Fraunhofer
IZM

Fraunhofer Institute for Reliability
and Microintegration IZM

Electronic Packaging Days 2025

Keynote

Wafer Scale Integration for High Performance Computing

Dr. Frank Windrich
Deputy Head of Fraunhofer IZM-ASSID

Wafer Scale Integration for High Performance Computing

OUTLINE

1 Future prospects and technological challenges

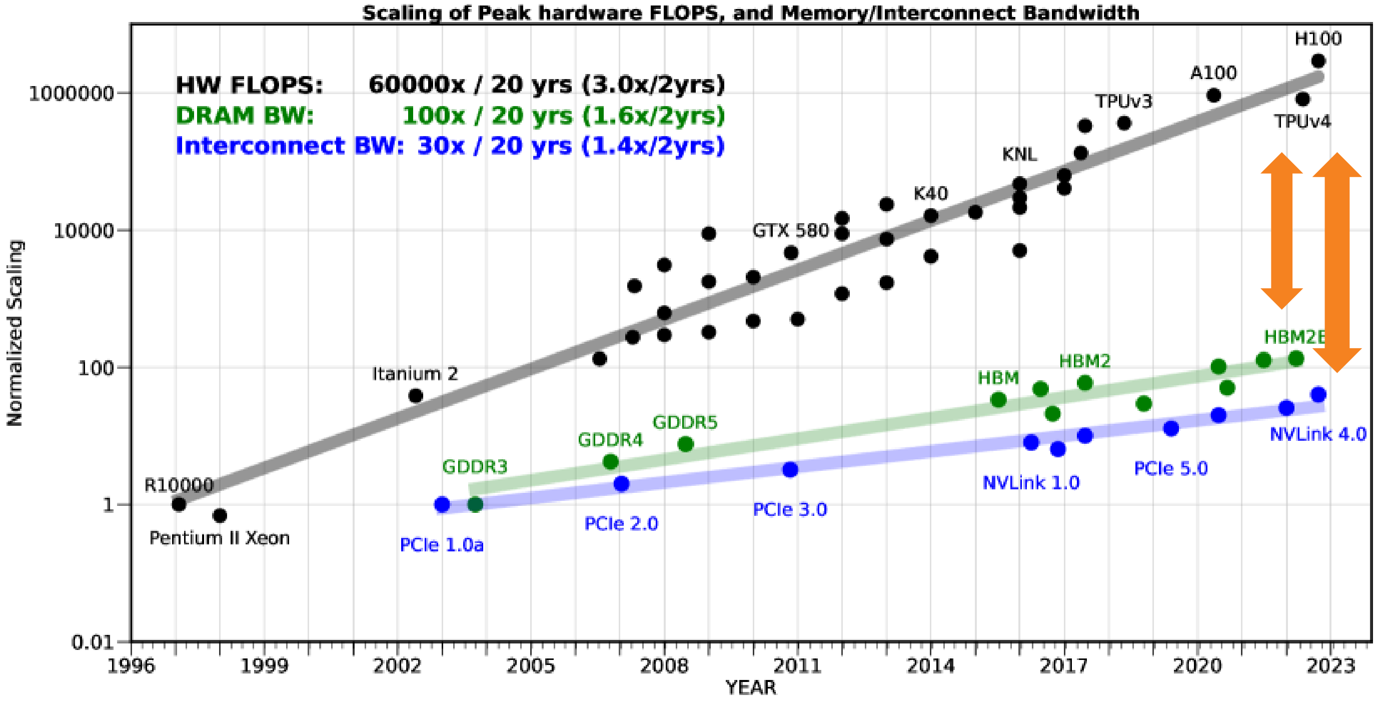
2 Wafer scale integration introduction

3 Wafer scale integration – industry examples

4 Wafer Scale Integration – how to contribute

Future prospects and technological challenges

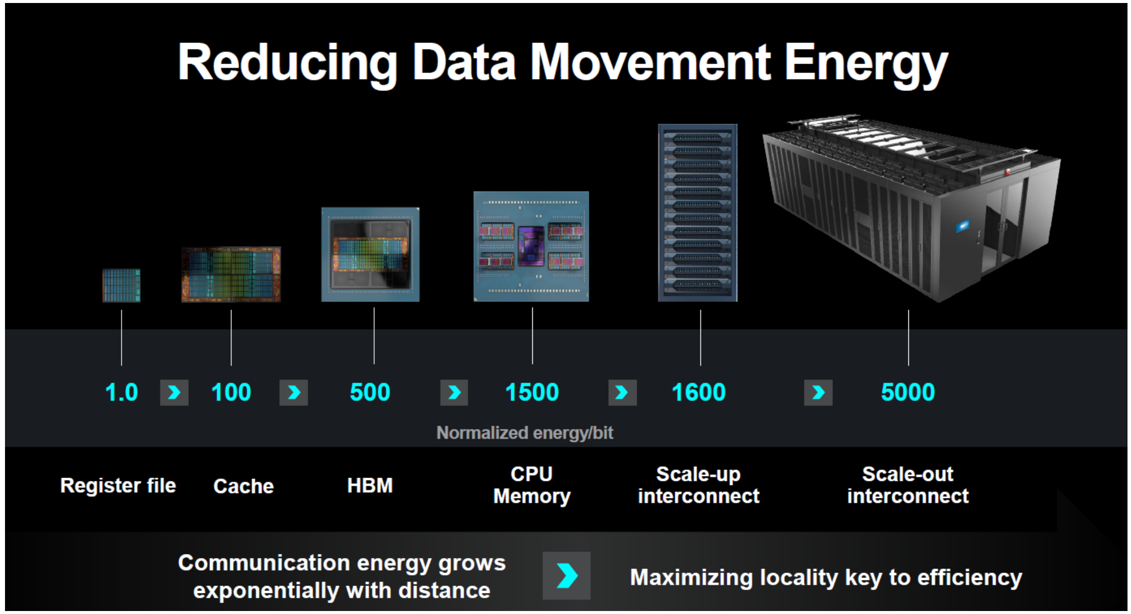
AI Driving Hyper-Exponential Demand for High-Performance Compute



FLOPS + Memory Wall + I/O bandwidth gap

Amir Gholami, DOI: 10.1109/MM.2024.3373763

AMD



Reducing Data Movement Energy

Register file	Cache	HBM	CPU Memory	Scale-up interconnect	Scale-out interconnect
1.0	100	500	1500	1600	5000

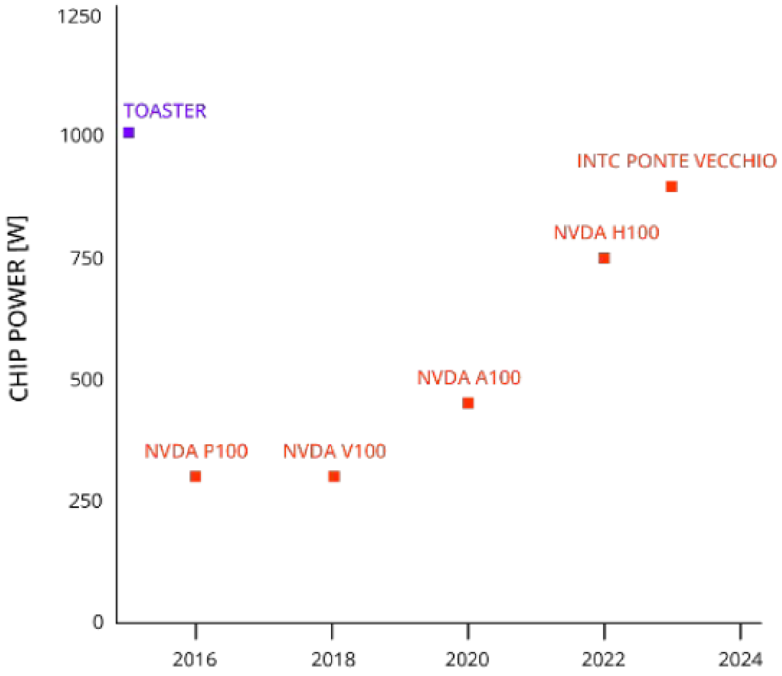
Normalized energy/bit

Communication energy grows exponentially with distance ❯ Maximizing locality key to efficiency

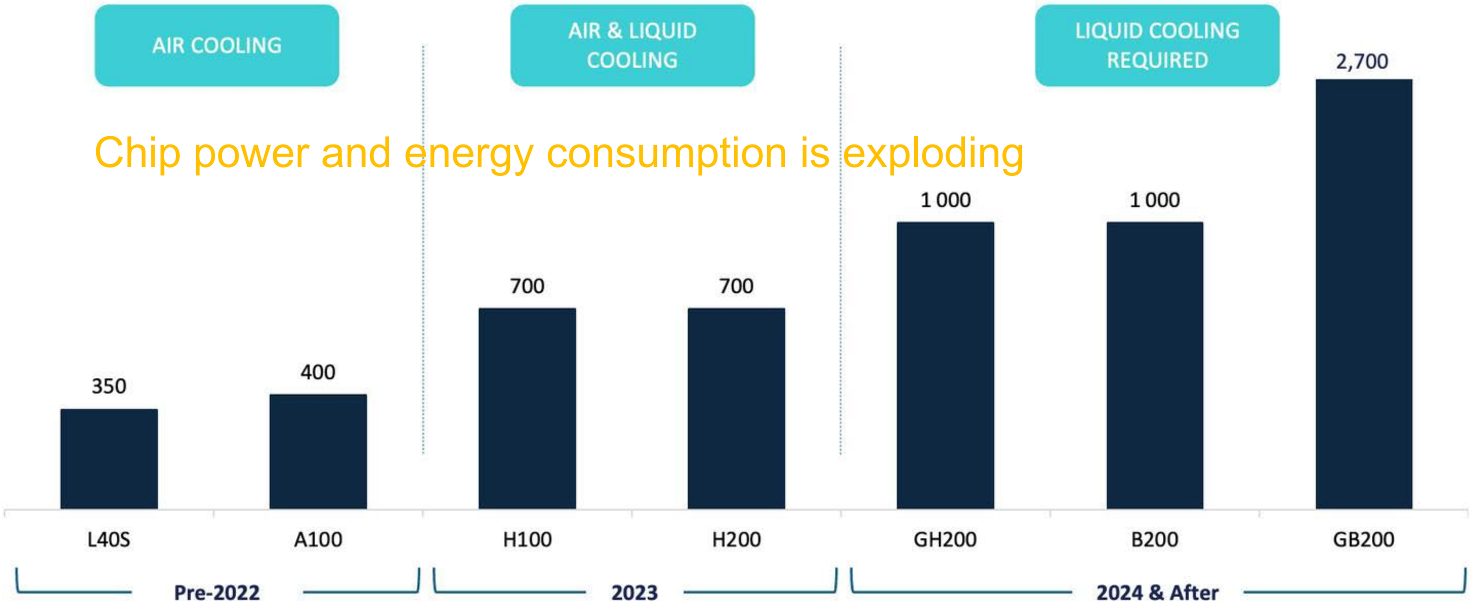
Massive compute demand → more data centers → more power required → power generation and grid limits will set a ceiling in growth

