Wafer Scale Computing

Tesla Dojo vs Cerebras WS2

System, hardware and packaging analysis

Tom DeBonis

With additional input fromYoann Foucher, Avishaii Abuhatzera, and many other experts

WW03 2022

Source: Tesla AI Day Livestream

intel®

Summary

- Tesla is building a custom solution for AI training to replace Nvidia GPU currently used to train AI models
  1. **Innovative packaging technology** build on new TSMC technologies like InFO-SoW (Wafer-scale, KGD) and InFO_SoIS (Chiplet level VRs)in addition to a custom thermal & power delivery system
  2. By breaking traditional system boundaries, Tesla introduces a massive and **unique system architecture** customized for AI Vision training

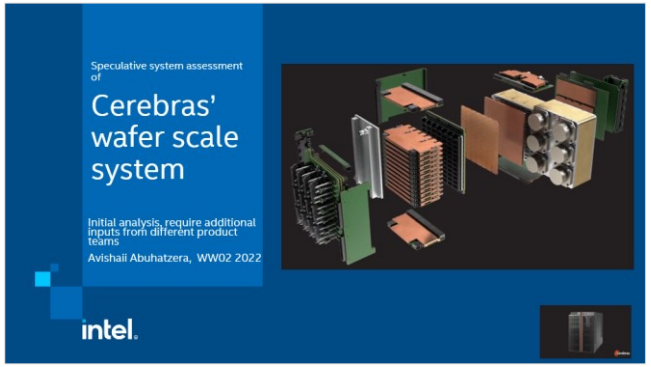
- Cerebras is shipping 2nd gen wafer-scale AI training system
  - ~215 x 215mm array of "die" cut from single wafer. Flexible membrane interfaces to PD and IO

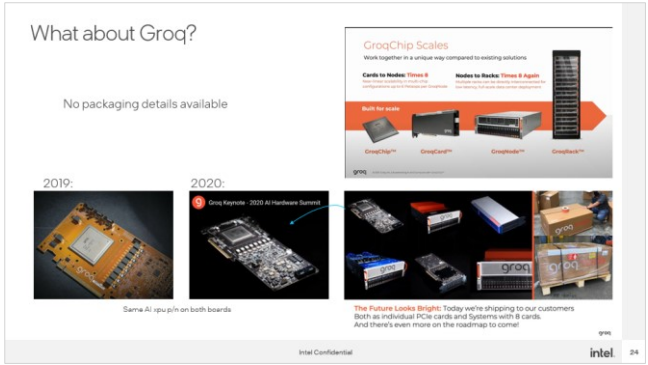
Contents


Dojo vs Cerebras Construction
Tom DeBonis, ATTD C/A
Intel Confidential
intel.
4


Dojo Assembly Deep Dive
Intel Confidential
intel.
25


2. Dojo Architecture overview
Yoann Foucher, DCAI CSO
intel.


Speculative system assessment of
Cerebras' wafer scale system
Initial analysis, require additional inputs from different product teams
Avishaii Abuhatzera, WW02 2022
intel.


What about Groq?
No packaging details available
GroqChip Scales
2019:
2020:
Intel Confidential
intel.
24

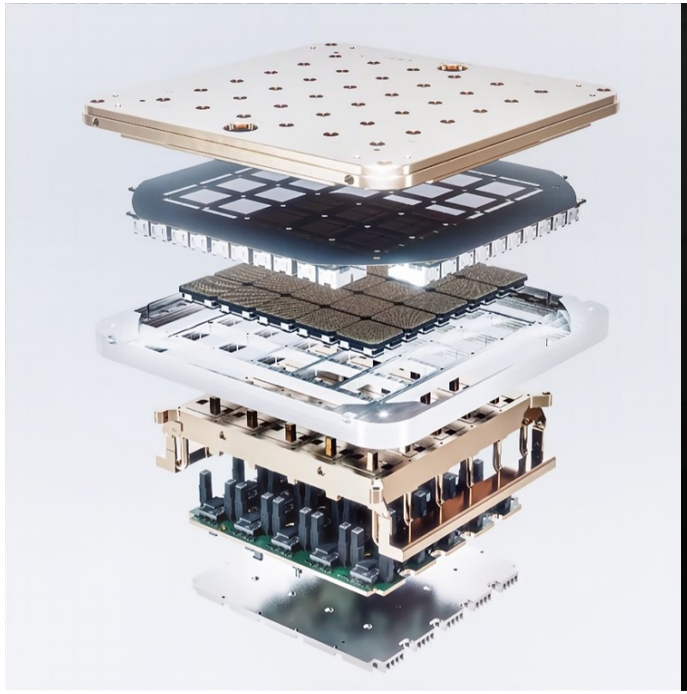
Dojo vs Cerebras Construction

Tom DeBonis, ATTD C/A

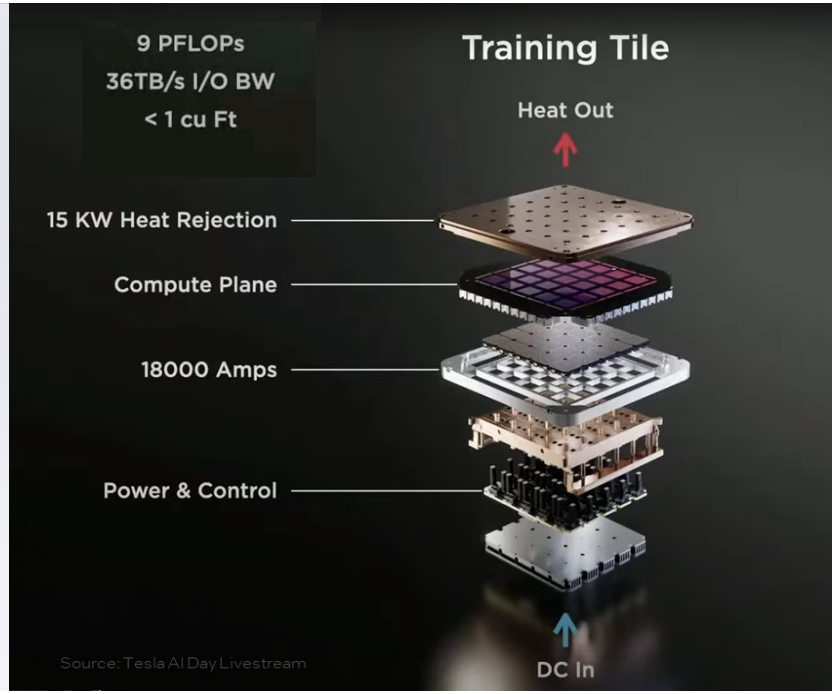
Dojo…

Sources:

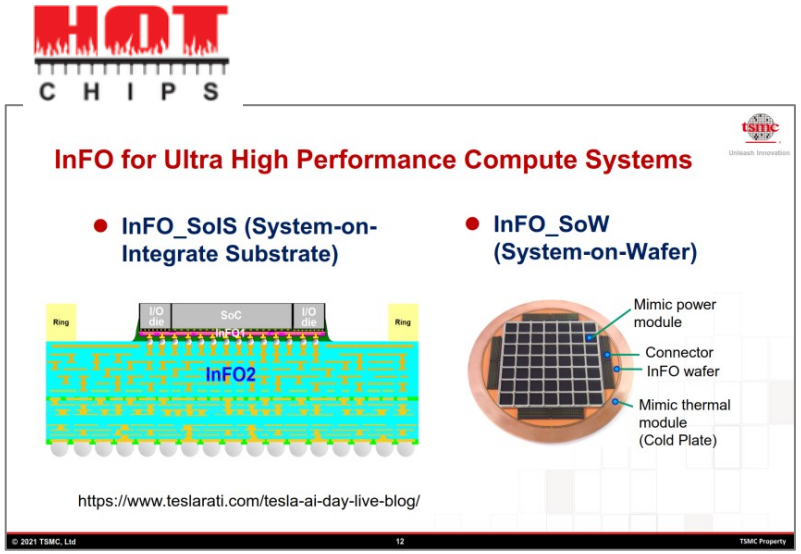
@DennisHongRobot,
Twitter post, 8/2



Ganesh Venkataramanan
Tesla AI Day, 8/19



Doug Yu, TSMC
Advanced Packaging Tutorial, 8/22



-
-
TSMC Papers/Conferences

Construction:

- Full-wafer InFO_SoW interposer connects 25 CPUs + IOs
- SoIS "PCBs" reflowed to SoW for local Power Management
- Vertical Power Delivery. Dual side cooling.

Cerebras CS-2 Website, YouTube, Datasheets

https://cerebras.net/system/

>$4B valuation (2H-21), $250M funding.
Main customer: Argonne Nat Lab
~300 employees (mostly SW)
SW Beta release only in mid-October 2021

cerebras

CS-2 DATA SHEET
PAGE 1

Cerebras CS-2

Purpose-built deep learning system delivering performance at unprecedented speeds and scale

AI Insights in Minutes, not Months

The CS-2 is the industry's fastest AI accelerator. It reduces training times from months to minutes, and inference latencies from milliseconds to microseconds. And the CS-2 requires only a fraction of the space and power of graphics processing unit-based AI compute.

The CS-2 features 850,000 AI optimized compute cores, 40GB of on-chip SRAM, 20 PB/s memory bandwidth and 220Pb/s interconnect, all enabled by purpose-built packaging, cooling, and power delivery. It is fed by 1.2 terabits of I/O across 12 100Gb Ethernet links. Every design choice has been made to accelerate deep learning, reducing training times and inference latencies by orders of magnitude.

Powered by the 2nd Generation Wafer-Scale Processor

The CS-2 is powered by the largest processor ever built — the industry's only 2.6 trillion transistor silicon device. The Cerebras Wafer Scale Engine 2 (WSE2) delivers more AI optimized compute cores, more fast memory, and more fabric bandwidth than any other deep learning processor in existence.

At 46,225 mm^2, WSE2 is 56 times larger than the largest graphics processing unit. The WSE2 contains 123x more compute cores and 1,000x more high performance on chip memory.

Seamless Software Integration

The Cerebras software platform integrates with popular machine learning frameworks like TensorFlow and PyTorch, so researchers can use familiar tools and rapidly bring their models to the CS-2.

The platform is fully programmable and provides both an extensive library of primitives for standard deep learning computations, as well as a familiar C-like interface for developing custom kernels and applications.

Unlock New Paths of ML Research

The performance and scale of the CS-2 unlocks entirely new classes of models, learning algorithms, and researcher opportunities. These include exceptionally sparse networks and very wide, shallow networks. The CS-2 provides faster time to solution, with cluster-scale resources on a single chip and with full utilization at any batch size, including batch size 1.

cerebras

CS-2 DATA SHEET
PAGE 2

Datacenter Deployment

Standard install, interface, and management redundancy and carrier-grade reliability built-in

Specifications

Sparse Linear Algebra Compute Cores
850,000

On-chip Memory
40GB SRAM

Memory Bandwidth
20 PB/sec

Core-to-Core Bandwidth
220 Pb/sec

Maximum Power Requirement
23 kW

System IO
12x 100 Gb Ethernet

Cooling
Air- or water-cooled

Dimensions
15 Rack Units (26.25")

Dimensions

15 RU x 445mm x 1005mm
- Fits in standard EIA 19" 1200 mm (47") deep rack
- Depth-adjustable support rails and brackets

300 kg (660 lbs)
- ~20kg/RU (44lb/RU)

Power

6+6 redundant 4kW power supplies

Inlets: 12x IEC 60320 C20

Inputs: 200-240 VAC, 16A, 50/60 Hz
- Independent single-phase inputs
- Protection: each inlet individually protected with external 16A (20A UL) circuit breaker

Network

Integrated optical multimode transceivers

12x 100GbE Data Ports (OM4 MPO/MTP-12)
- 100GBase-SR4 link
- Accepts MPO/MTP-12 fiber strand push-on cables

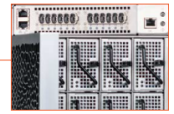
Use Type-B cross-over OM4 MPO/MTP-12 50/125µm multi-mode fiber patch cable to plug into industry standard 100GBase-SR4 optical module

Management

1x 1GbE Management Port (RJ45)

1x Console Port (RJ45)

1x Power Management Port (RJ45)

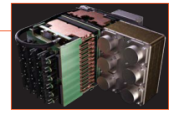


Cooling

Internal closed-loop, direct-to-chip liquid cooling

Can be deployed with external liquid coolant loop or liquid-to-air cooling

Internal coolant loop: 1+1 redundant hot-swappable pumps



CEREBRAS SYSTEMS, INC.
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET

cerebras

WSE-2 DATA SHEET

Wafer-Scale Engine: The Largest Chip Ever Built

The Wafer-Scale Engine (WSE-2), which powers the Cerebras CS-2 system, is the largest chip ever built. The WSE-2 is 56 times larger than the largest GPU, has 123 times more compute cores, and 1000 times more high-performance on-chip memory. The only wafer scale processor ever produced, it contains 2.6 trillion transistors, 850,000 AI-optimized cores, and 40 gigabytes of high performance on-wafer memory all at accelerating your AI work.

Cerebras Wafer-Scale Engine

Fabrication process
7nm

Silicon area
46,225mm²

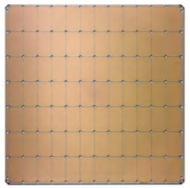
Transistors
2.6 Trillion

AI-optimized cores
850,000

Memory (on-chip)
40GB

Memory bandwidth
20PB/s

Fabric bandwidth
220Pb/s



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

Largest GPU
54.2 Billion Transistors
826 mm² Silicon

Compute Designed for AI

Each core on the WSE-2 is independently programmable and optimized for the tensor-based, sparse linear algebra operations that underpin neural network training and inference for deep learning. The WSE-2 empowers teams to train and run AI models at unprecedented speed and scale, without the complex distributed programming techniques required to use a GPU cluster.

Cluster-Scale in a Single Chip

Unlike traditional devices with tiny amounts of on-chip cache memory and limited communication bandwidth, the WSE-2 features 40GB of on-chip SRAM, spread evenly across the entire surface of the chip, providing every core with single-clock-cycle access to fast memory at an extremely high bandwidth of 20PB/s. This is 1,000x more capacity and 9,800x greater bandwidth than the leading GPU.

High Bandwidth, Low Latency

The WSE-2 on-wafer interconnect eliminates the communication slowdown and inefficiencies of connecting hundreds of small devices via wires and cables. It delivers an astonishing 220 Pb/s interconnect bandwidth between cores. That's more than 45,000x the bandwidth delivered between graphics processors. The result is faster, more efficient execution for your deep learning work at a fraction of the power draw of traditional GPU clusters..

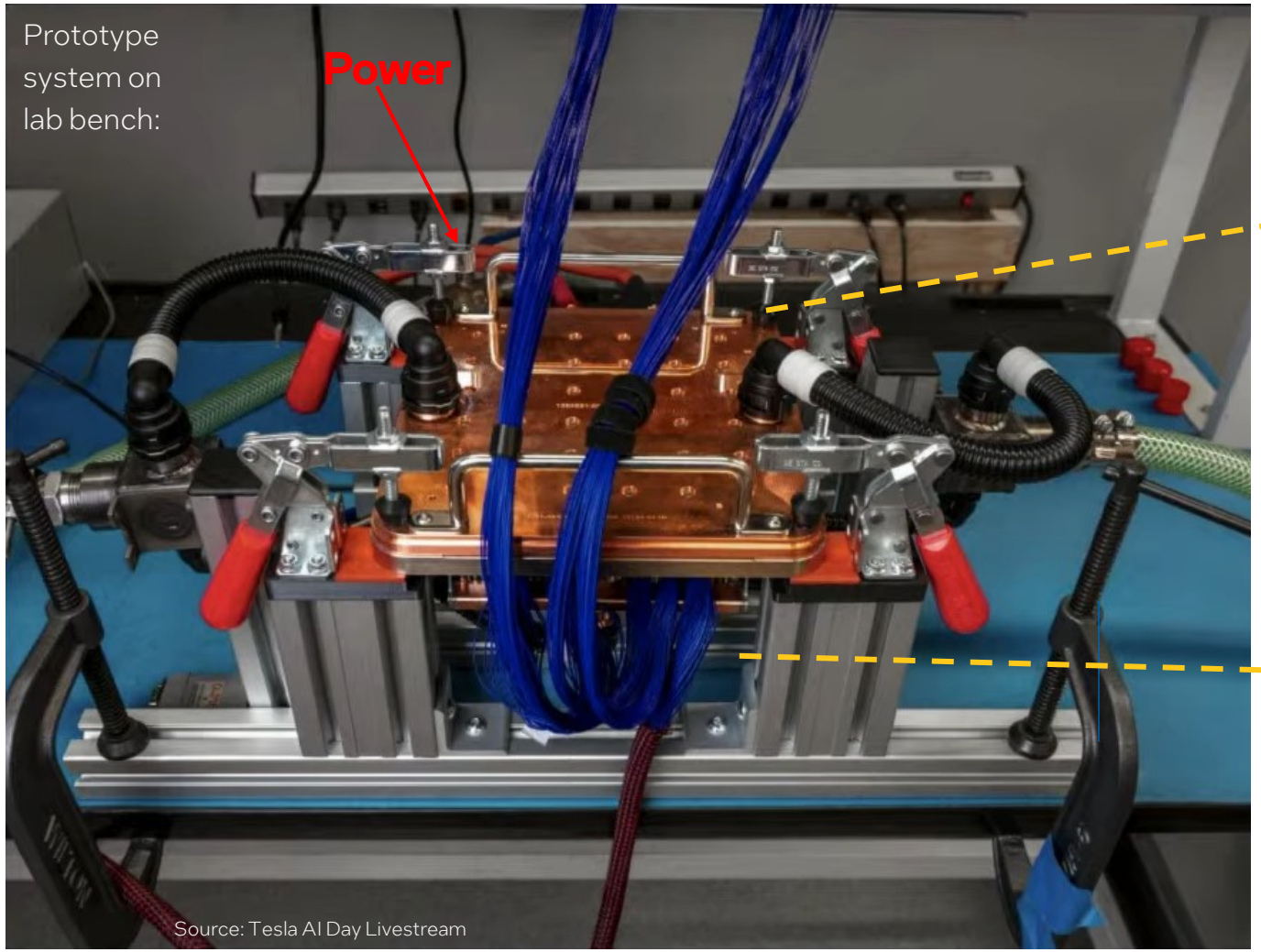
For more information about the Cerebras CS-2 system click here.

CEREBRAS SYSTEMS, INC.
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET

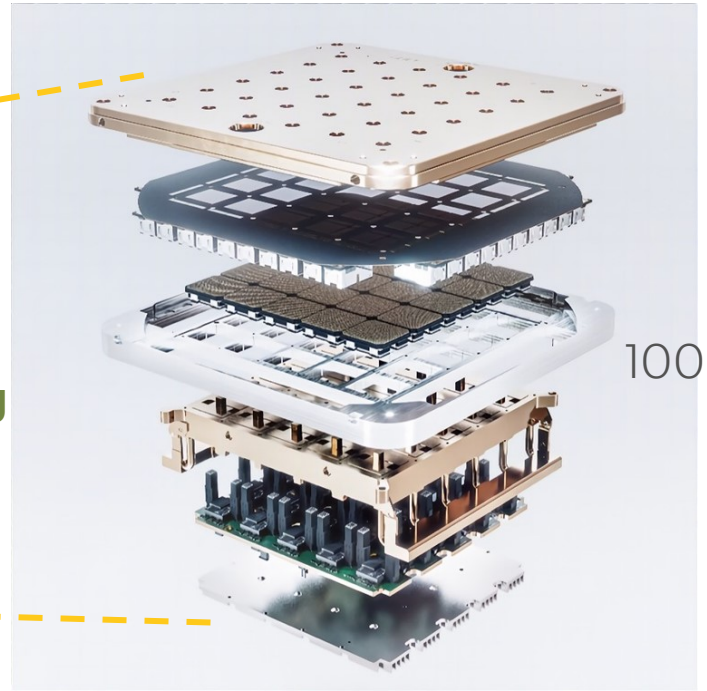
Dojo Lab Prototype

HSIO

275mm x,y



Cooling



100mm z

Control

Cerebras

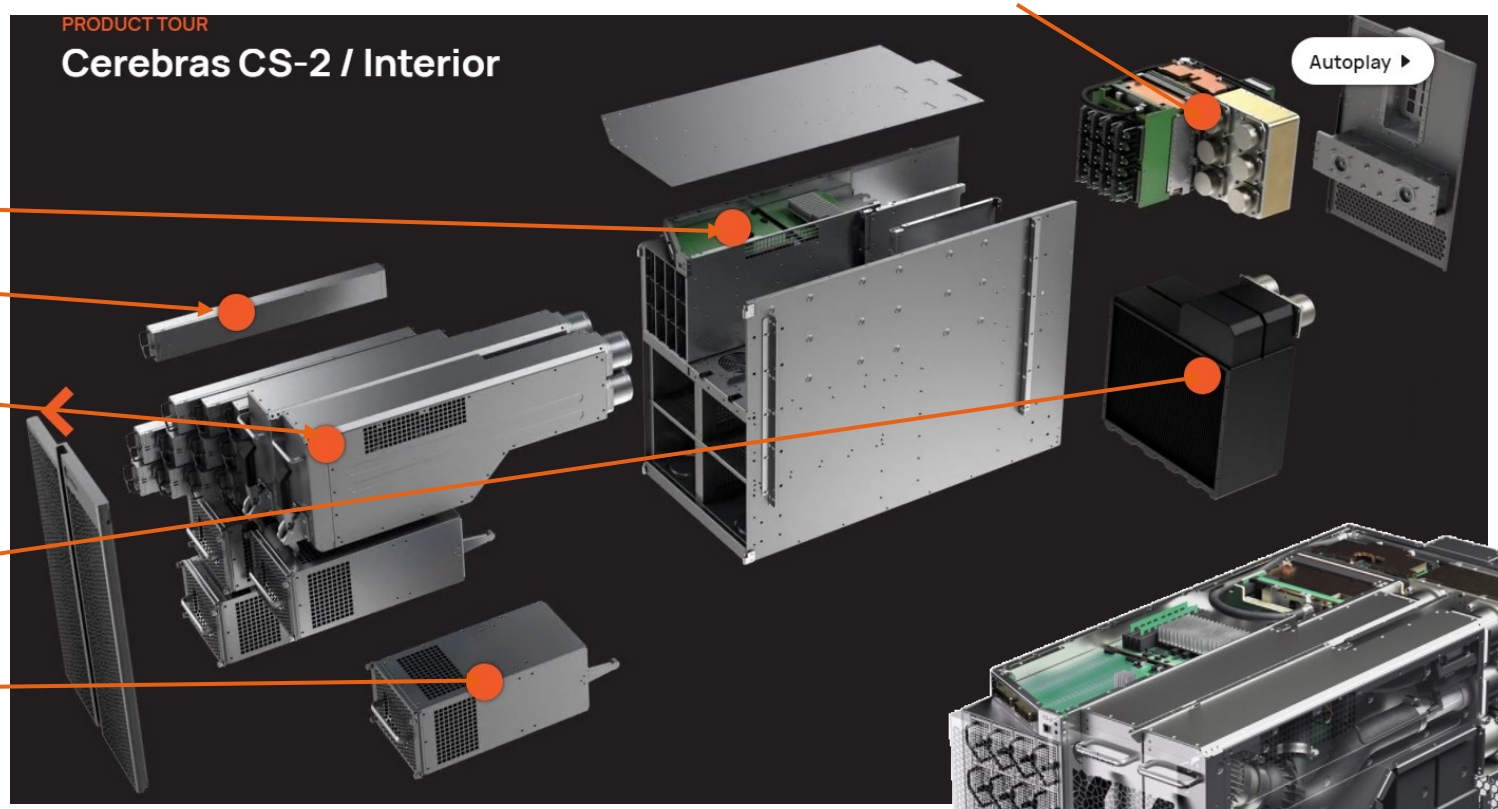
Engine Block

System Manager Board

- 240VAC-48V DC Power Supplies (12)
- 10gpm Pump Modules (2)
  - hot swappable

Air Heat Exchanger

- Fan Modules (4)

