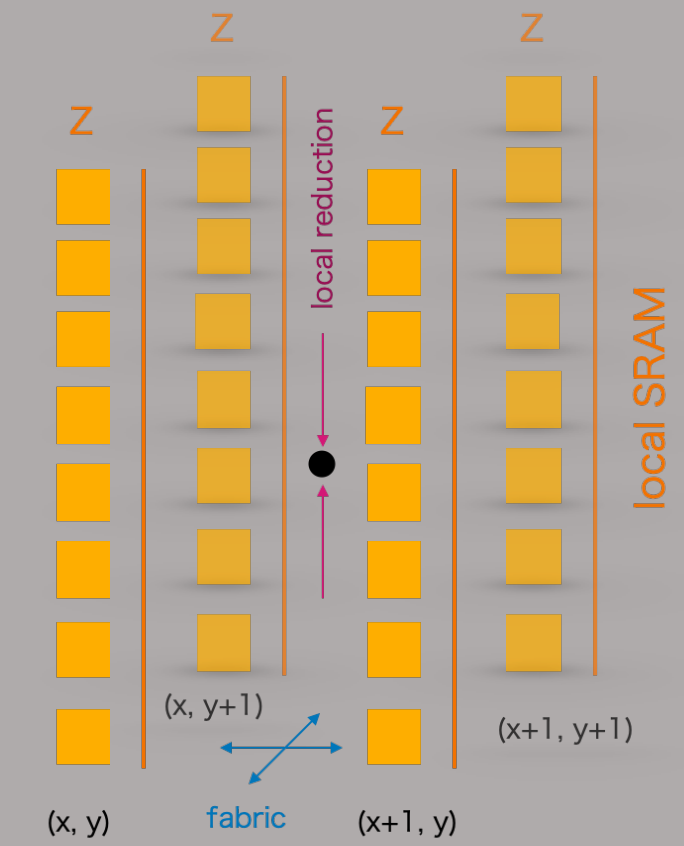


JOULE

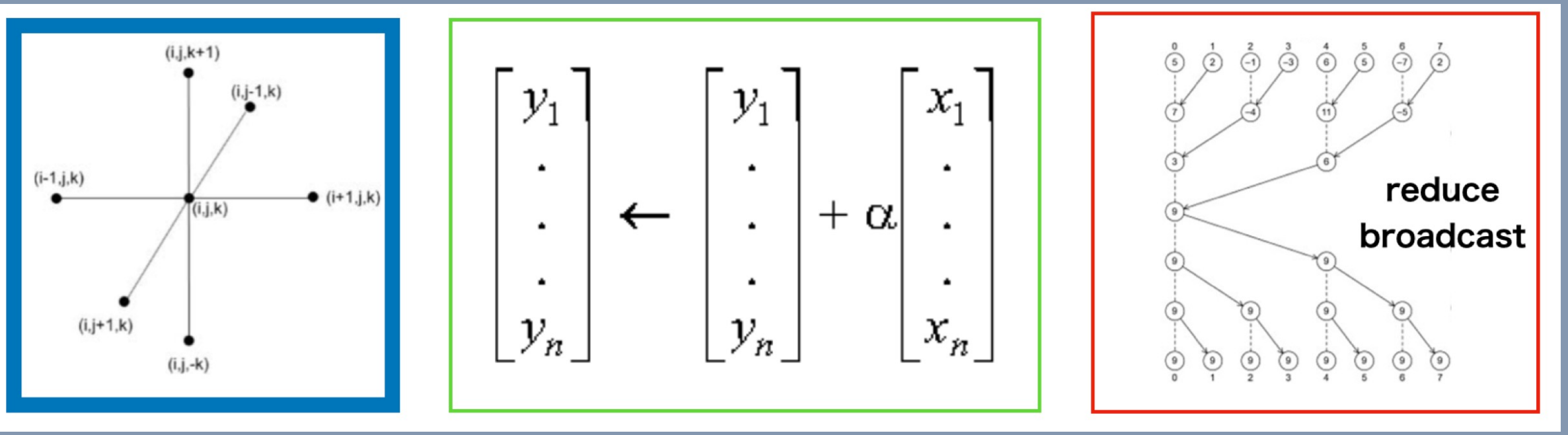
NETL SUPERCOMPUTER

On the CS-1: 3D mesh --> 2D machine





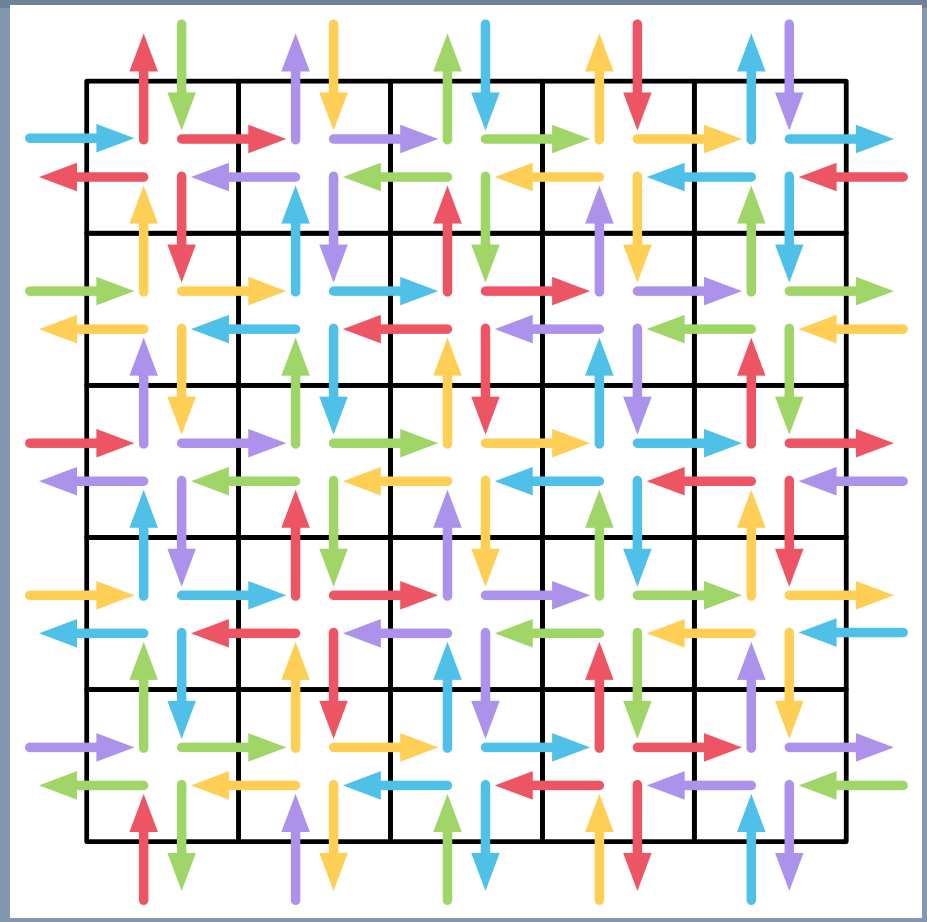
