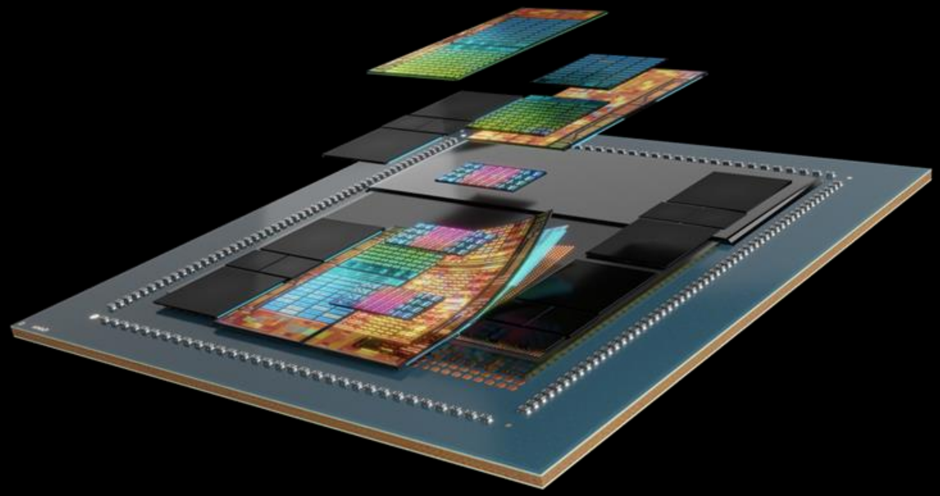
ENABLING HETEROGENOUS INTEGRATION THROUGH CHIPLET ARCHITECTURES



HEMANTH DHAVALESWARAPU

ADVANCED PACKAGING ARCHITECT, AMD

Cautionary statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products; TAM for data center, PCs, embedded and gaming; and technology trends, innovation and roadmaps, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD’s Securities and Exchange Commission filings, including but not limited to AMD’s most recent reports on Forms 10-K and 10-Q.

AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

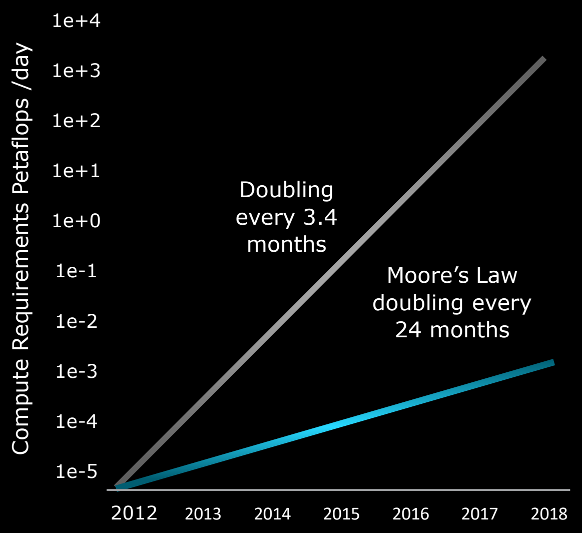
FUNDAMENTAL INFLECTION OF AI COMPUTING

Benefits are fundamental to solving the world's most pressing problems

AI DRIVING COMPUTE & MEMORY GROWTH

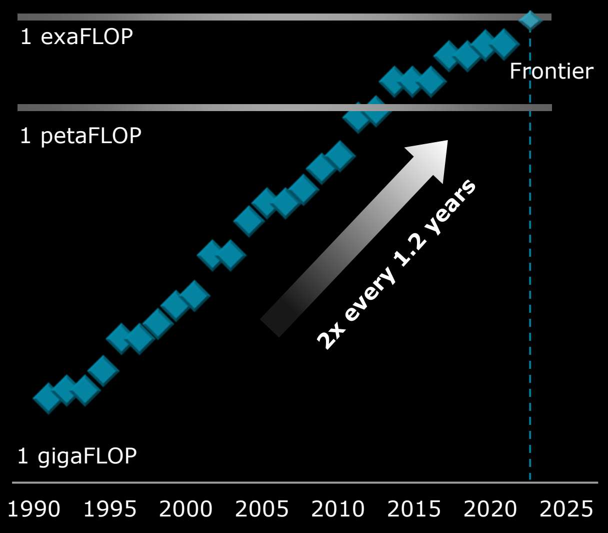
AI COMPUTE

AlexNet to AlphaGo Zero: A 300,000X Increase in Compute



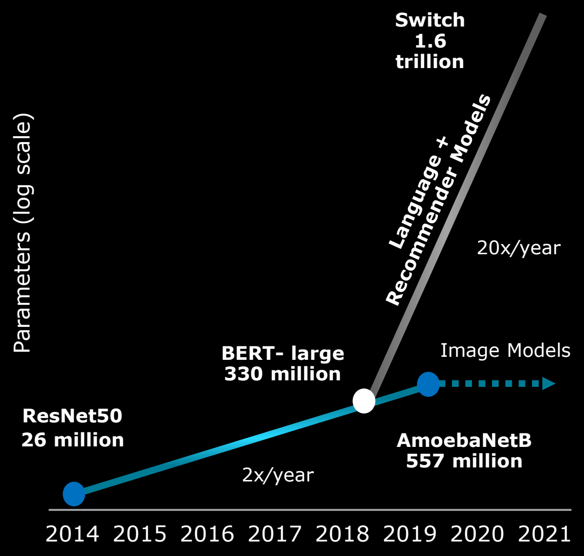
Source: https://blog.openai.com/ai-and-compute

LEADING SUPERCOMPUTER PERFORMANCE



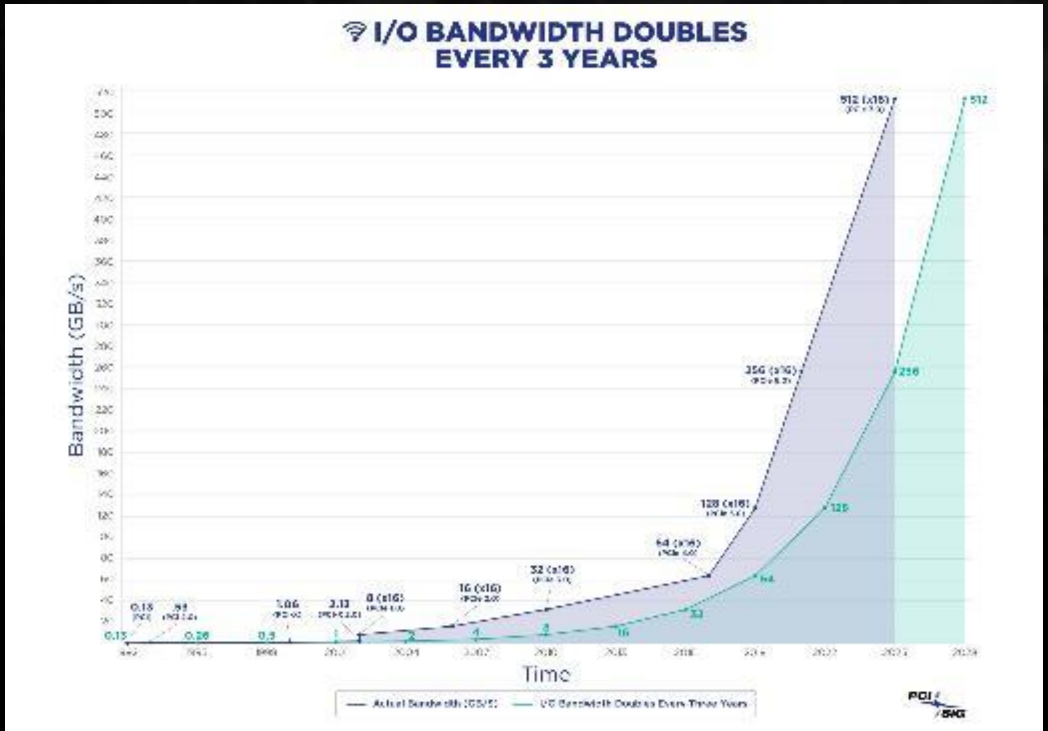
Source: https://www.top500.org

AI MEMORY



Source: TechInsights (2022)

IO BANDWIDTH GROWTH



[1]

Signal speeds upwards of 100G for next gen accelerators. Driving adequate SI at these bandwidths is requiring several disruptions:

- Substrate routing resources taxed → larger modules, thinner cores to optimize via resistance losses
- Advancing PCB capabilities to reduce in board signal loss
- Avoiding through core routing either through cabling or optical interconnects

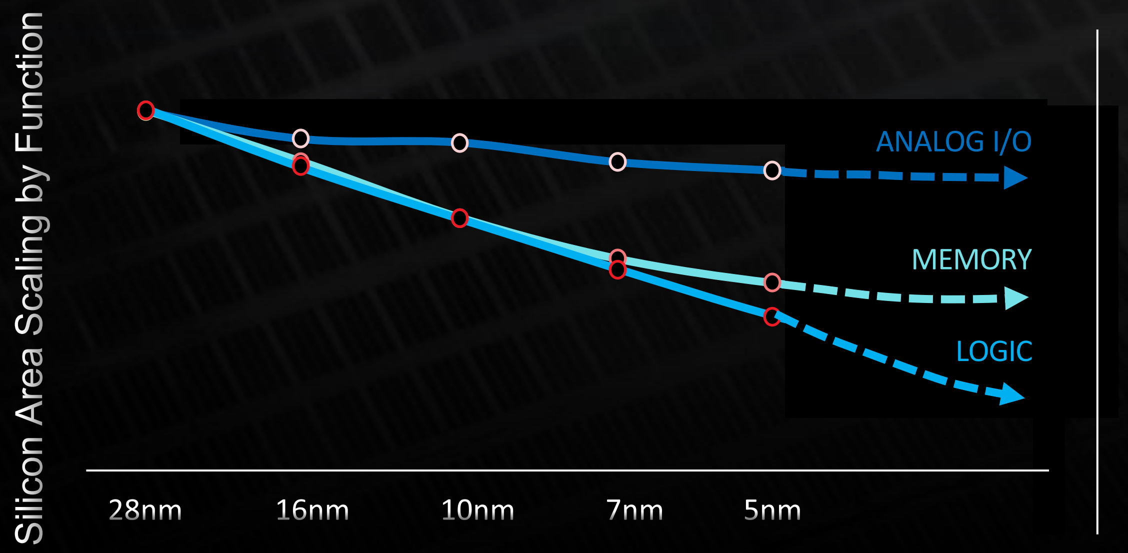
All these solutions bring new quality & cost challenges

[1] PCI-SIG, or Peripheral Component Interconnect Special Interest Group 2022

- Silicon growth, memory integration and high-speed signal needs require large substrate/module sizes
- Substrate size growth will come with capacity, yield, and other tooling challenges

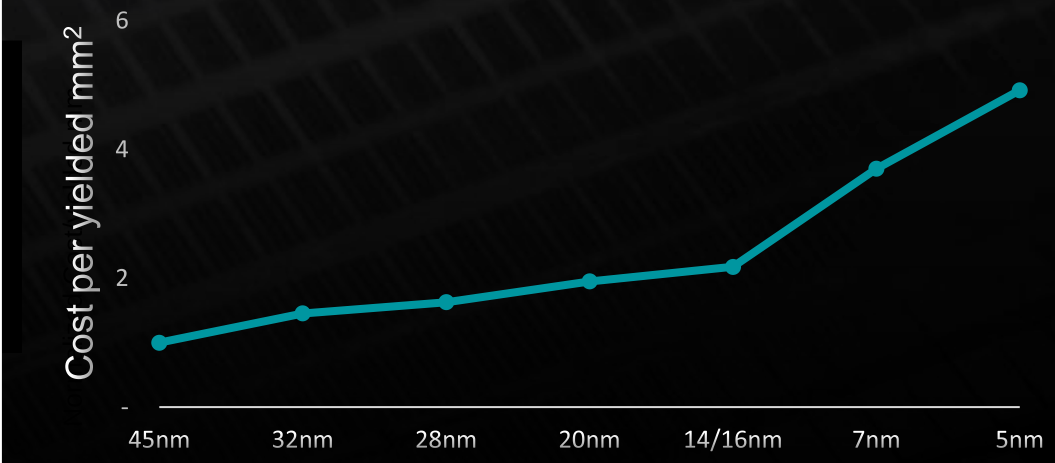
TRADITIONAL SCALING CHALLENGES

Limited and Divergent Scale Factors [1]