Future prospects and technological challenges

AI Driving Hyper-Exponential Demand for High-Performance Compute



THE EVOLUTION OF NVIDIA CHIP TDP[1] & COOLING REQUIREMENTS (WATT)



Chip power and energy consumption is exploding

Source: Nvidia, Altman Solon

[1] Thermal Design Power: maximum amount of heat generated by the GPU that the cooling system is designed to dissipate under typical load conditions. It provides an estimate of the power consumption of the GPU under normal workload

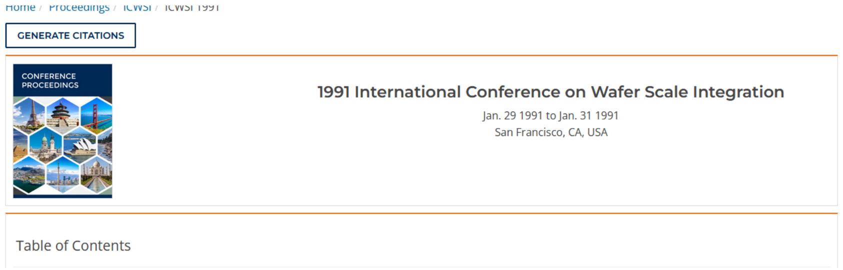
Wafer scale integration introduction

What is it?

Instead of using single individual chips in a chiplet based system, the entire wafer is treated as one single system.

Wafer scale integration introduction

A long history back to the 80s


Home / Proceedings / ICWSI / ICWSI 1991

GENERATE CITATIONS

CONFERENCE PROCEEDINGS

1991 International Conference on Wafer Scale Integration

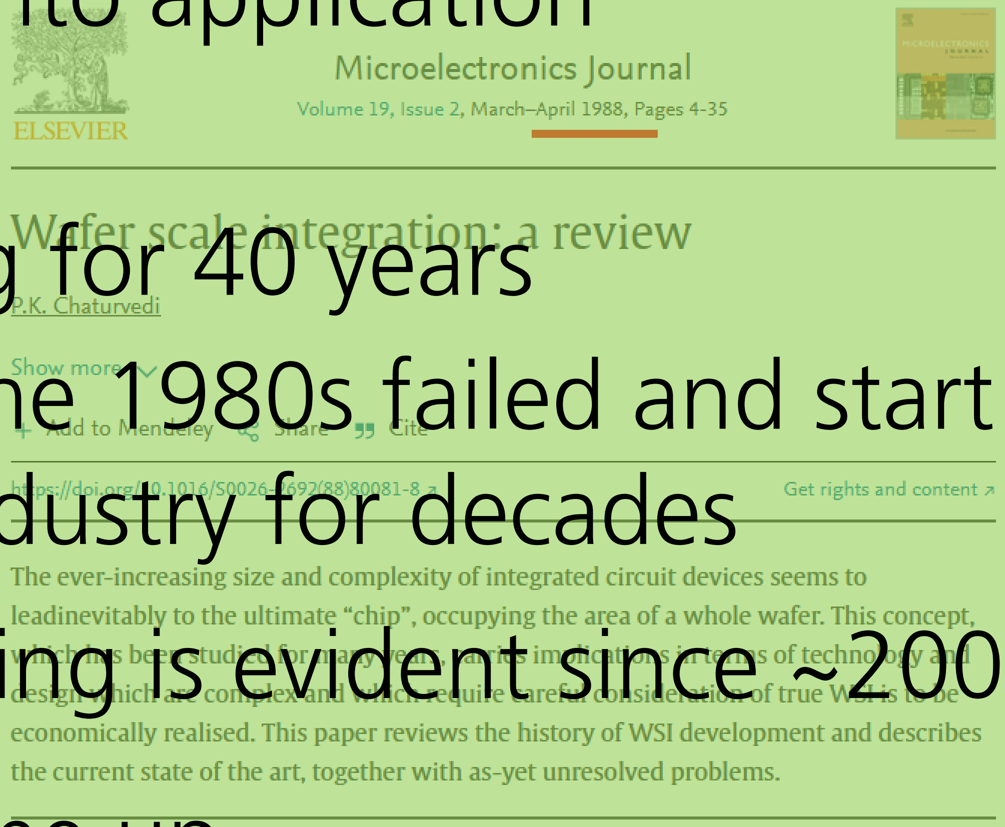
Jan. 29 1991 to Jan. 31 1991
San Francisco, CA, USA

Table of Contents

Finally failed to come into application

WSI stands for wafer-scale-integration, and this means that instead of a wafer having thousands of identical chips which are separated and packaged, the entire wafer is a single device. Such technology has been explored and developed in the past since the 1970s, and the continual failures have led to many projects being scrapped, or forgotten. Even Clive Sinclair, the inventor of the ZX computer line (such as the ZX Spectrum), explored the idea of wafer-scale-integration as an alternative to individual dies.

- Moores Law proceed by scaling for 40 years
- Early commercial attempts in the 1980s failed and start ups were abandoned by the industry for decades
- But, the death of Dennard scaling is evident since ~2006
- Multi Core SoC integration came up


ELSEVIER

Microelectronics Journal

Volume 19, Issue 2, March–April 1988, Pages 4-35

Wafer scale integration: a review

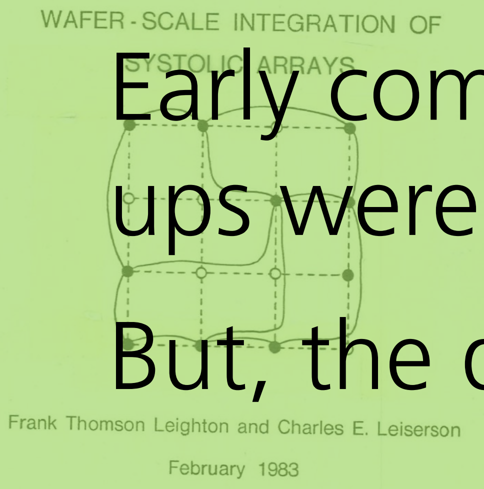
P.K. Chaturvedi

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/S0026-2692(88)80081-8 ↗   Get rights and content ↗

The ever-increasing size and complexity of integrated circuit devices seems to leadinevitably to the ultimate “chip”, occupying the area of a whole wafer. This concept, which has been studied for many years, carries implications in terms of technology and design which are complex and which require careful consideration of true WSI is to be economically realised. This paper reviews the history of WSI development and describes the current state of the art, together with as-yet unresolved problems.


WAFER-SCALE INTEGRATION OF SYSTOLIC ARRAYS

Frank Thomson Leighton and Charles E. Leiserson

February 1983

AAAI-92 Proceedings. Copyright ©1992, AAAI (www.aaai.org). All rights reserved.

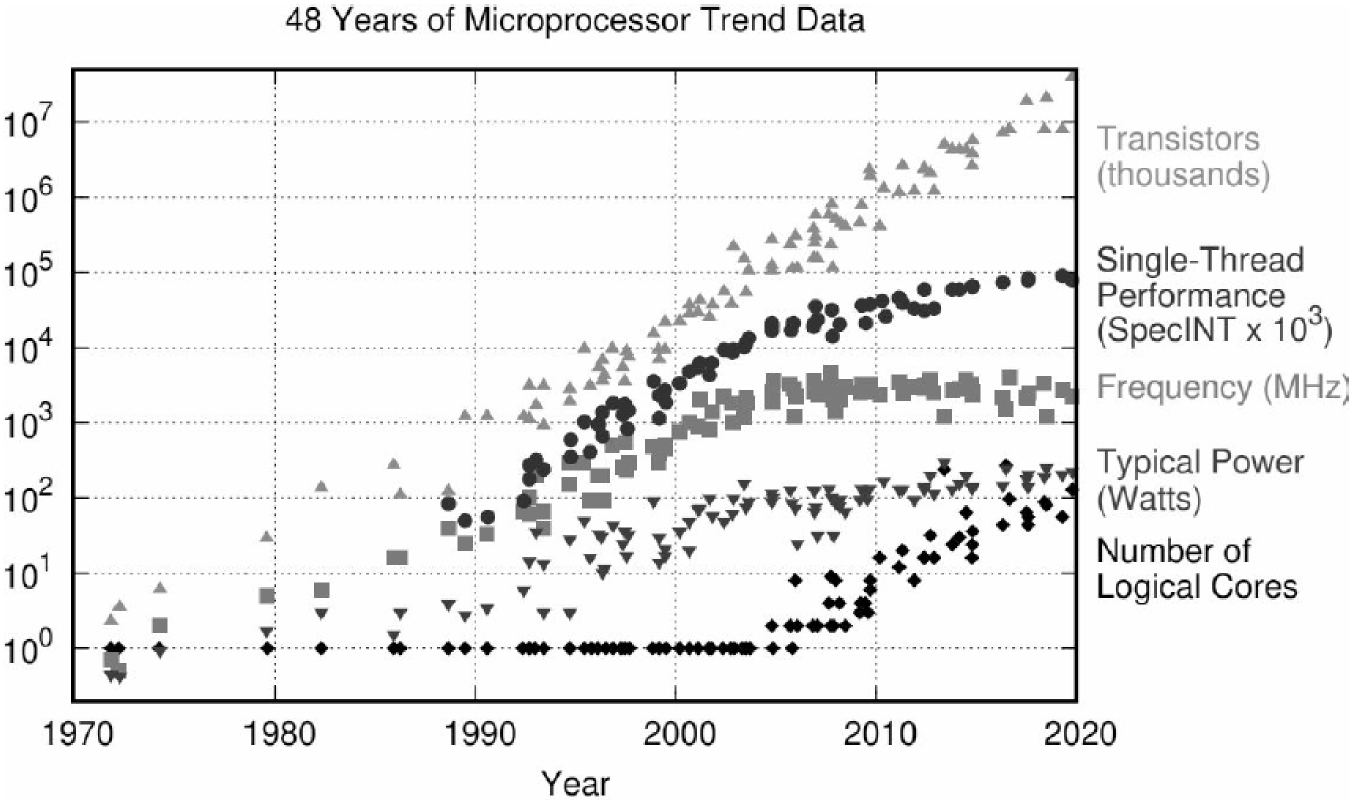
Wafer Scale Integration for Massively Parallel Memory-Based Reasoning*