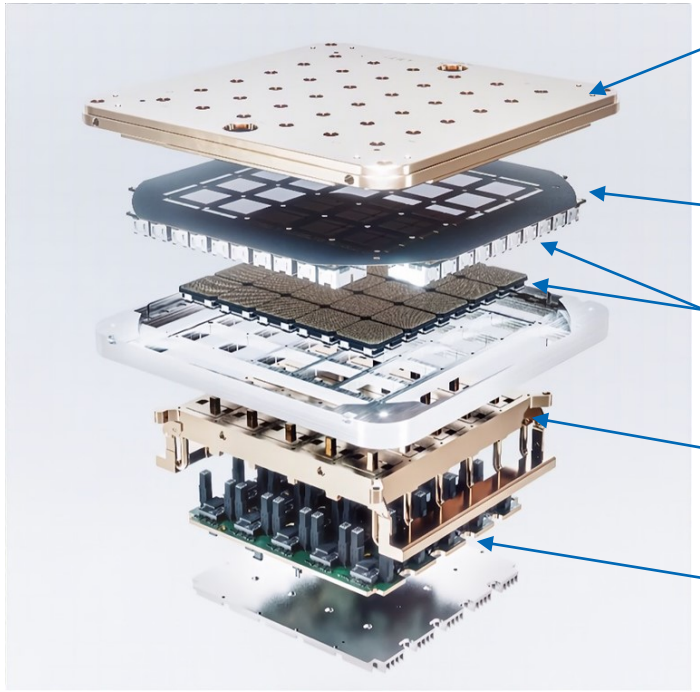


https://cerebras.net/system/

Back

Front

Construction Overview …



Tesla Dojo

Water cooled cold plate for NPUs
Ethylene glycol cooling manifold
U-channel cold plate

InFO_SoW Compute plane
Wafer-scale engine

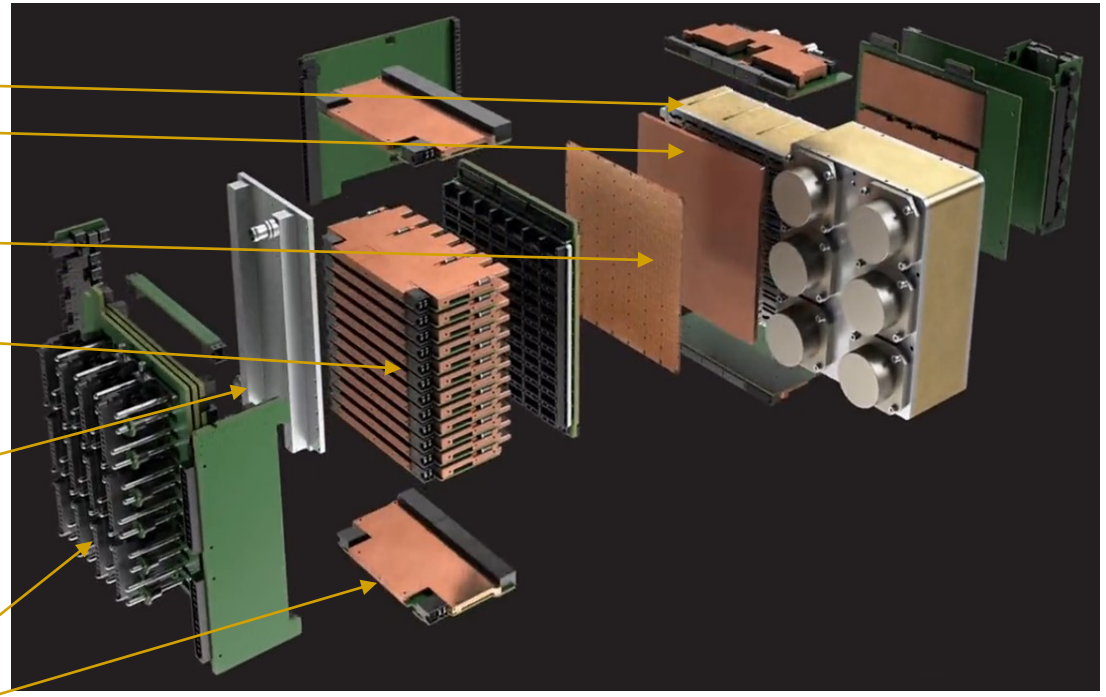
InFO_SoIS with IO & local VRMs
Hot swappable 54V to 0.9V PSUs

Power supply cooling manifold

Power & control module
Power distribution module

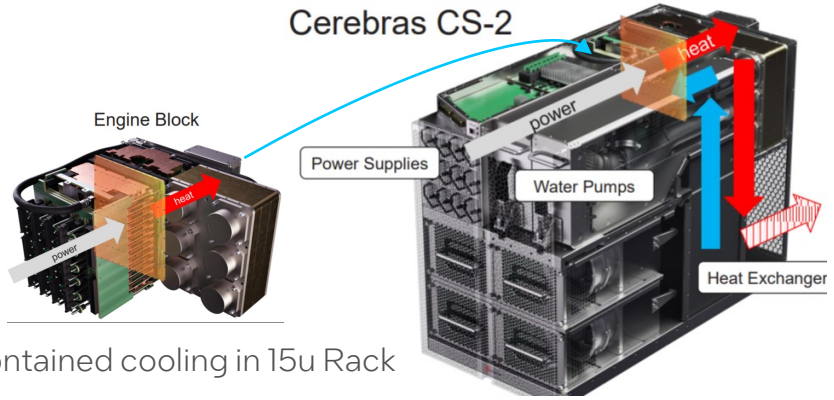
(2) Optical IO Boards

CS-2 Engine Block



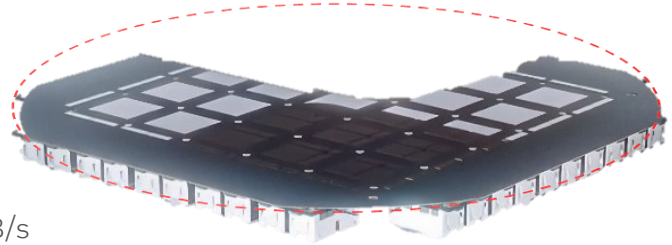
<1 Cu Ft, requires external cooling
6 tiles/tray, 2 trays in a cabinet

Cerebras CS-2



Self contained cooling in 15u Rack

Dojo vs Cerebras: Wafer-scale, Die level power delivery



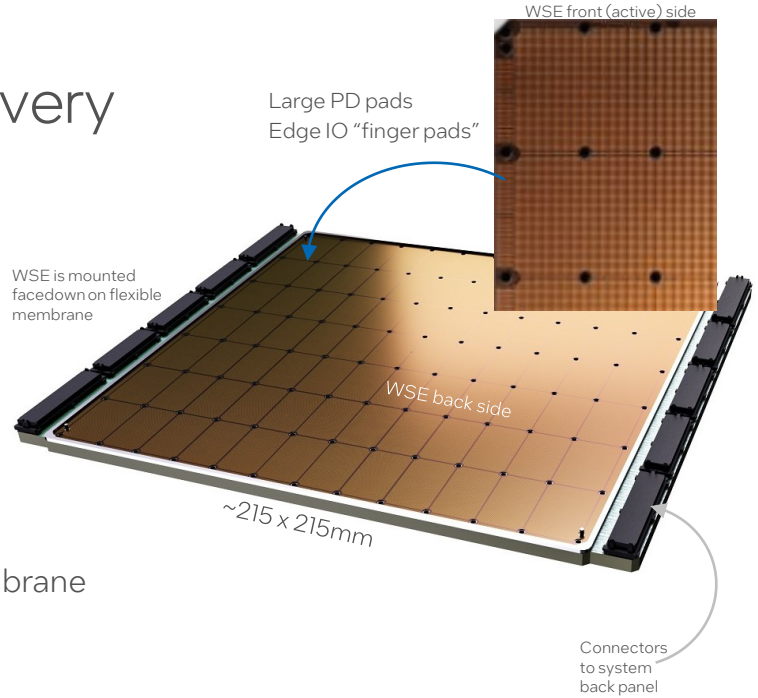
Dojo:

- Known Good Die
  - 40 IO Chips, 36 TB/s
  - 25 D1 NPU, ~730mm² die, 425MB SRAM, 400W
- Reconstituted with InFO
  - InFO_SoW: 6 Layers, 3@ 5/5, 3 @ 15/20 L/S
- Local TIM BLT control, ~screwed to cold plate

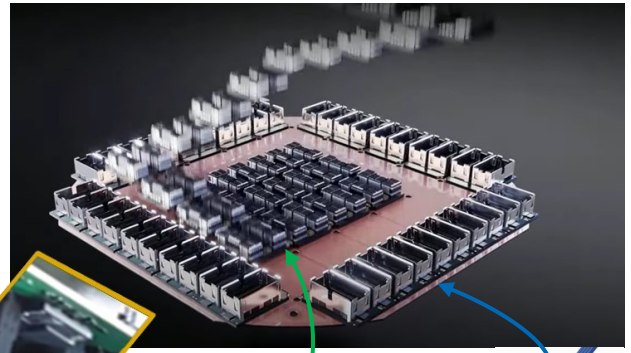
- IO's and VRs SMTd to "PSB side"
  - InFO_SoIS: 4+RDLs @ > 8/10 L/S

Cerebras:

- 1 Wafer , 46,225mm²
- Redundancy
  - 84 reticles, ea 525mm²
  - 40GB SRAM, 20PB/s bw
  - 220 Pb/s Fabric bw
- Floats on cold plate
- PD and IO via flexible membrane



Connectors and power modules are mounted on InFO_SoIS and SMT'd to bottom of InFO-SoW



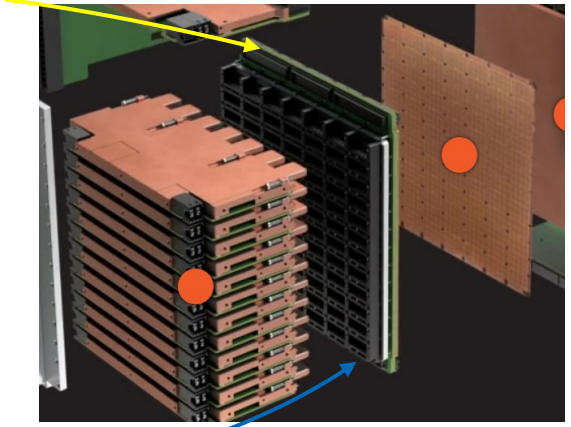
Vicor spring connector from bottom PCB

One VR module

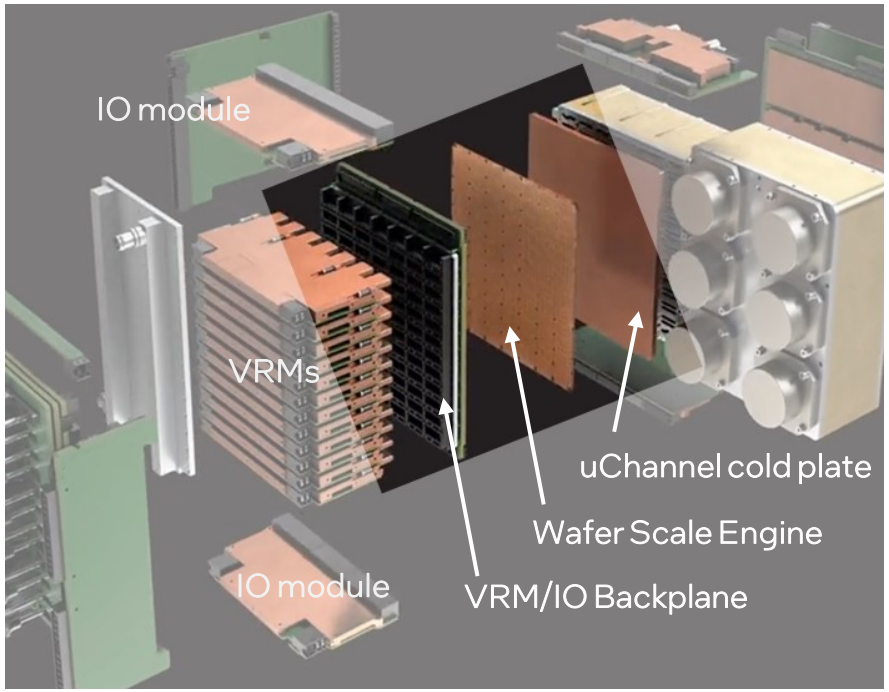
PCIe connectors?

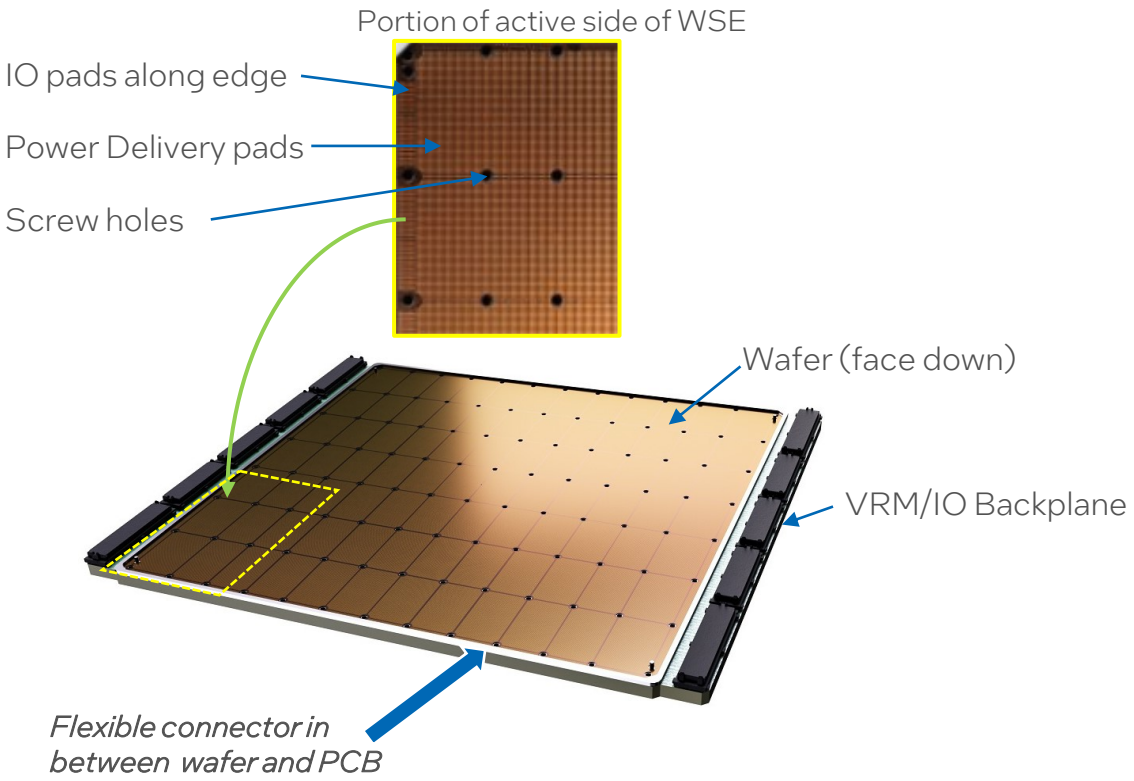
Flexible membrane connects WSE to bank of 12 water cooled VRM modules (48V → 1.2v, 1.0V, 0.8V).

12x 100GbE VCSEL Optical IO modules are above and below bank of VRM modules. Connection to WSE is not disclosed.

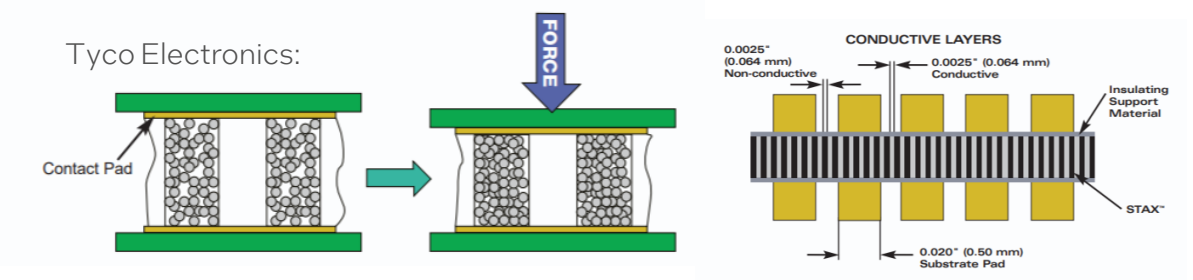


Cerebras: Wafer-scale, Die level power delivery





Tyco, Fujipoly Zebra, or similar elastomeric connector or flexible membrane provides electrical connection, and accommodates thermal expansion, etc. between wafer scale engine and PSU and IO backplane



Cerebras

System Manager Board

- 240VAC-48V DC PSUs (8+4)
- 10gpm Pump Modules (2)
  - hot swappable

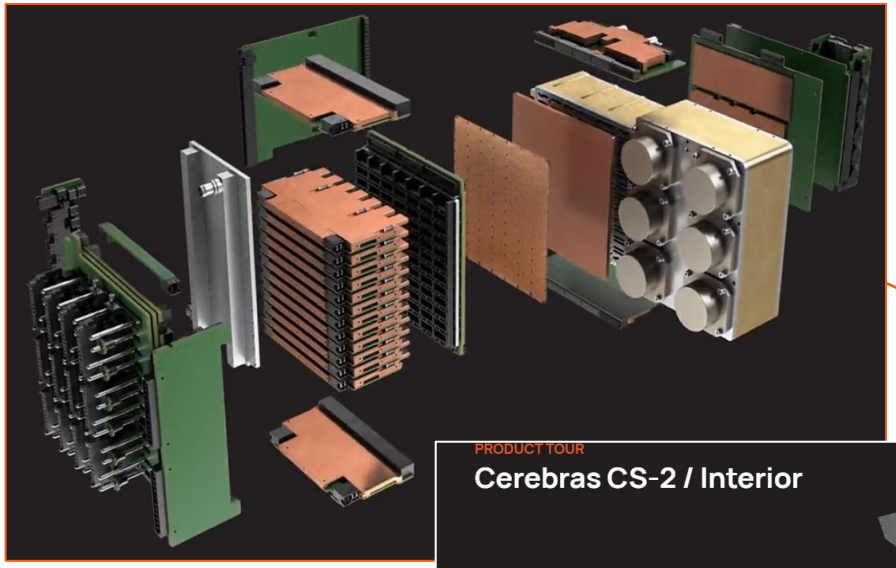
Air Heat Exchanger

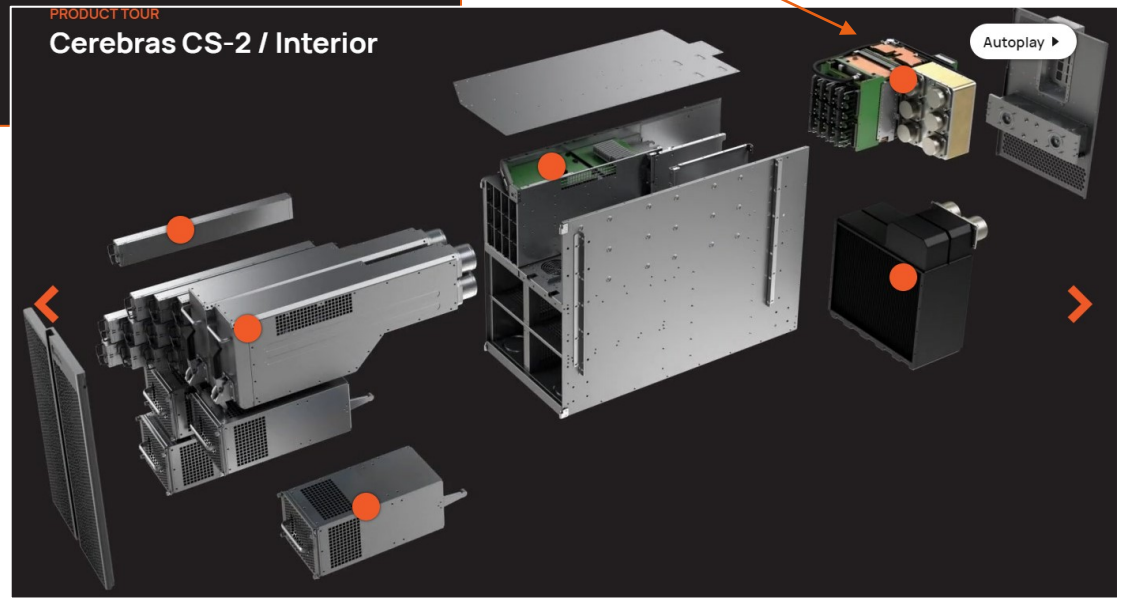
- Fan Modules (4)

Back

Front

Engine Block

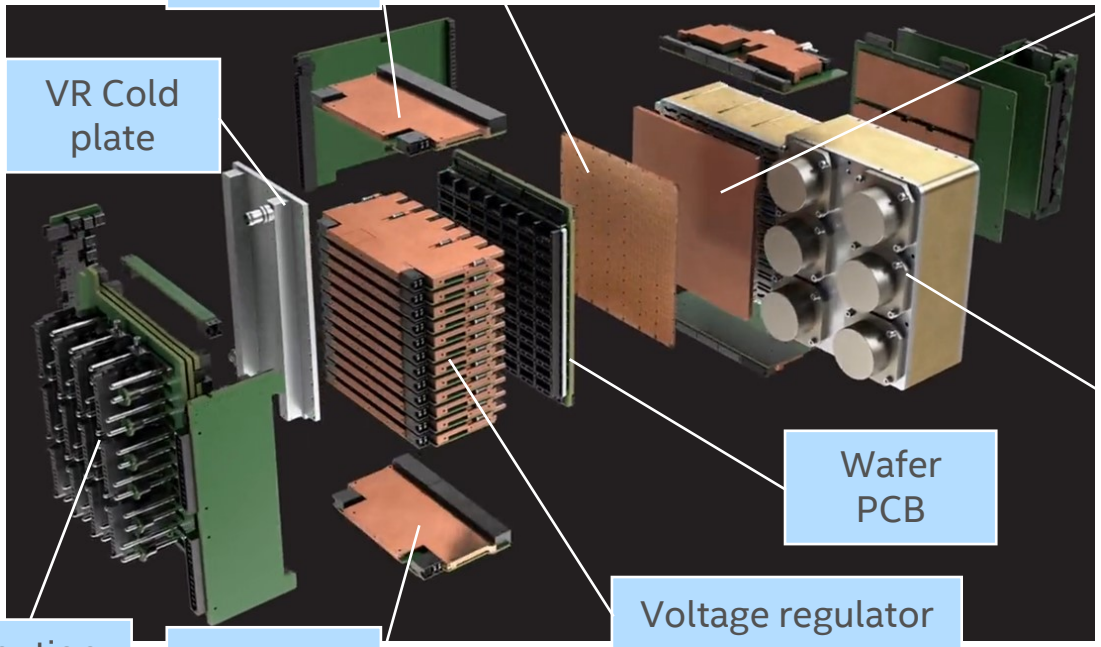




https://cerebras.net/system/

Cerebras Engine block

For holding the assembly together while maintaining electrical contacts for power and IO, Cerebras "sandwiched" the wafer in an assembly comprising a thick PCB, a flexible membrane, the WSE and its heat exchanger. An array of clamping fasteners distribute the packaging force evenly across the assembly while still allowing the wafer to expand and contract. This was solved by having the wafer "float" on a large copper heat exchanger. Water is pumped through micro-fins on the backside of the heat exchanger to remove heat from the powered wafer; the wafer is allowed to expand and contract while remaining in contact with the polished front side of the exchanger. The ability to float the wafer to maintain thermal connection to the heat exchanger despite the different coefficients of thermal expansion of copper and silicon is crucial.



Power supply connector

Liquid cooled voltage regulators delivering 20KA at less then 1v vertically

Cold plate

Copper heat exchanger design to >20kw of heat from wafer. Wafer is "floating" on the cold plate, allowing it to expend and contract with temperature while remining in contact with the plate

Cooling manifold

Transfer heat between cerebra's internal coolant and the facility water system. (separated from data center liquid system)
Upper 4 connector – connected to the water pump , the lower two used for heat exchange.

Cerebras Power delivery

- Supplying the >15kw into the wafer at while maintaining good regulation.
- Cerebras' solution employs more than 300 water cooled voltage regulation modules (VRMs) distributed over the wafer.
- Multiple VRMs per reticle ensures redundancy in the power distribution and gives individual control of each reticle's power domain.
- 12 power supplies (PSUs), in a 9+3 redundant configuration

Power challenge

Delivering >15kW to a single chip is hard!

Power is traditionally delivered from the periphery of a chip.

Power regulators are traditionally placed between large chips.

Traditional multi-stage power conversion topologies are inefficient.