Density Gains Diminishing

Increasing Costs[2]



Costs Increasing

Moore’s Law slowing down and cost to add number of transistors/chip is increasing

[1] Naffziger, VLSI short Course, 2020, [2] Cost per yielded mm2 for a 250mm2 die

CHIPLET TECHNOLOGY

AMD leadership

2015 — 2.5D HBM

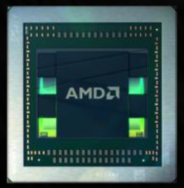
2017 — MULTICHIP MODULE

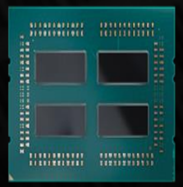
2019 — CHIPLETS

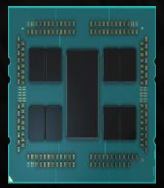
2021 — 3D CHIPLETS — 2.5D EFB

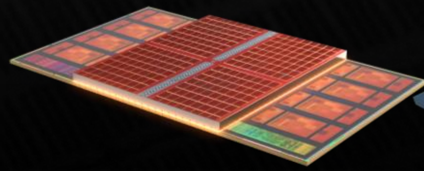
2022 — 2.5D WLFO

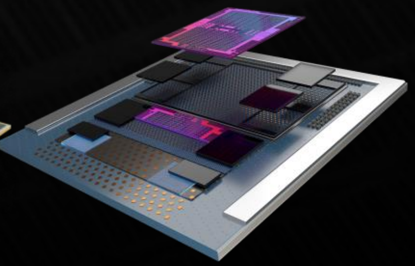
2023 — 3.5D

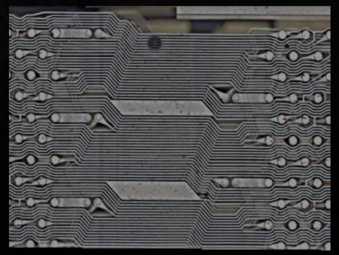










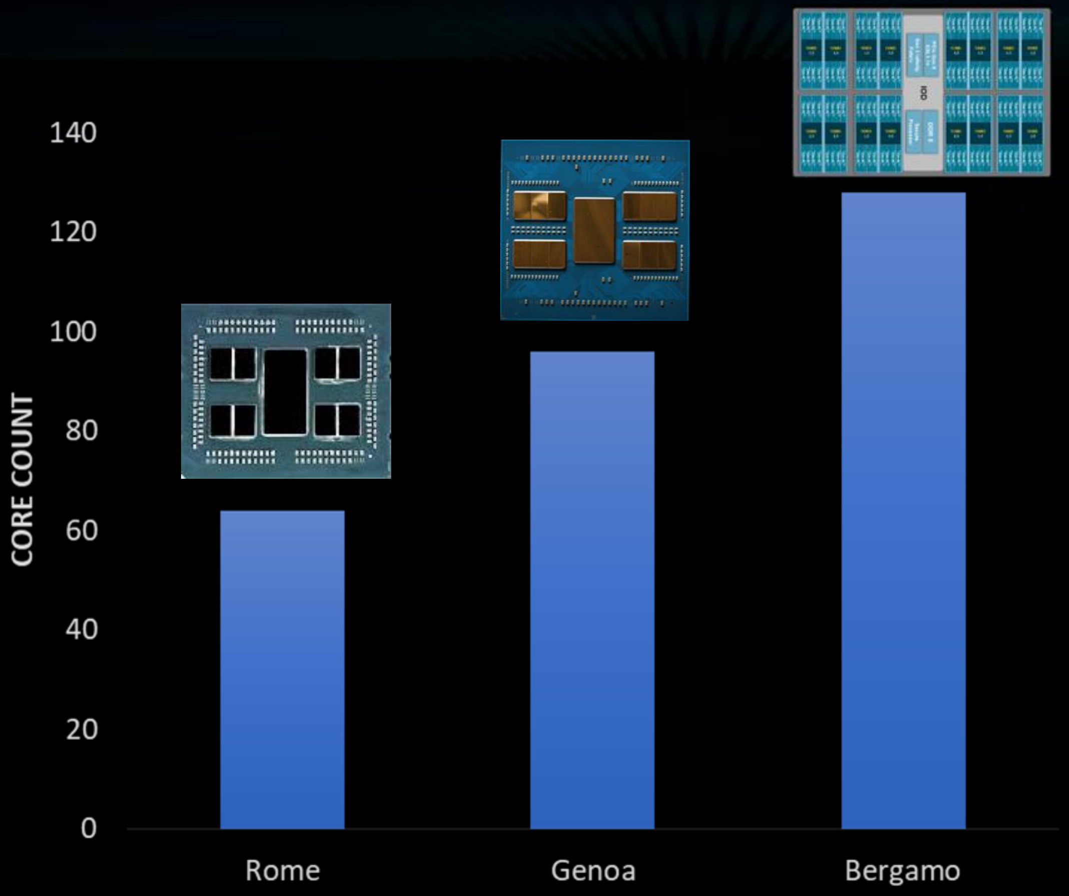


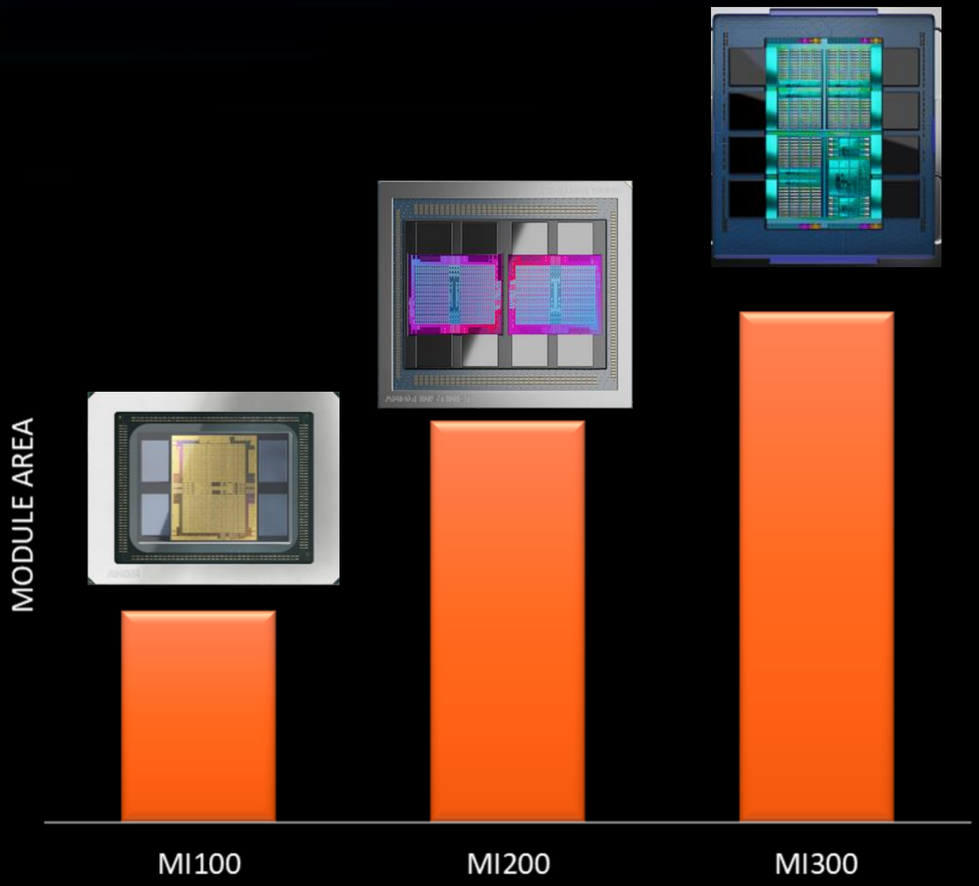
Led Industry in Chiplet Architecture

New Era of Products driven by Aggressive 3D/2.5D Integration Roadmap

Continuous innovation in horizontal and vertical scaling to meet compute needs

INNOVATION AROUND MOORE’S LAW

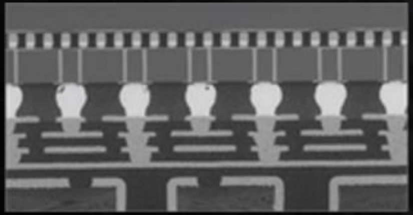




Chiplet technology allows to drive higher core count in CPUs and more heterogenous integration in GPUs