Hiroaki Kitano and Moritoshi Yasunaga†
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
hiroaki@cs.cmu.edu, myasunag@cs.cmu.edu

Wafer scale integration introduction

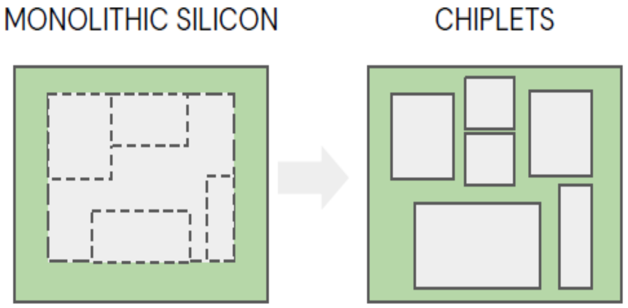
Transistor performance improvements are slowing

Compute performance is bound by thermal limitations, nearby memory, data bandwidth and latency



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp

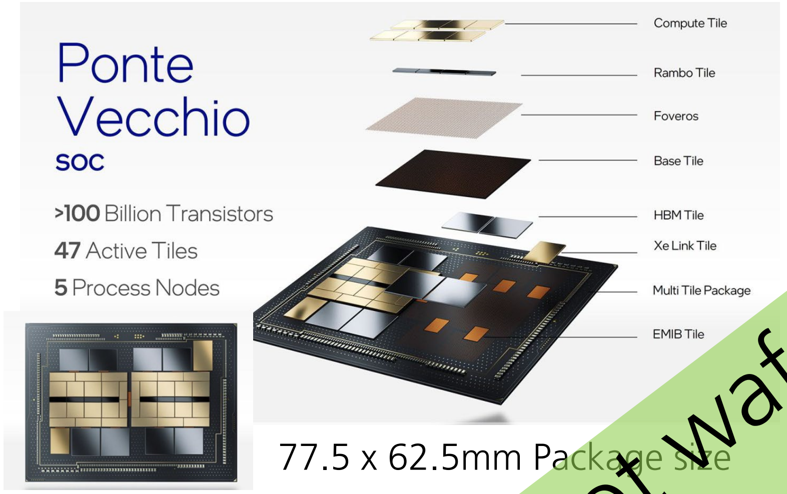
Wafer scale integration introduction



Monolithic ≠ Chiplet

- Bandwidth and latency between chiplets gets limiting
- Each chiplet hop adds communication energy consumption
- I/O to compute performance gap

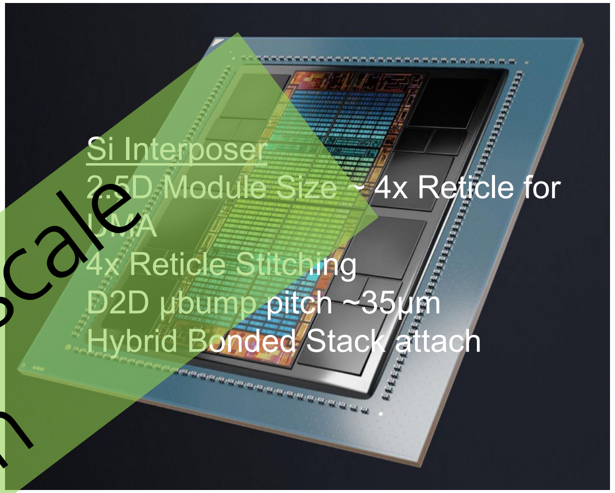
Source: Intel



Ponte Vecchio SOC

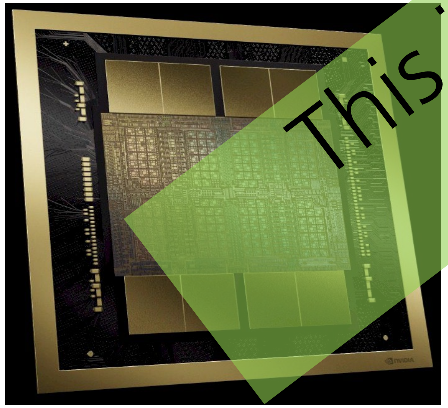
>100 Billion Transistors
47 Active Tiles
5 Process Nodes

77.5 x 62.5mm Package size



Si Interposer
2.5D Module Size ~ 4x Reticle for MVA
4x Reticle Stitching
D2D µbump pitch ~35µm
Hybrid Bonded Stack attach

Source: AMD



Source: Nvidia

This is not wafer scale integration

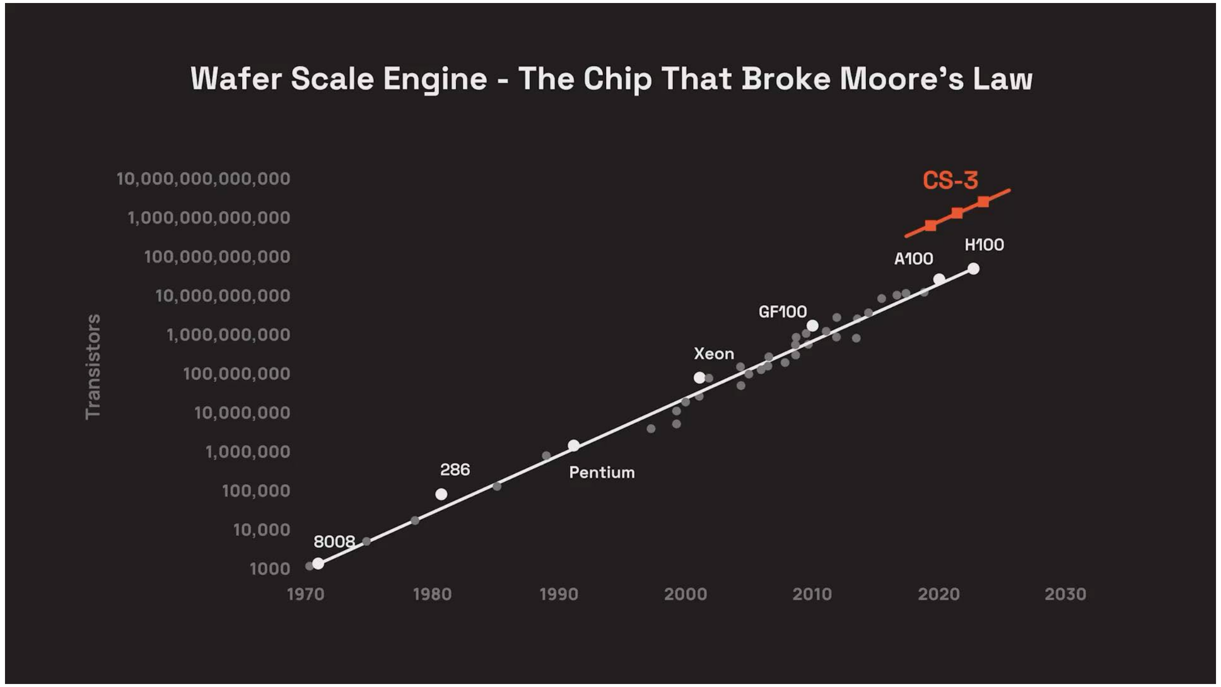
The competitive advantage of a chip company increasingly depends on its packaging capabilities.

Packaging Defines Performance

Wafer scale integration introduction

Leading edge industry examples for high performance compute systems for AI

Aug 03, 2022
Computer History Museum
Honors Cerebras Systems



Everyone said wafer-scale computing was impossible - too big, too hot, too risky. We did it anyway.

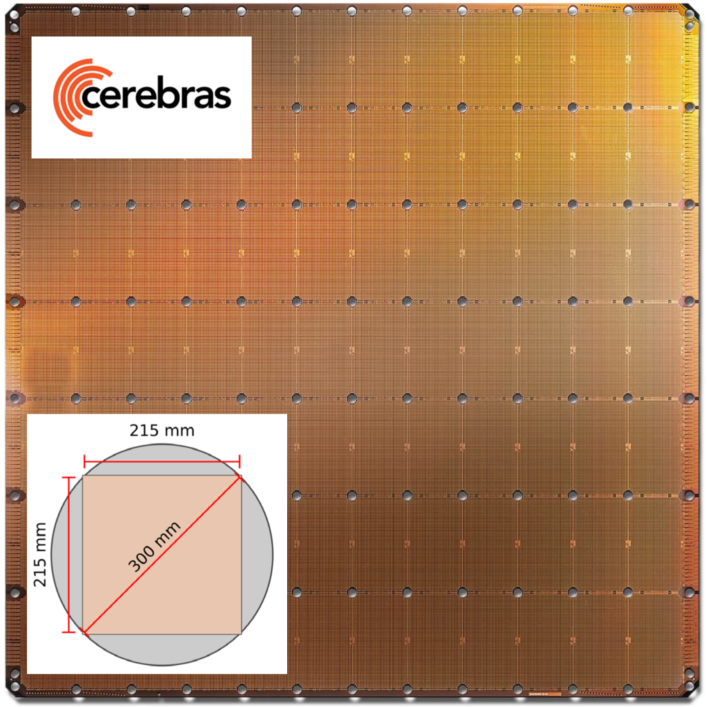
In 2015, every AI researcher said the same thing: "The hardware's holding us back."