Established power distribution and redundancy expectations in datacenters.

cerebras

Power solution

Fit power conversion within footprint of WSE

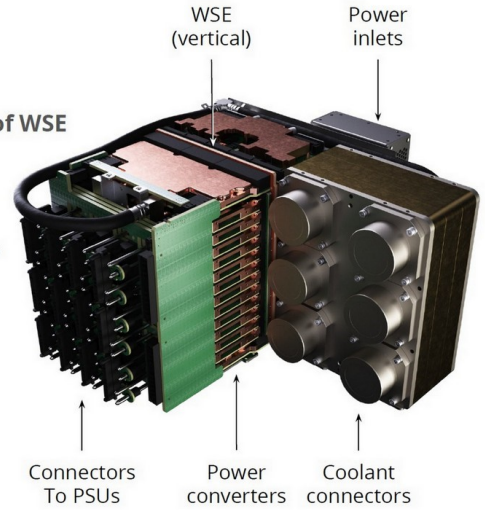
Place converters on opposite side of PCB

Bring in power through IEC C20 16A inlets

12x 4kW hot-swappable universal PSUs

Universal high-voltage AC to 54V DC

Direct conversion from 54V to 0.9V

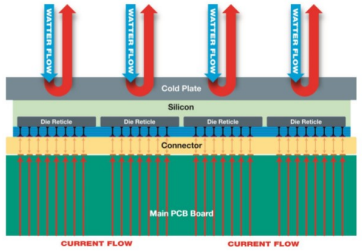


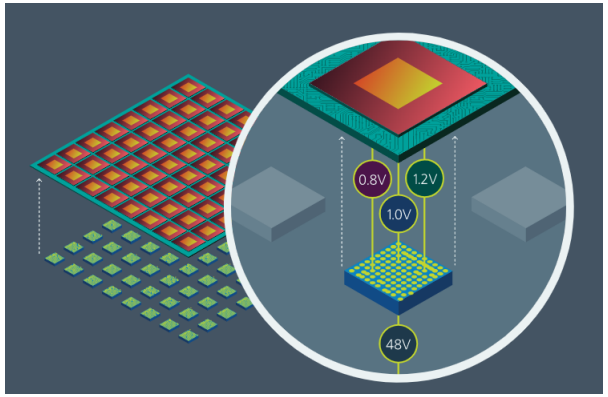
cerebras

Using the 3rd Dimension

- Power delivery
  - Current flow distributed in 3rd dimension perpendicular to wafer
- Heat removal
  - Water carries heat from wafer through cold plate

Co-designed with system

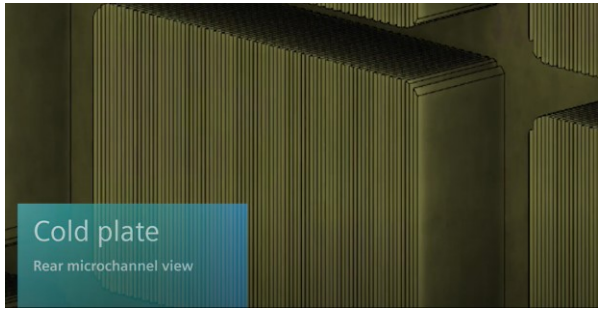


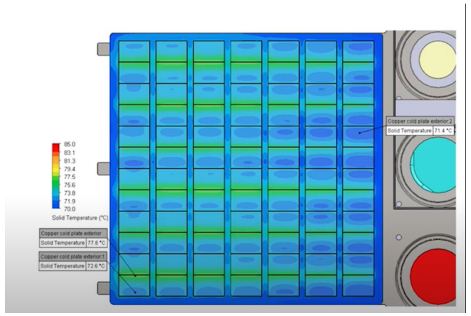


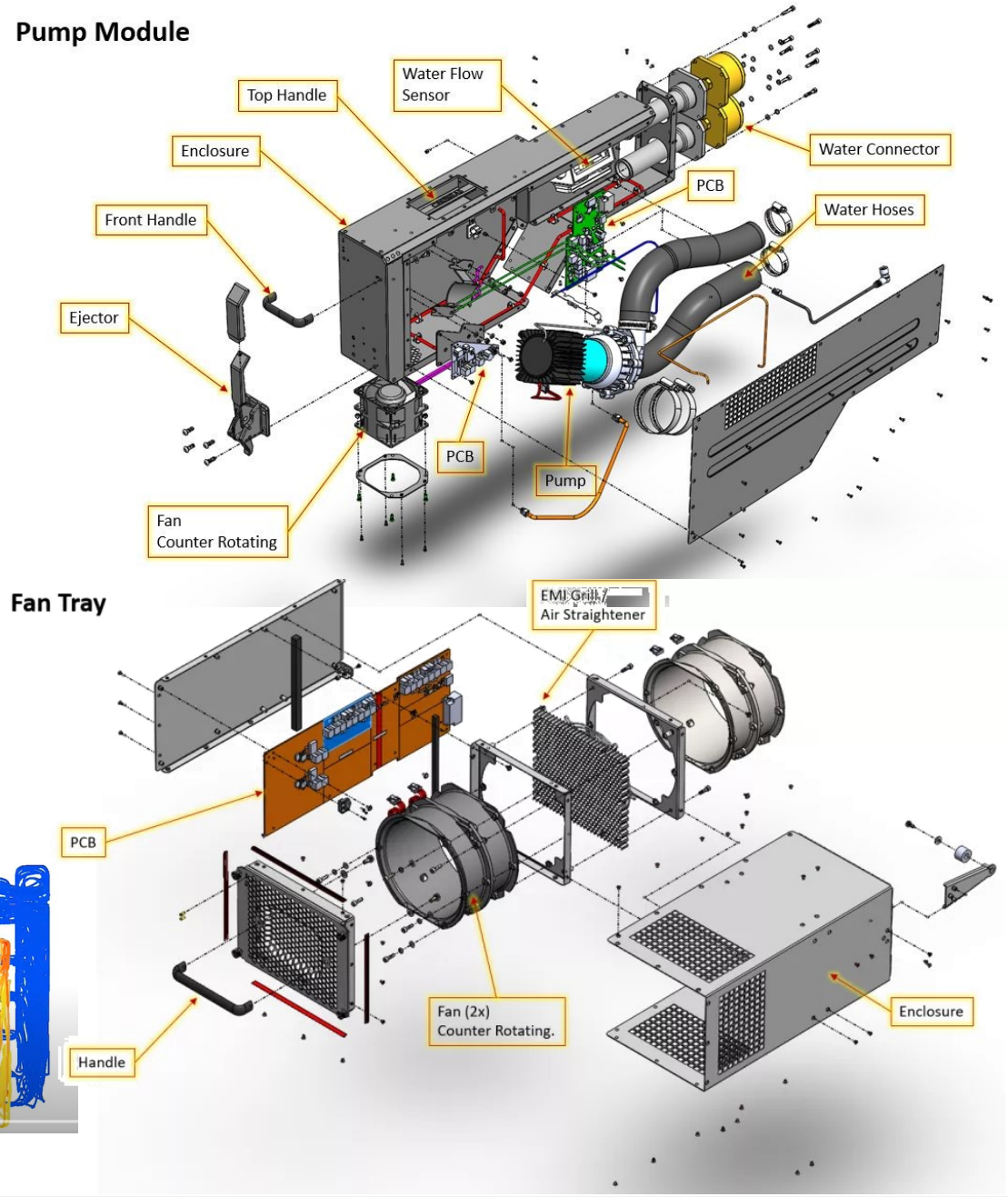
Cerebras Cooling system

- Water cooled cooper plate stacked on top of silicon.
  - Vendor : Motivair MCDU-25, which has 625 kW of thermal capacity
  - Water and Glycol
- The system involves two cooling loops. A primary loop mixes water with glycol coolant that traverse the CS-1s using pumps, then transfers heat onto a secondary loop (i.e., the datacenter's chilled water supply) via a heat exchanger.
- The cold plate receives water from a manifold to the right, which then delivers cooled water to several individual zones on the surface of the cooling plate.
- The heated water is then extracted, again from the small zones that ensure consistent thermal dissipation and pumped down to the heat exchanger at the bottom of the unit.
- The exchanger consists of an EMI grill and is cooled by powerful fans that employ air straighteners. Overall, the chip runs at half the junction temperature of a standard GPU, which increases reliability.
- CS2 cooling system occupies ~70% of the system size

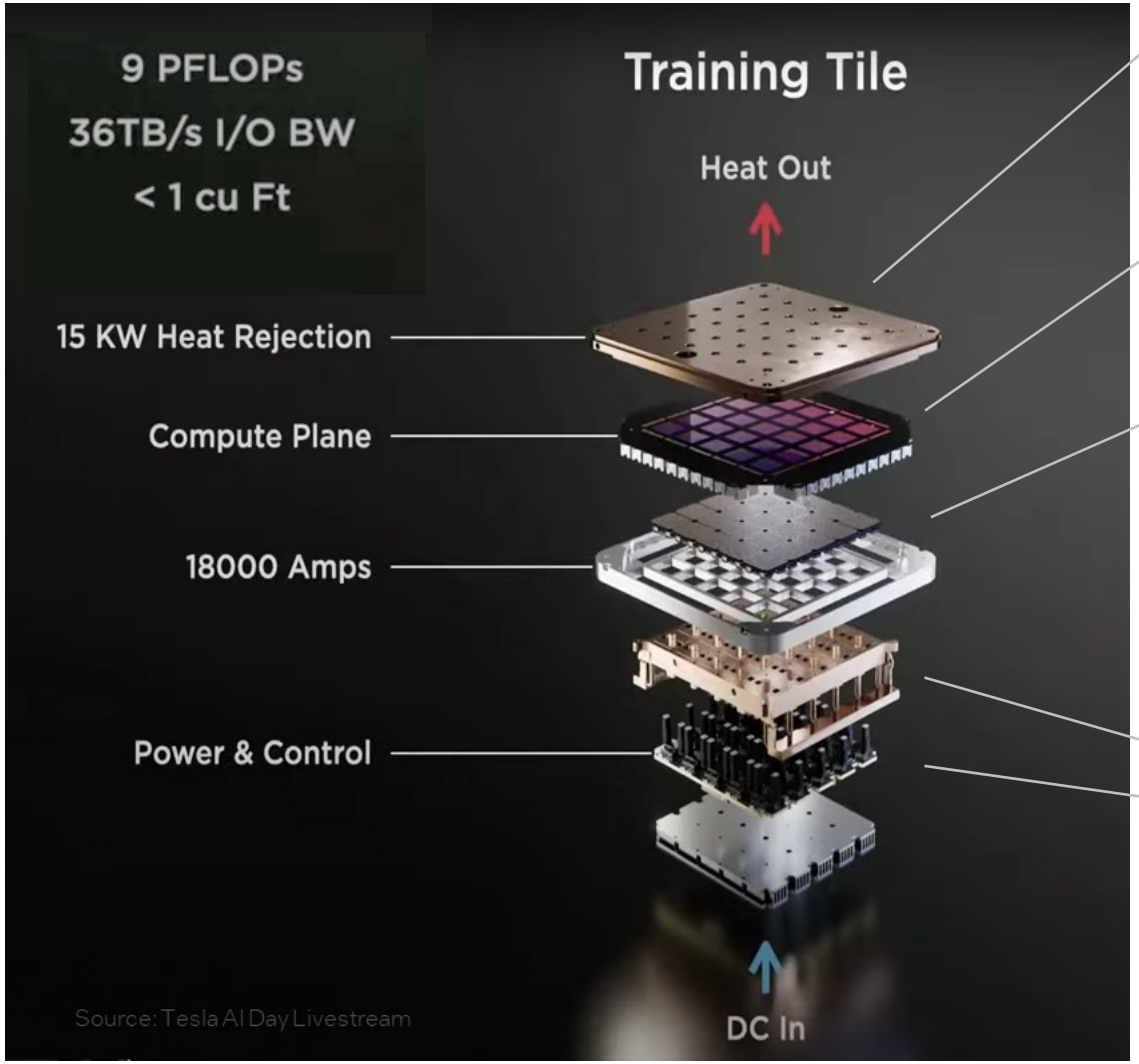
https://www.youtube.com/watch?v=dACywU1YCpc







Dojo Overview…



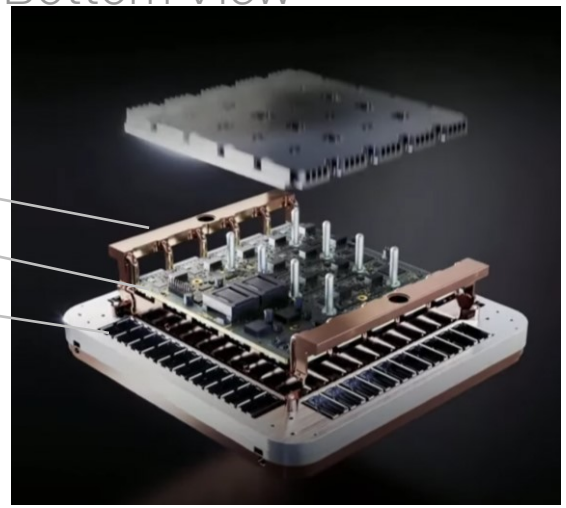
Water cooled "cold plate" for compute plane
Local TIM BLT Control

"Compute Plane"

- 300mm wafer Fan Out interposer
  InFO_SoW: 6 Layers, 3@ 5/5, 3 @ 15/20 L/S
- Local power delivery, no PCB
  InFO_SoIS: 4+RDLs @ > 8/10 L/S

More Water Cooling
Power Supply Board
HSIO Connectors

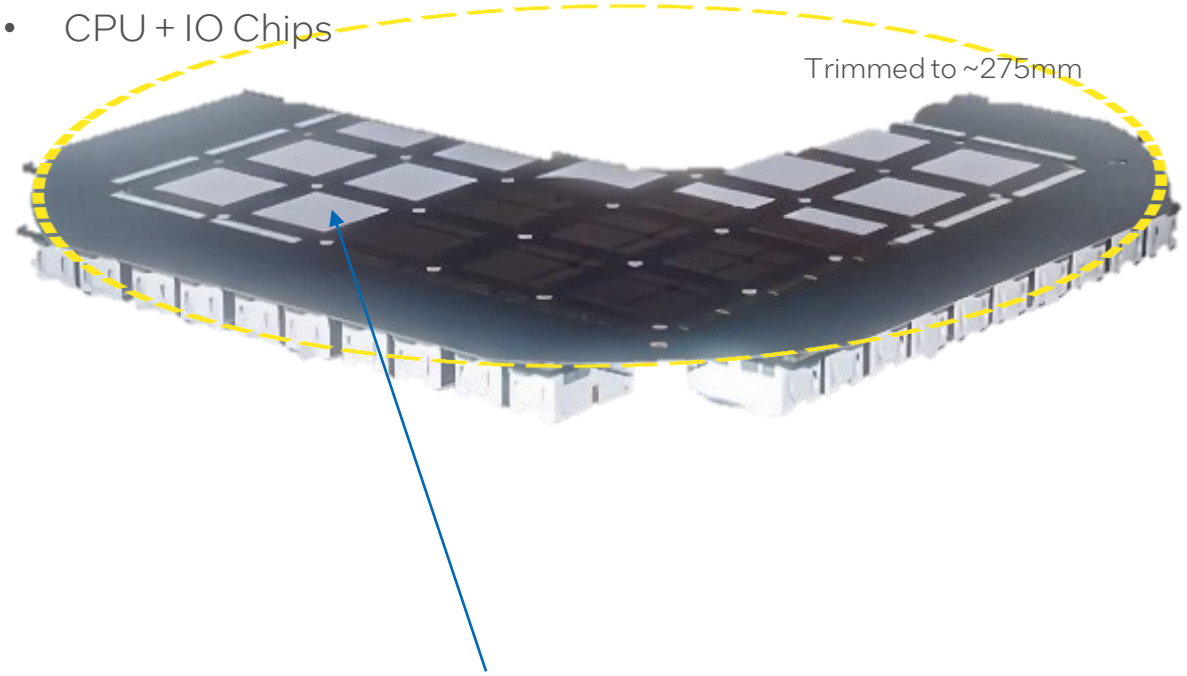
Bottom View



Dojo InFO-SOW Compute Complex

Dojo Compute Plane:

- Reconstituted with InFO
- Known Good Die
- CPU + IO Chips



25 D1 chips, 27x27mm CPUs, 362 teraflops/tile

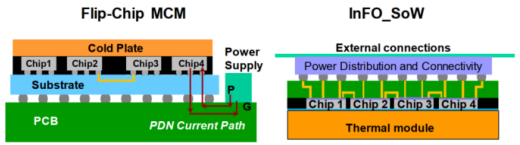
40 ~4x16mm IO tiles

Initial production ~2000 wafers

--China Times, 2020

InFO_SoW (System-on-Wafer) for High Performance Computing

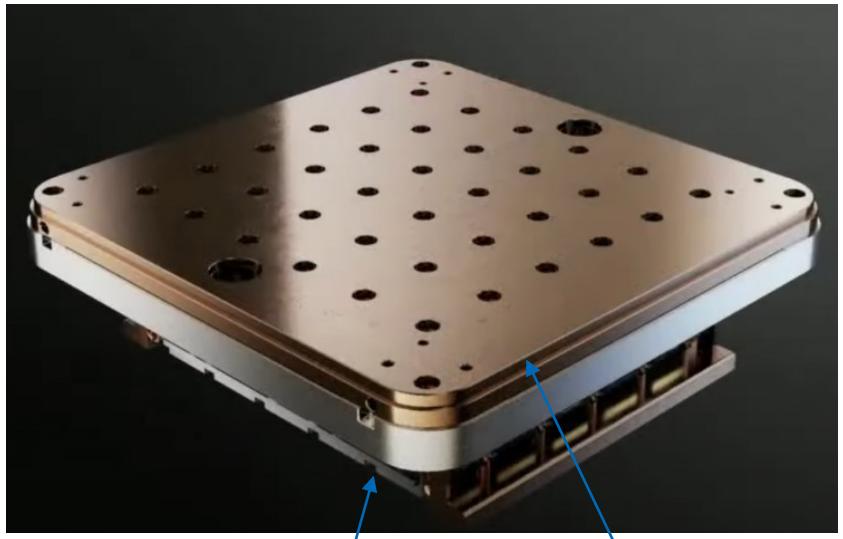
Shu-Rong Chun, Tin-Hao Kuo, Hao-Yi Tsai, Chung-Shi Liu, Chuei-Tang Wang, Jeng-Shien Hsieh, Tsung-Shu Lin, Terry Ku, Douglas Yu
Research and Development
Taiwan Semiconductor Manufacturing Company
Hsinchu, Taiwan, R. O. C
srchun@tsmc.com



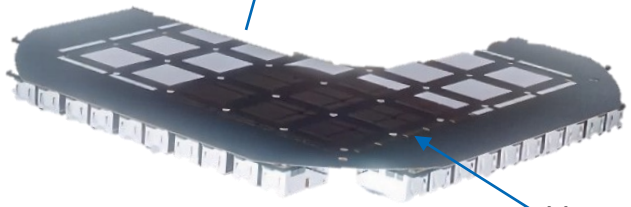
	Flip-Chip MCM	InFO_SoW
Line width / space (μm)	10 / 10	5 / 5
Line density	1x	2x
Bandwidth density	1x	2x
PDN impedance	1x	0.03x

- Wafer-scale InFO demo
  - No PCB - Connectors and power modules are soldered to InFO wafer followed by assembly of thermal module.
  - 6 RDL: 3 - 5/5, 3 - 15/20um L/S
  - Chip-first
  - 60% lower ELK stress vs flip chip, due to thick compliant RDL
  - 97% lower PDN impedance vs Flip-Chip MCM
  - 15% interconnect power savings due to lower Cu RDL surface roughness vs Substrate or PCB. 0.7dB/30mm

InFO-SoW Thermal Solution



Water cooled "cold plate"

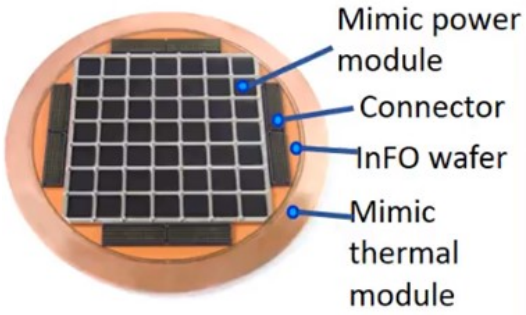


Mounting through holes to control BLT
"Localized modulation"

InFO_SoW (System-on-Wafer) for High Performance Computing

Shu-Rong Chun, Tin-Hao Kuo, Hao-Yi Tsai, Chung-Shi Liu, Chuei-Tang Wang, Jeng-Shien Hsieh, Tsung-Shu Lin, Terry Ku, Douglas Yu
Research and Development
Taiwan Semiconductor Manufacturing Company
Hsinchu, Taiwan, R. O. C
srchun@tsmc.com

- 7000 W (1.2 W/mm^2) Thermal Solution
  - 2x5 array heater and cooling system
  - Localized modulation for HS/SoC contact
  - 4 LPM DI water 16C inlet, 90C outlet



Closer look at "Compute Plane" InFO_SoW

Outer edges of 300mm dia. interposer trimmed to ~200mm square

25 CPU chiplets 27x27mm

Exposed die molded Interposer, est. 300mm dia

40 I/O (SerDes?) chiplets, 4x16mm

I/O connectors sub-board

Mounting through holes

SerDes Twinax or Fiber Optic connectors ~18x30x9 mm

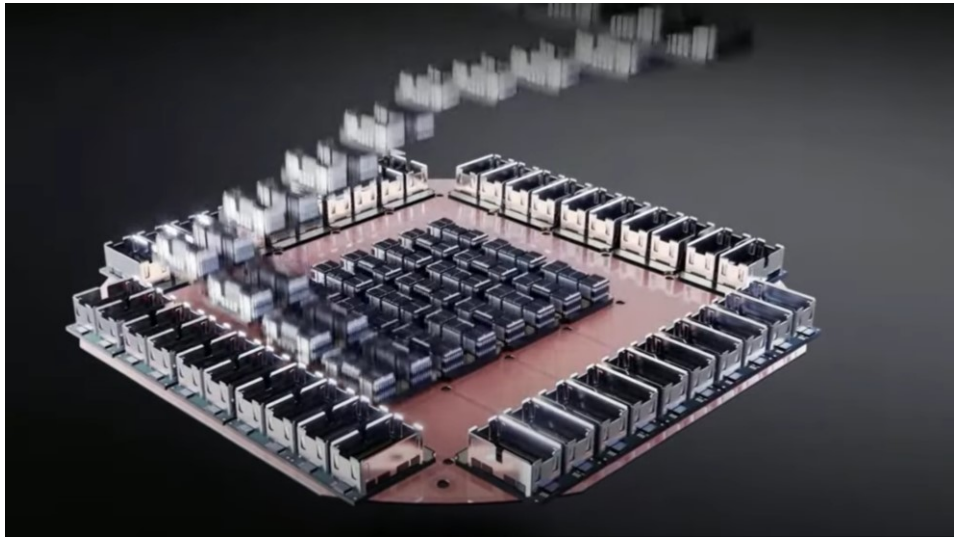
Each chiplet has a different size/shape for underlying "die" indicating this is a reflection of pedestals on the bottom of the cold plate.