BiCGStab: Building Blocks



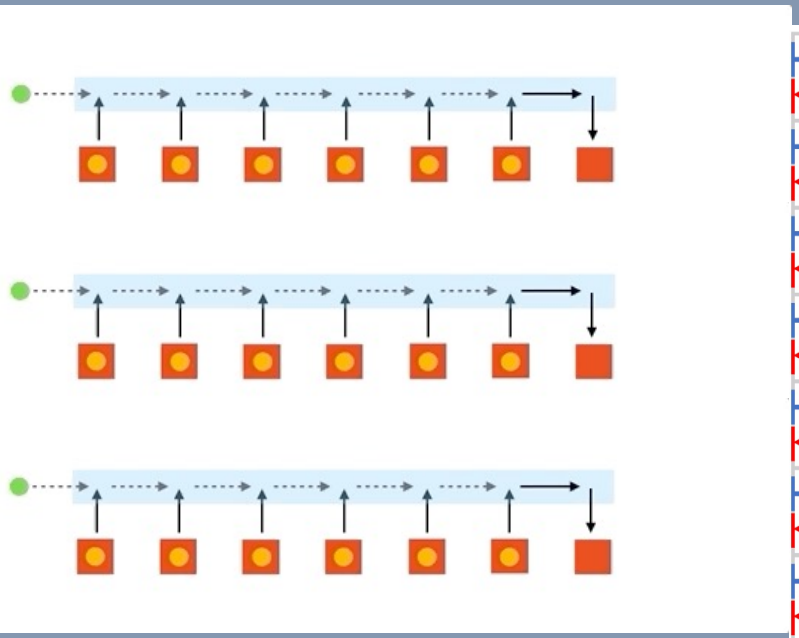
Interprocessor Communication

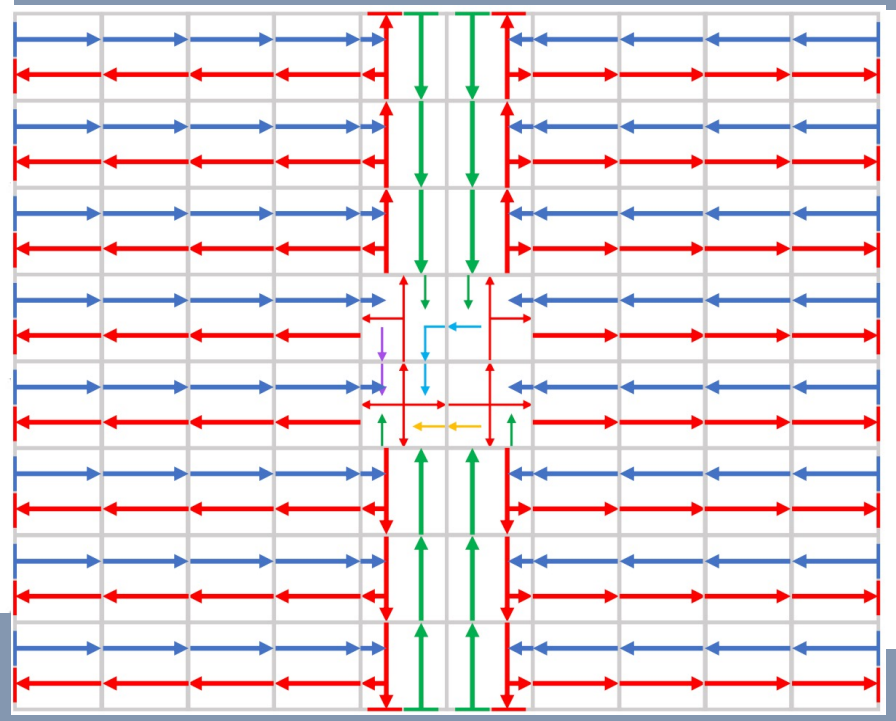
- The wafer is a dataflow computer:

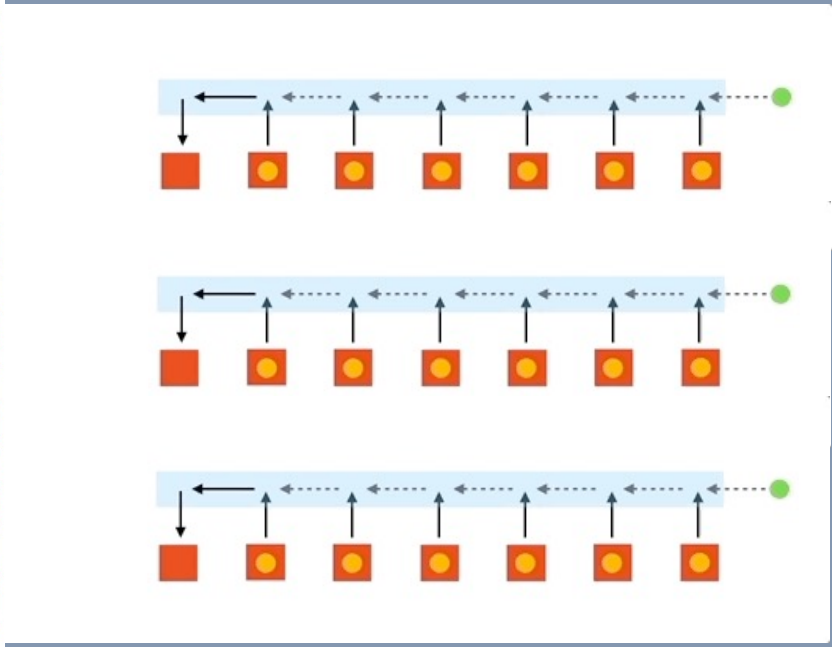
- Pre-routed virtual channels (“colors”)
- Single word packets
- Single clock latency
- Arrival triggers a task
- Data arrives in registers
- 24 colors
- Link level flow control
- Communication in the ISA