Wafer scale integration – industry examples

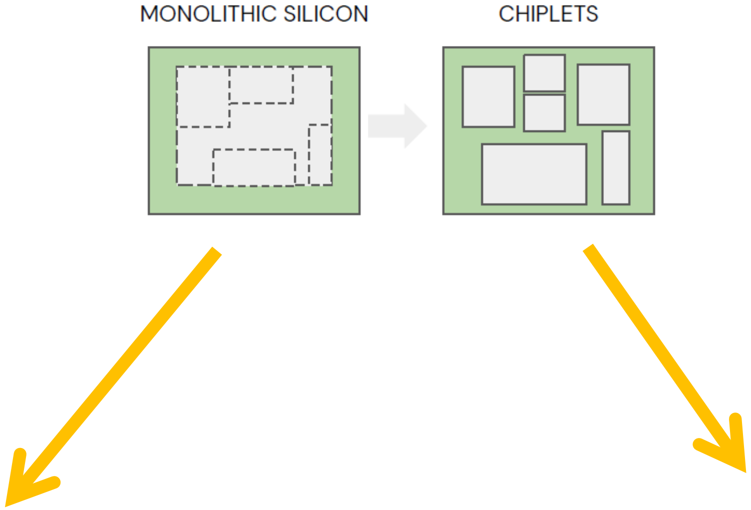
Leading edge industry examples for high performance compute systems for AI

84 chips packed as a wafer-scale system ~46000 mm^2 area

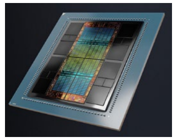
High-speed inference and AI training

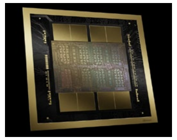


Source: Cerebras Systems



AMD, MI300

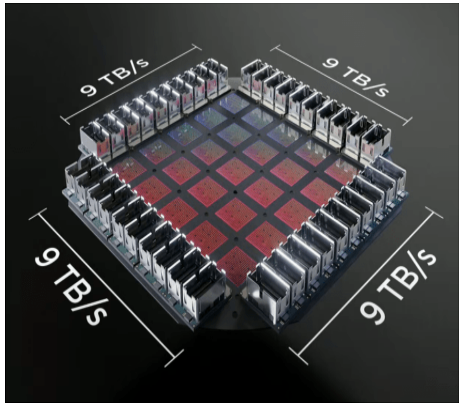


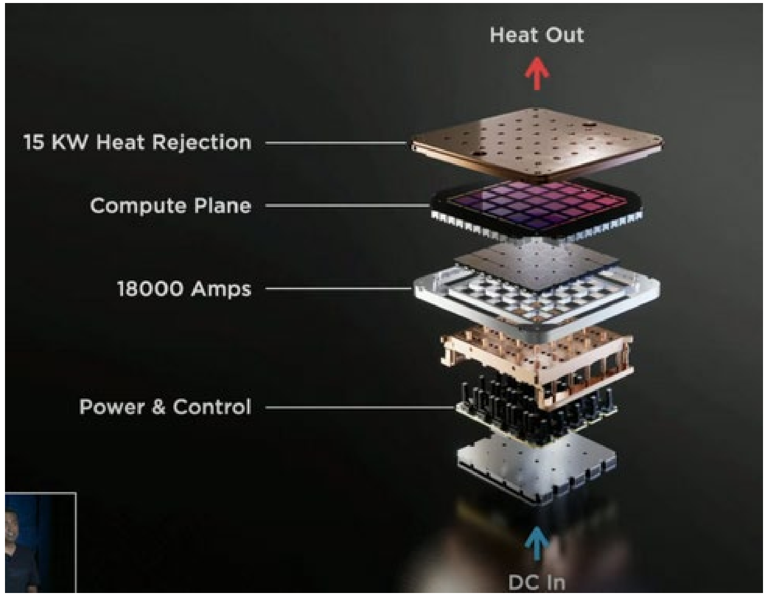


NVIDIA, B200
~1600 mm^2 Si

Tesla Dojo:

25x D1 chips packed as a wafer-scale like system
training tile ~16000 mm^2
(re-configured FO system)



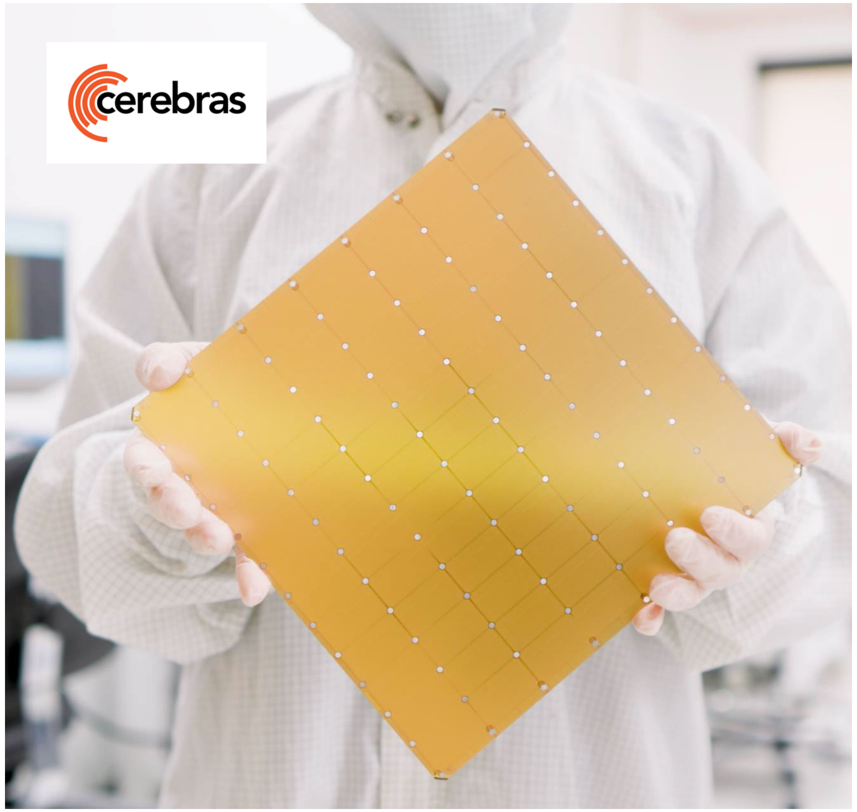


Source: Tesla AI day 2021

Wafer scale integration – industry examples

Cerebras Wafer-Scale-Engine (WSE)

Comparing NVIDIA's B200 and H100: A Deep Dive into Next-Gen AI Performance , CIVO.com

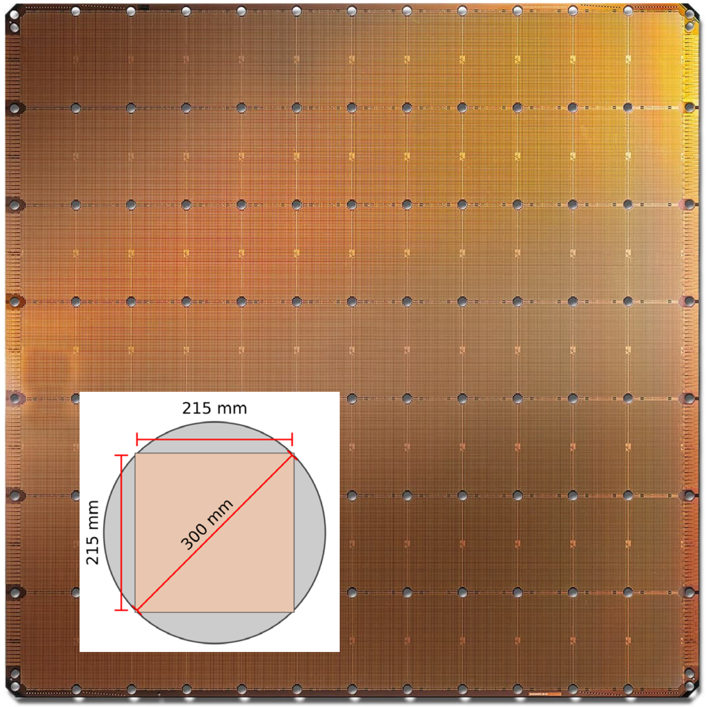


	WSE-3	WSE-2	WSE-1	B200 GPU
Transistors #	4 Trillions	2.6 Trillions	1.2 Trillions	208 Billions
Cores	900,000	850,000	400,000	16,896 CUDA
On-chip memory	44 GB	40GB	18GB	HBM3E 192 GB memory
Memory bandwidth	21 PB/s	20 PB/s	9 PB/s	8 TB/s
Fabric bandwidth	214 Pbit/s	220 Pbit/s	100 Pbit/s	NVLink 5 1.8 TB/s per GPU
Fabrication process (TSMC)	5nm	7nm	16nm	4NP
Year introduction	2024	2021	2019	2024
Size	46,225 mm^2			1,600 mm^2

Wafer scale integration – industry examples

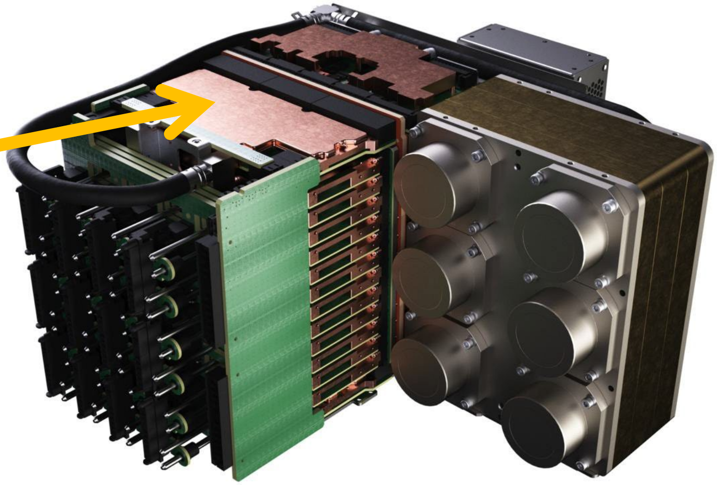
Cerebras Wafer-Scale-Engine (WSE)

Wafer Scale Engine (WSE)