Full-wafer InFO-SoW fan out interposer.
Each CPU chiplet is ~725mm^2 (27x27mm)

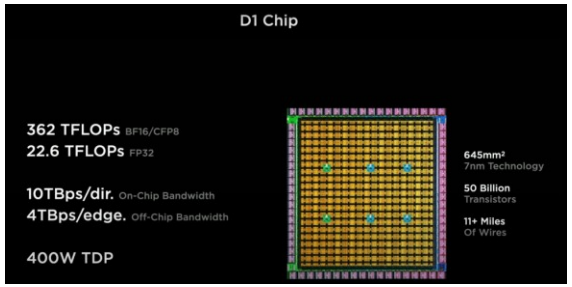
Connectors and power modules are mounted on **SoIS** and SMT'd to bottom of InFO-SoW

Source: Tesla AI Day Livestream

Close-up view of InFO_SoIS modules beneath InFO_SoW



Connectors and power modules are SMT'd to bottom of InFO-SoW

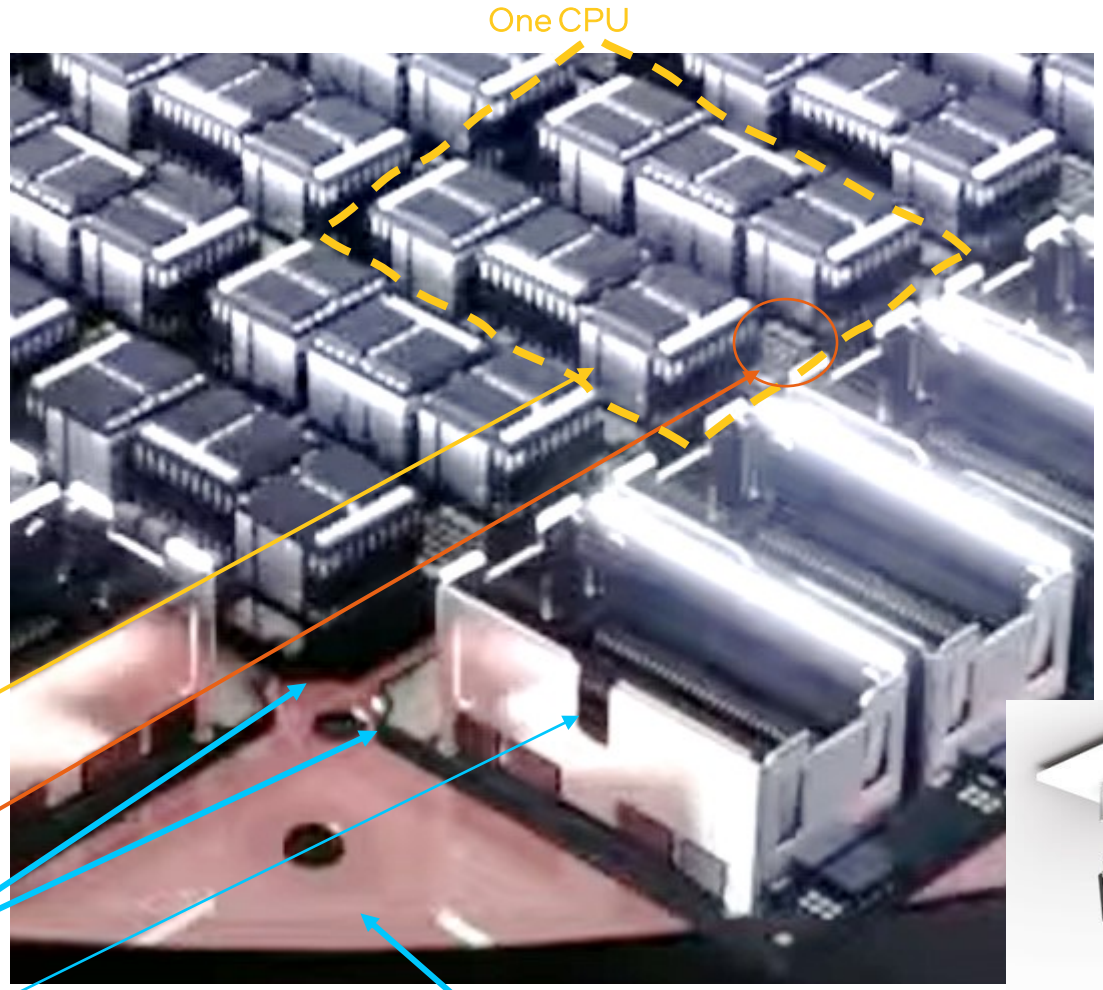


Multi-level power modules

Power connector landing pad

InFO_SoIS "boards"

PCIe connectors on SoIS boards

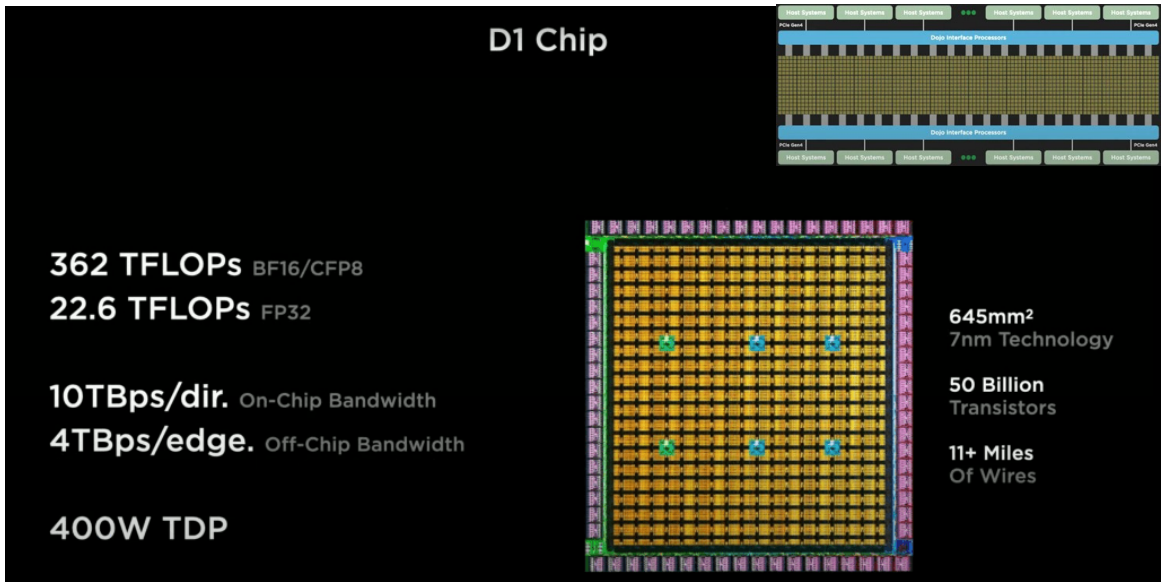


Source: Tesla AI Day Livestream

InFO_SoW

Typical MPO connector

IO capability outpacing traditional switch solutions

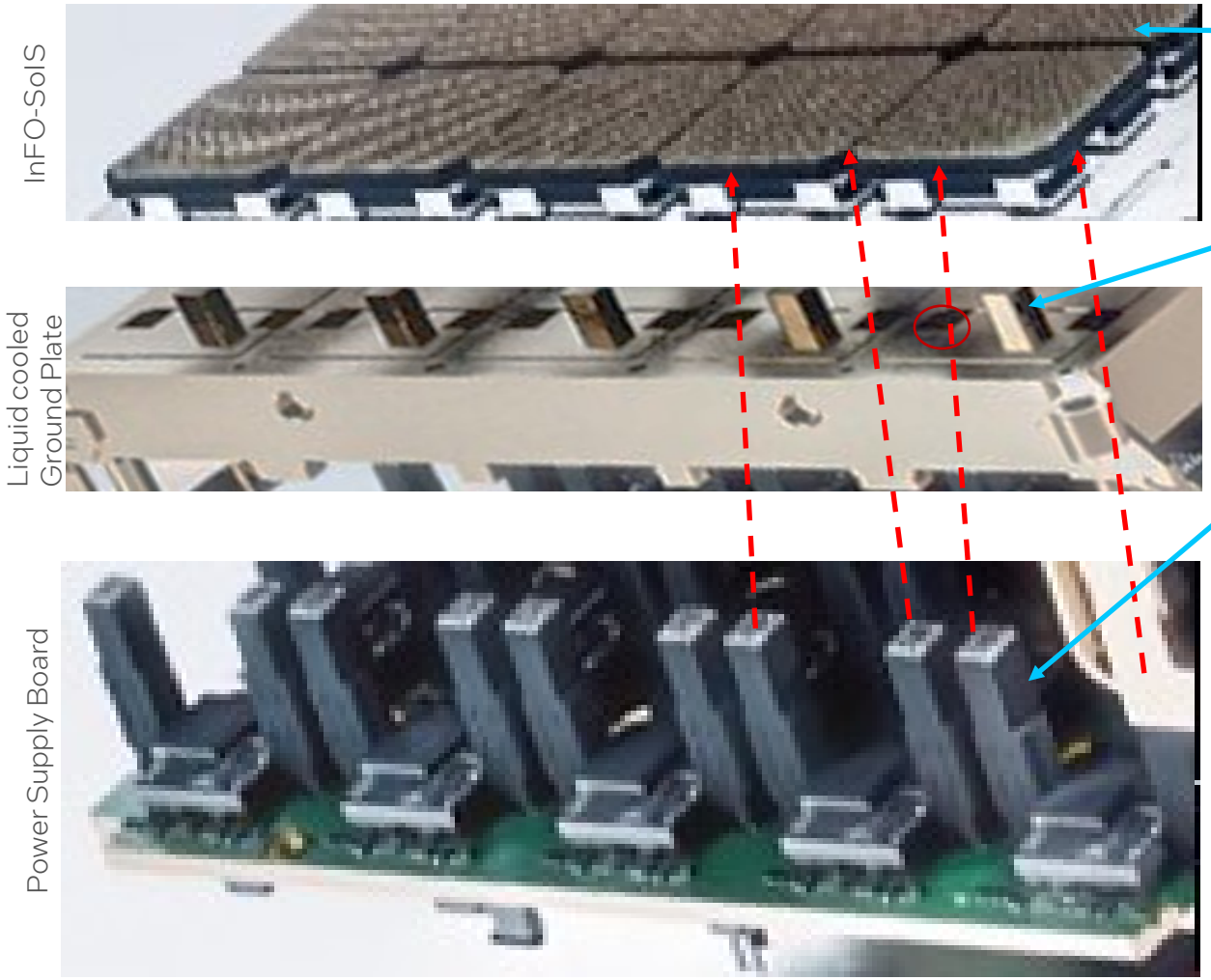


- The architecture choice require a massive interconnect to connect all the chips together and to host/memory.
- Without any external memory, each D1 chip includes 572 Serdes at 112Gb speed.
- By using a Silicon substrate, TSMC claims 2x more bandwidth, 15% less power
- *Note: Tesla has used interchangeably Gb or GB in slides creating confusions…*

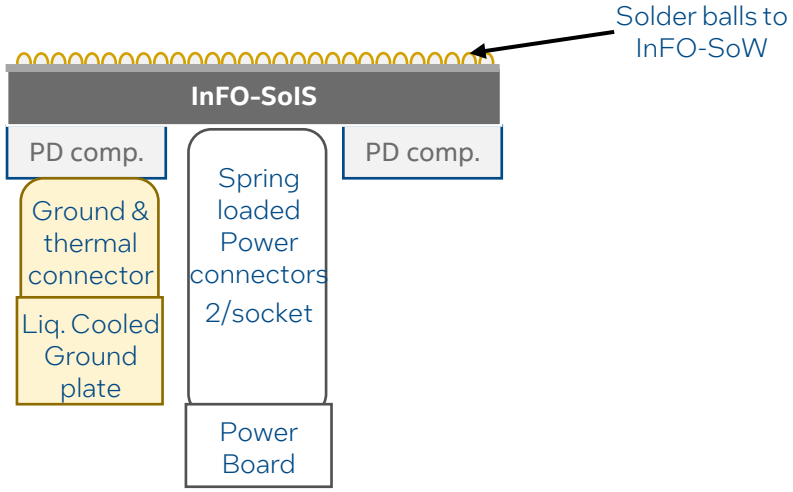
	Nber of Serdes	SerDes Speed	Total BW	Power (TDP)	Die details
Tesla Dojo D1 Chip	572	112G (SR?)	64Tb/s 8TB/s	400W	645mm2, 7nm TSMC 50B transistors
Broadcom Tomahawk 4-50G (Shipping)	512	51.6G (LR)	25Tb/s	350W	7nm TSMC 32B transistors
Tomahawk 4-100G (2021)	256	100G (LR)	28Tb/s	400W	
Innovium Teralynx 8	256	112G (LR)	25.6Tb/s		
Tofino 2	260	56G (112G with TF3)	12.9Tb/s	510W	7nm TSMC, up to 4 28nm SerDes tiles

Switch products

Closer look at Dojo PD & mechanical assembly details

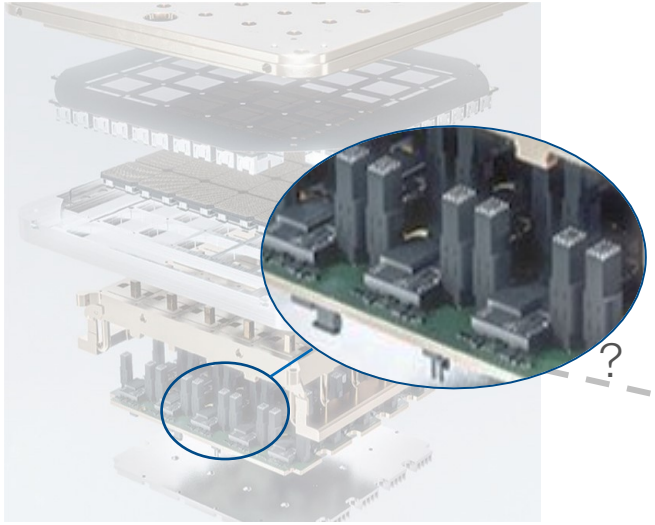


- "InFO_SoIS "Substrates" preassembled with stacks of components (PD, caps? Memory?) ready to solder onto InFO-SoW CPUs.
- Connector stud from copper plate for ground & thermals. Tubes beneath plate also cool power supply.
- Power delivered from bottom board via 2x Vicor-like spring loaded connecters landing on each side of the socket after passing thru ground/cooling plate.



Vertical Power Delivery and Control

- VICOR claims that the PDN resistance was reduced by 10X (vs lateral power delivery)
- Intel AIPG's 2019 assessment: space and thermal challenges prevented implementation for AI chip in OCP FF
- Dojo is likely using a customized PD solution with water-cooling
- "Control" path from bottom board to compute plane is undetermined



Top (Compute) Side

?

Performance loss analysis

	Vicor Vertical	Vicor Lateral	Conventional
PDN resistance	5μΩ	50μΩ	400μΩ
PDN loss @ 500 Amps	1.25W loss 99.7% efficiency	12.5W loss 96.8% efficiency	100W 75% efficiency
PDN loss @ 1000 Amps	5W loss 99.4% efficiency	50W loss 93.75% efficiency	400W 50% efficiency

PDN Power Loss, due to circuit board copper resistance = I^2R

VICOR ©2019 Vicor

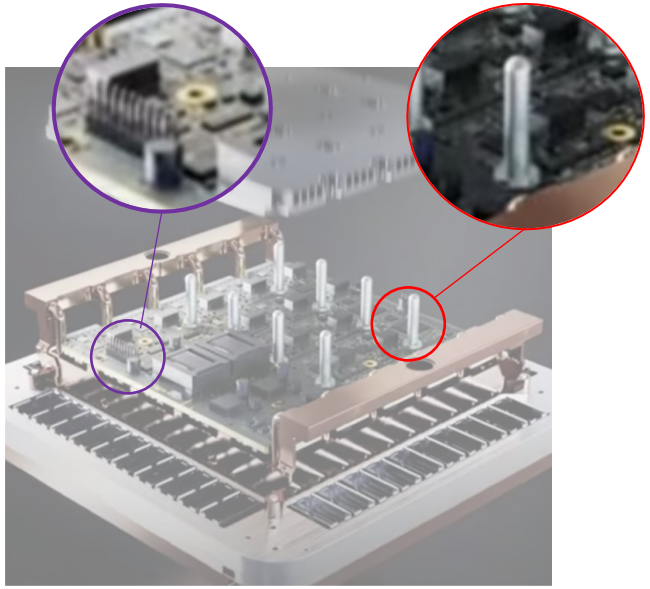
Monolithic

Vicor

Example modules

"Control"
12 pin Ribbon connector

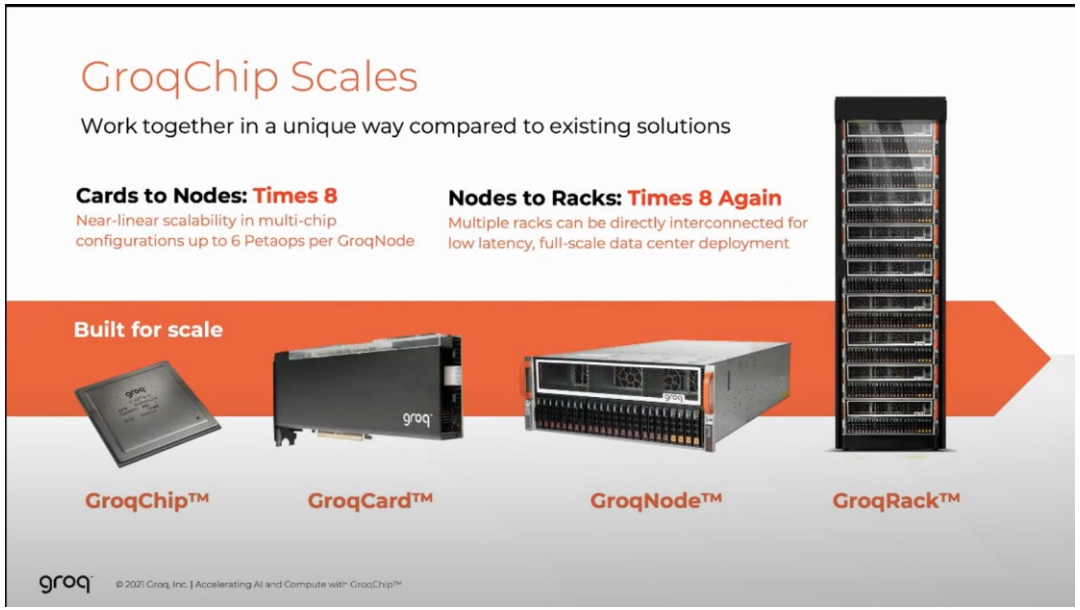
8 Power Lugs



Bottom Side: 52V in

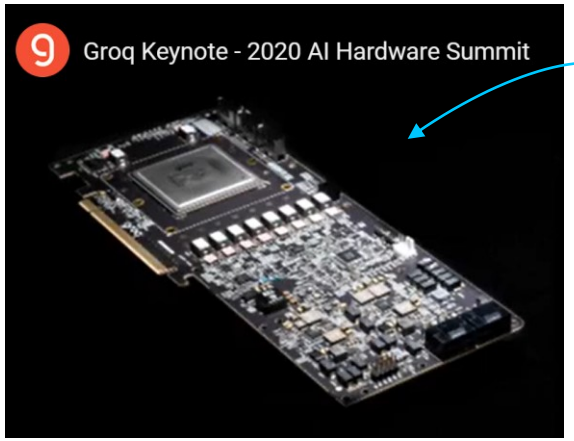
What about Groq?

No packaging details available



2019:

2020:



Same AI xpu p/n on both boards

The Future Looks Bright: Today we're shipping to our customers
Both as individual PCIe cards and Systems with 8 cards.
And there's even more on the roadmap to come!

groq

Dojo Assembly Deep Dive

InFO_SoIS - Complex Substrate

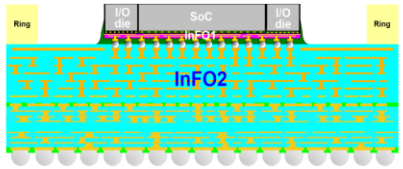
2021 IEEE 71st Electronic Components and Technology Conference (ECTC)

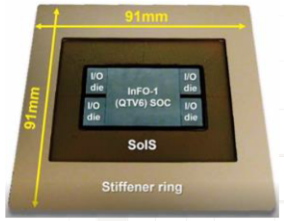
SoIS- An Ultra Large Size Integrated Substrate Technology Platform for HPC Applications

13

SoIS (System on Integrated Substrate)

- Leverage InFO to build organic substrate for FC/InFO/CoWoS stacking with KGDs- chip, passives, components, PKGs and supporting substrate
- Achieve high yield- > 95% (91mmSQ), 100% yield 110mmSQ





2021 IEEE 71th Electronic Components and Technology Conference 2021 Virtual Conference

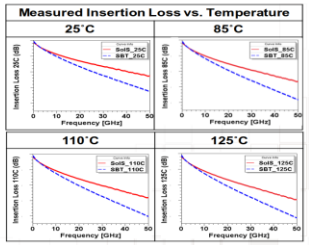
© 2021 TSMC, Ltd 13 TSMC

14

SoIS Interconnect Performance

- SoIS exhibits ~25% and ~30% lower insertion loss than organic substrate over 25˚C to 125˚C at 28GHz and 50GHz, respectively.
- The varied temperature conditions (over 25˚C to 125˚C) will not impact the measured insertion loss.

Insertion Loss vs Temperature		Diff Impedance= 90Ω			
		25˚C	85˚C	110˚C	125˚C
Insertion Loss @ 28 GHz	SoIS	~ 0.75x	~ 0.75x	~ 0.75x	~ 0.76x
	SBT (GL102)	1.0x	1.0x	1.0x	1.0x
Insertion Loss @ 50 GHz	SoIS	~ 0.7x	~ 0.7x	~ 0.7x	~ 0.71x
	SBT (GL102)	1.0x	1.0x	1.0x	1.0x

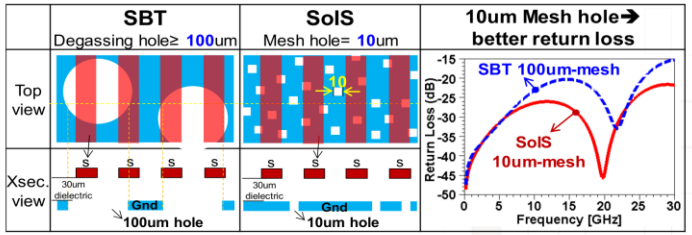


© 2021 TSMC, Ltd 14 TSMC

15

SoIS Design Rule & Power Performance

- High-density rouitng capability with finer line pitch (< 10um pitch) & via (25um CD) to gain more SerDes pairs and mitigate signal crosstalk.
- Samll mesh hole (10*10um) on P/G planes showed significantly better return loss (< - 45dB).



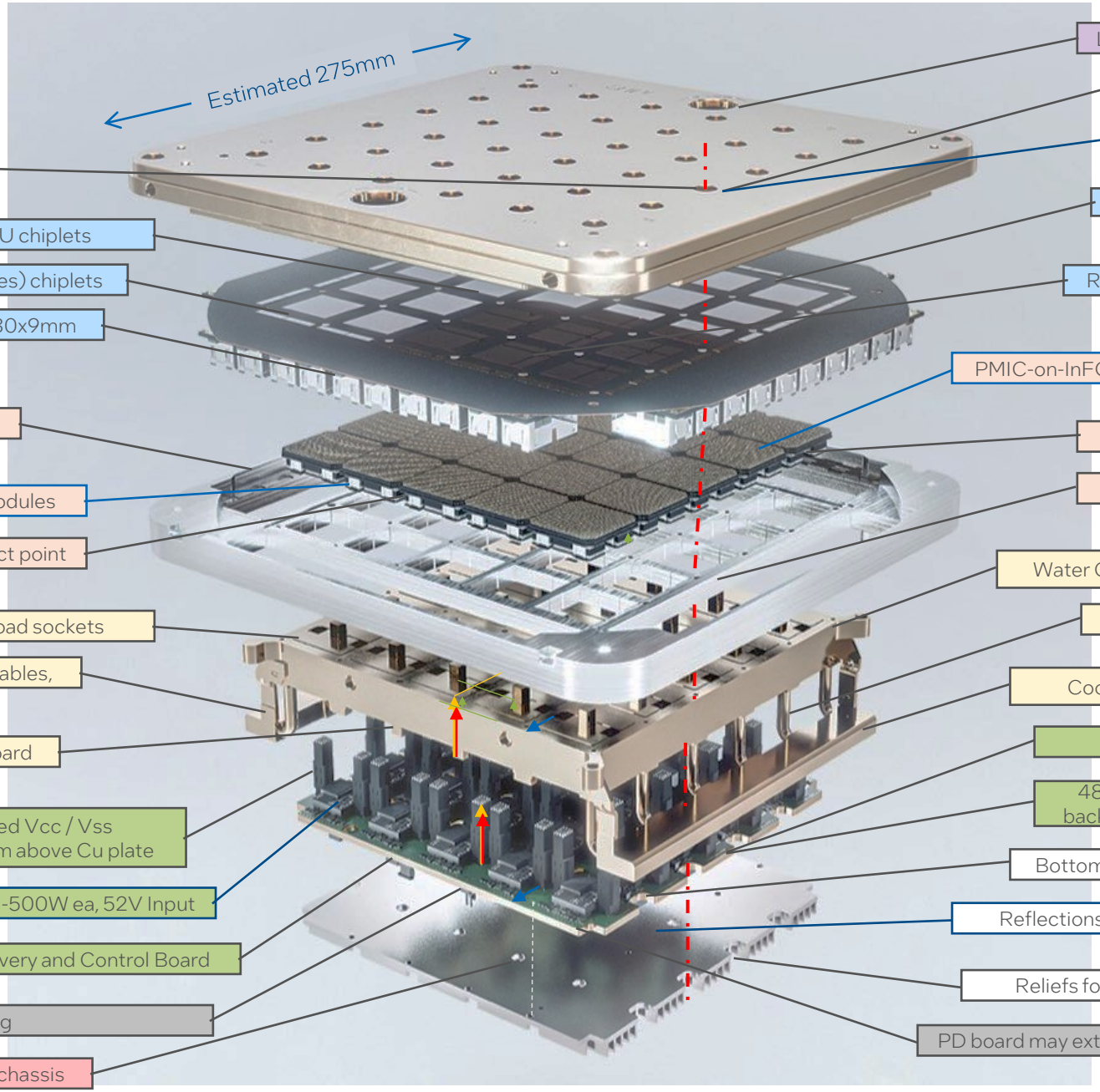
© 2021 TSMC, Ltd 15 TSMC Property

	SBT	SoIS
90Ω Differential SerDes L/S occupied area	1.0x	~1.0x for 75% IL < 0.6x for equal IL
50Ω Single-end L/S occupied area	1.0x	< 0.3x
Via Pad size	1.0x	< 0.6x

- 8 / 10 um L/S, >95% yield. "Leveraging wafer fab [InFO] process"
- No IO's going to PCB; 25-30% Improvement in Insertion and Return Losses is useful.
- Provides vertical PMIC to CPU connections to InFO_SoW on Dojo

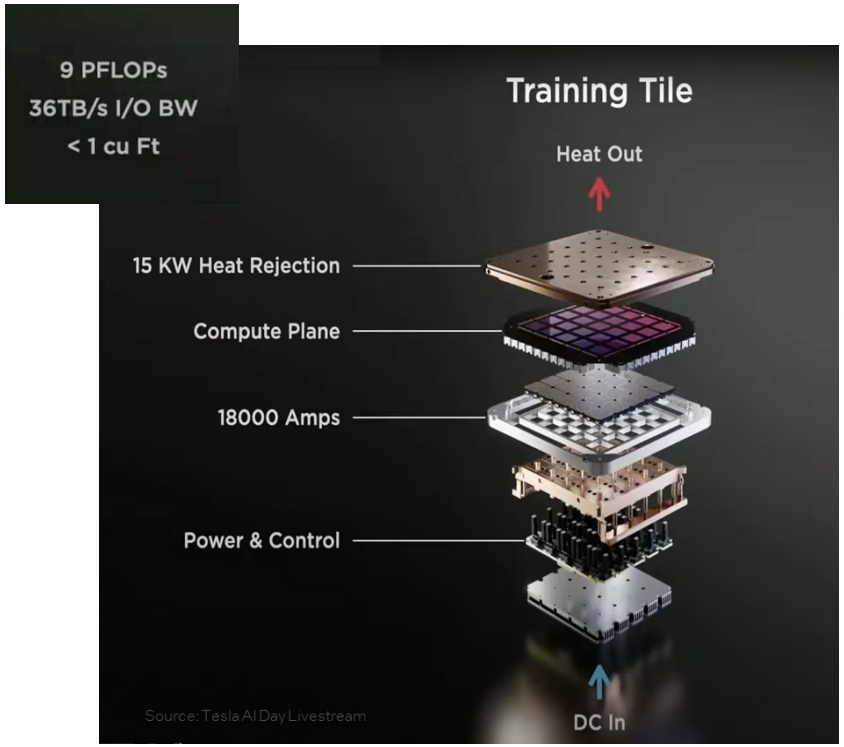
InFO-SoIS is a costly alternative to HDP for typical package-on-board applications. Dojo leveraged its advantages.