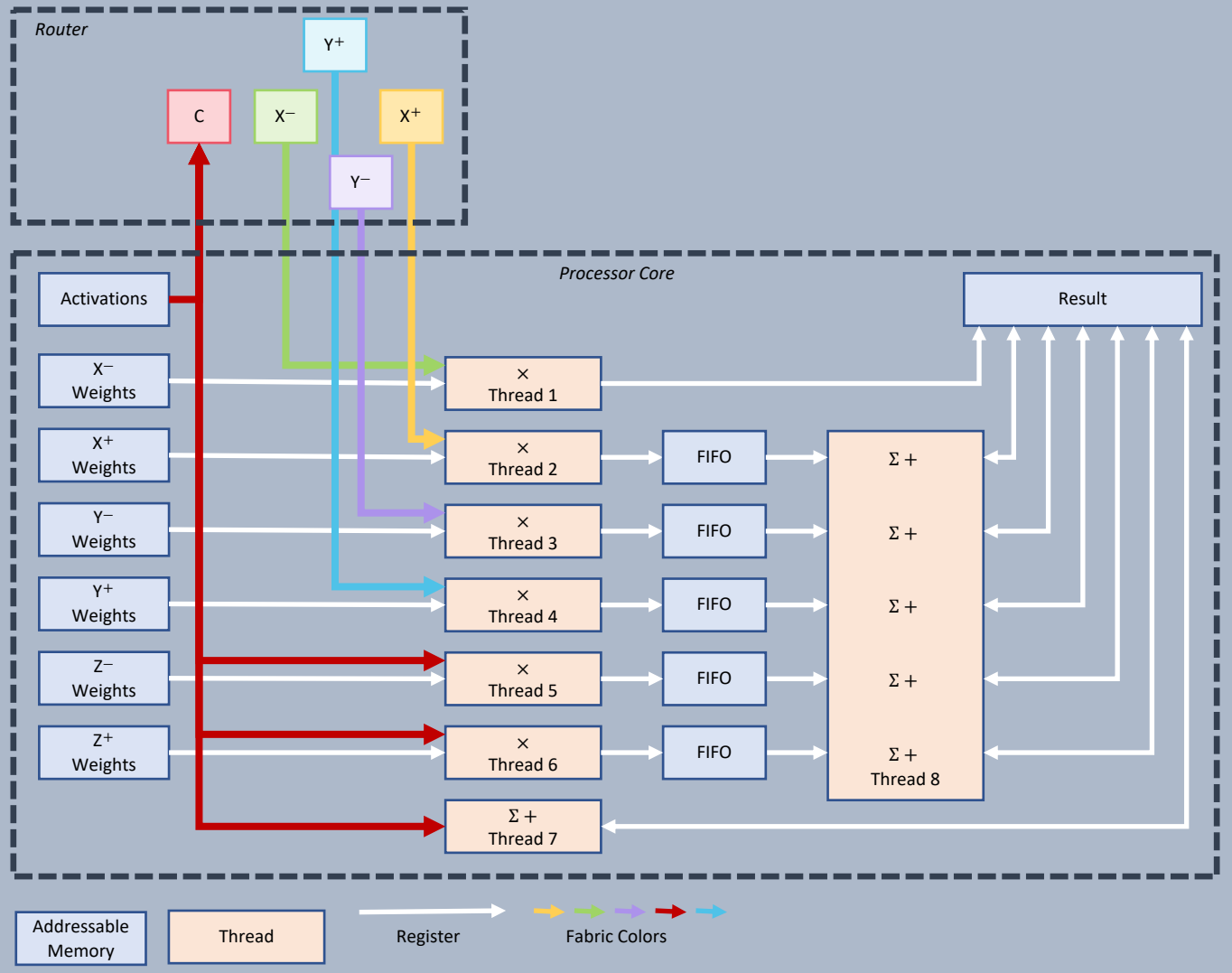
Allreduce in 1.3 μs







Sparse matrix vector product via vector operations and dataflow



Writing your own code: The SDK

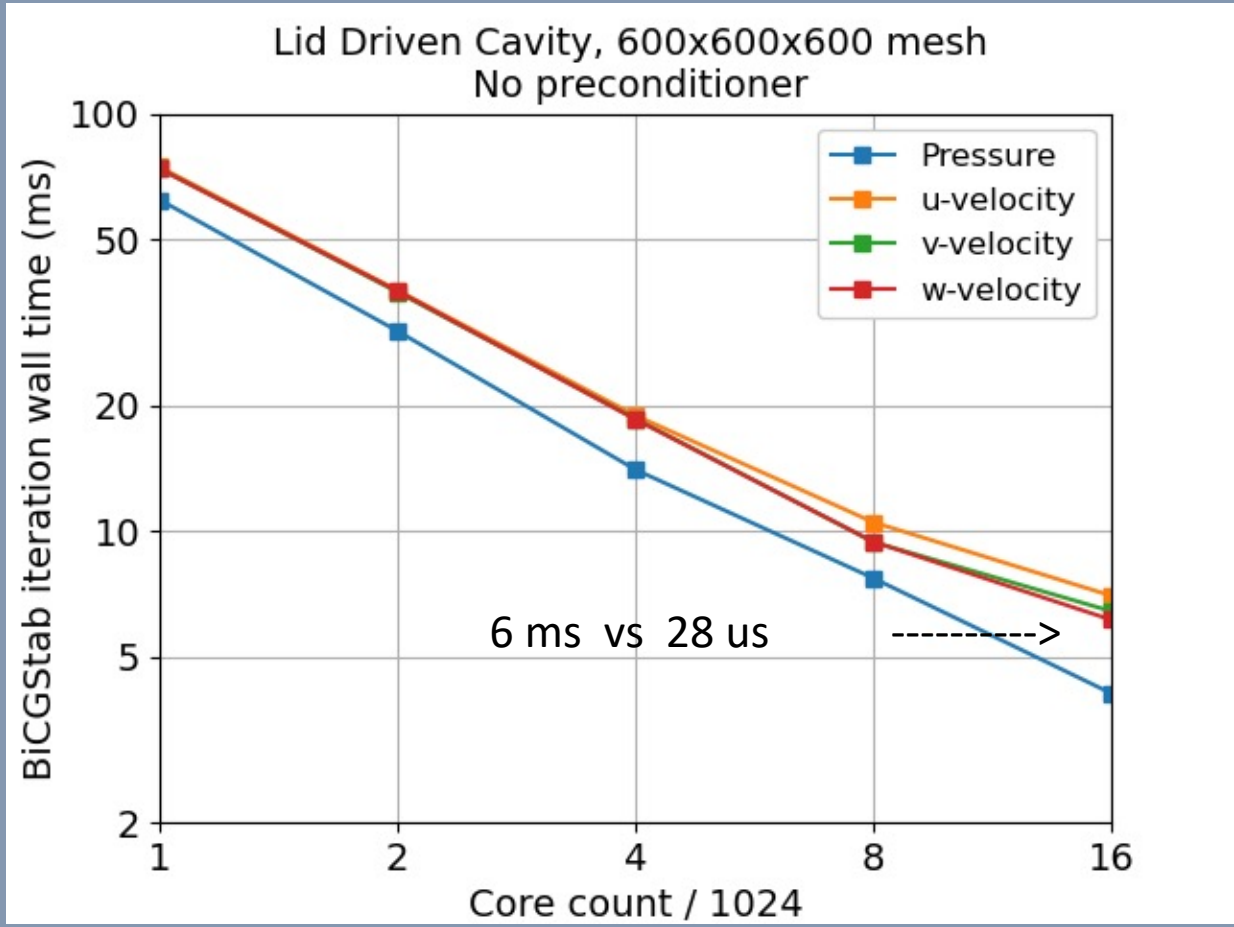
- The SDK: Low-level programming for creating custom kernels
  - DSL with abstractions for the ***lower-level*** constructs of the WSE architecture
  - Libraries for common primitives, such as communication, BLAS, rand, etc
  - Debugger and performance profiling tools
  - Hardware simulator
  - Examples and documentation: language specification, sample code, and programming guides