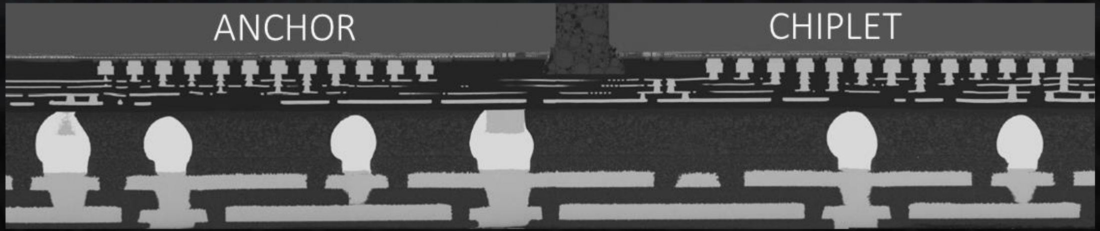
ADVANCED PACKAGING TECHNOLOGIES

2.5D Si INT



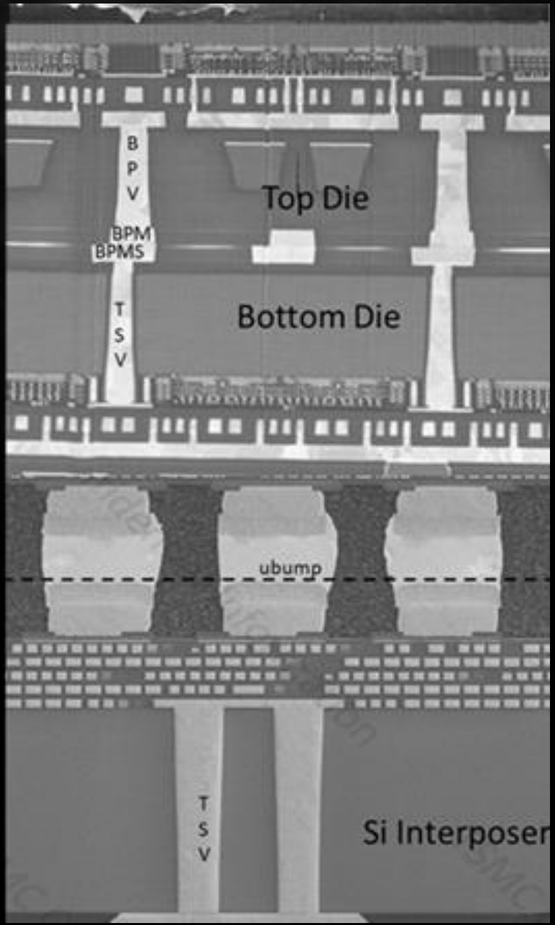
AMD Fiji GPU

2.5D Die First WLFO



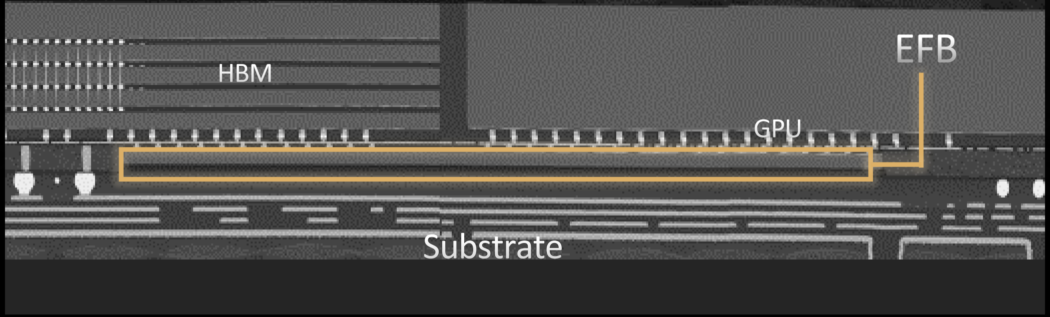
AMD Navi31

3.5D



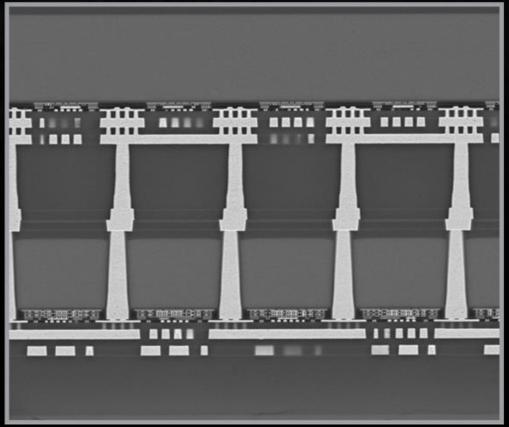
AMD MI300

2.5D Die Last EFB



AMD MI200

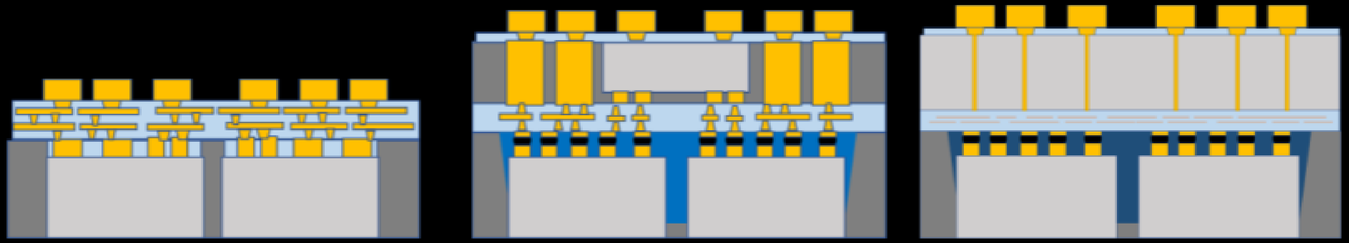
3D



AMD 3D V-cache

Optimal choice based on product Performance, Power, Area and Cost

2.5D ARCHITECTURE COMPARISON



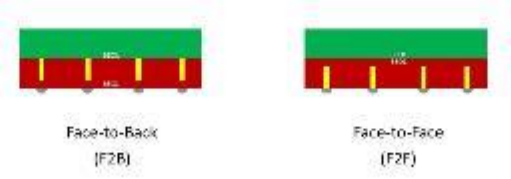
Architecture	2.5D Die First WLFO	2.5D Die Last EFB	Si Interposer
Fan Out	RDL only	RDL+Si Bridge	Si only
Min. Pad Pitch	35-45um		
Areal Interconnect Density (wires/mm2)	500-800		
Linear Density (wires/mm/layer)	220-250 (RDL feature limited)	500-1000 (silicon scales better)	
Interconnect Power (pJ/bit)	~0.2-0.3	~0.2-0.3	~0.3
Cost			

AMD
together we advance_packaging

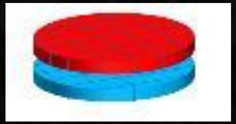
3D HYBRID BONDING

Direct bonding (Fusion/Hybrid)

F2F, F2B (B2B in multi-hi stacks)



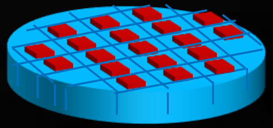
Wafer on Wafer
200k-500k wires/mm²



Wafer to wafer (W2W)

Chip on Wafer
12k-30k wires/mm²

Die to wafer (D2W)

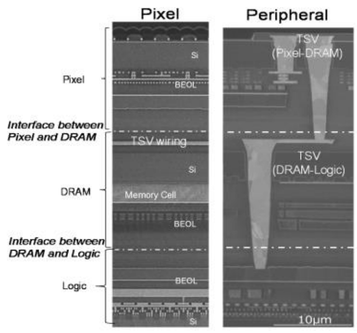


CIS (Sony, TSMC)

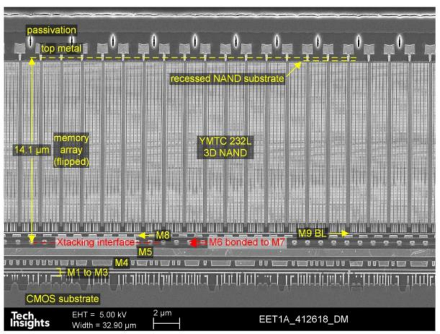
3D NAND (YMTC)

3D DRAM (Tezzaron)