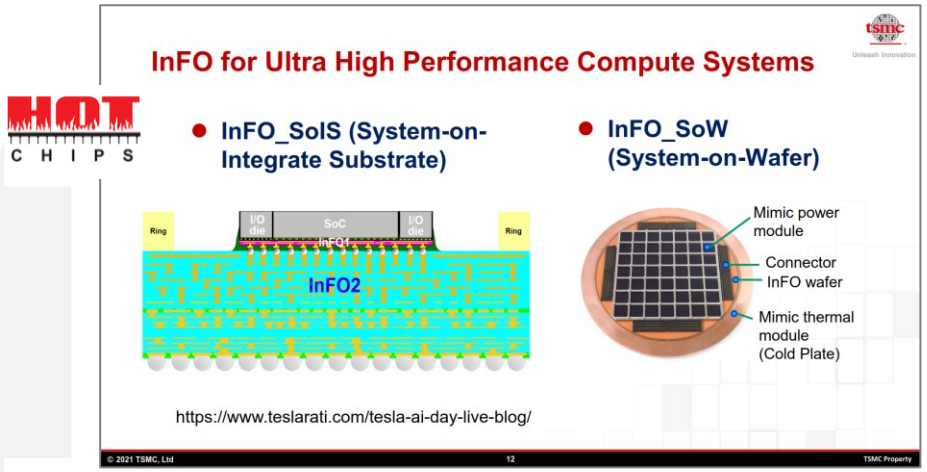
Tesla Dojo



- Countersunk holes, screws connect to chassis
- 25 27x27mm CPU chiplets
- 20 pairs 4x16mm I/O (SerDes) chiplets
- 40 I/O (SerDes? TwinAx/Fiber?) connectors 18x30x9mm
- Cable connector guide (only 1 of 4 shown)
- Inductors/caps SMT'd to bottom of PMIC modules
- Inductors/caps may be chassis/socket load contact point
- 25 Spring loaded pins to load sockets
- Chassis provides clearance for IO cables,
- Pedestals contact power delivery board
- 25 pairs 3x3x26mm Spring Loaded Vcc / Vss connectors. Connectors extend ~6mm above Cu plate
- 25 18x24mm 52V-0.7Vpower delivery modules, ~400-500W ea, 52V Input
- 2-sided Power Delivery and Control Board
- Possible 52 Volt DC Power Lug
- Holes for screws extending up to chassis
- Liquid Cooling Inlet/Outlet ~25mm dia.
- Top plate / Microchannel Heat Sink
- 15KW Heat Rejection
- Fan-Out 300mm InFO_SoW Interposer
- Reflected image of Top plate pedestals
- PMIC-on-InFO_SoIS modules reflowed to InFO-SoW.
- Power connector landing pads between ?
- Socket Frame
- Water Cooled Chassis dissipates <15% of heat
- 6 cooling tubes run beneath chassis
- Cooling manifold, connection on bottom
- Power Delivery Board
- 48-12V power delivery components on backside cooled by bottom plate heat sink
- Bottom plate / Heat Sink, dissipates ~3% of heat
- Reflections of bottom-side 52V power delivery
- Reliefs for pipe manifold tubes
- PD board may extend 5mm beyond bottom plate

Dojo architecture amplifies advanced packaging benefits.

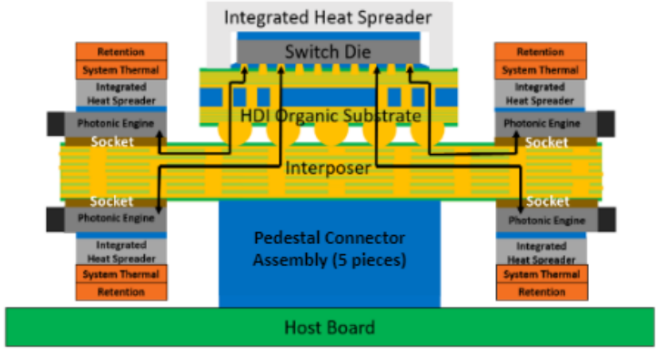




It is likely SoIC stacked-SRAM (up to 25GB/layer) compatible

- Full-wafer InFO_SoW interposer connects 25 CPUs + IOs
- SoIS "PCBs" reflowed to SoW for local Power Management
- Novel construction
- Vertical Power Delivery

Overall assembly bears similarities to Intel "Pathforward" co-packaged optical architecture from ATTD & DPG-SPPD: http://iattj.intel.com/#/article/19502

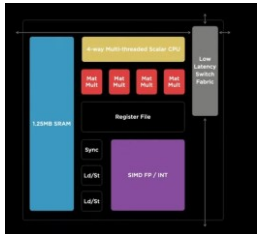


2. Dojo Architecture overview

Yoann Foucher, DCAI CSO

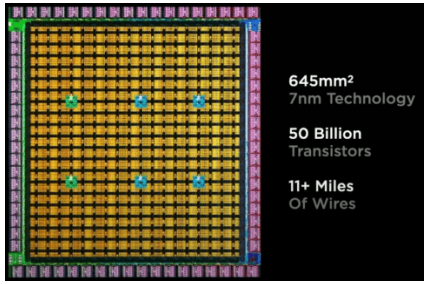
intel®

AI Dojo complete system view



Training node compute unit

354

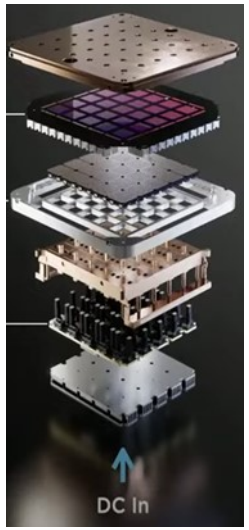


D1 Chip

362"TFLOPS"
(BFP16 or CFP8)

22.6 TFLOP FP32

25



Training tile

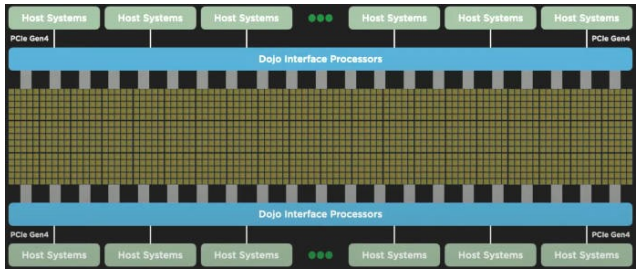
9 "PetaFlops"
(BF16/CFP8)

12

2 trays server (cabinet)

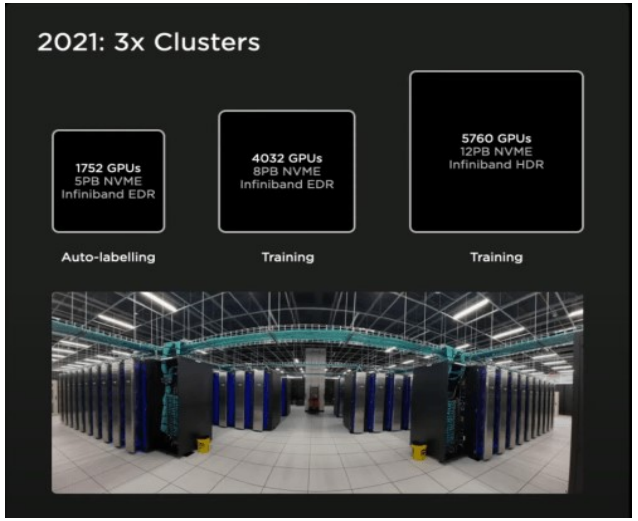
180kW and >100 "PetaFlops"

125

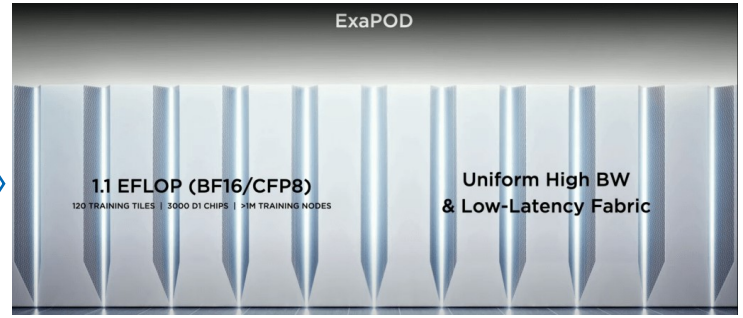


Processing plane
15 x100 array
>500k training nodes

2

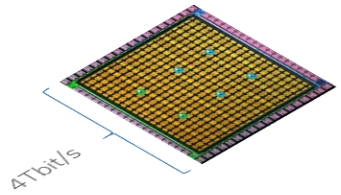


4x perf
1.3X perf/W
5x smaller footprint

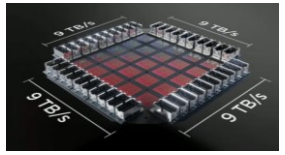


Tesla Dojo ExaPOD
10 racks x12 tiles - 3000 D1 chips
1.1 EFLOPS (BF16/CFP8)
~68 PFLOPS (FP32)

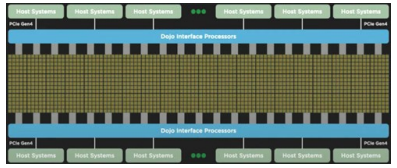
Tesla D1: Creating disruption in 3 key areas



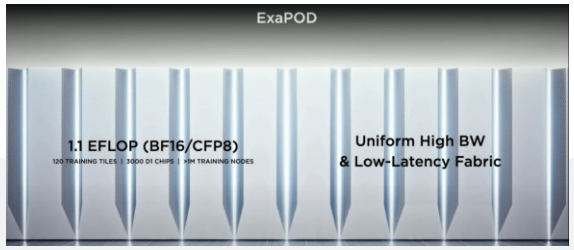
- <u>Industry-leading I/O Bandwidth</u>
  - At its edges, the D1 integrates 576 lanes of low-power serdes, delivering up to 2Tbit/s of bidirectional bandwidth per edge for a total of 16Tbits/s of chip-to-chip data.
  - 36000 Gbps (4x 9Tbps) of network BW available per tile (25 chips)



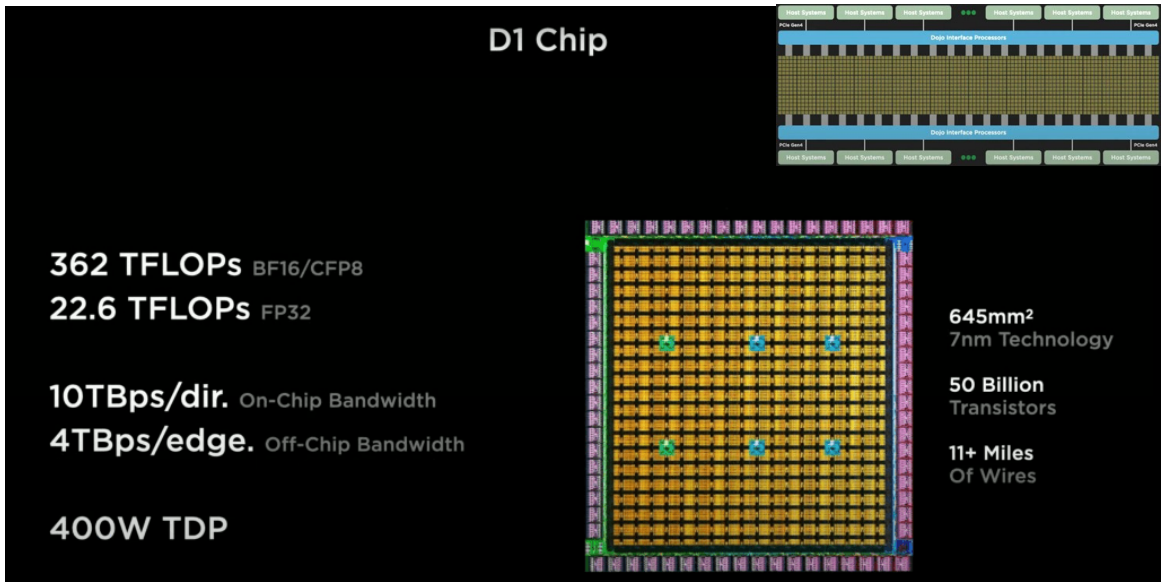
- <u>SRAM based</u>
  - Specific memory to compute ratio – with nearly 450MB of data SRAM on each chip to store parameters. Each functional unit with 1.25MB (Similar arch to Tenstorrent)
  - Host/DRAM pool available on each end of a 5 rack row.



- <u>Extreme power density achieved</u>
  - Each chip consumes 400W with vertical power delivery allowing extreme high efficiency. Reduced power on IO at system level.
  - 15KW per system (288A @52V) (10Kw for D1 chips, 5kW for voltage regulator/IO).
  - 180KW/rack ➔ 1.8MW for a 10 racks x12 tiles ExaPOD system.



IO capability outpacing traditional switch solutions

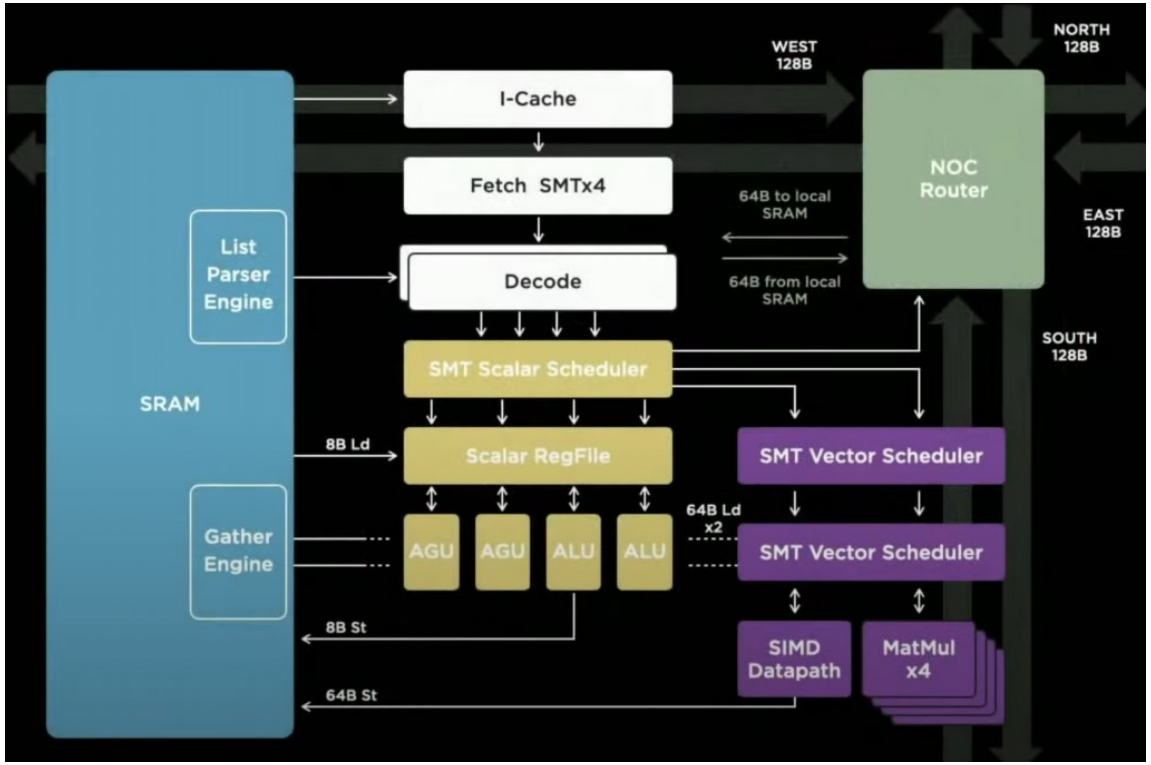


- The architecture choice require a massive interconnect to connect all the chips together and to host/memory.
- Without any external memory, each D1 chip includes 572 Serdes at 112Gb speed.
- By using a Silicon substrate, TSMC claims 2x more bandwidth, 15% less power
- *Note: Tesla has used interchangeably Gb or GB in slides creating confusions…*

	Nber of Serdes	SerDes Speed	Total BW	Power (TDP)	Die details
Tesla Dojo D1 Chip	572	112G (SR?)	64Tb/s 8TB/s	400W	645mm2, 7nm TSMC 50B transistors
Broadcom Tomahawk 4-50G (Shipping)	512	51.6G (LR)	25Tb/s	350W	7nm TSMC 32B transistors
Tomahawk 4-100G (2021)	256	100G (LR)	28Tb/s	400W	
Innovium Teralynx 8	256	112G (LR)	25.6Tb/s		
Tofino 2	260	56G (112G with TF3)	12.9Tb/s	510W	7nm TSMC, up to 4 28nm SerDes tiles

Switch products

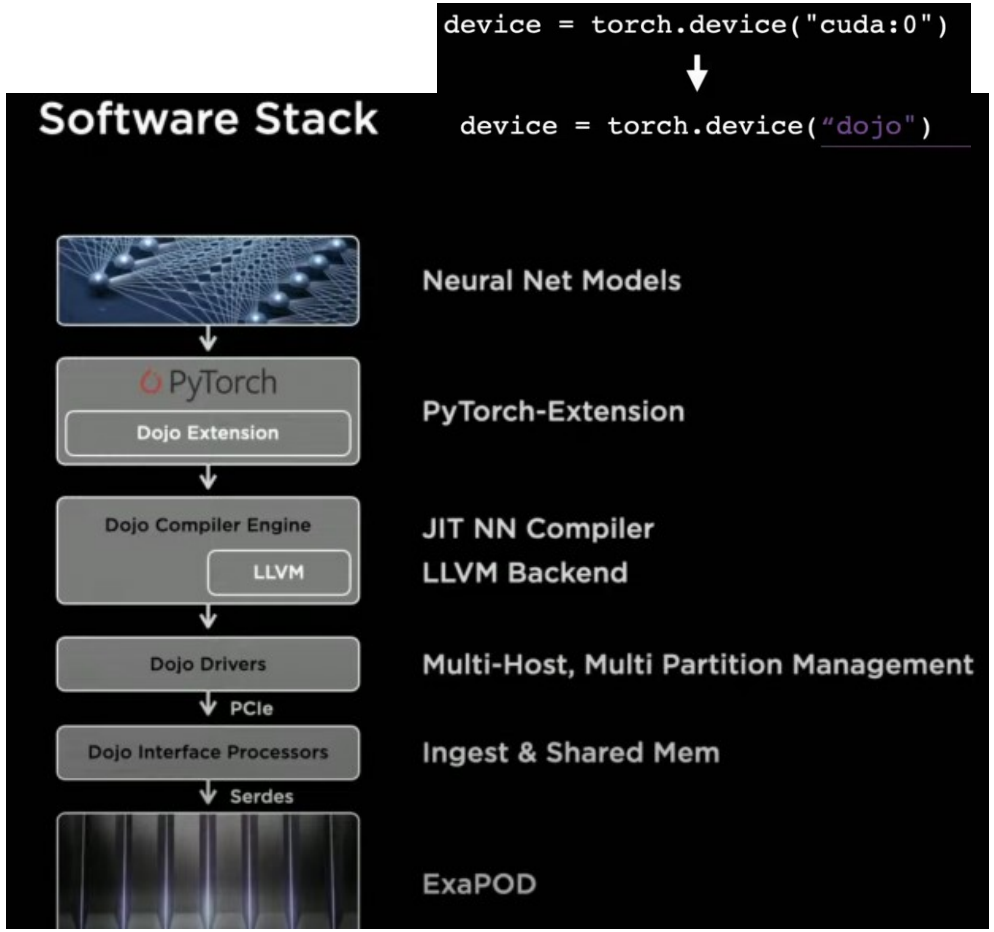
Distributed compute architecture for AI Vision training workloads

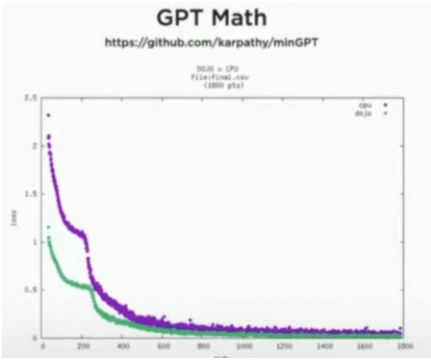


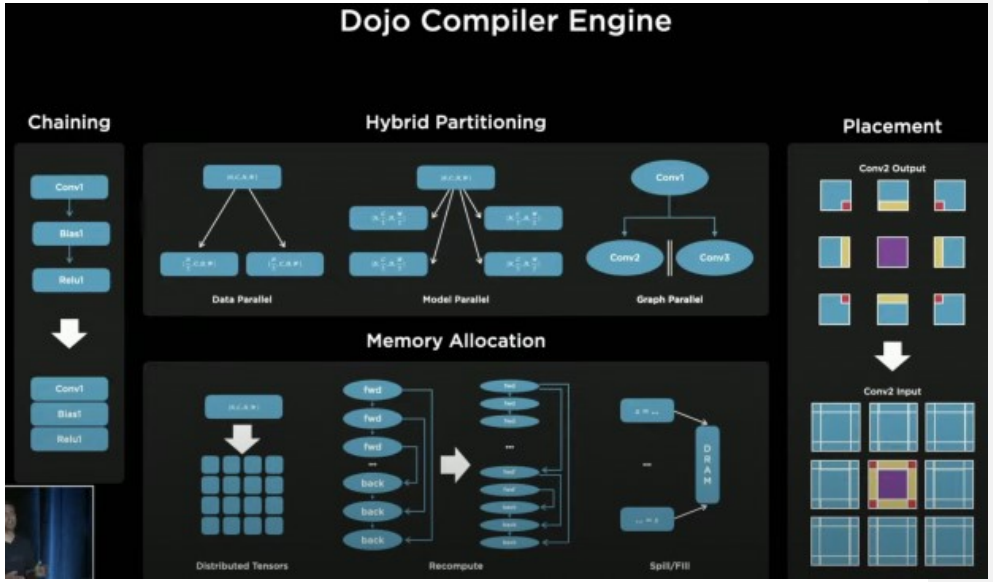
Training node architecture diagram
2GHz Superscalar in-order CPU
4-way multithreaded
Custom ISA (optimized for ML kernels)

- Compiler extracts spatial and temporal locality and maps the models onto the chip with data moving to the next compute element
  - Training node size and frequency determined by the number of cycles to access the next node (1 cycle/hop)
- Compute based on a 64b Superscalar CPU with 4 matrix multiplier units (8x8) and vector SIMD.
  - Datatypes supported are FP32, BF16, CFP8 (Configurable FP8), INT32/16/8
- 1.25MB ECC SRAM per compute node (include i-cache?)
  - 424.8 MB per chip
- Custom ISA for AI with transfer, gather, link traversals and broadcast.