Beta --- September 2021

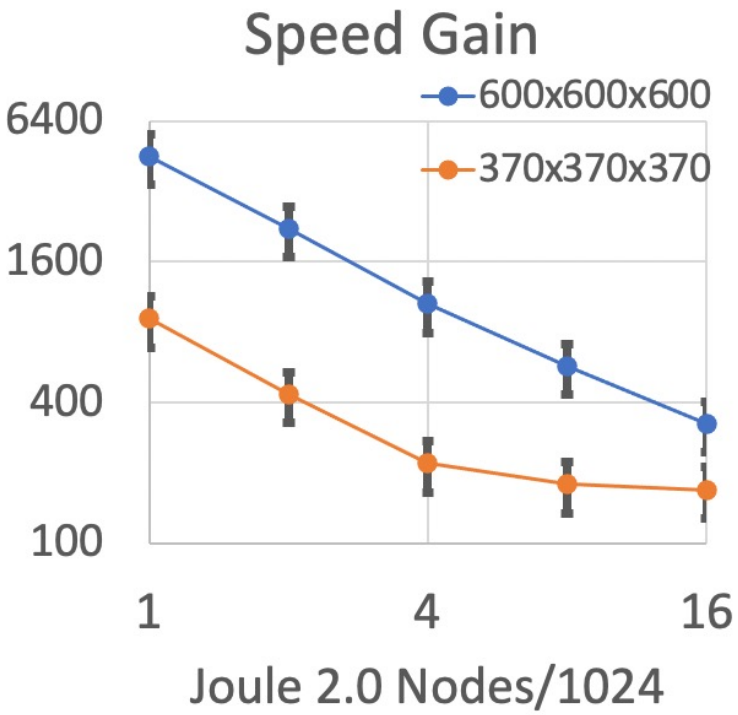
Results

On NETL Xeon Cluster

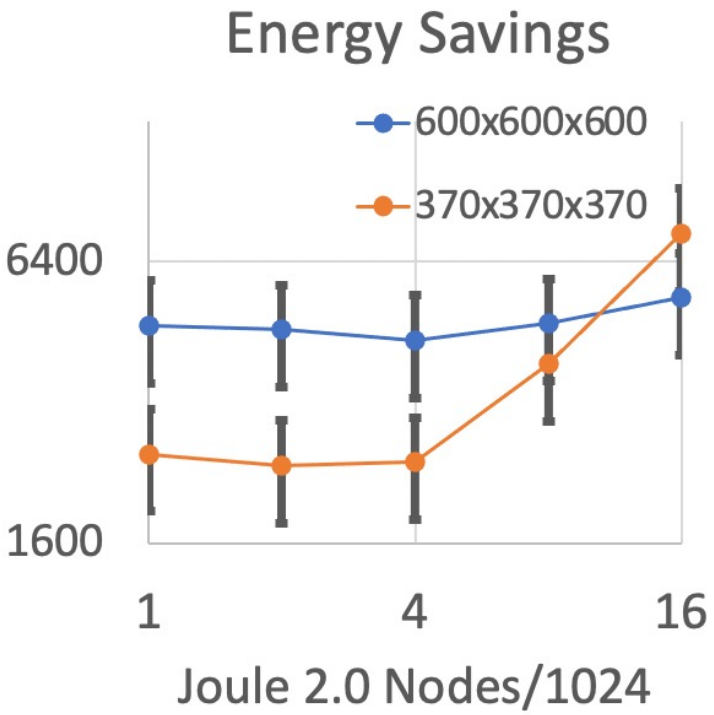
- 0.86 PF/s (600x600x1536)
- Over 30 % of peak
- 28 usecs / iteration
- ~200 X cluster