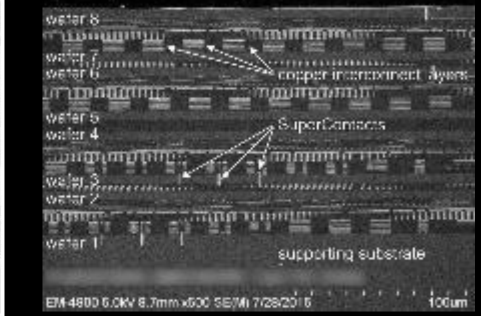
AMD/TSMC



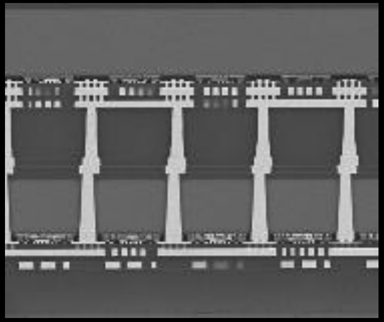
3-hi stack



2-hi stack

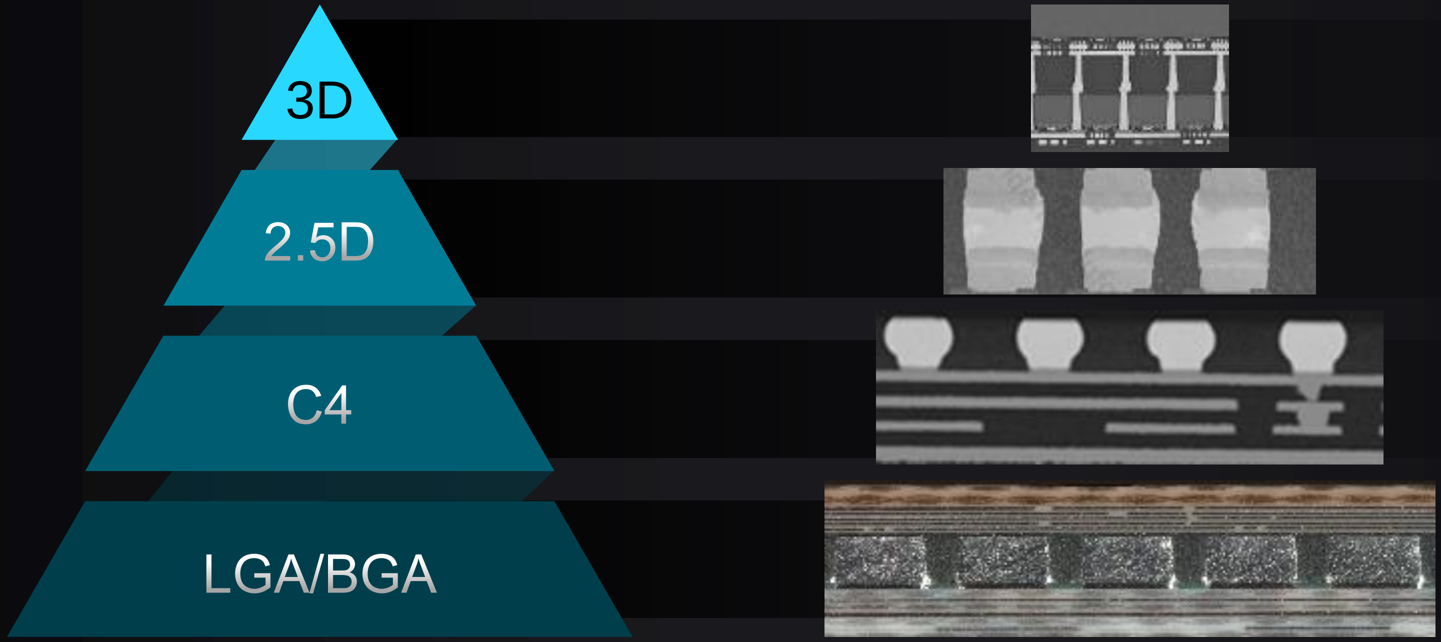


8-hi stack



2-hi stack

INTERCONNECT SCALING



<10um Pitch

<50um Pitch

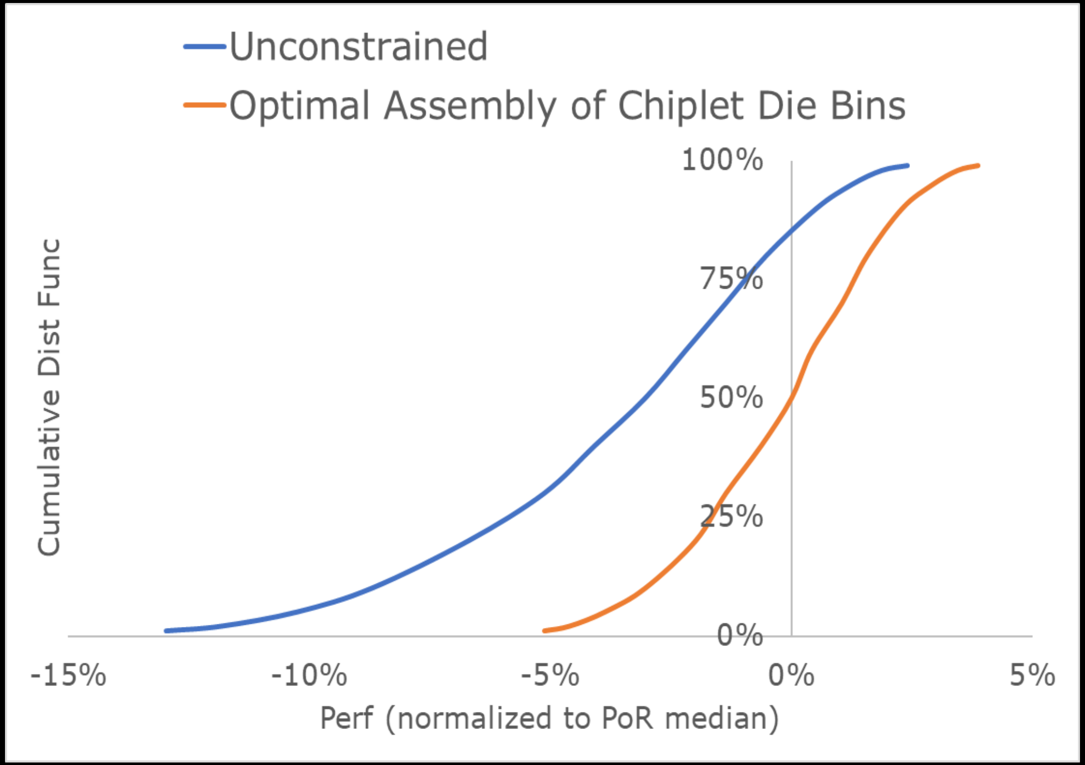
<150um Pitch

<1000um Pitch

MULTI-SCALE DESIGN/PROCESS OPTIMIZATION NEEDED TO MANAGE WARPAGE AND INTERCONNECT RELIABILITY

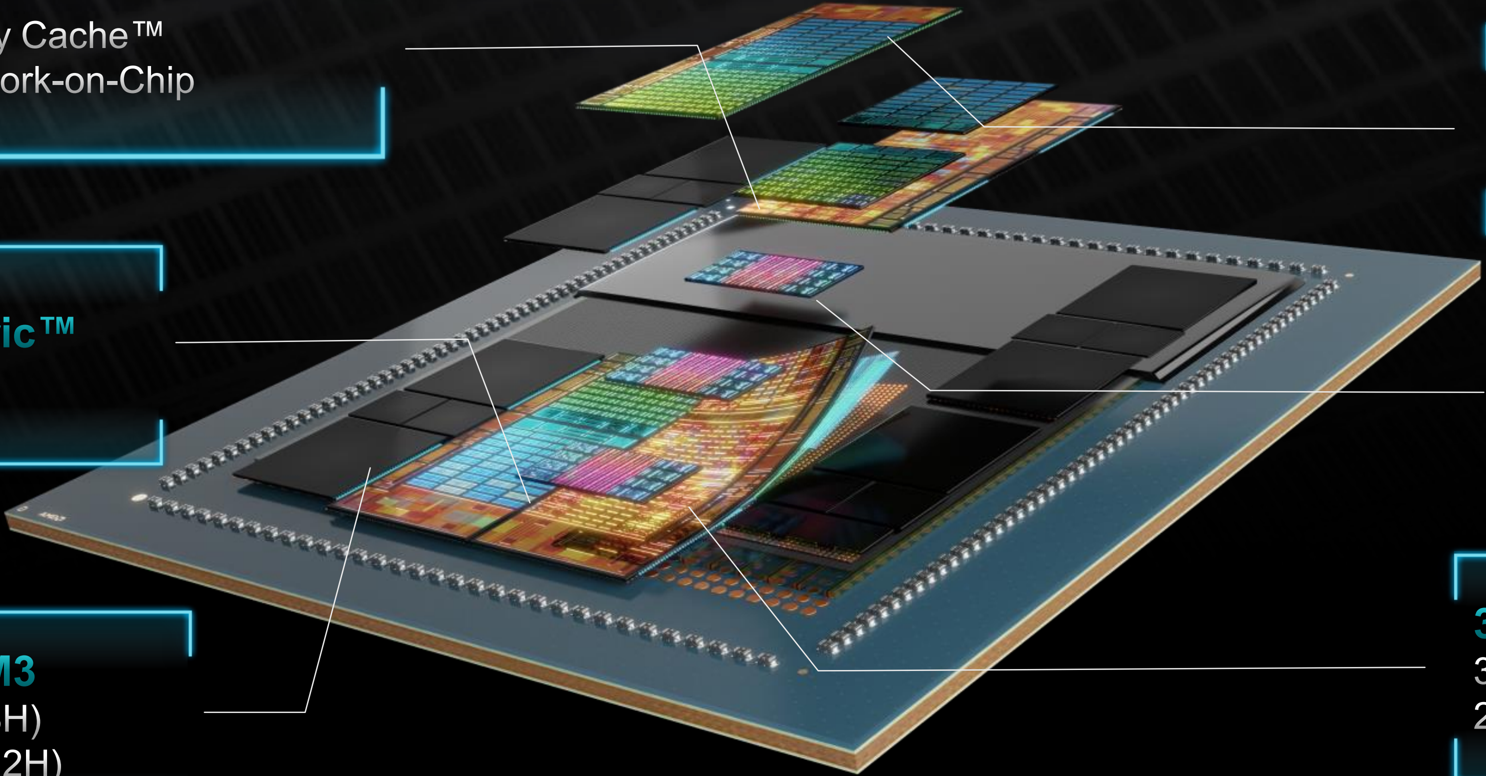
TECHNOLOGY, PACKAGING, AND TEST

- Chiplet and advanced packaging technologies have enabled us to build large and complex data center processors at high yields.
- Traditional product test and binning at die-chiplet level is enhanced with innovative assembly instruction viz., optimize packaging of dies from similar or different bin(s) depending on ease of voltage scaling at component level, power sloshing between components, etc.
- Significant improvement in final processor delivered performance and tightening of performance variation across process/manufacturing distribution window is achieved.



AMD INSTINCT™ MI300 SERIES

Key Innovations



I/O Die (IOD)
256MB AMD Infinity Cache™
Infinity Fabric Network-on-Chip

AMD Infinity Fabric™
AP Interconnect

8 stacks of HBM3
MI300A: 128 GB (8H)
MI300X: 192 GB (12H)

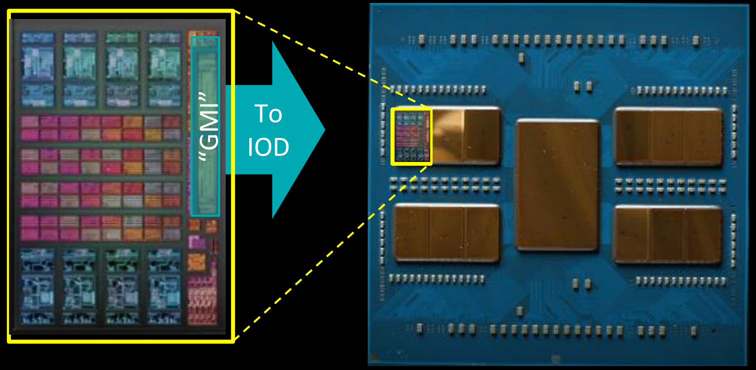
Accelerator Complex Die (XCD)
6X38 AMD CDNA™3 Compute die

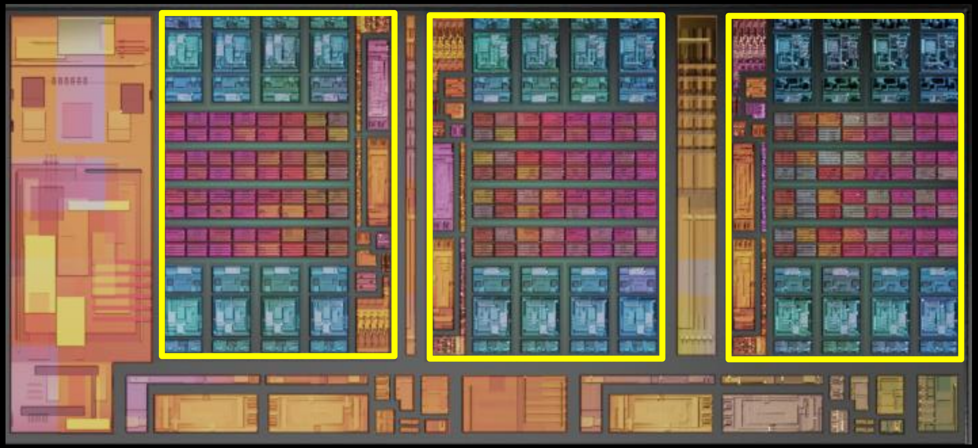
CPU Complex Die (CCD)
3 x 8 “Zen 4” Cores

3.5D Package
3D hybrid bonding
2.5D silicon interposer

CHIPLET REUSE AND MODULARITY BENEFITS EXEMPLIFIED

- Same CCD adapted to work for 4th Gen EPYC™ CPUs and AMD Instinct™ MI300A 3D stack
  - EPYC™ MCM uses "GMI" SerDes interface through package substrate
  - AMD Instinct™ MI300A vertical stack uses dense TSV interface from IOD to CCD in two-link 'wide' mode
  - Dramatically higher 3D signal density enabled virtually no die size increase with simple interface multiplexing

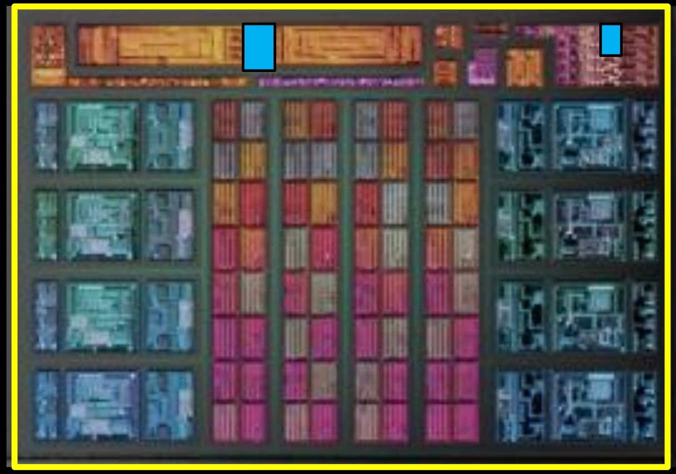




Original CCD for Genoa only

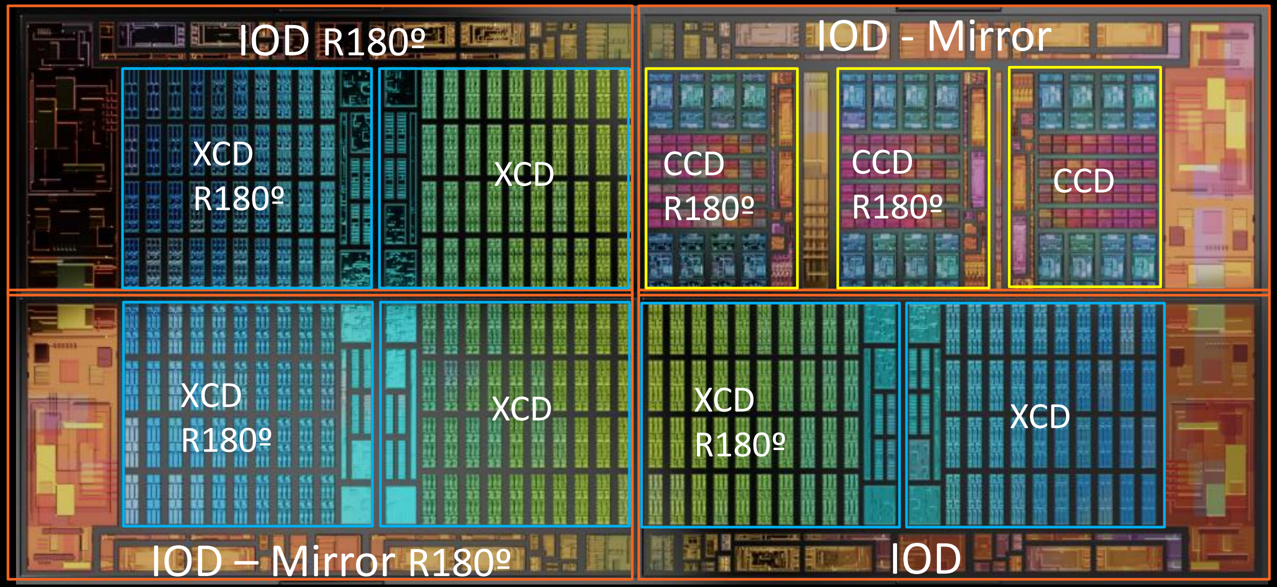
Same CCD for Genoa + MI300A

3D Interface to IOD



AMD INSTINCT™ MI300 ACCELERATOR

MODULAR CONSTRUCTION

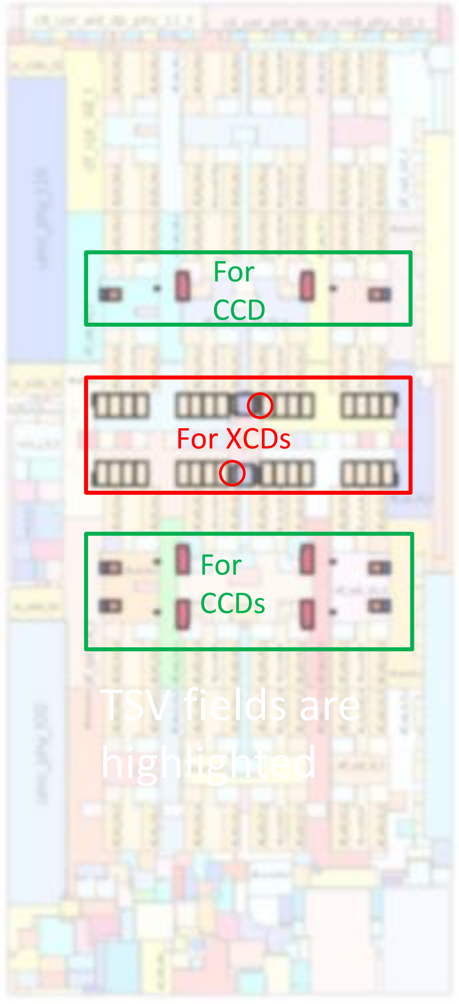


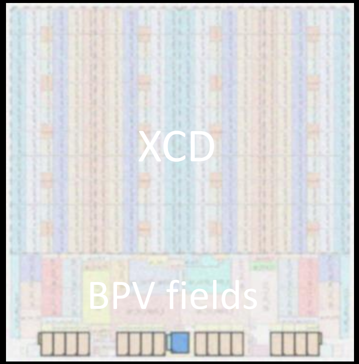
- Multi-variant (APU/XPU) architecture requires all chiplets to act as if they are LEGO blocks
- Many new construction and analysis tools needed to be developed to enable this capability
- Mirrored versions of the IODs enable symmetric construction