Tesla SW compiler and SW stack







- Each tile can be partitioned with multiple tasks (not lower than chip level)
- Software Stack:
  - Claimed that the compiler map directly PyTorch models to the HW
  - Compiler extracts parallelism and optimizes for memory footprint so large models can be mapped (at low batch sizes)
  - Tesla has modified LLVM backend to handle ISA
  - Driver stack to support multi-host & storage/DDR.
- Status:
  - Demonstrated one algo running on one tile, claimed to be running at 2GHz. No other details in performance.
  - While Tesla can narrow its development to a specific WL, building a complete SW stack takes significant resources/time

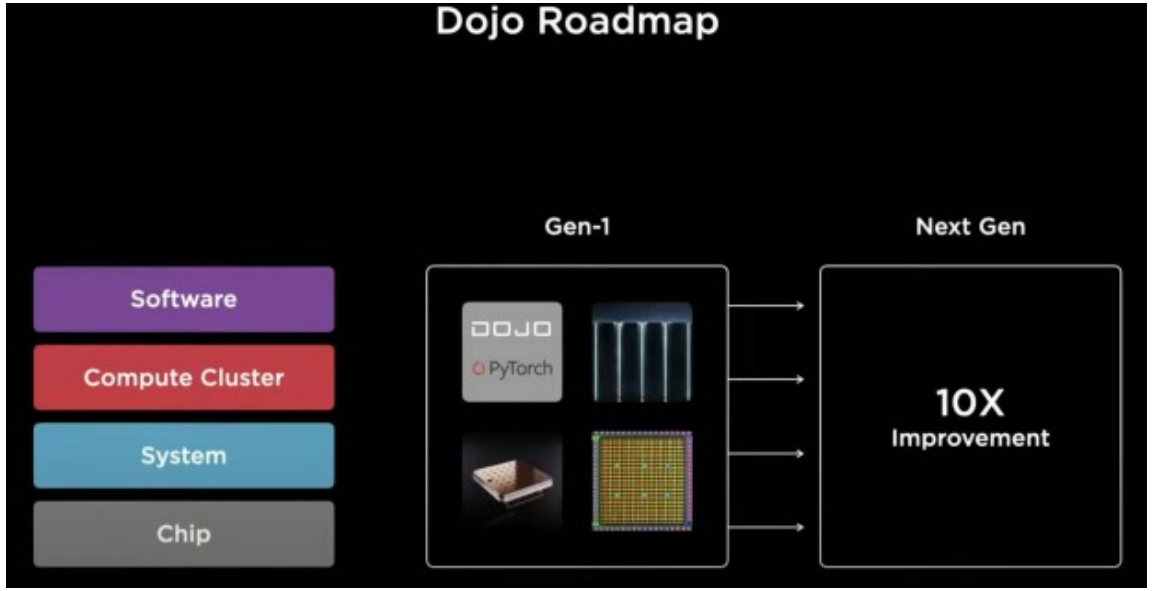
How does Tesla D1 compare to existing solutions?

- **Narrower application domain**: The system is optimized for one type of workloads – Vision
  - Within a tile, total memory is limited to 11GB. Compute/mem ratio specific
- The system is rather simplified by having few key architecture elements
- By pipelining the entire workload through the system leveraging a very high BW proprietary network, the system doesn't require external DDR memory (Host/DRAM pool at then end of the racks).
- More focus on the overall SW stack - toolchain/orchestration.
  - Partitioning at chip level
  - One framework supported, one type of WL

	Nvidia A100	Tesla D1	Tesla D1 Package
Architecture	Ampere	Custom	Custom
Core Count	108 SMs	354 cores	8,850 cores
Peak Clock Speed	1.4GHz	2.0GHz	2.0GHz
Peak BF16 Perf	312Tflop/s	362Tflop/s	9,050Tflop/s
Peak FP32 Perf	20Tflop/s	23Tflop/s	575Tflop/s
On-Chip Memory	40MB	442MB	11,050MB
DRAM Capacity	80GB HBM	None	None
DRAM Bandwidth	2TB/s	None	None
High-Speed I/O	12x NVLink v3	Serdes	Serdes
I/O Bandwidth	0.6TB/s	16TB/s	36TB/s
Die Size	826mm^2	645mm^2	25x 645mm^2
Power (TDP)	400W	400W	15,000W
IC Process	TSMC 7nm	TSMC 7nm	TSMC 7nm
Production	2Q20	2Q22*	2Q22*

Linley deep dive – Tesla D1:
https://linleygroup.com/mpr/article.php?id=12528

Timeline and What's next?



- No clear timeline – Musk indicated that "it could be next year" (estimated Q2'22)
  - Tesla has shown a picture of a tile system running – but full system integration has yet to be completed
- Focusing only on Vision AI workloads (autonomous driving)
  - By choosing a CPU approach, Tesla maintains some flexibility to train new models.
- Announced as well major recruiting effort to staff the program
  - Also developing a Tesla bot…

- Talked about offering *training as a service* in the future…
- Next generation targeting 10x more performance
  - Improvements expected with 3D stacking
  - Expecting more algo optimizations (Sparse matrix?)

Summary

- Tesla is building a custom solution for AI training
  - The overall solution has been teased first in early August then unveiled during Tesla AI day on August 19th
  - Additional details have been shared from TSMC during HotChips'21
  - Tesla has been engineering custom chips and custom systems to replace Nvidia GPU currently used to train AI models

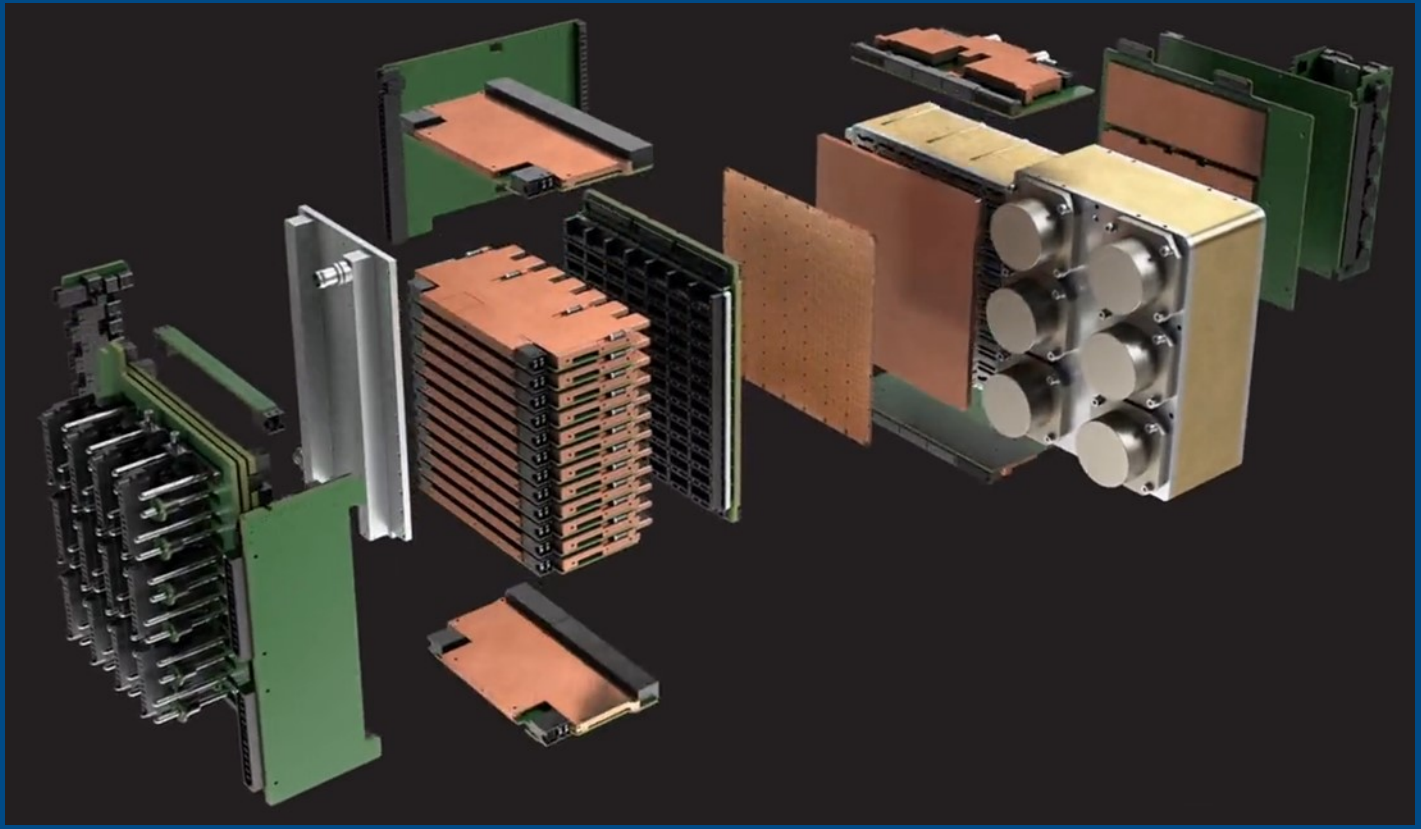
- Tesla is bringing a **disruptive solution** with:
  1. **Innovative packaging technology** build on new TSMC technologies like InFO-SoW and InFO_SoIS in addition to a custom thermal & power delivery system
  2. By breaking traditional system boundaries, Tesla introduces a massive and **unique system architecture** customized for AI Vision training

Speculative system assessment of

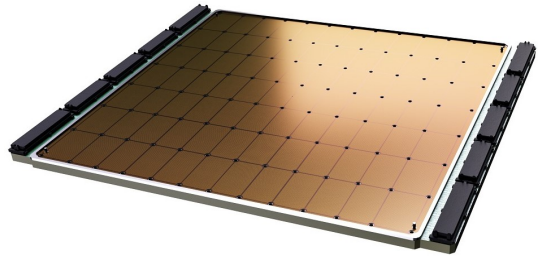
Cerebras' wafer scale system

Initial analysis, require additional inputs from different product teams

Avishaii Abuhatzera, WW02 2022



Cerebra system

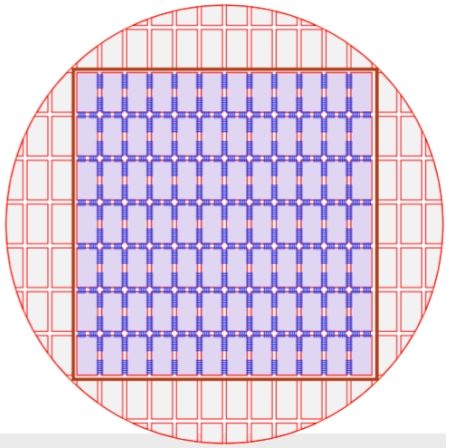


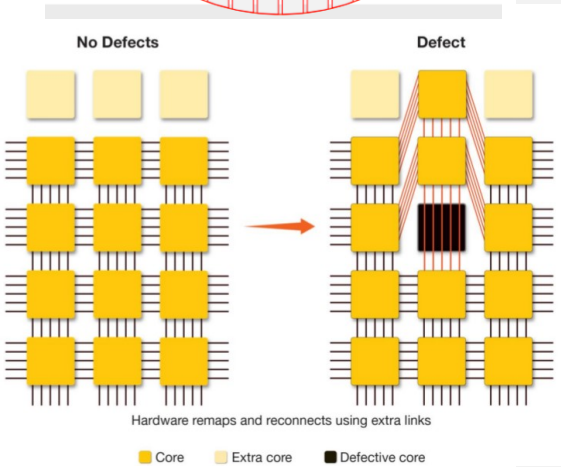
- >$4B valuation (2H-21), $250M funding.
- 2nd gen of wafer scale design (area 46,225mm²)
- Uniform small core architecture each with 9 threads.
  - 7x12=85 “cores” each 525mm²
  - Array of 800x1060 identical tiles
- Wafer (core, fabric links) and system (power, fans and pumps) redundancy
- Main customer: Argone Nat Lab
- ~300 people (mostly SW)
- System size 15RU , 23kW.
- SW Beta release only in mid-October, 2021

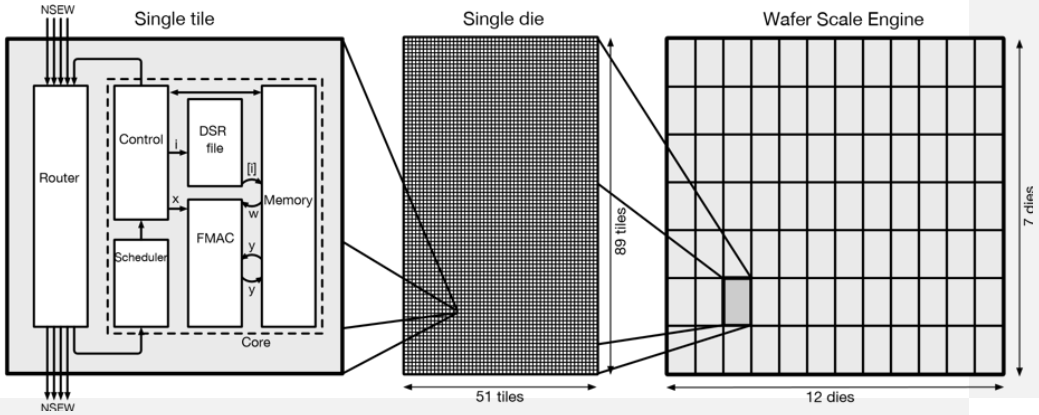
	WSE	WSE-2	per tile estimate
IC process	16nm	7nm	7nm
Production date	4Q-19	3Q-21	3Q-21
Core count	400,000	850,000	1
On-Chip SRAM	18 GB	40 GB	48 kB
Memory bandwidth	9 PB/s	20 PB/s	23GB/s (24B per clock)
Fabric bandwidth	12.5 PB/s	27.5 PB/s	32 GB/s (~8G per direction)
Transistor count	1.2 Trillion	2.6 Trillion	~3.1 M
Silicon area	46,225 mm²	46,225 mm²	~0.05 mm²
Power consumption	~17-20 kW	23 kW	~20-23 mw
Cross wafers BW		12x100Gbe	
Ops (FP16) est.	~3.1 Pflops	~6.8 Pflops	8 GFlops

Architecture

- Uniform small core architecture each with 9 threads.
  - 7x12=85 “cores” each 525mm²
  - Array of 800x1060 identical tiles
- Architecture redundancy for cores and fabric
- Router: 24 virtual channels w/ HW queues – configure in compilation time
- Extend 3D mesh cross dies (BW) enable low power (>0.15pj/b communication)
- External connection in the chip boundary
- SIMD : upto 4 for 16b operation
- Memory : 24B/cycle







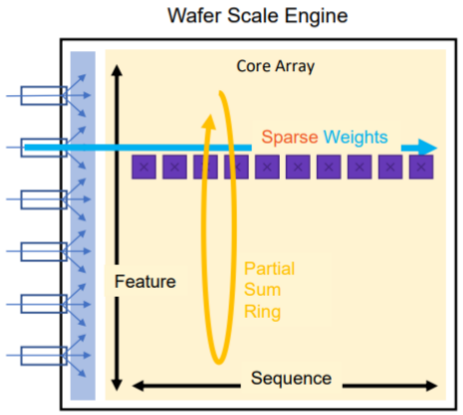
Architecture: sparsity

- Design for sparsity.
  - Each tile “remove” zeros before transfer, receive tile skip zero computation.
- Fine grain execution
  - Each core is independent.
  - Non-uniform work
- Dataflow scheduling in Hardware.
  - Data and control receive from fabric

Sparse MatMul Kernel

The Wafer is the Matrix-Multiply array

- **High-capacity local memory** stores all activations across compute fabric
- **Large compute core array** receives sparse weight stream and triggers multiplies with activations
- **Massive memory bandwidth** enables full performance of operands to the datapath
- **High BW interconnect** enables partial sum accumulation across wafer at full performance
- **No matrix blocking** or partitioning required

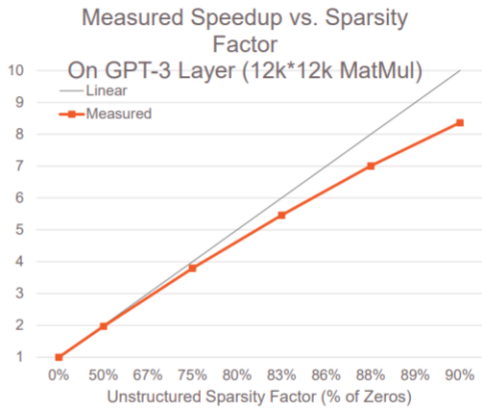


Same flow supports dense and sparse

Demonstrated Unstructured Sparsity Speedup

Sparsity reduces time-to-accuracy

- WSE runs AXPY at full performance
- Limited only by low fixed overheads
  - Minimized by high bandwidth interconnect
  - Reduced as networks grow larger
- Accelerates all unstructured sparsity
  - Fully dynamic and fine-grained
  - Even fully random patterns



Near-linear sparsity acceleration

Architecture: Memory issue

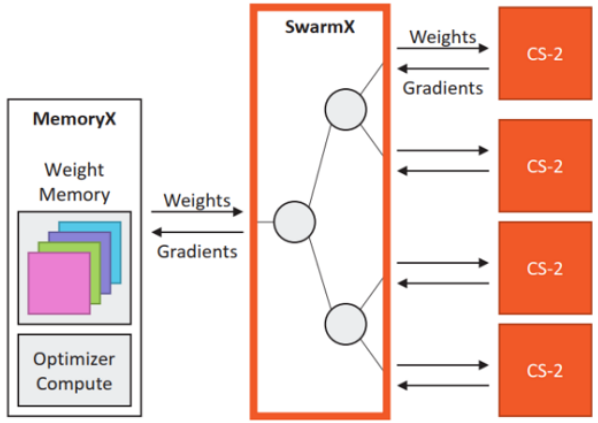
- The memory size on the WSE restricts the batch size and/or model size that can be used.
- Inputs to WSE (memory/data) enter from the edges. This makes connecting the WSE to an external memory even more unreasonable.
- Had to develop Memory/storage appliance to overcome memory ineefciency

Cerebras' MemoryX appliance

- Mix of DRAM and flash storage
- Scales from 4-TB to 2.4-PB

Weight streaming approach

- Scales from 4-TB to 2.4-PB

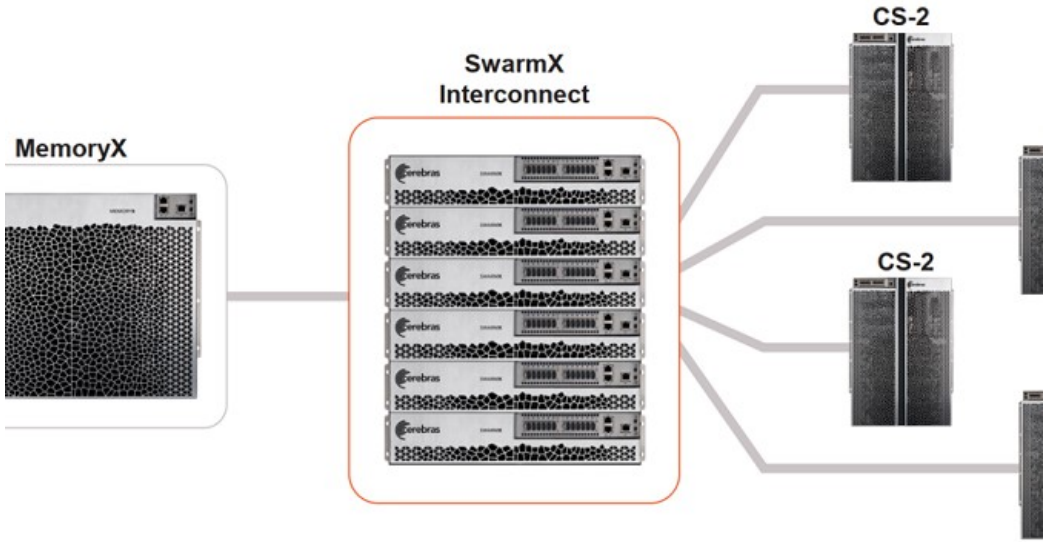


- Data parallel training across CS-2s
- Weights are **broadcasted** to all CS-2s
- Gradients are **reduced** on way back
- **Multi-system scaling with the same execution model as single system**
  - Same system architecture
  - Same network execution flow
  - Same software user interface

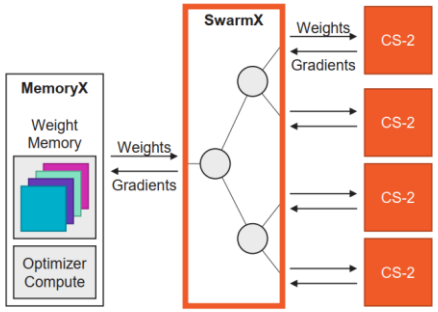
Scalable to extreme model sizes

Compute scaling independent from capacity

System



SwarmX Fabric Connects Multiple CS-2s



- Data parallel training across CS-2s
- Weights are **broadcast** to all CS-2s
- Gradients are **reduced** on way back
- **Multi-system scaling with the same execution model as single system**
  - Same system architecture
  - Same network execution flow
  - Same software user interface

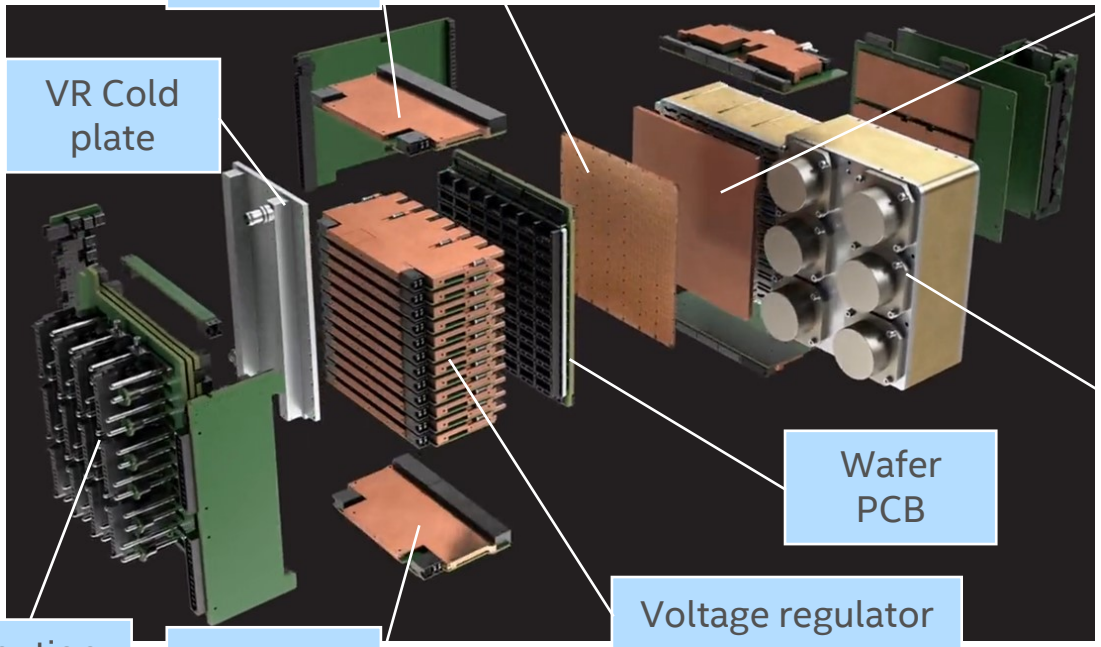
Scalable to extreme model sizes

Compute scaling independent from capacity

- can build clusters with up to 192 CS-2 systems, comprising an aggregate 163 million cores
- Key motivation, large data set, small models

Cerebras Engine block

For holding the assembly together while maintaining electrical contacts for power and IO, Cerebras "sandwiched" the wafer in an assembly comprising a thick PCB, a flexible membrane, the WSE and its heat exchanger. An array of clamping fasteners distribute the packaging force evenly across the assembly while still allowing the wafer to expand and contract.