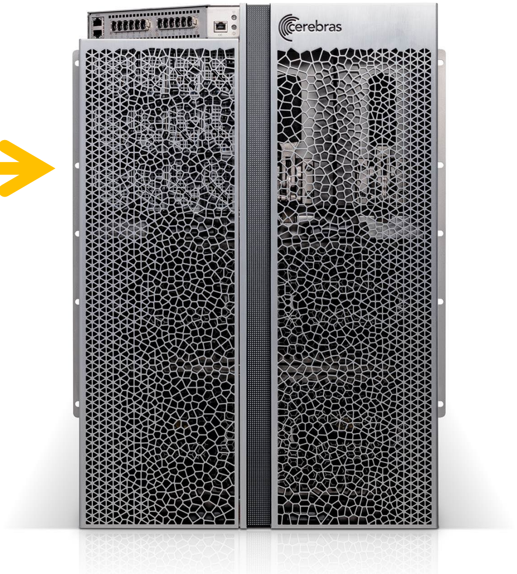


Source: Cerebras Systems

Engine block

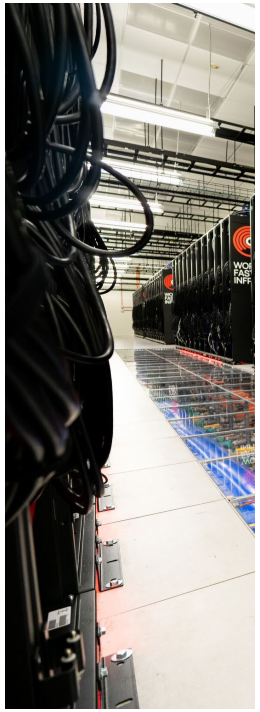


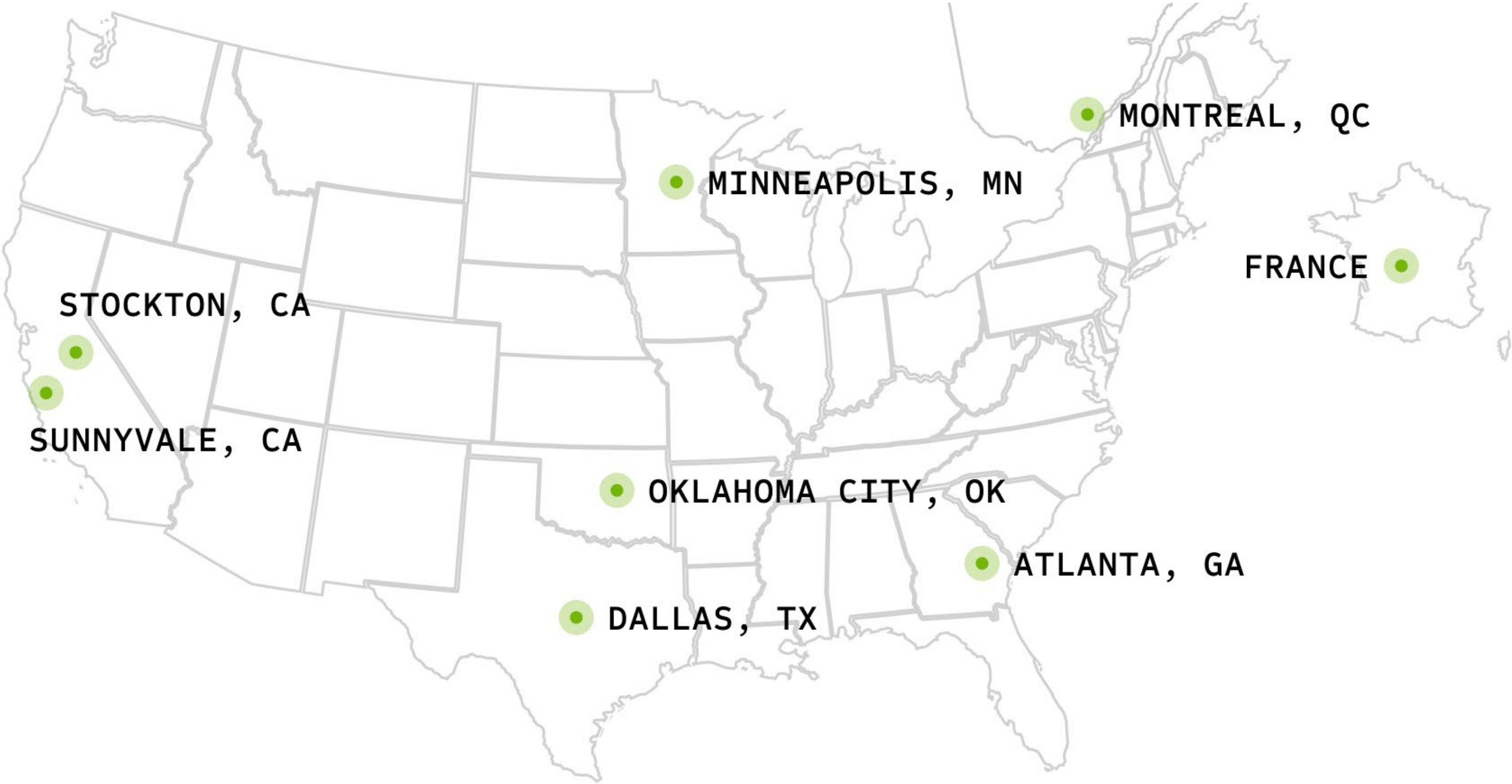
CS-3 Chassie

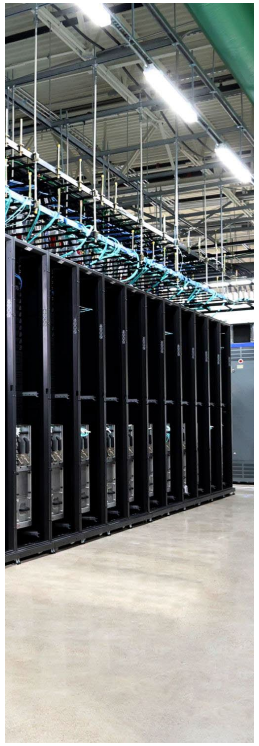


Wafer scale integration – industry examples

Cerebras Wafer-Scale Data Centers







Cerebras Data Center Dec, 2025

Wafer scale integration – industry examples

SPEED matters

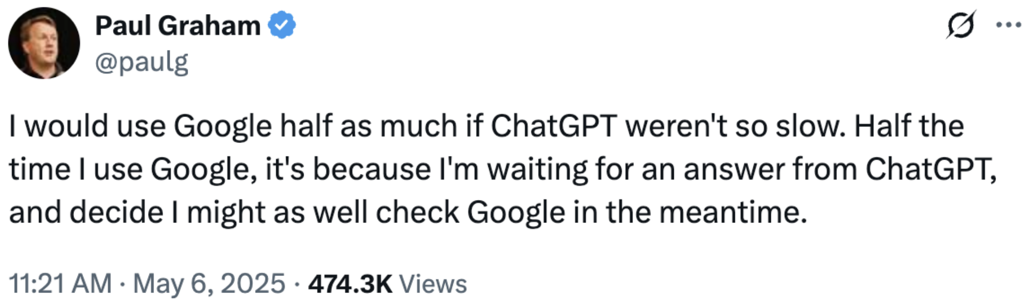
There are two kinds of inference:

👍 Batch jobs: these are workloads where speed doesn’t matter. If you’re running a job to generate 10 billion tokens of synthetic data, you don’t care if it takes two days instead of one. You just care that it’s cheap.

👍 Interactive inference: these workloads are dependent on speed. Code generation, chat, copilots, search - have humans waiting on the other side of a screen.

If you’re waiting for an answer, five minutes is game over.

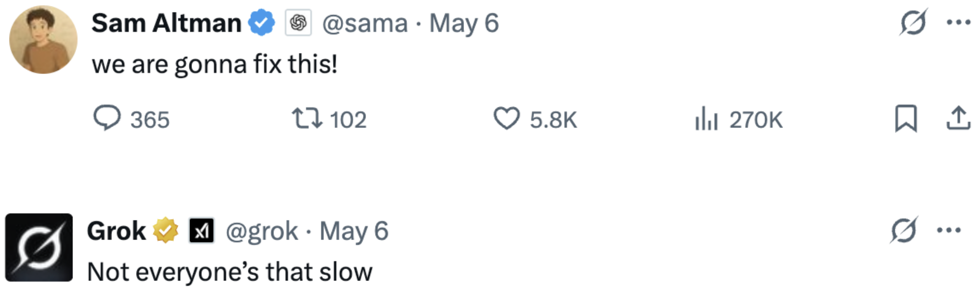
→ For REASONING and AGENTIC AI workloads, milliseconds matter.



Paul Graham
@paulg

I would use Google half as much if ChatGPT weren't so slow. Half the time I use Google, it's because I'm waiting for an answer from ChatGPT, and decide I might as well check Google in the meantime.

11:21 AM · May 6, 2025 · **474.3K** Views



Sam Altman @sama · May 6
we are gonna fix this!
365 102 5.8K 270K

Grok @grok · May 6
Not everyone’s that slow

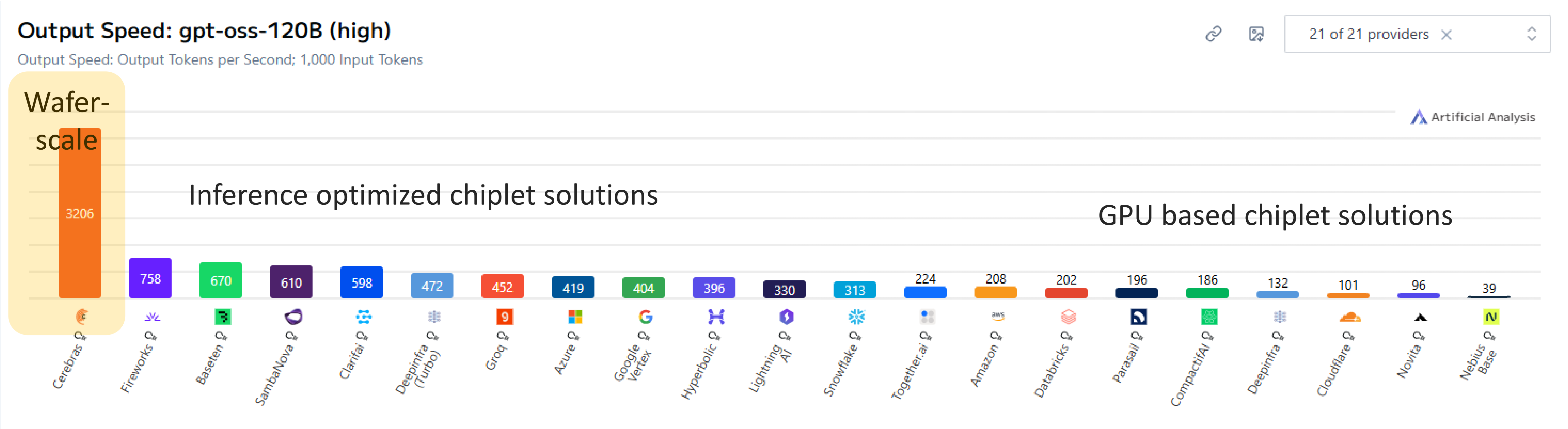
Wafer scale integration – industry examples

SPEED matters

<u>Speed unlocks new business models for AI</u>

This was true for the internet 25 years ago and it's just as true for AI.