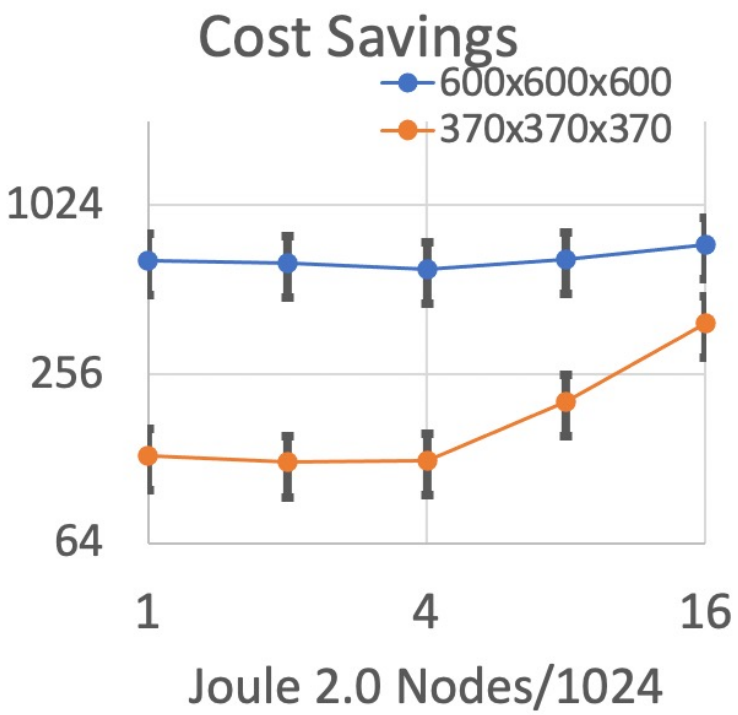
National Energy Technology Lab



200x Faster



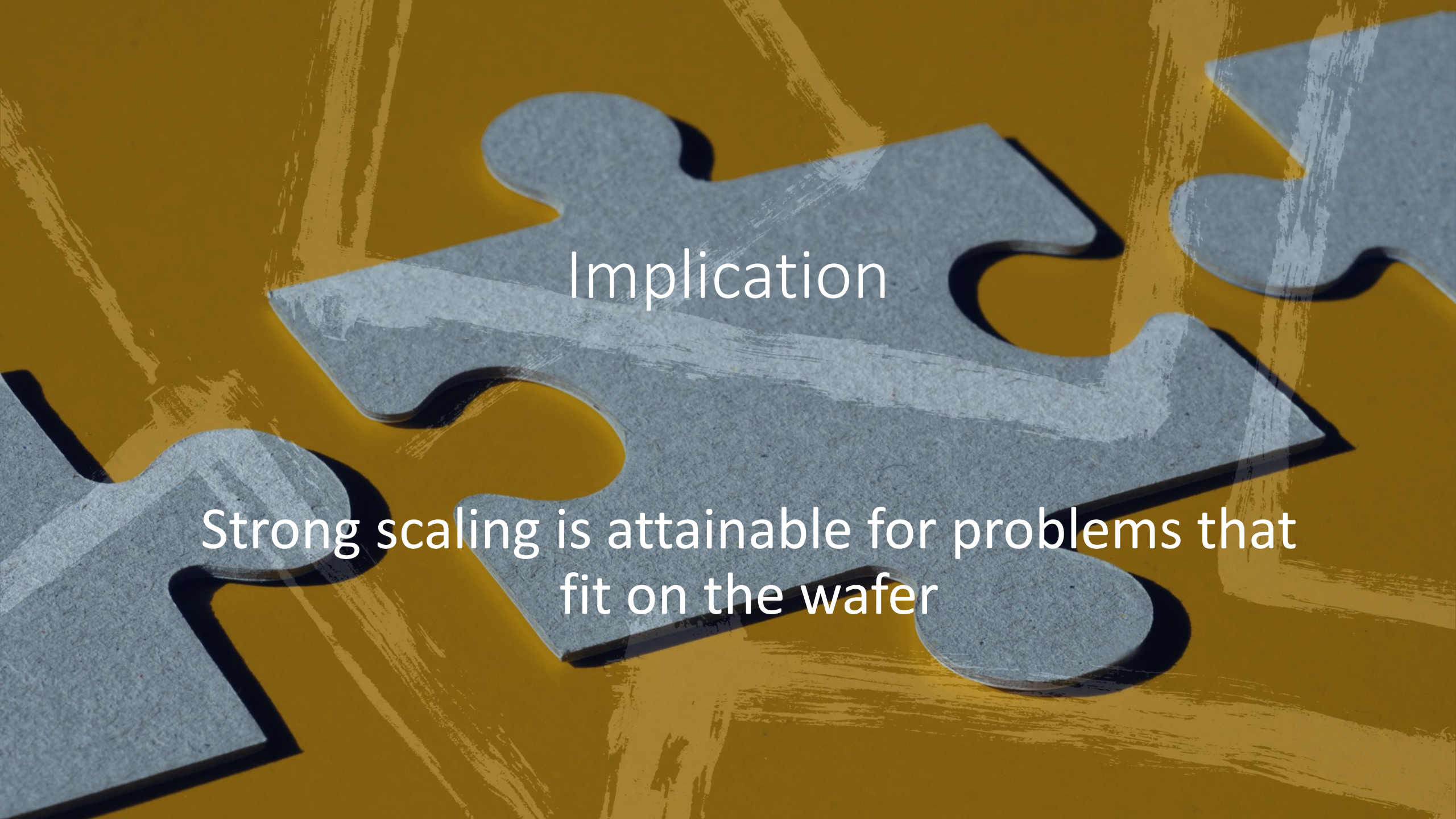
4,600x Less Power



650x Less Cost

cerebras

*Plots are from customer's cost models.