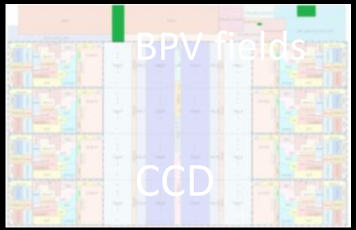
CONNECTING CHIPLETS IN 3.5D

MIRRORED HETEROGENEOUS CHIPLET INTERFACES

- BPV: Bond Pad Via. The landing site on the stacked die that is aligned with TSV in IOD
- IOD Supports 2 separate landing sites for CCD BPVs to enable IOD mirroring while CCDs can only be rotated (not mirrored)
- Similarly, XCD/IOD interface also had extra TSVs to support IOD mirroring (red circle)



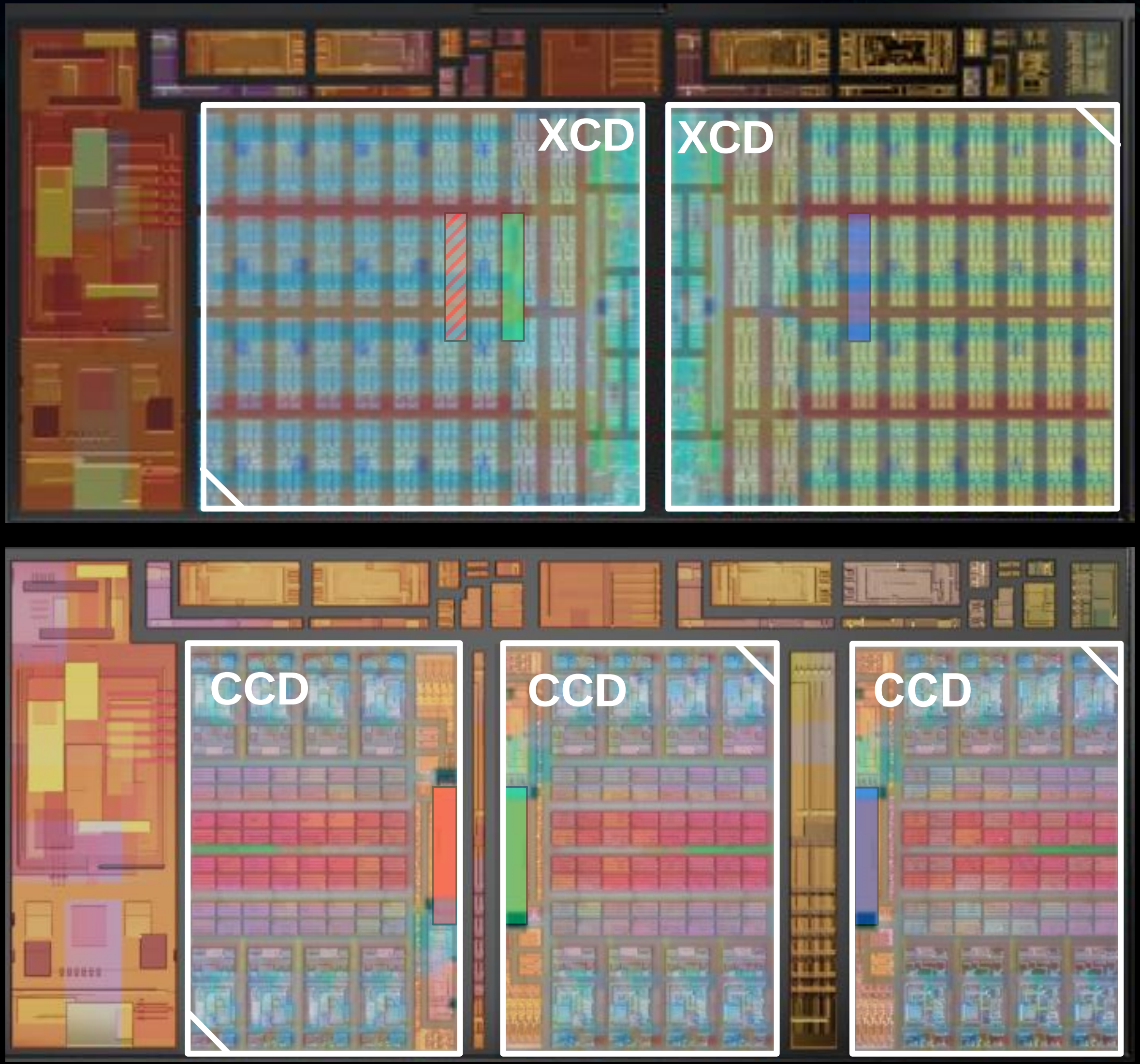


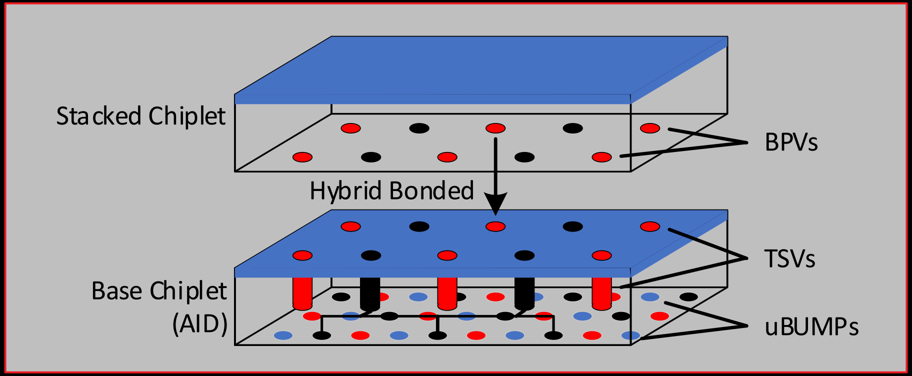


AMD Instinct™ MI300 Accelerator

Floorplan – Power TSVs

- Power delivery to top die must support
  - IOD mirroring
  - XCD /CCD rotation (0 and 180 degree)
  - Different stacked die (CCD and XCD)
- This placed new symmetry requirements on power grid
- Significant advanced planning to ensure exact alignment of all power and ground TSV+BPVs

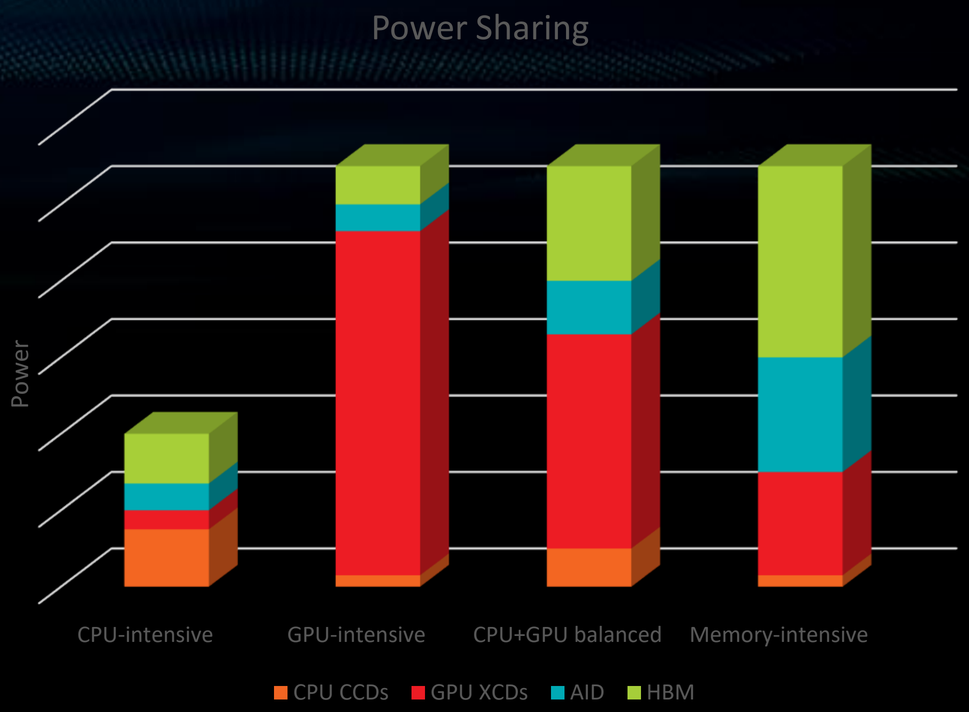


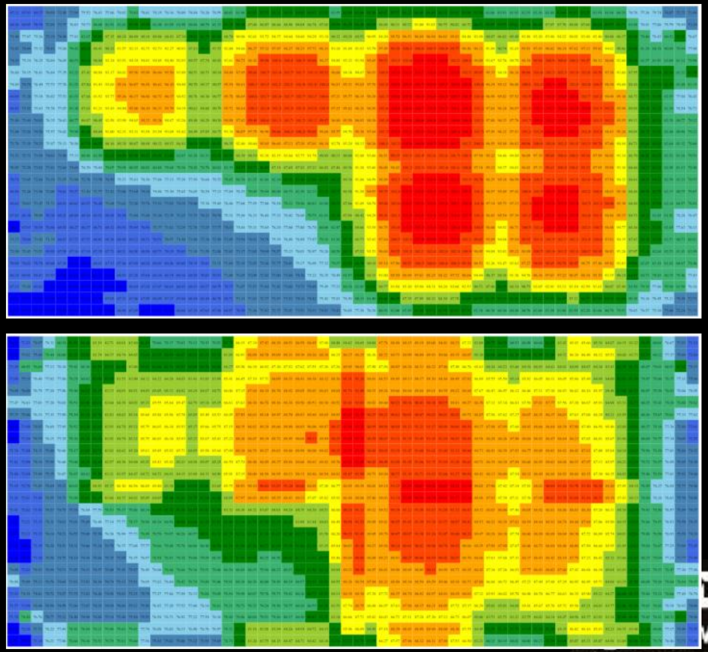


AMD INSTINCT™ MI300 ACCELERATOR

POWER MANAGEMENT AND HEAT EXTRACTION

- Key to MI300 power efficiency is the ability to dynamically “slosh” power between fabric (IOD), GPU (XCD) and CPU (CCD)
- Massive HBM and Infinity Cache bandwidth can drive high data movement power in the SOC domain
- Compute capability can similarly consume high power
- This creates 2 types of extreme operating conditions -- GPU-intensive and Memory-intensive
- Both thermal and power delivery must support the full range – careful engineering of TSVs and power map