When the internet was slow, Netflix mailed DVDs in envelopes,
today it’s a movie studio.



Output Speed: gpt-oss-120B (high)

Output Speed: Output Tokens per Second; 1,000 Input Tokens

21 of 21 providers

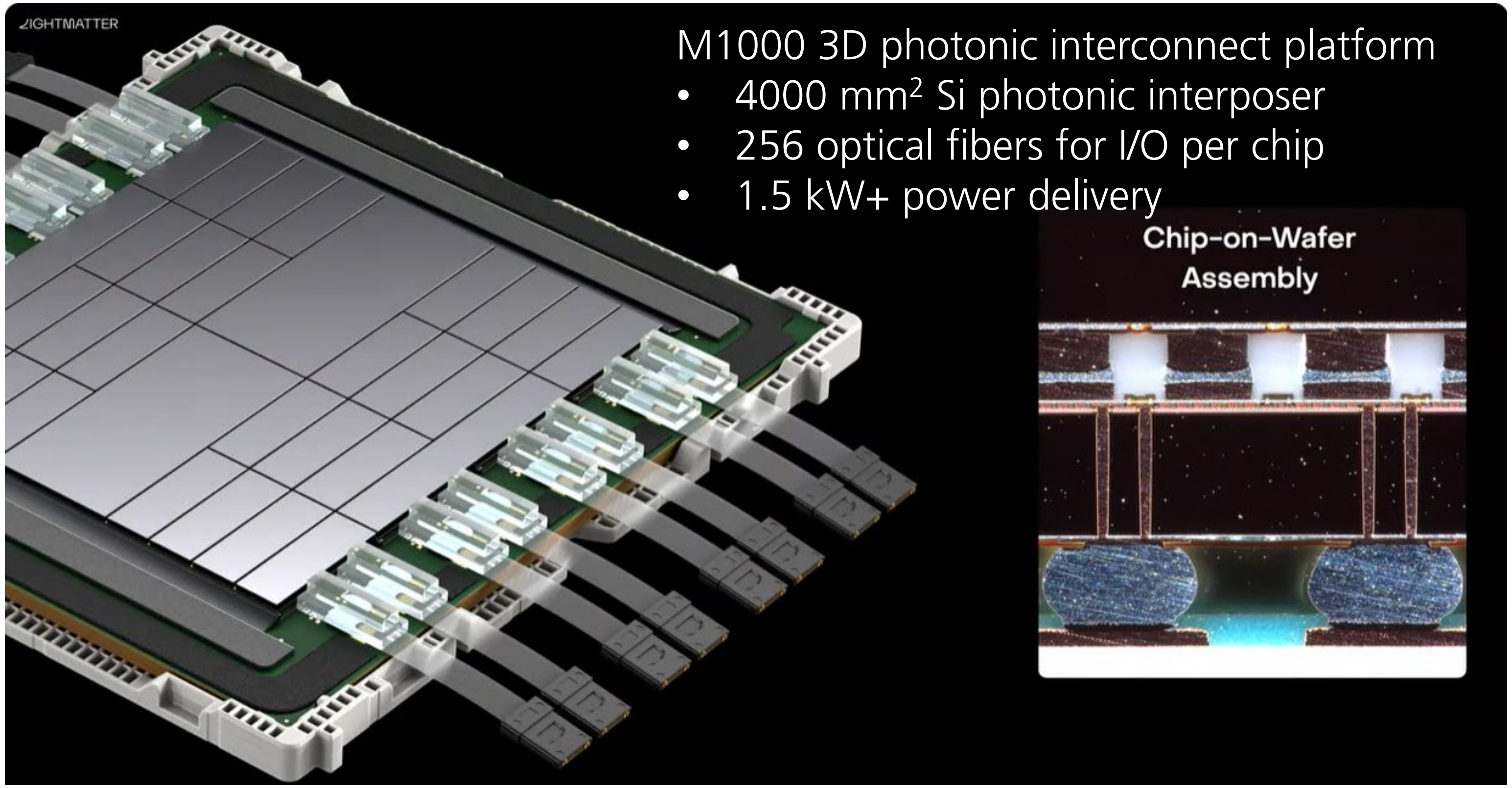
Wafer-scale

Inference optimized chiplet solutions

GPU based chiplet solutions

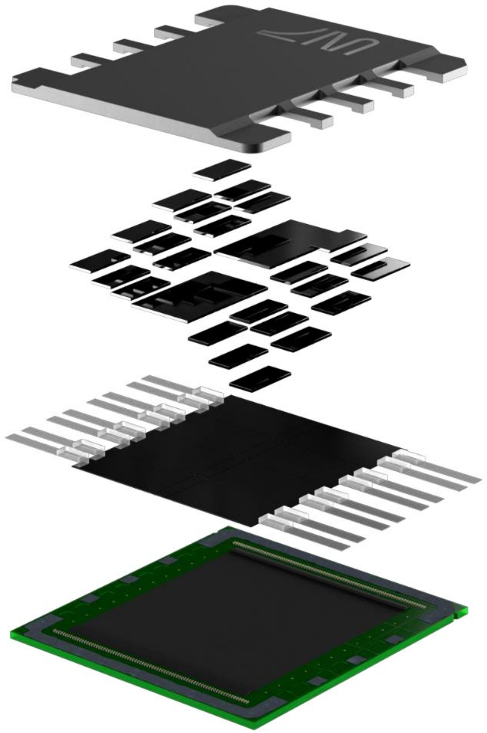
Wafer scale integration – industry examples

3D Co-packaged silicon photonics - Passage^{TM} by Lightmatter



M1000 3D photonic interconnect platform

- 4000 mm^2 Si photonic interposer
- 256 optical fibers for I/O per chip
- 1.5 kW+ power delivery