Implication

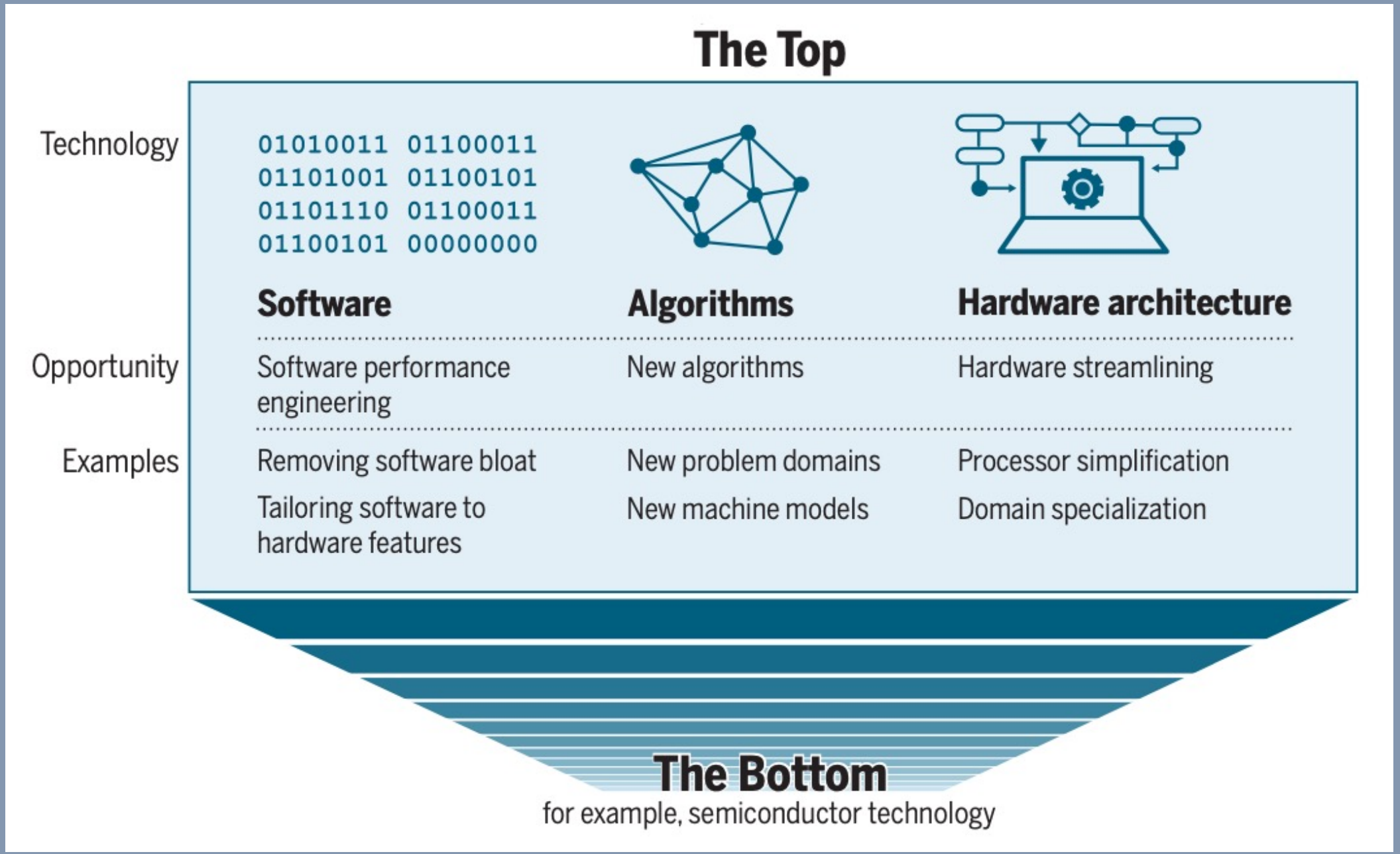
Strong scaling is attainable for problems that fit on the wafer

Making an Impact: Real Time CFD

Real Time CFD

- Online Equipment Monitoring
- Cyber-Physical Security
- Failure Prediction
- Renewable Integration
- Dynamic Baseload Power
- Higher Efficiencies
- Safer Operation
- Better Command and Control

Conclusion


The Top
Technology
01010011 01100011
01101001 01100101
01101110 01100011
01100101 00000000
Software
Algorithms
Hardware architecture
Opportunity
Software performance engineering
New algorithms
Hardware streamlining
Examples
Removing software bloat
New problem domains
Processor simplification
Tailoring software to hardware features
New machine models
Domain specialization
The Bottom
for example, semiconductor technology