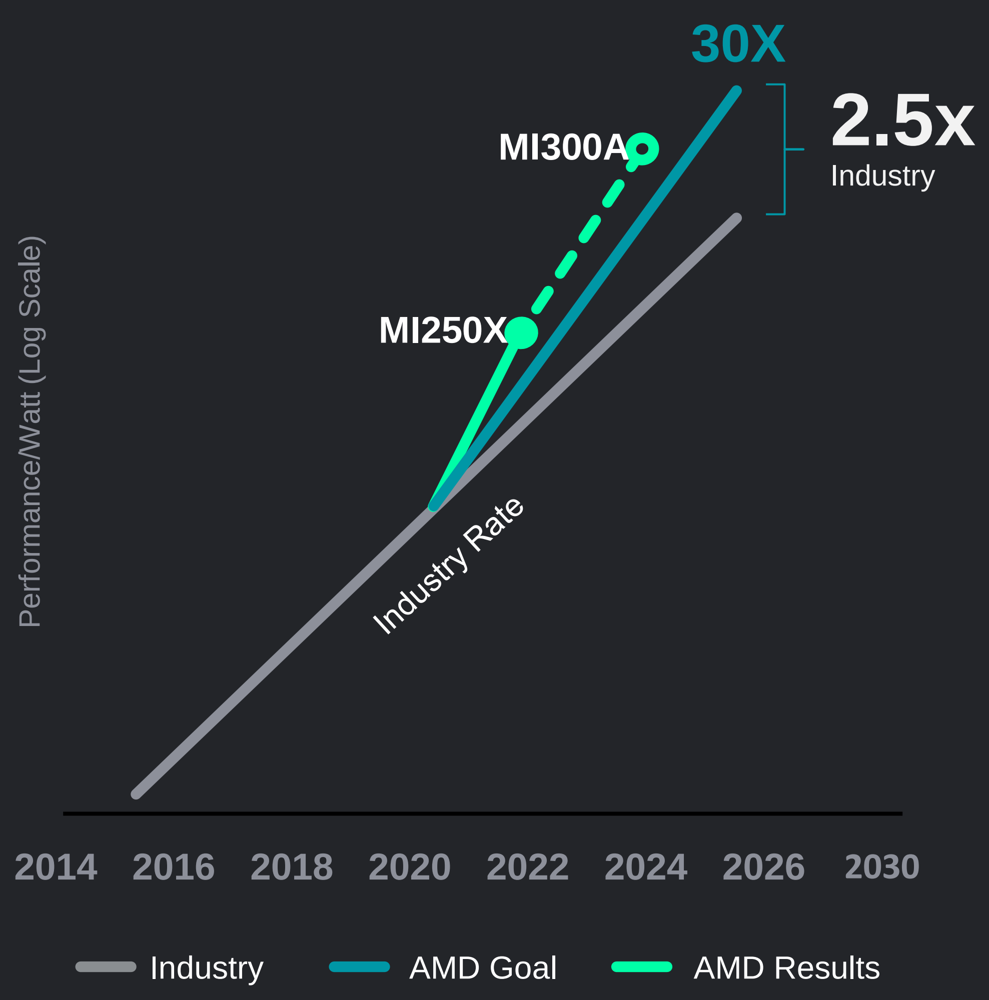




Resulting Node-Level Efficiency Gain

- AMD roadmap on track to exceed aggressive 30x goal
- Architecture, packaging and interconnect innovations pay off
- Chiplet and 3D-enabled architecture put AMD Instinct™ products on a path to exceed 30x goal

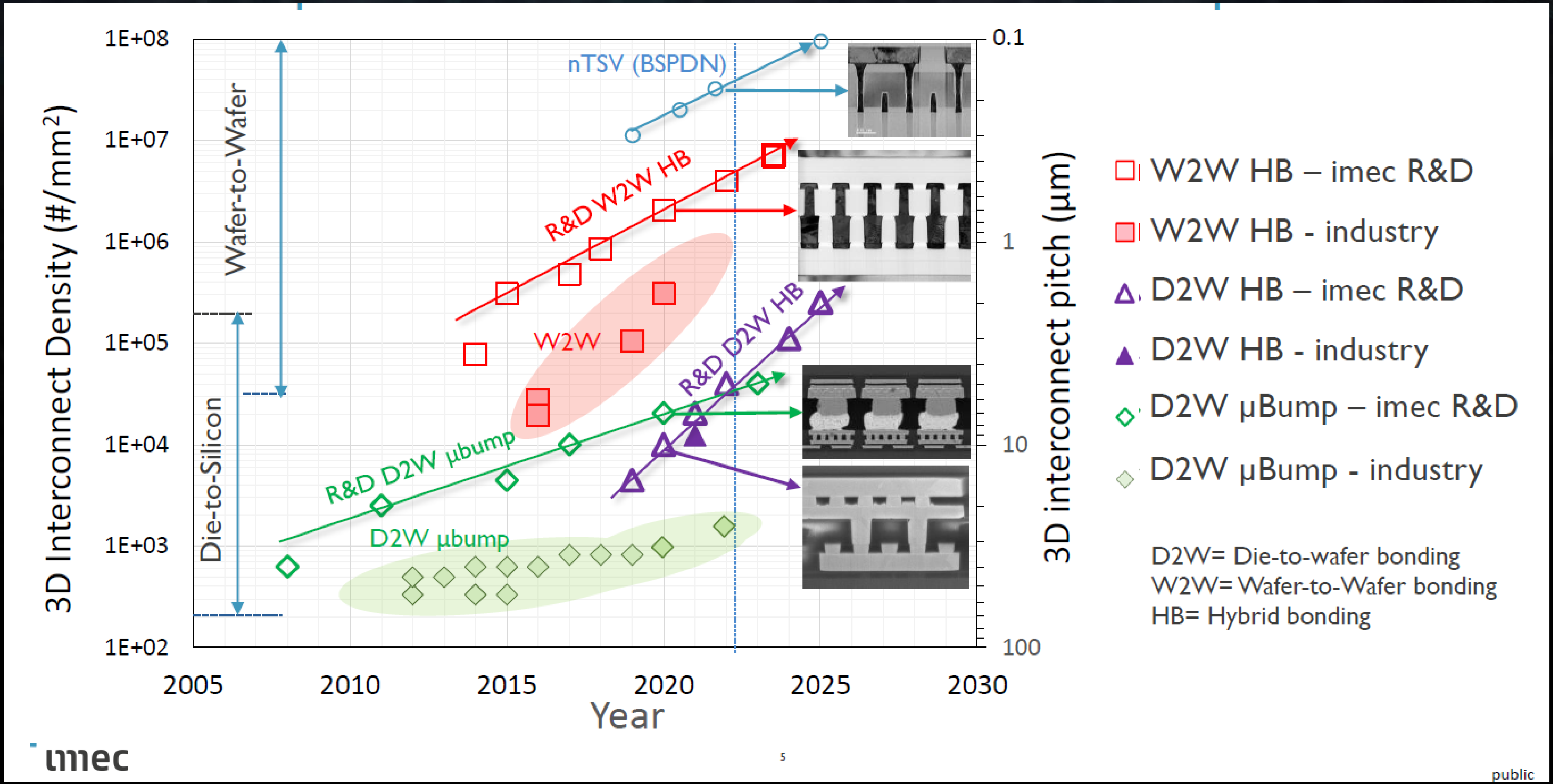
Accelerated computing performance/watt trends



Based on 2015-2020 industry trends in energy efficiency gains and data center energy consumption in 2025.

*Includes AMD high performance CPU and GPU accelerators used for AI training and High-Performance Computing in a 4-Accelerator, CPU hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size. AI training: lower precision training-focused floating-point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.

3D Interconnects Roadmap



D2W= Die-to-wafer bonding
W2W= Wafer-to-Wafer bonding
HB= Hybrid bonding

imec

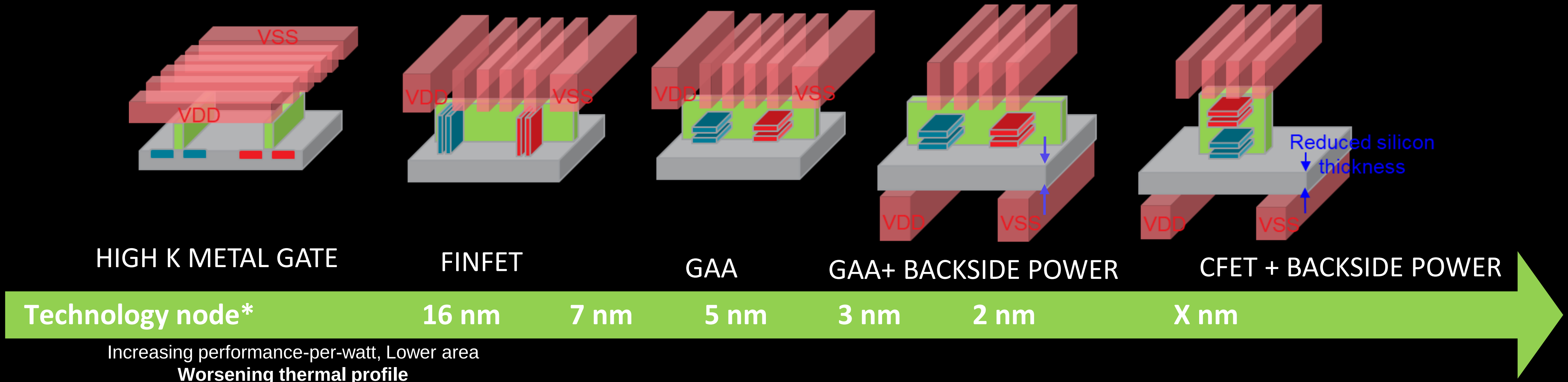
5

public

Source: IMEC public (https://www.imec-int.com/en/articles/view-3d-technology-landscape)

FUTURE OF TRANSISTOR

- More transistors on die
- Higher frequency of operation for more demanding applications
- More heat generated per transistor

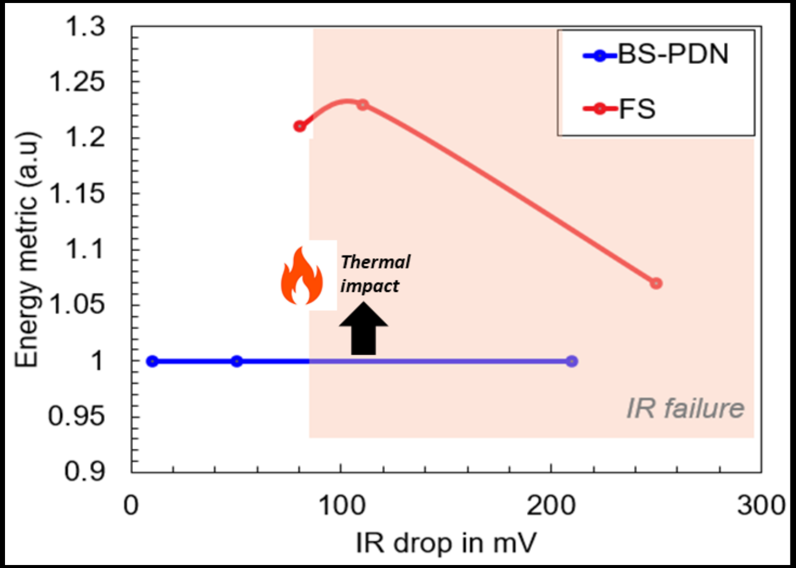


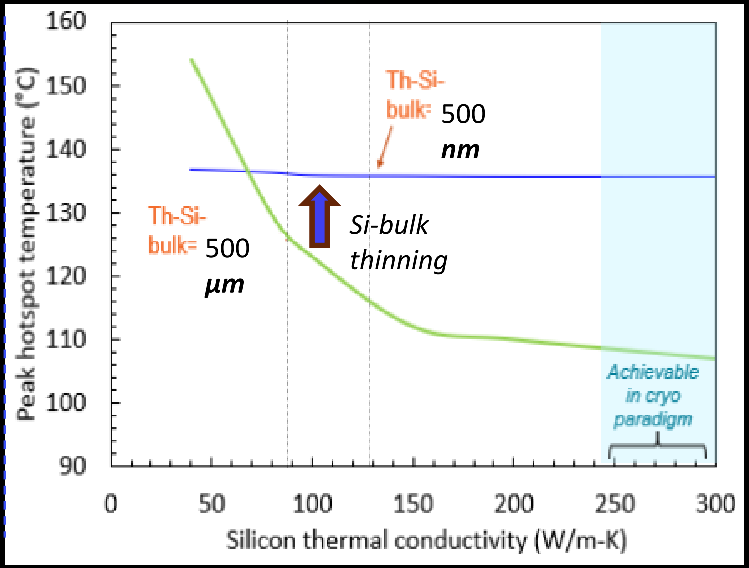
Increasing performance-per-watt, Lower area
Worsening thermal profile

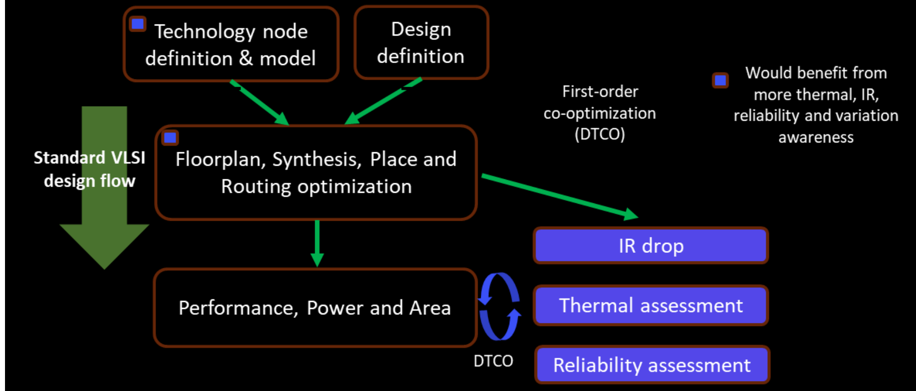
[1] IMEC chip scaling roadmap: smaller, better, faster
[2] J. Kim et Al., DAC 2021
[3] A. Shilov, AnandTech, Jun 16 2022
[4] J. Ryckaert, et Al., IEEE VLSI T

BACKSIDE POWER

- Assuming a heterogenous 2D design with back-side power with 500 nm Si-bulk
- The temperature hotspot can increase by ~1.2X, reducing performance by ~2-5%






Technology node definition & model
Design definition
Standard VLSI design flow
Floorplan, Synthesis, Place and Routing optimization
Performance, Power and Area
DTCO
First-order co-optimization (DTCO)
Would benefit from more thermal, IR, reliability and variation awareness
IR drop
Thermal assessment
Reliability assessment

FS: Front-side Power Delivery Network
BS-PDN: Back-side Power delivery network

[7] Y. Yu, et. Al., ITherm, 2017
[8] A. Choudhury et. Al, ASME
[10] D. Prasad, et. Al, IEDM 2019

Tightly coupled device and technology modelling and VLSI design flows are imperative going forward

TRANSISTOR TRENDS

- µArchitecture shift to multicore with the end of Dennard scaling, power limit, circa 2003.
- Circa 2020, enhancements via multi-design (efficiency vs Perf cores), multi-process node and flexible configurations (memory & core) in the new era of 3D chiplet packaging.