LID

XPU ASICs

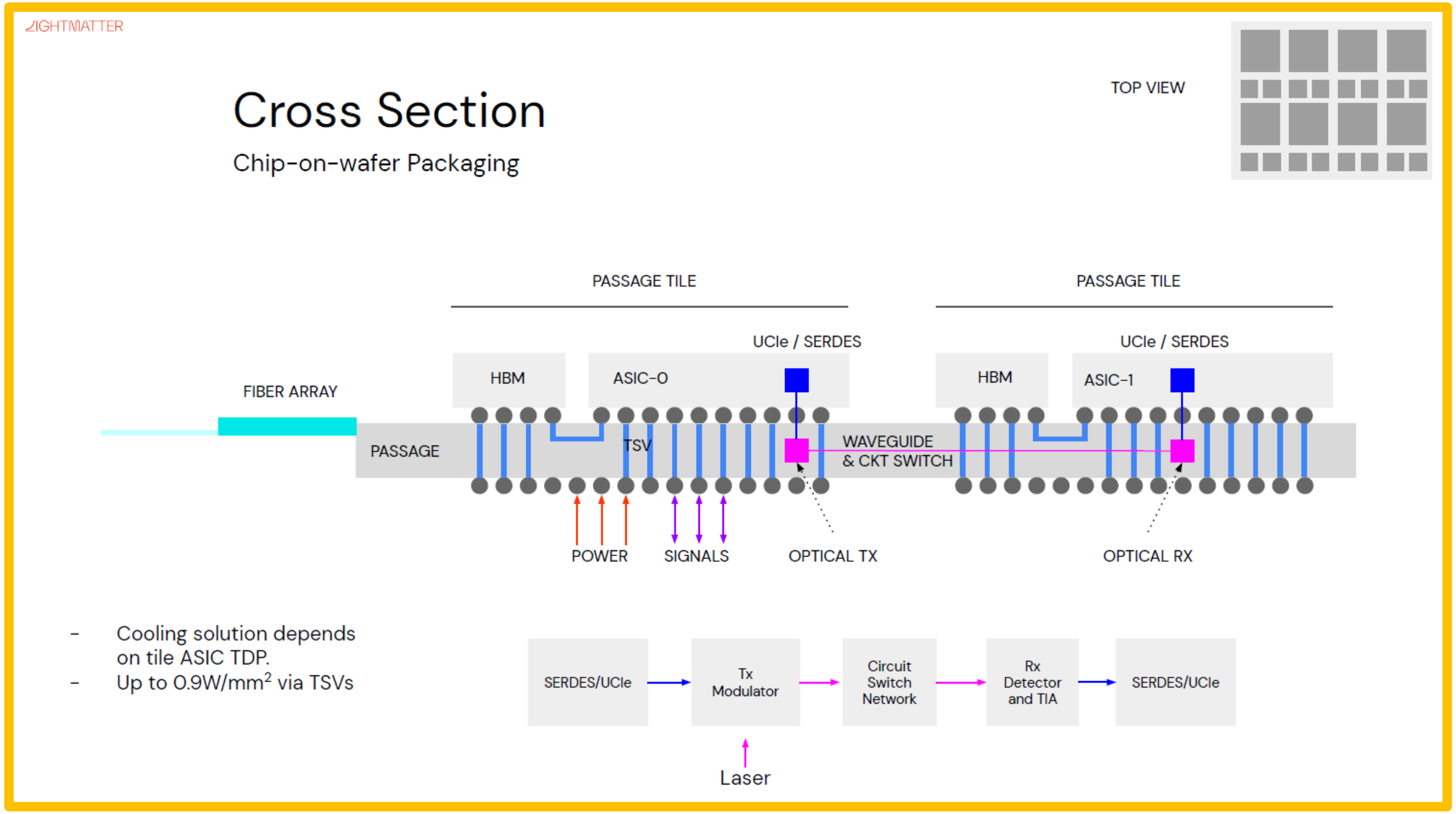
Fiber array Passage silicon

Organic substrate to power

Source: Lightmatter 04/2025

Wafer scale integration – industry examples

3D Co-packaged silicon photonics - Passage^{TM} by Lightmatter



LIGHTMATTER

Cross Section

Chip-on-wafer Packaging

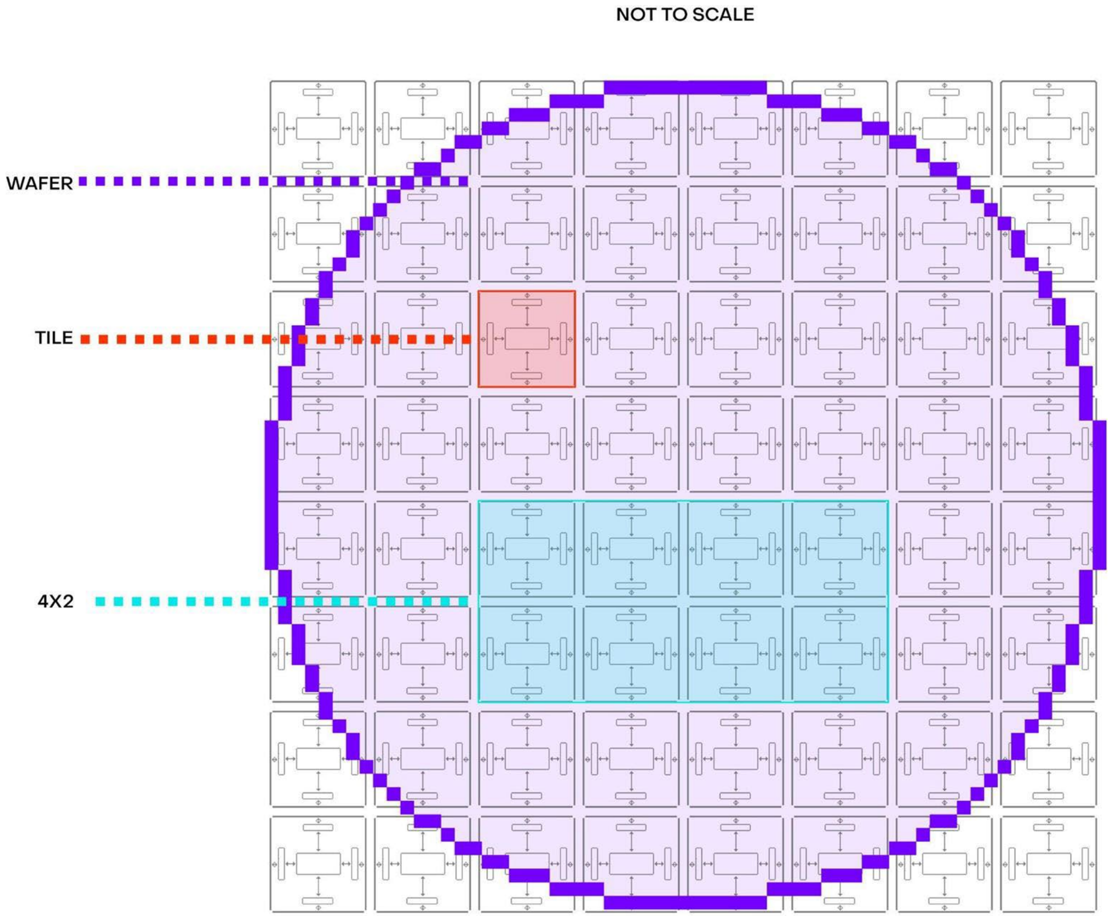
- Cooling solution depends on tile ASIC TDP.
- Up to 0.9W/mm^2 via TSVs

Source: Lightmatter

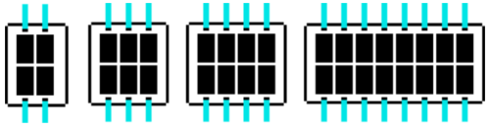
- Built-in solid state optical circuit switching
- Cross-reticle stitching
- Integrated transistor and photonics control technology to work with custom XPUs.

Wafer scale integration – industry examples

Wafer-scale silicon photonics interposer - Passage^{TM} by Lightmatter



Uniform architecture allows flexible dicing based upon end application

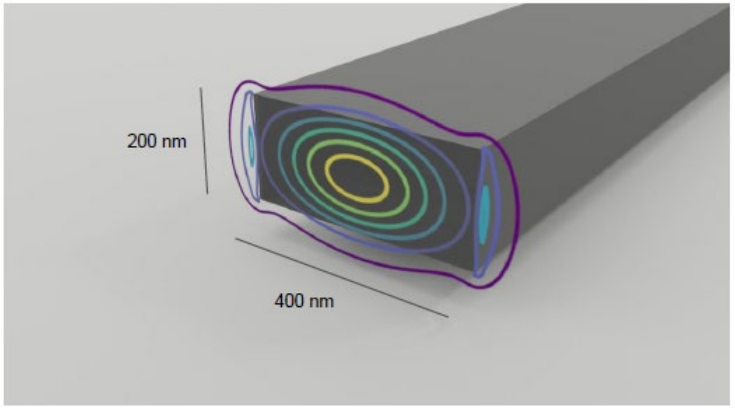


Inter-reticle optical die to die communication

&

Die to wafer packaging

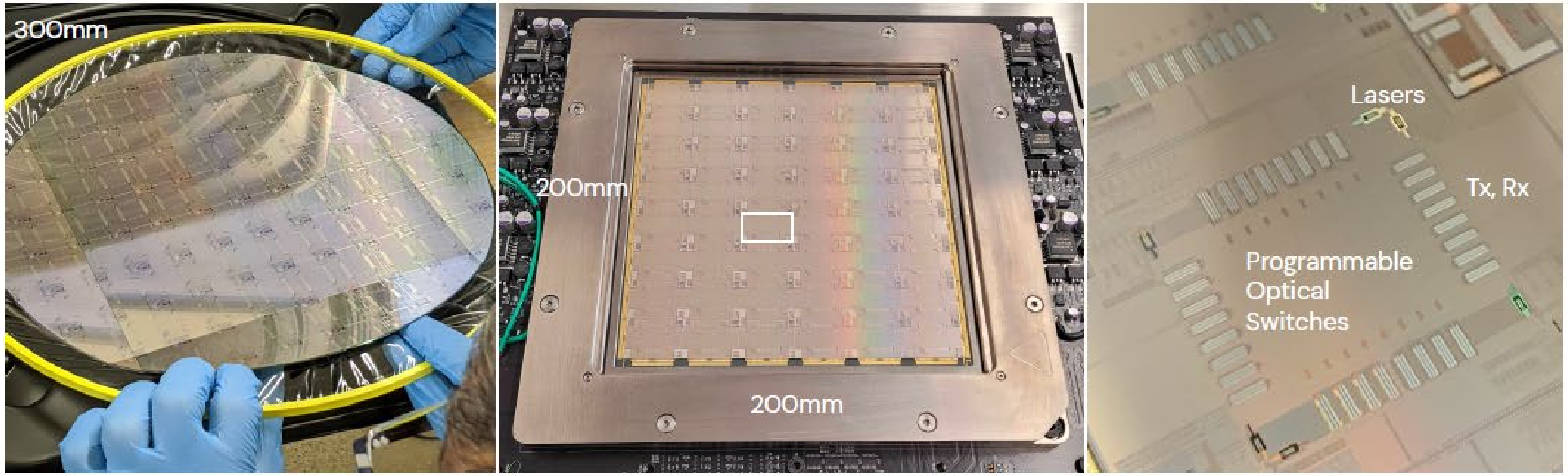
300mm CMOS Fab



Source: Lightmatter

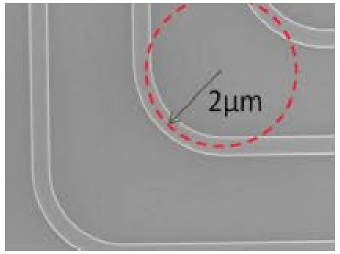
Wafer scale integration – industry examples

3D Co-packaged silicon photonics - Passage^{TM} by Lightmatter



Passage™ Alpha Silicon

- <50 Watts
- 32 channels per site, 1.024 Tbps
- 32 Gbps per channel NRZ
- 48 x 800mm^2 tiles
- 288x 50 mW Lasers
- 6,144 DACs
- 6,144 MZIs
- 150,000 photonic components
- JTAG interface
- Integrated Lasers, transistors, photonics
- Programmable interconnect topologies

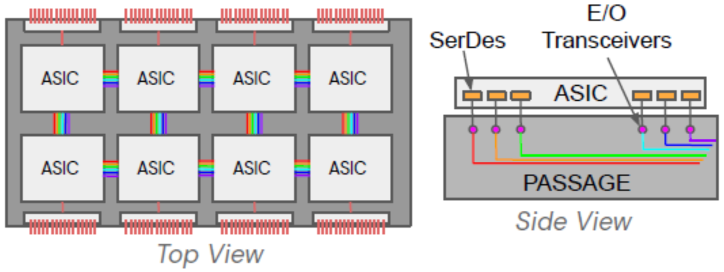


Photonic waveguides with ~4 µm pitch.

Worlds 1st wafer scale programmable photonic interconnect fabric

- Enables low power (<2.3 pJ/bit) and latency (<5 ns) for data communication between custom compute ASICs
- Enables very high bandwidth for communication (~114 Tbps for 2x4 passage tiles)
- Leverages 3D D2W approaches and kind of wafer scale integration