Cold plate

Copper heat exchanger design to >20kw of heat from wafer. Wafer is "floating" on the cold plate, allowing it to expend and contract with temperature while remining in contact with the plate

Cooling manifold

Transfer heat between cerebra's internal coolant and the facility watter system. (separated from data center liquid system)
Uper 4 connector – connected to the water pump , the lower two used for heat exchange.

Power supply connector

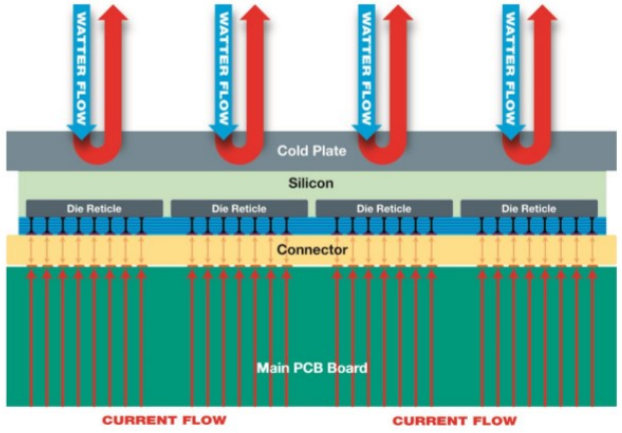
Liquid cooled voltage regulators delivering 20KA at less then 1v vertically

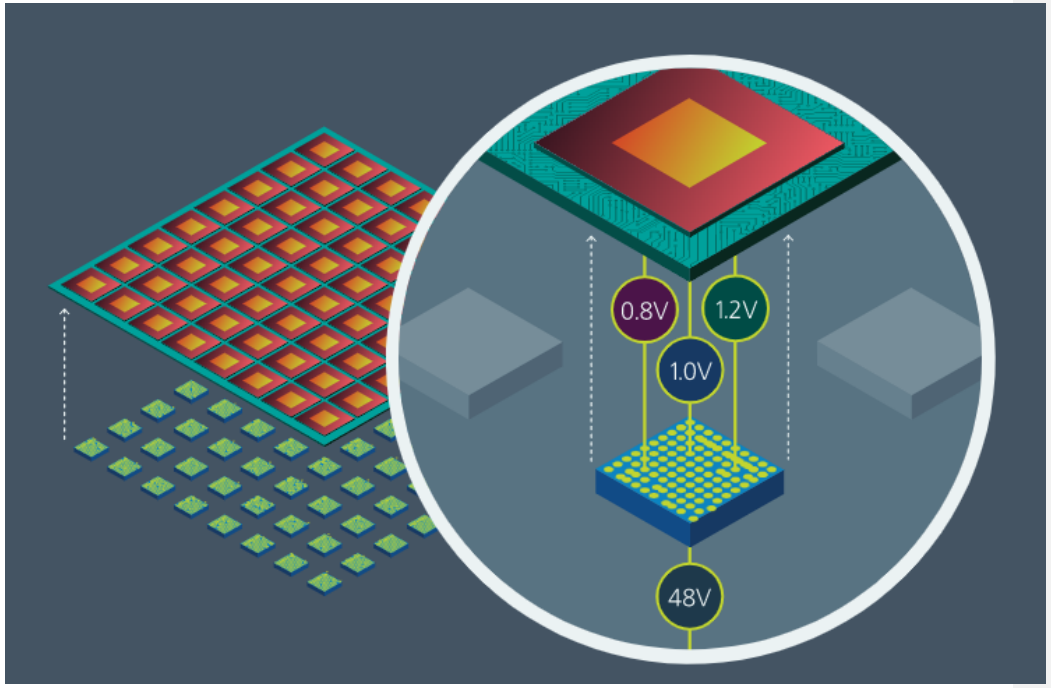
Cerebras Engine block

Using the 3rd Dimension

- Power delivery
  - Current flow distributed in 3rd dimension perpendicular to wafer
- Heat removal
  - Water carries heat from wafer through cold plate

Co-designed with system

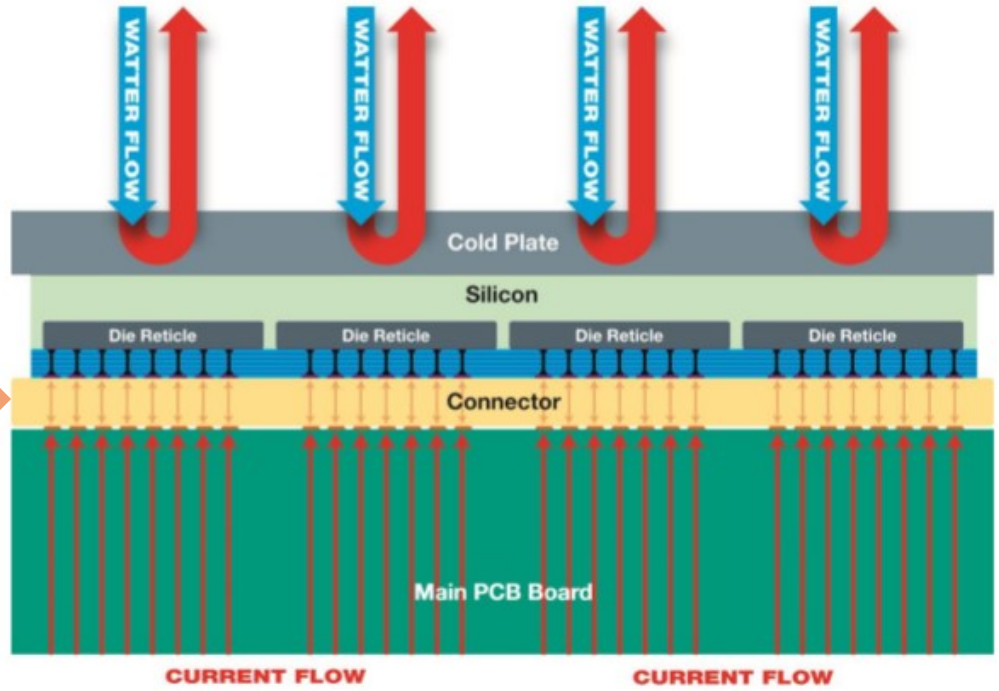




The power-delivery system sits above the chip with a watercooled cold plate below.
Cerebras implemented Vicor's Vertical Power Delivery (VPD) architecture that reduces power delivery network (PDN) resistance by more than 50%, thereby achieving higher overall density and power system efficiency.
At the macro level using 48V power supply which reduce datacenter energy losses by over 30%, with >98% peak efficiency

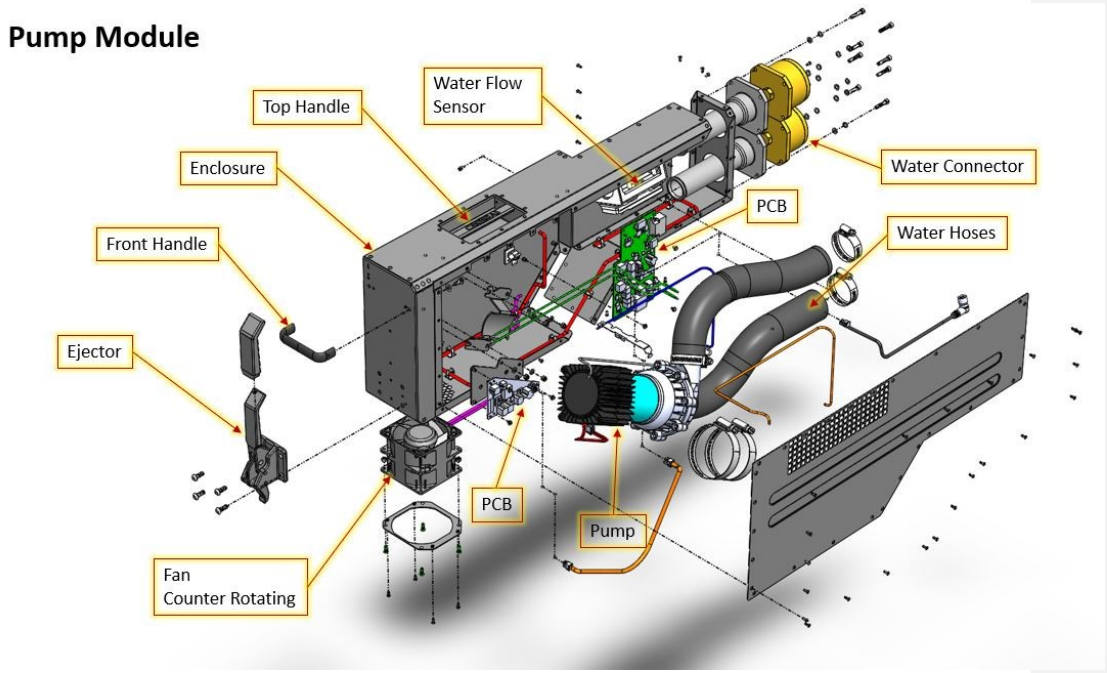
Engine block: connecting wafer to PCB

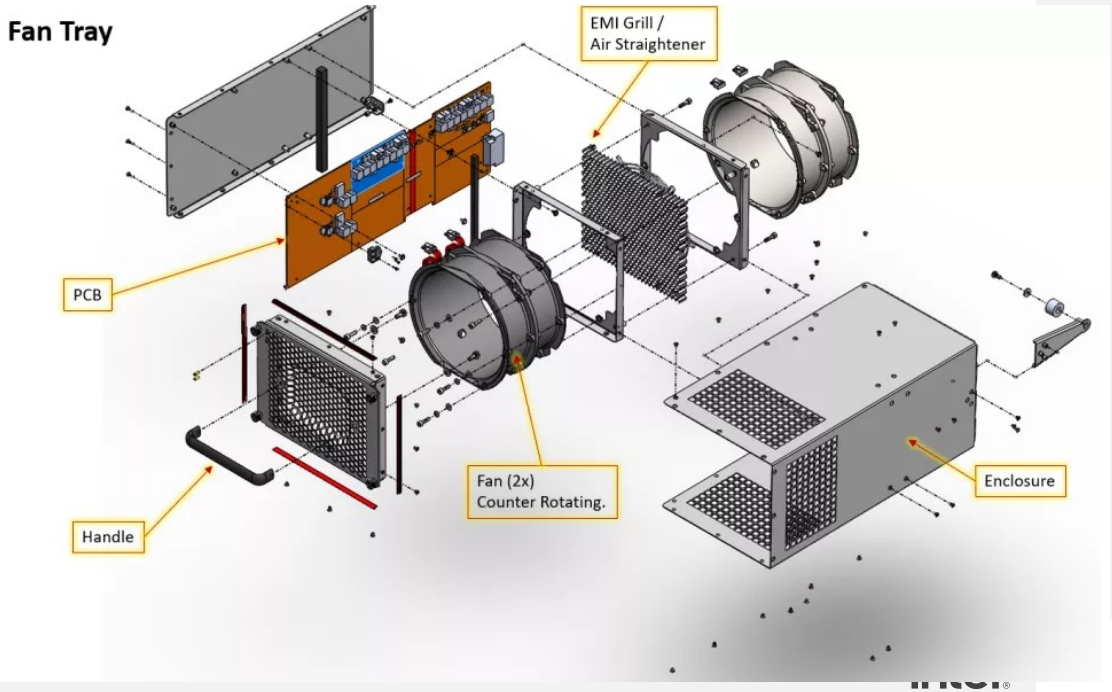
- The "engine block" sits up front.
  - A sandwiched design that has the power subsystem, motherboard, chip, and cold plate mounted as one assembly (left).
- Develop costume connector to connect wafer to PCB.
- The connector absorbs the (temperature) size variation while maintaining connectivity



Cooling system

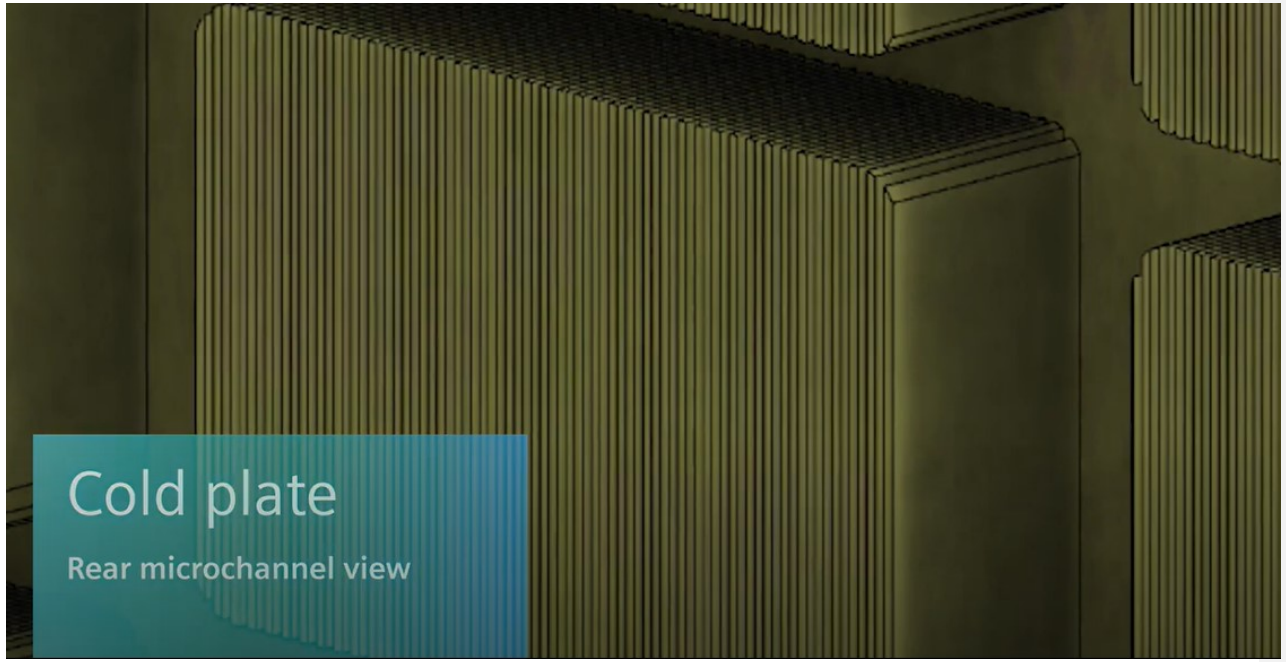
- Water cooled cooper plate stacked on top of silicon.
  - Vendor : Motivair MCDU-25, which has 625 kW of thermal capacity
  - Water and Glycol
- The system involves two cooling loops. A primary loop mixes water with glycol coolant that traverse the CS-1s using pumps, then transfers heat onto a secondary loop (i.e., the datacenter's chilled water supply) via a heat exchanger.
- The cold plate receives water from a manifold to the right, which then delivers cooled water to several individual zones on the surface of the cooling plate.
- The heated water is then extracted, again from the small zones that ensure consistent thermal dissipation and pumped down to the heat exchanger at the bottom of the unit.
- The exchanger consists of an EMI grill and is cooled by powerful fans that employ air straighteners. Overall, the chip runs at half the junction temperature of a standard GPU, which increases reliability.
- CS2 cooling system occupy ~70% of the system size

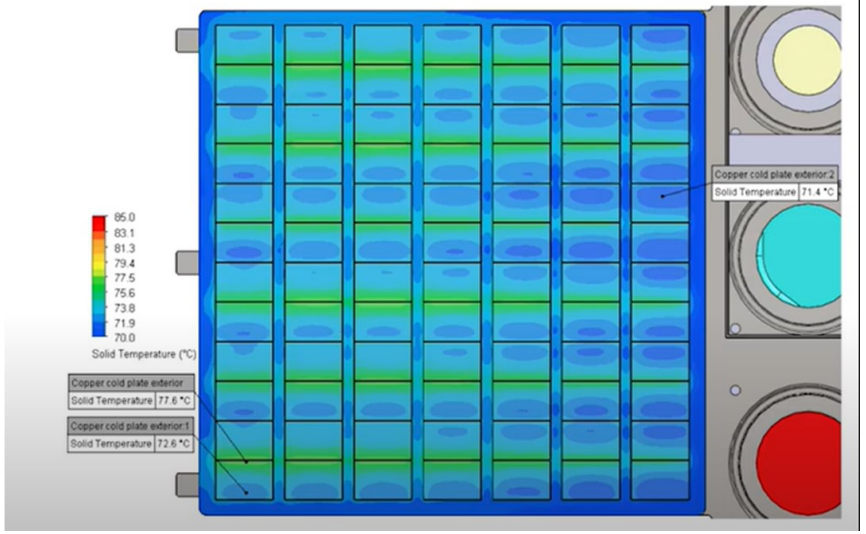


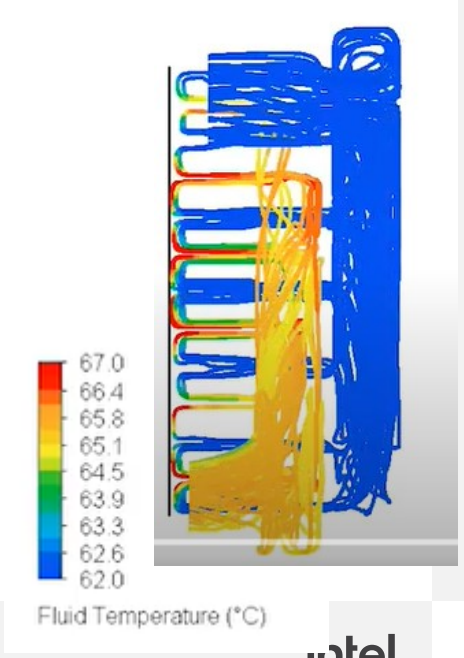


Cooling system

- Cold plate has microchannels in the back
- 30% ethylene glycon water cooling
- 20 gallon per minute







https://www.youtube.com/watch?v=dACywU1YCpc

Power delivery

- Supplying the >15kw into the wafer at while maintaining good regulation.
- Cerebras' solution employs more than 300 water cooled voltage regulation modules (VRMs) distributed over the wafer.
- Multiple VRMs per reticle ensures redundancy in the power distribution and gives individual control of each reticle's power domain.
- 12 power supplies (PSUs), in a 9+3 redundant configuration

Power solution

Fit power conversion within footprint of WSE

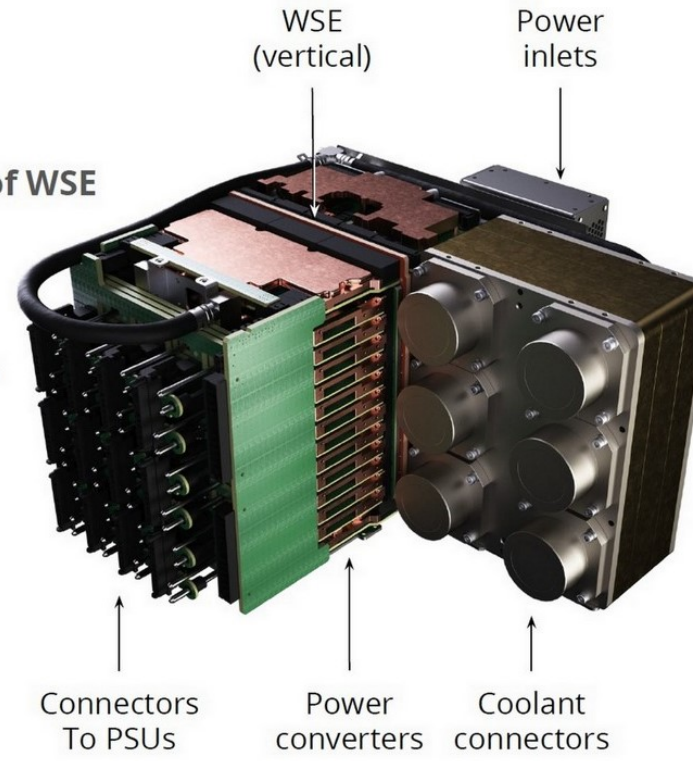
Place converters on opposite side of PCB

Bring in power through IEC C20 16A inlets

12x 4kW hot-swappable universal PSUs

Universal high-voltage AC to 54V DC

Direct conversion from 54V to 0.9V

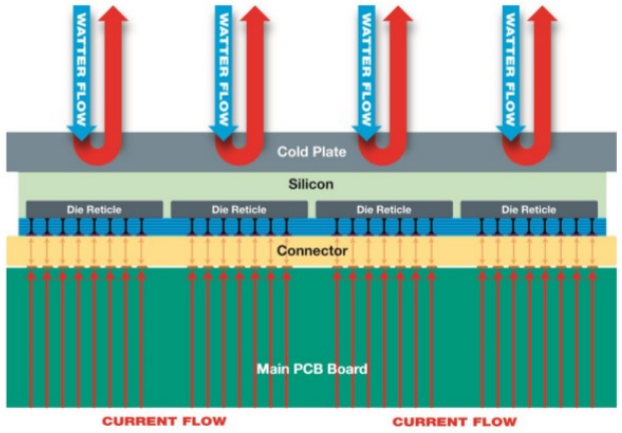


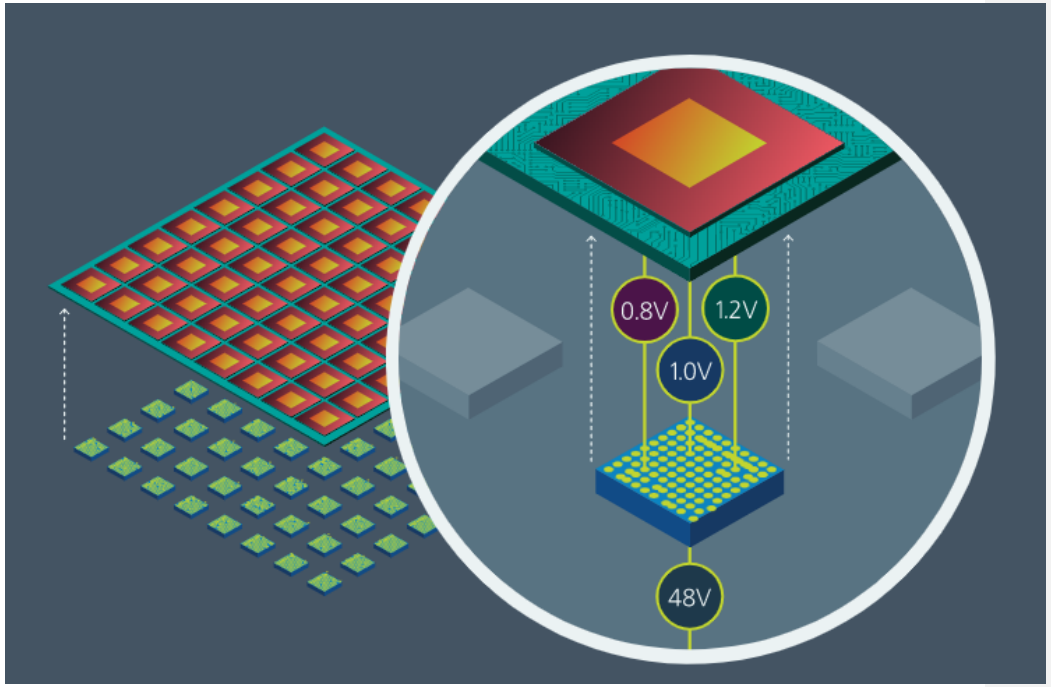
Power delivery

Using the 3rd Dimension

- Power delivery
  - Current flow distributed in 3rd dimension perpendicular to wafer
- Heat removal
  - Water carries heat from wafer through cold plate

Co-designed with system

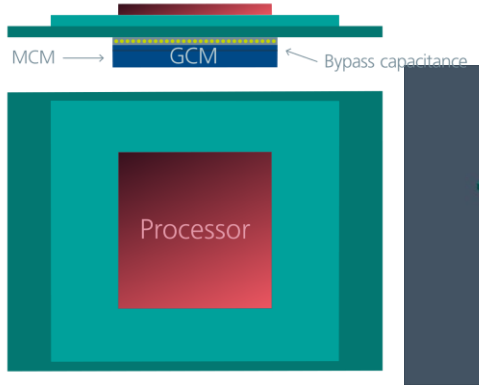


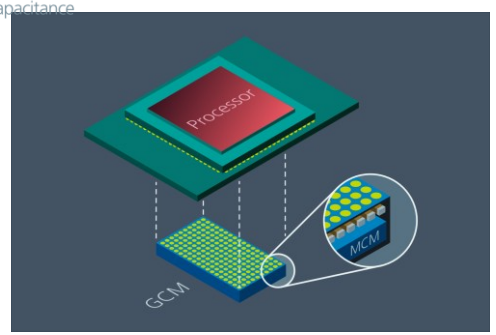


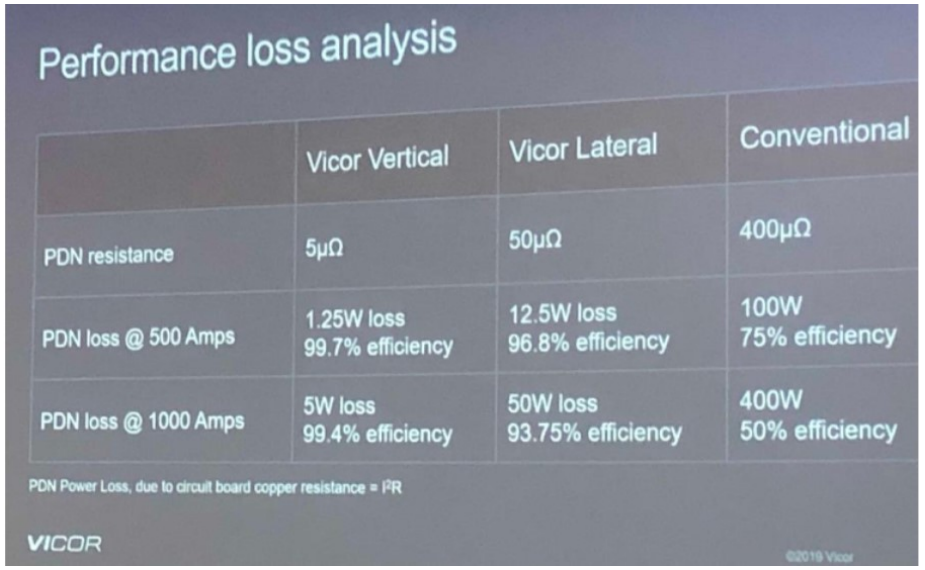
The power-delivery system sits above the chip with a water cooled cold plate below.
Cerebras implemented Vicor's Vertical Power Delivery (VPD) architecture that reduces power delivery network (PDN) resistance by more than 50%, thereby achieving higher overall density and power system efficiency.
At the macro level using 48V power supply which reduce datacenter energy losses by over 30%, with >98% peak efficiency

intel®

Vertical power delivery – VICOR example







Performance loss analysis

	Vicor Vertical	Vicor Lateral	Conventional
PDN resistance	5μΩ	50μΩ	400μΩ
PDN loss @ 500 Amps	1.25W loss 99.7% efficiency	12.5W loss 96.8% efficiency	100W 75% efficiency
PDN loss @ 1000 Amps	5W loss 99.4% efficiency	50W loss 93.75% efficiency	400W 50% efficiency

PDN Power Loss, due to circuit board copper resistance = I^2R

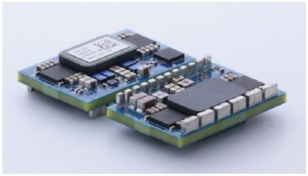
VICOR

©2019 Vicor

- In 2019, AIPG evaluated VICOR Vertical Power Delivery (VPD)
  - VICOR claimed that the PDN resistance was reduced by 10X (vs lateral power delivery)
  - In 2019, the assessment was showing other challenges like thermal was preventing such solution with Intel AI training chip. It could not be possible on OCP card FF.
  - Tesla Dojo chip has included a water-cooled copper plate between the chip and the power delivery
  - Example power delivery modules shown below; Dojo is likely using a customized solution.