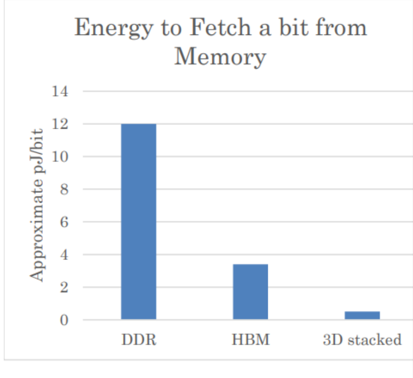


Fig. 6. Energy per bit for different memory types and integration approaches (AMD internal estimates)

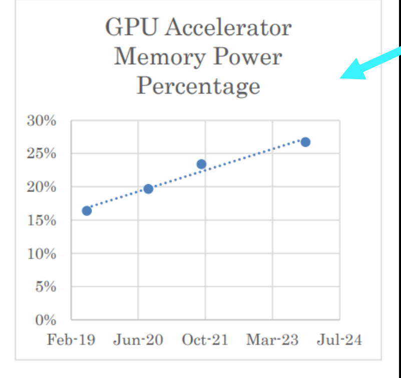
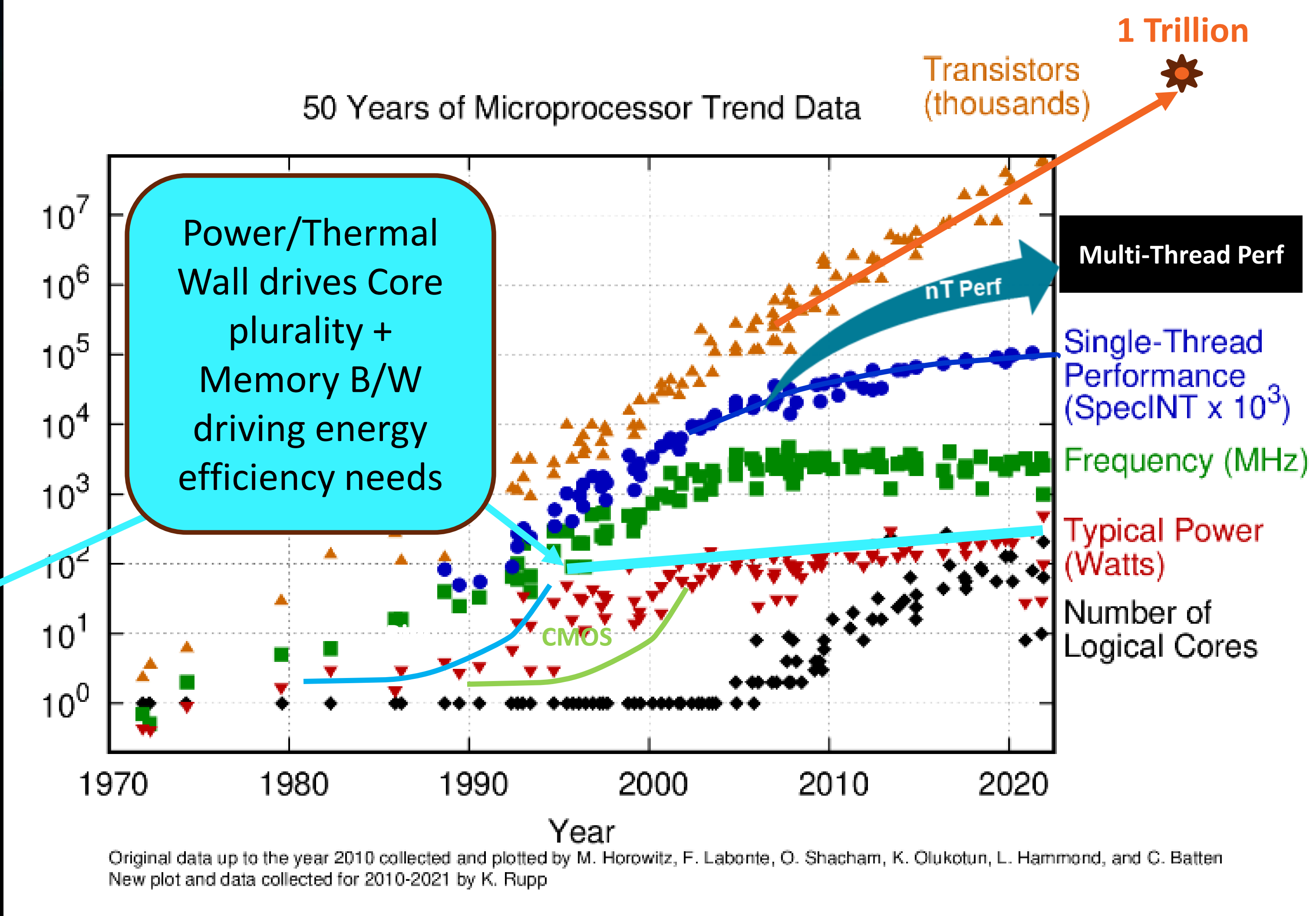


Fig. 7. Approximate HBM memory power percentage relative to total GPU package power

Source: IEDM '23 15-4: "Innovations For Energy Efficient Generative AI" S. Naffziger (AMD)

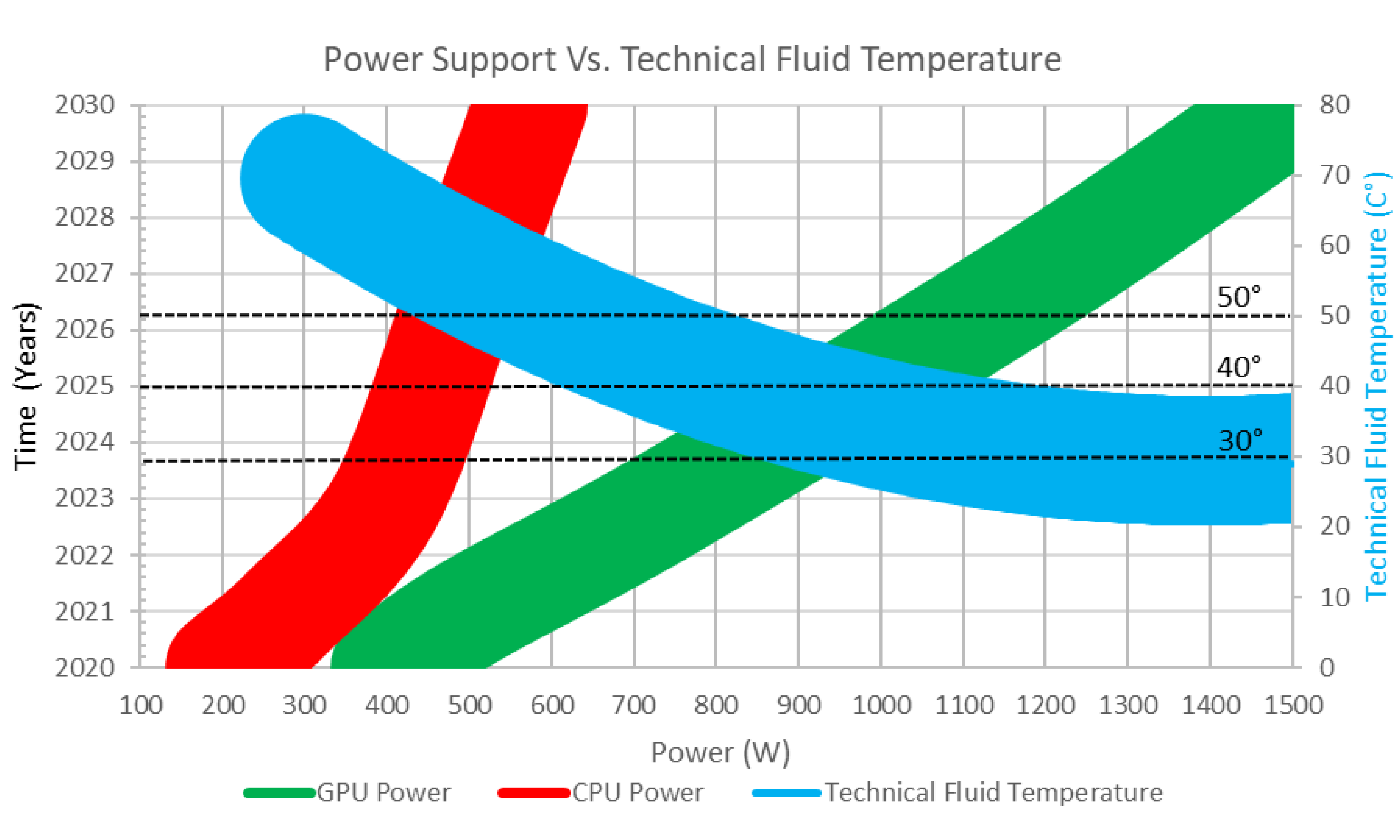


Power/Thermal Wall drives Core plurality + Memory B/W driving energy efficiency needs