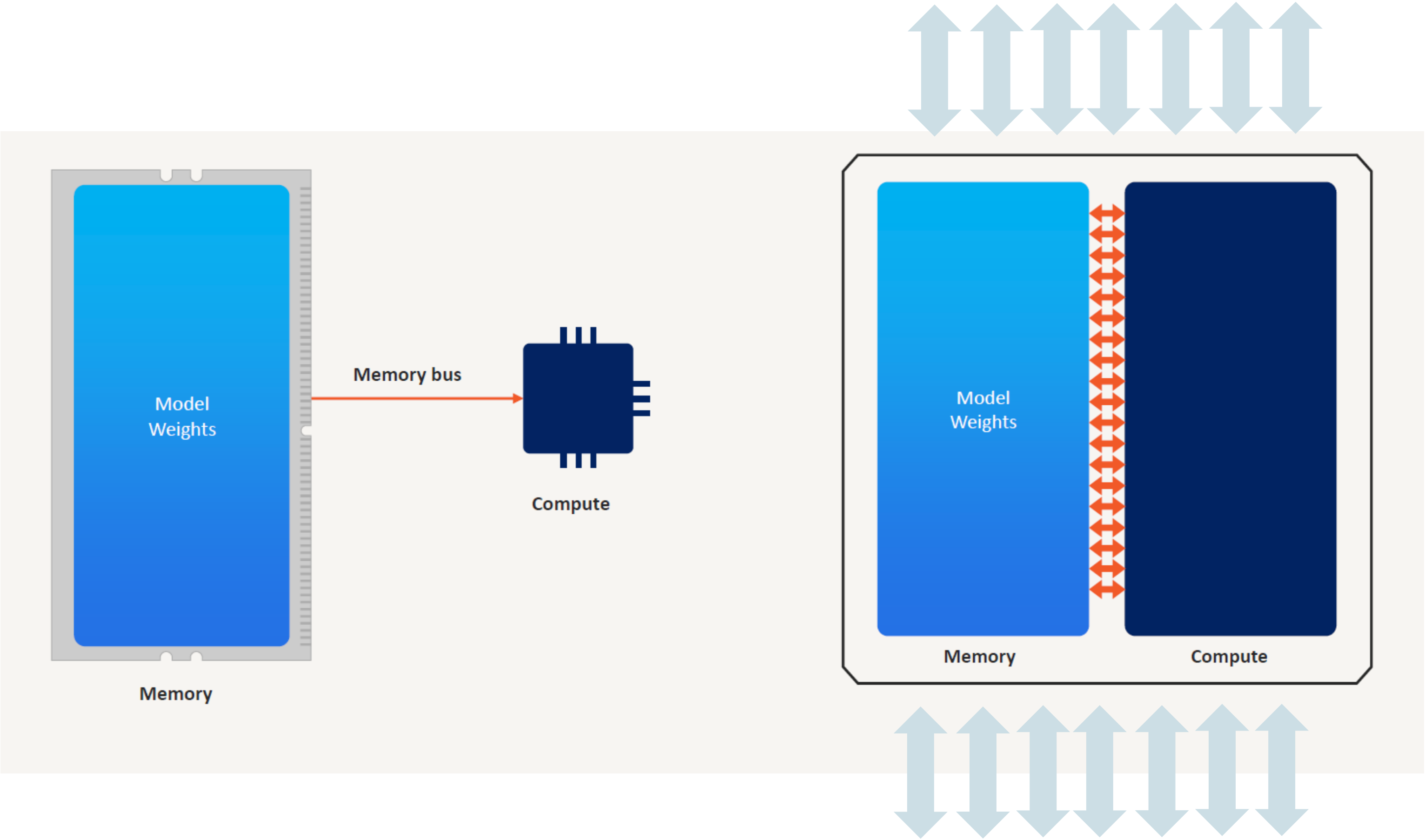
Passage



Source: Lightmatter

Wafer scale integration

Hyperscaling connectivity



Communication happens at the chip perimeter, **but there is not enough shore line**

FLOPS	∝ Chip Area (L^2)
I/O Bandwidth	∝ Chip Perimeter (L)

↓

I/O to Compute Performance Gap

↓

Explore the 3rd dimension

Wafer scale integration – how to contribute

Hyperscaling connectivity

How to overcome the I/O to compute performance gap → 3D wafer-on-wafer packaging

1. Think the entire system	→ STCO: system technology co-optimization
2. Provide on-silicon inter-reticle connectivity on large areas	→ Advanced i-line lithography for 3D-packaging
3. Provide solutions to integrate power delivery	→ 3D power delivery (e.g. by TSVs, eDTC, eIVR)
5. Provide solutions to enable 3D co-packaged optics	→ explore 3rd dimension with 3D silicon photonics
6. Provide solutions to integrate cooling	→ explore 3rd dimension for 3D cooling in the stack
7. Provide solutions to extend memory	→ explore 3rd dimension and stack memory wafers → use advanced memory technologies
8. Next technological hurdle ….	

Wafer scale integration – how to contribute

300mm wafer-level building blocks

→ 3D wafer-level processing

- 3D quasi-monolithic integration – 3D-QMI
- Large area advanced-line lithography (sub µm scale)
- Through silicon via integration (TSV mid, TSV last)
- Double sided process
- Deposition technologies CVD, PVD, ALD

→ Advanced wafer thinning, pre-assembly and singulation

- <10 µm thin wafer and thin die processing
- BEOL layer transfer technology
- Bonding / De-Bonding
- Stress free chiplet singulation

→ Multi die assembly at state of the art pitches

- D2W hybrid bonding sub-µm accuracy
- D2W re-configured wafer for 3D-QMI
- Damascene processing
- Mix-pitch assembly with mix interconnect technologies

APECS CEASAX PREVAIL

→ In-line metrology and test

- Defect inspection at CDs <1 µm
- CD and OVL characterization at <1 µm
- Planarity, material properties
- Electrical characterization

→ 3D Wafer-Stacking

- Compute / Memory / Power delivery wafer stacks
- Hybrid bonding W2W <200nm accuracy
- Integrated cooling
- Optical I/O communication

Co-funded by the European Union

Bundesministerium für Forschung, Technologie und Raumfahrt

STAATSMINISTERIUM FÜR WISSENSCHAFT KULTUR UND TOURISMUS

Freistaat SACHSEN

Conclusion

Andrew Feldman, Founder and CEO - Cerebras Systems

“Progress doesn’t come from safety rails.

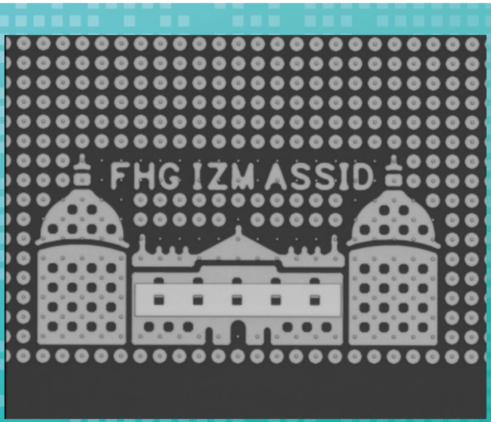
It comes from putting sharp tools in people’s hands
- and trusting them to use them.”

Fraunhofer IZM can help with using very sharp tools for 3D over the entire system value chain from 200/300mm wafer to advanced substrates

Contact

Fraunhofer IZM

Fraunhofer Institute for Reliability
and Microintegration IZM



Dr. Frank Windrich
Deputy head of IZM-ASSID
Wafer Level System Integration (WLSI)

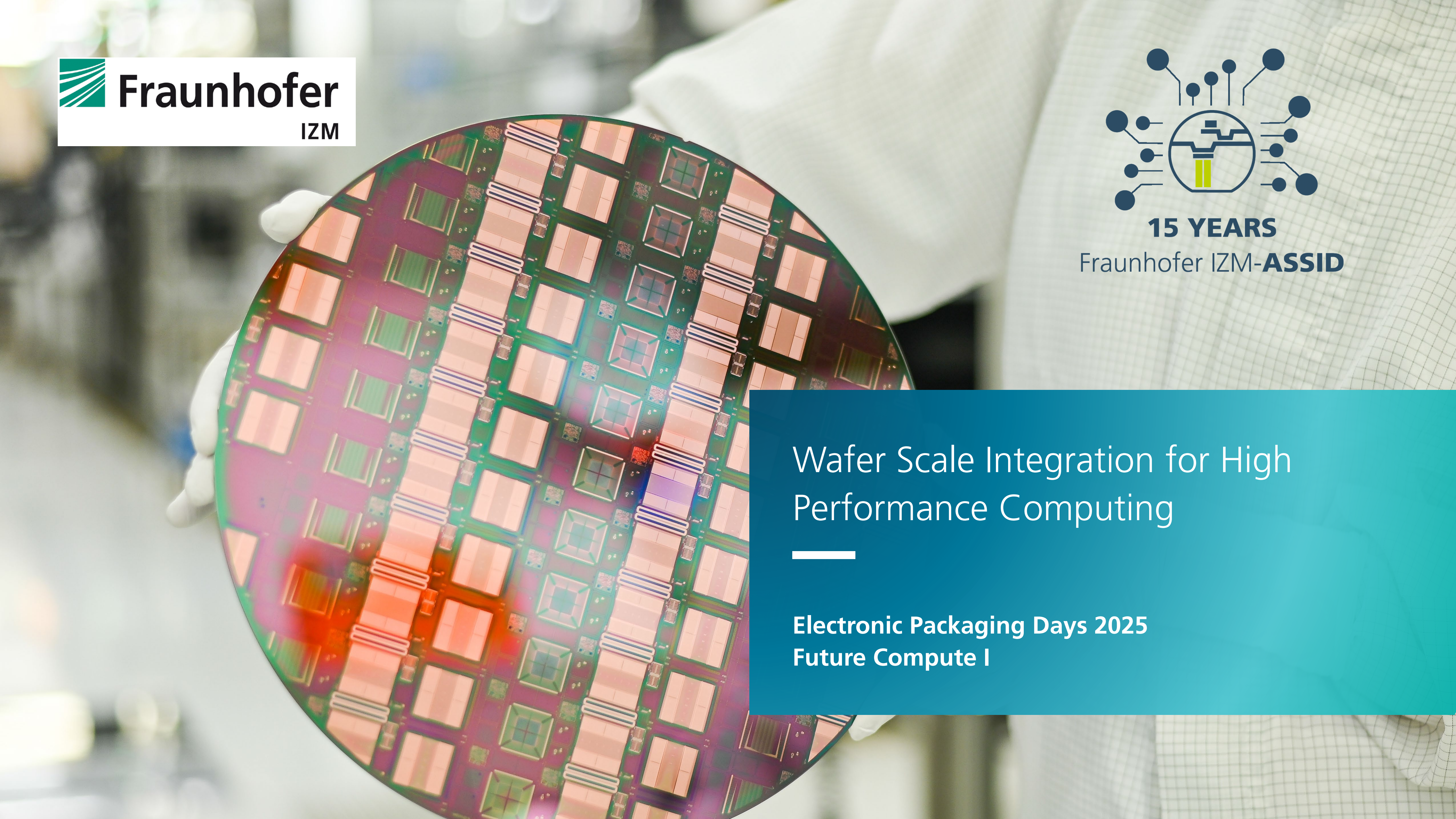
Phone: +49 351 795572 - 49
E-Mail: frank.windrich@assid.izm.fraunhofer.de

Fraunhofer IZM Berlin
Gustav-Meyer-Allee 25
13355 Berlin
Germany
+49 30 46403-100

Fraunhofer IZM-ASSID
Ringstraße 12
01468 Dresden-Moritzburg
Germany
+49 351 795572-12

Fraunhofer IZM Außenstelle Cottbus
Karl-Marx-Straße 69
03044 Cottbus
Germany
+49 355 383 770-12

Fraunhofer
IZM



15 YEARS
Fraunhofer IZM-ASSID

Wafer Scale Integration for High Performance Computing

Electronic Packaging Days 2025
Future Compute I