PI31xx

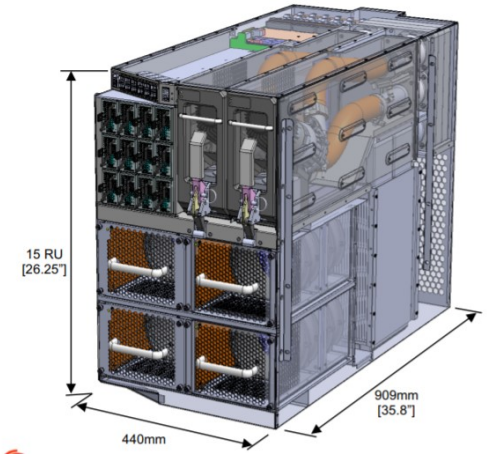
Maxi, Mini, Micro

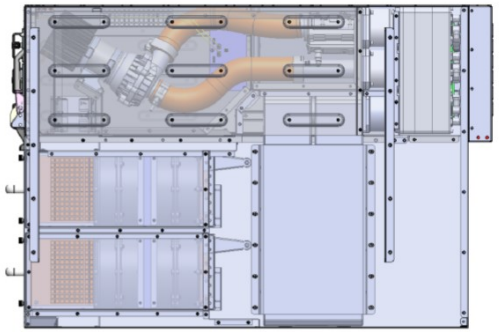


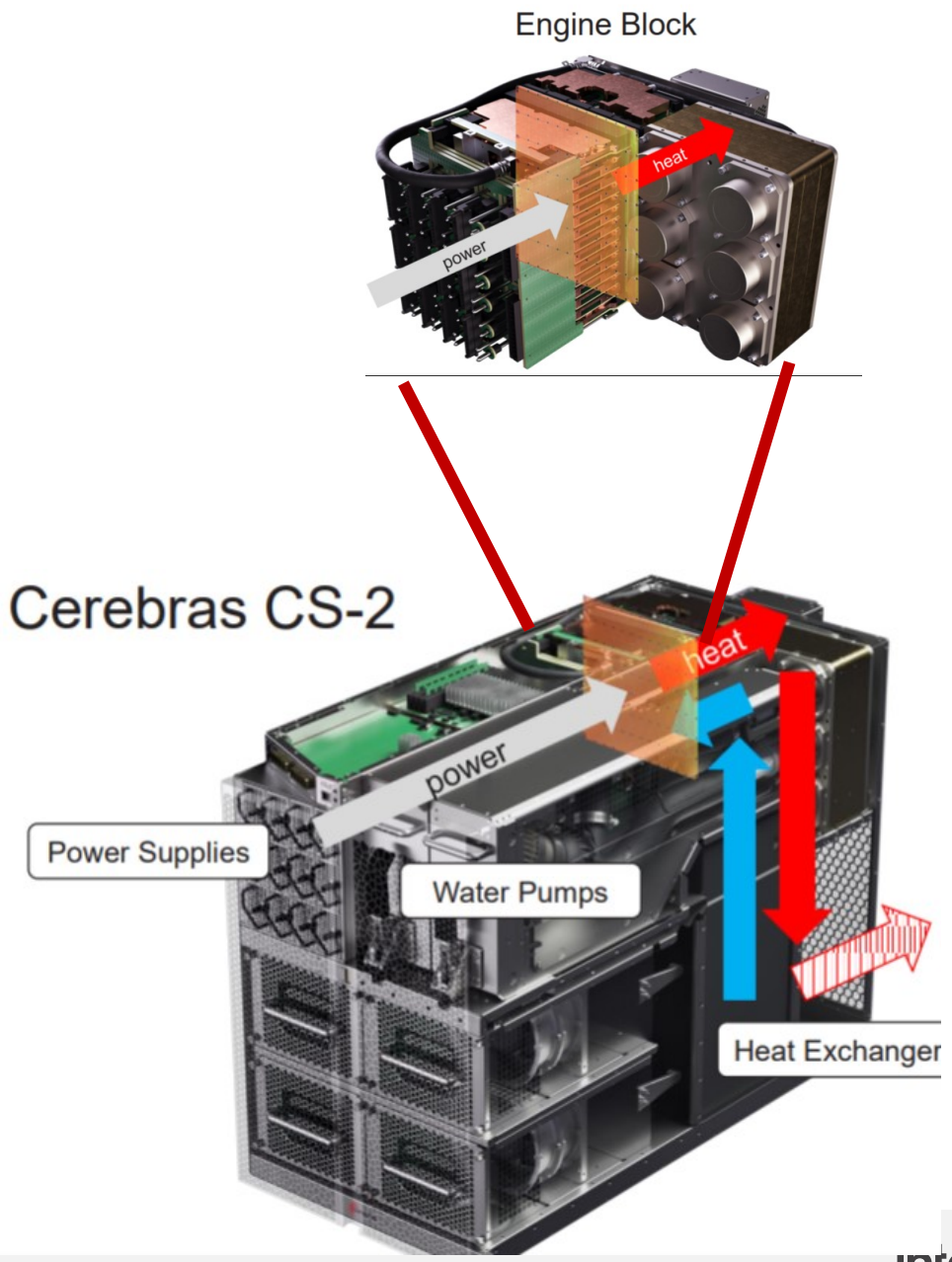
Dojo: 

CS1 – system view

CS-1 System View







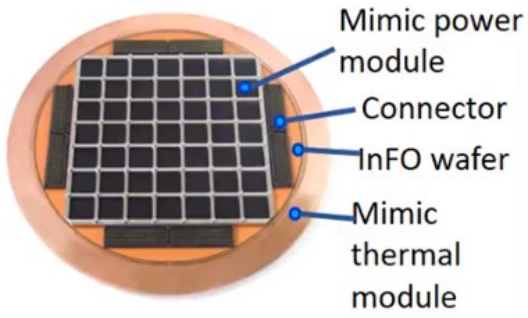
WSE 2 housing is primarily devoted to cooling.
tubes, pumps, fans and a heat exchanger
housing takes ~15RU or about a third of a standard rack.

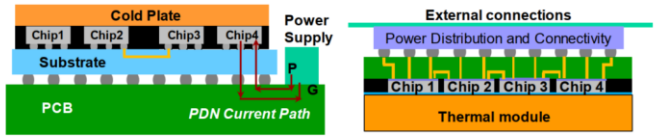
TSMC InFO_SoW (System –on-Wafer) for HPC

- Wafer-scale InFO demo
  - No PCB - Connectors and power modules are soldered to InFO wafer followed by assembly of thermal module.
  - 6 RDL: 3 - 5/5, 3 - 15/20um L/S, ***Chip-first?***
  - 60% lower ELK stress vs flip chip, due to thick compliant RDL
  - 97% lower PDN impedance vs Flip-Chip MCM
  - 15% interconnect power savings due to lower Cu RDL surface roughness vs Substrate or PCB. 0.7dB/30mm
- 7000 W (1.2 W/mm^2) Thermal Solution
  - 2x5 array heater and cooling system
  - Localized modulation for HS/SoC contact
  - 4 LPM DI water 16C inlet, 90C outlet

InFO_SoW (System-on-Wafer) for High Performance Computing

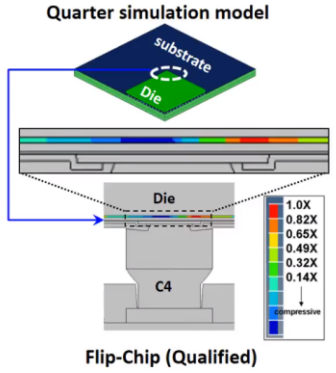
Shu-Rong Chun, Tin-Hao Kuo, Hao-Yi Tsai, Chung-Shi Liu, Chuei-Tang Wang, Jeng-Shien Hsieh, Tsung-Shu Lin, Terry Ku, Douglas Yu
Research and Development
Taiwan Semiconductor Manufacturing Company
Hsinchu, Taiwan, R. O. C
srchun@tsmc.com

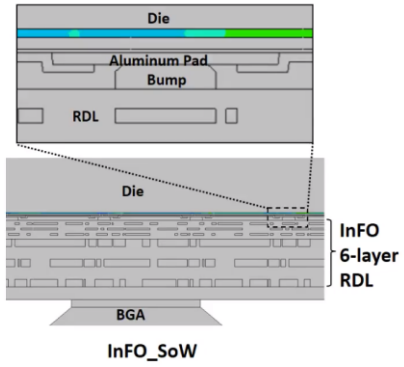




Line width / space (μm)	10 / 10	5 / 5
Line density	1x	2x
Bandwidth density	1x	2x
PDN impedance	1x	0.03x

Full TSMC paper embedded in document:



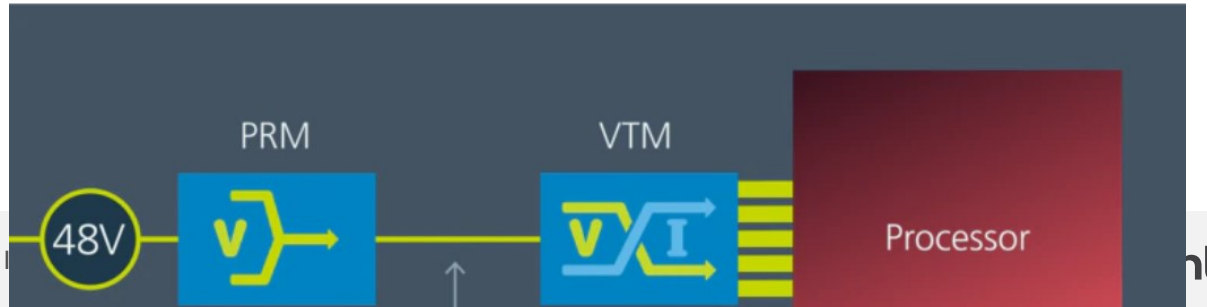


Dojo interposer resembles TSMC InFO_SoW test vehicle

Vicor

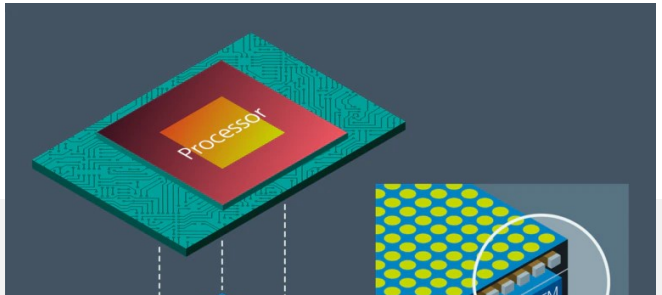
- Vicor 48V direct-to-load (<1V) Factorized Power Architecture (FPA™) is a major departure from the common 48V intermediate bus architecture (IBA) consisting of an intermediate bus converter followed by multiphase PoL regulators.
- Factorized Power Architecture™ is based on the fundamental principle of dividing a power converter into two primary functions, optimizing each separately and then implementing those functions as a system. The two functions are regulation and current multiplication.
- Regulation
- The efficiency of a regulator is inversely proportional to the work performed — the more work, the lower the efficiency. The closer the input and output voltages of a regulator are to each other, the less work is performed and a higher efficiency is achieved. By virtue of its position in the system, FPA™ minimizes the regulator's input-to-output voltage differential. The PRM™ regulator is implemented using a zero-voltage switching (ZVS) buck-boost topology, which features high efficiency where the input and output voltage difference is small. ZVS greatly reduces switching losses, enabling high-frequency operation and greatly reducing converter size. The PRM typically regulates an input between 40 and 60V to an output voltage between 30 and 50V.
- Soft switching and current multiplication
- The PRM is followed by a second stage performing a voltage step-down and current step-up function. This is implemented using the Sine Amplitude Converter (SAC™) topology in a device called a VTM™ Current Multiplier. The VTM's behavior can be realized as an ideal transformer, where the input and output voltage are related by a fixed ratio and the device impedance remains low (hundreds of μΩ) beyond 1MHz.
- Since there is no energy storage in the VTM, it can provide large amounts of power if it is kept sufficiently cool. This allows for matching the power capability of the VTM with the thermal capability of the processor.
- The SAC topology uses a zero-voltage and zero-current switching control system, further reducing switching noise and power losses.
-

- Together, the PRM and VTM form the building blocks of FPA. One is dedicated to regulation and the other dedicated to transformation and current multiplication.
- SM-ChiP package reduces noise and improves thermals
- While the topology and architecture used to implement a high-performance regulator are important, of equal importance is the packaging technology. The Vicor SM-ChiP™ package integrates everything—passives, magnetics, FETs and control—into a single device. Moreover, this package is engineered to enable the most efficient extraction of current at the lowest thermal impedance to facilitate cooling. Many SM-ChiPs also include grounded metal shielding over a significant surface of the device. This serves not only to facilitate cooling but also to localize high-frequency parasitic currents to keep them from propagating outside the device.
- Vertical power delivery cuts PDN losses by 95%
- Lateral power delivery for clustered processor arrays is almost impossible with large arrays. The better solution for cluster-processor power delivery is vertical power delivery (VPD). In VPD, the current multiplier is located directly underneath the processor on the opposite side of the board, significantly reducing PDN losses by reducing the distance the current travels through the motherboard. VPD needs two key features to achieve this function.



Vicor

- Figure 2: Vertical power delivery (VPD) with GTM™ Geared Current Multiplier placed underneath processor maximizing power delivery performance. The VPD solution also relieves the processor top-side periphery for options including higher I/O routing, onboard memory or tighter processor clustering.
- First, the area directly under the processor contains high-frequency capacitors which are necessary to decouple very high-frequency currents (>10MHz) from the rest of the system. Secondly, for maximum efficiency the physical location and pattern of the current exiting the VPD solution must exactly mirror the location and pattern of the processor core power inputs. This enables the high-current flow to achieve a true “vertical” profile.
- To achieve these features, the Vicor VPD solution is an integrated module consisting of three layers: a VTM Current Multiplier array implemented with a gearbox below and a PRM Regulator mounted above to provide a completely regulated 48V-to-load solution for each processor, a DCM™. The gearbox performs two functions: it incorporates high-frequency decoupling capacitance and redistributes the current from the VTM into a pattern mirroring the processor above it. The VTM array is sized based on the processor output current requirement and PRM is sized based on the power requirement. If the GPU or ASIC requires multiple power rails then the VTM and PRM layers can be implemented with independent PRMs and VTMs sized to meet the current and power voltage requirements for each specific rail.

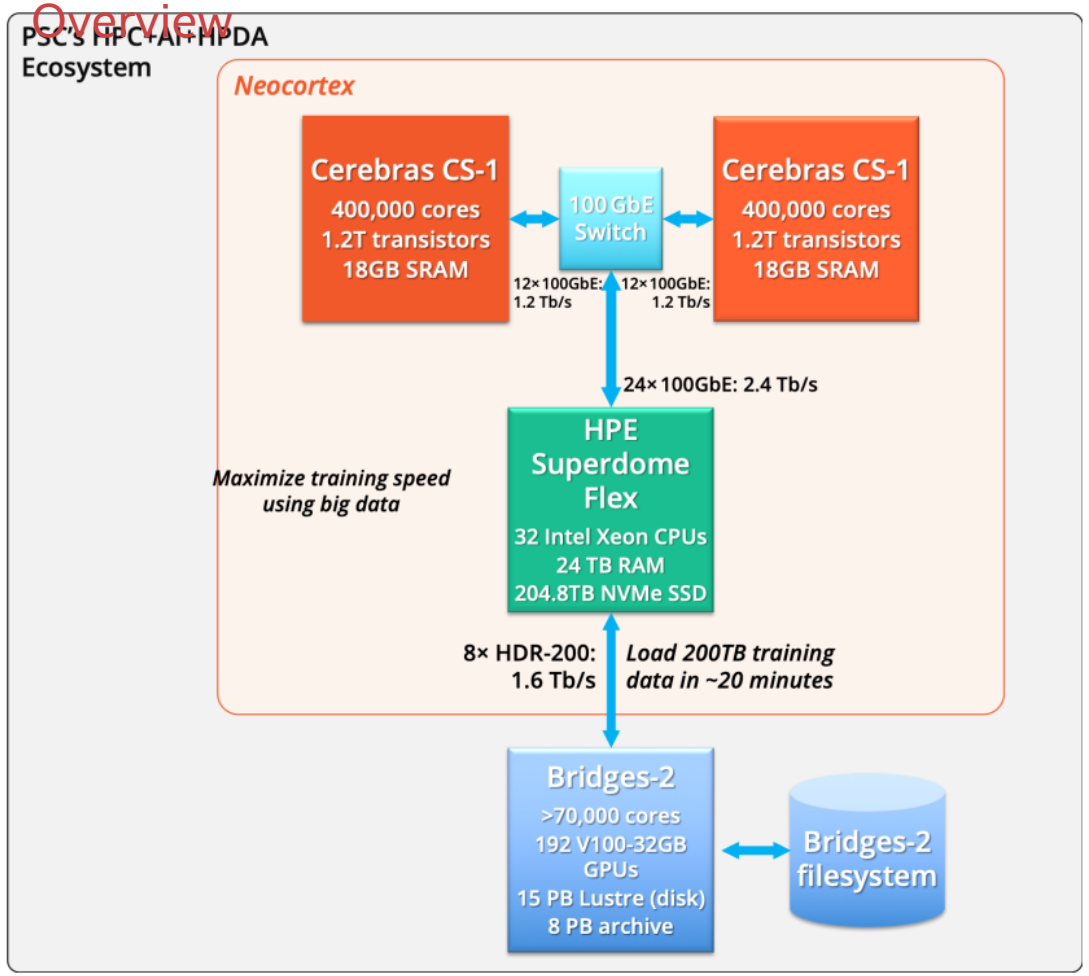


- Figure 3: The Vicor DCM is a complete 48V-to-load VPD solution in one advanced package for clusters of ASICs. The PRM, VTM and gearbox layers of the module provide regulation, current multiplication, decoupling capacitance and pin-to-pin footprint matching.
- Vicor FPA™ architecture, ZVS and ZCS control system, high-frequency SAC current multiplier topology and SM-ChiP packaging technology provide all of the elements for perfecting VPD. It solves the low-noise, clustered power delivery challenge while easing the cooling and thermal management mechanical design with high efficiency and a thermally-adept power module package. The VPD solution is a true enabler for higher-performance AI systems by allowing high-speed massive data analytics via the cluster to perfect training models and advance machine learning to significantly higher levels.
- A better way for high performance computing power
- AI and machine learning are in their infancy of growth. This train will only pick up speed as the years go by. This acceleration is going to require faster processing for more complex solutions. AI ASIC processor based supercomputers will demand more power than conventional methods can possibly deliver. A new, innovative approach to power delivery is the only way the promise of AI can come to fruition. It will require power system architectures, topologies, control systems and packaging working in concert to deliver ever-increasing high currents. Vertical Power Delivery, leveraging current multiplication, is the solution of choice. It is a proven approach that meets the demands to high performance computing today and can easily scale to to keep pace with future needs. It is compact, efficent and can reduce PDN losses by up to 50%.

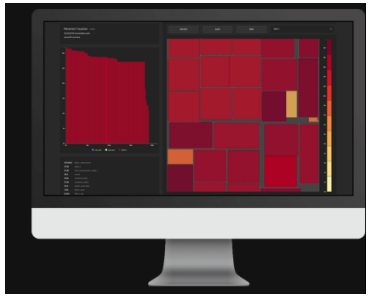
Costumers & partners

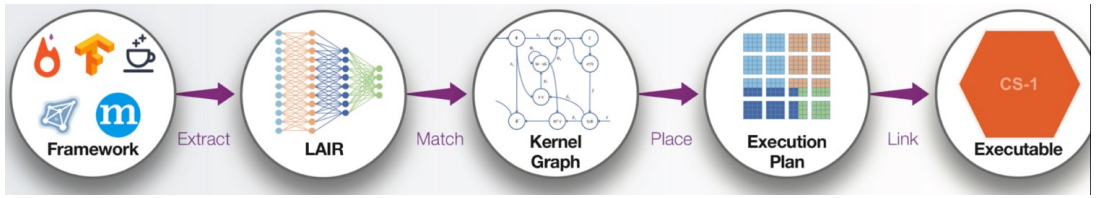
- Cirrascale (partner):
  - cloud hosting for Cerebras systems
- CMU/PSC, Neocortex:
  - research-based 2xCS-1

Neocortex (CS1): System Overview



Cerebras Software platform





Extract graph representation of model from framework, convert to LAIR (**L**inear **A**lgebra **I**ntermediate **R**epresentation)

Match computational subgraphs to kernels that implement portions of model

- For missing kernels, Cerebras **Kernel Compiler generates one dynamically** from the IR

Place & Route allocates compute and memory, assigns kernels to fabric sections, configures on-chip network

- Balance resources and throughputs

Link creates executable output that can be loaded and run by CS

Deep learning:

- High-level programming via ML frameworks (TF, PyTorch), with Cerebras Graph Compiler
- Ability to create custom kernels with Cerebras Kernel SDK

HPC:

- C-level programming interface with Cerebras Kernel SDK

Two execution modes:

Pipelined :

- Non von Neuman approach.
  - Resources need to be balance
  - Good for inference (Similar to Hailo, Unthter.ai, Graphcore).
    - Duplication like in–memory
  - Training: low utilization

    See Graphcore as example

Weight streaming

- “von neuman” like – will suffer from external access which is significantly slower then conventional