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 2010-2021 by K. Rupp

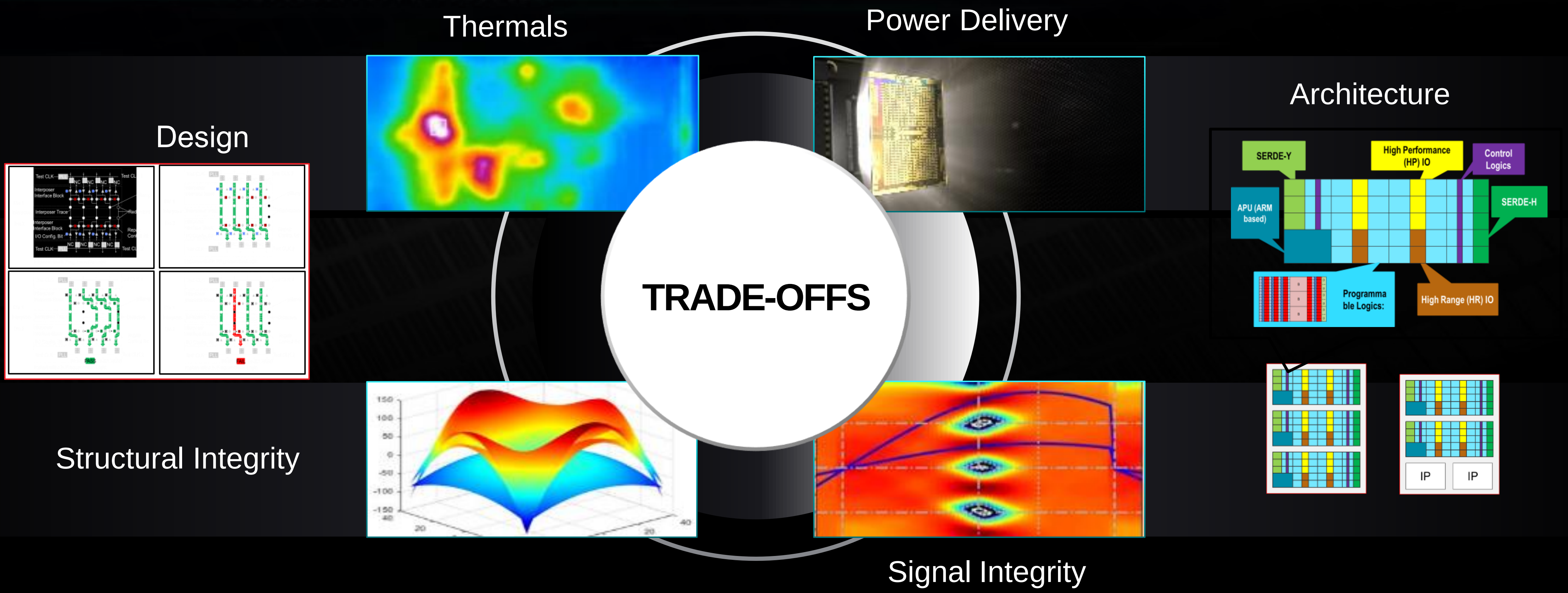
AMD
together we advance_packaging

TDP TRENDS



Increasing transistor count drives growth in TDP of CPUs and GPUs

Advanced Packaging→ Multi disciplinary focus



Thermals

Power Delivery

Architecture

Design

TRADE-OFFS

Structural Integrity

Signal Integrity

FULL MULTIPHYSICS SIMULATION TOOLS NEEDED FOR CO-OPTIMIZATION

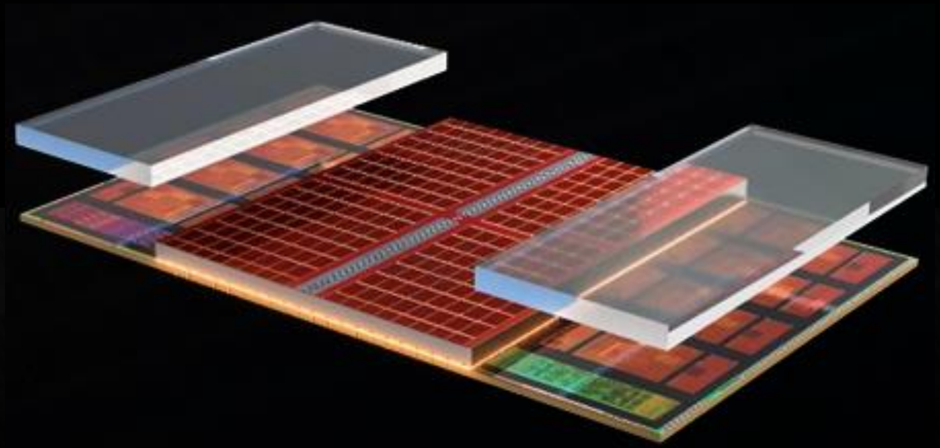
Next Gen 3D Arch Requires Next Gen EDA Tools

1) STANDARDIZE DRC TOOL SET

DRC decks spanning all design components in 2.5D/3D architectures

2) ENABLE TRUE SILICON-PACAKGE CO-DESIGN

Proliferate common tool platform across silicon and package designs



3) STANDARDIZE FILE FORMATS

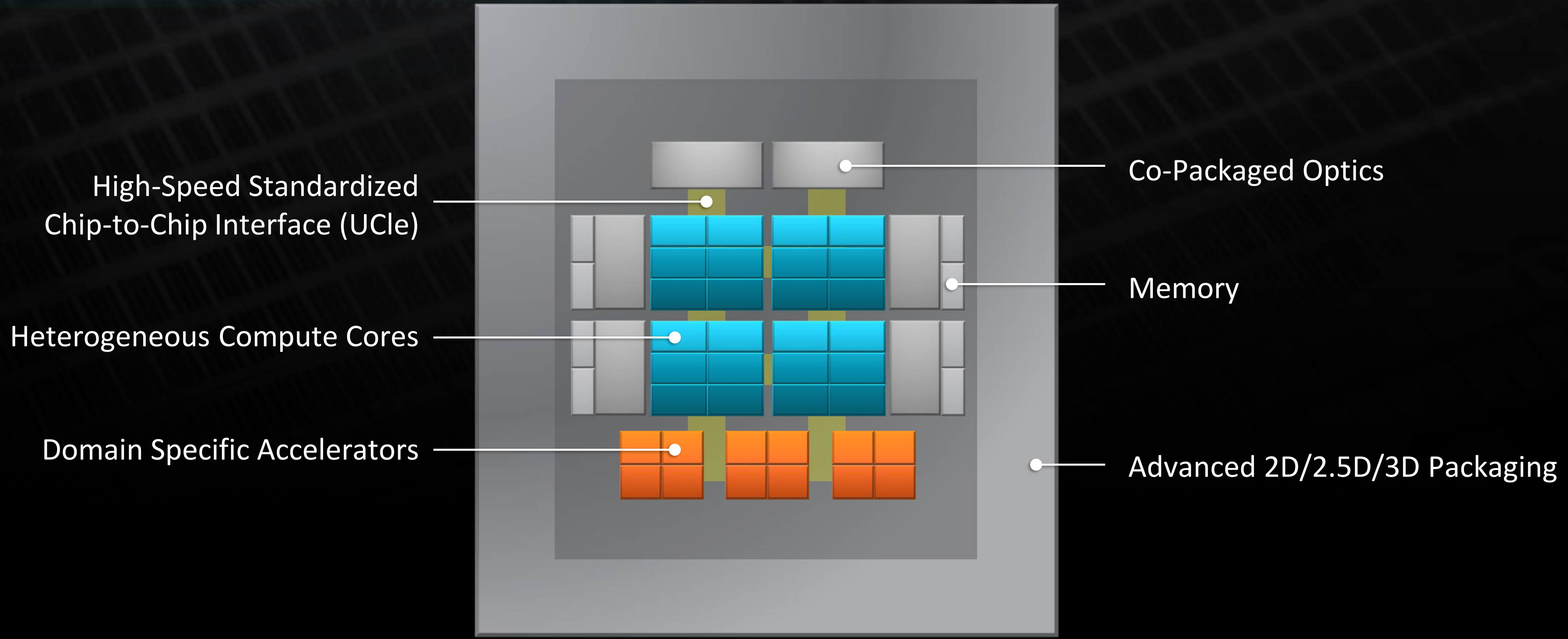
Create a seamless tool-to-tool interaction from design to analysis

4) INCREASE TOOL CAPACITY

EDA tools to stay lockstep with design pin count increases

EDA TOOL ECOSYSTEM NEEDS TO MERGE SILICON AND PACKAGE TOOLCHAINS

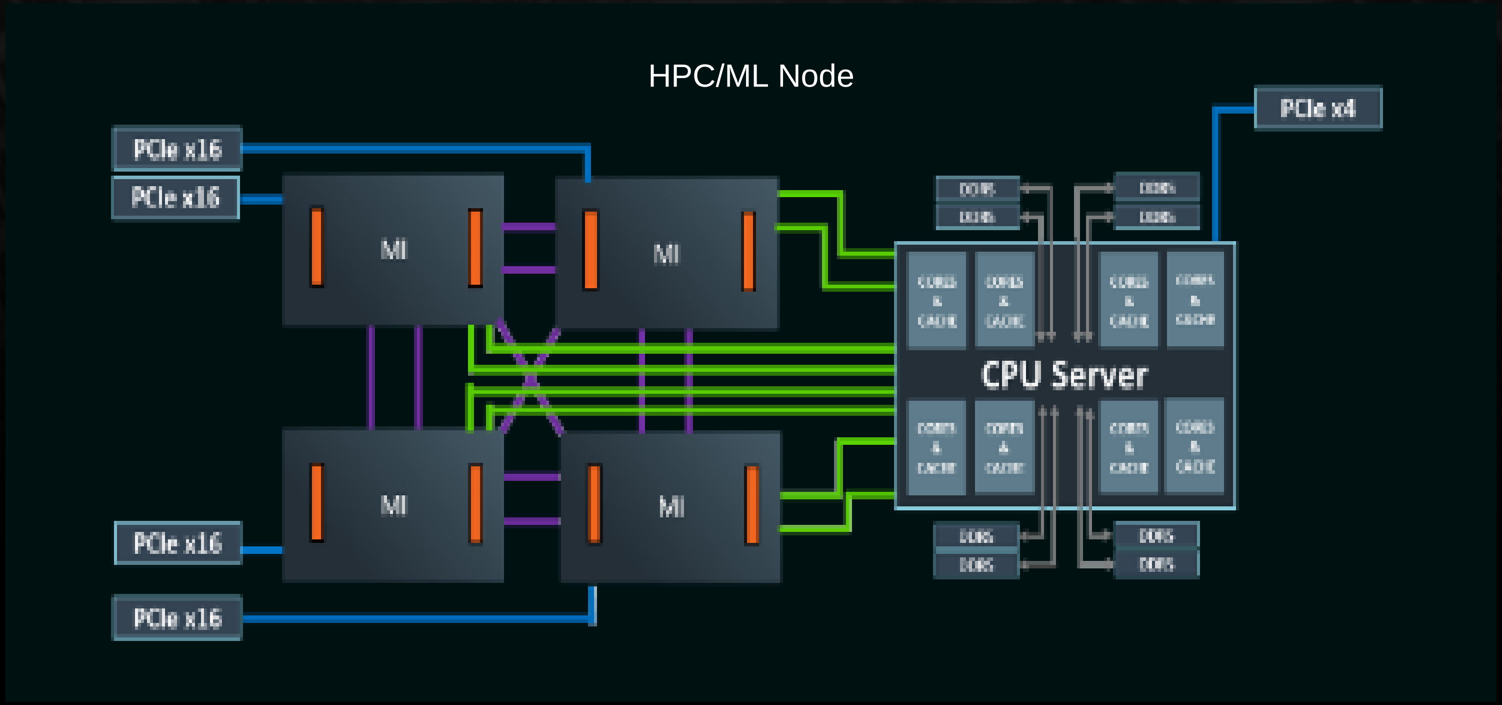
FUTURE SIP ARCHITECTURE


High-Speed Standardized Chip-to-Chip Interface (UCIe)
Heterogeneous Compute Cores
Domain Specific Accelerators
Co-Packaged Optics
Memory
Advanced 2D/2.5D/3D Packaging

- Advanced packaging enables maximally efficient integration of compute elements and memory
- System level communication accomplished with low-power and high-bandwidth optical connections

System Level Optimization

Accelerated compute nodes using AMD CPUs and GPUs



Based on 2015-2020 industry trends in energy efficiency gains and data center energy consumption in 2025.
* Includes AMD high performance CPU and GPU accelerators used for AI training and High-Performance Computing in a 4-Accelerator, CPU hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size. AI training: lower precision training-focused floating point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.

THANK YOU!

Endnotes

SP5-013D: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 05/31/2023. Comparison of published 2P AMD EPYC 9654 (1800 SPECrate®2017_int_base, 720 Total TDP W, $23,610 total 1Ku, 192 Total Cores, 2.500 Perf/W, 0.076 Perf/CPU$, http://spec.org/cpu2017/results/res2023q2/cpu2017-20230424-36017.html) is 1.80x the performance of published 2P Intel Xeon Platinum 8490H (1000 SPECrate®2017_int_base, 700 Total TDP W, $34,000 total 1Ku, 120 Total Cores, 1.429 Perf/W, 0.029 Perf/CPU$, http://spec.org/cpu2017/results/res2023q1/cpu2017-20230310-34562.html) [at 1.75x the performance/W] [at 2.59x the performance/CPU$]. Published 2P AMD EPYC 7763 (861 SPECrate®2017_int_base, 560 Total TDP W, $15,780 total 1Ku, 128 Total Cores, 1.538 Perf/W, 0.055 Perf/CPU$, http://spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30148.html) is shown for reference at 0.86x the performance [at 1.08x the performance/W] [at 1.86x the performance/CPU$]. AMD 1Ku pricing and Intel ARK.intel.com specifications and pricing as of 6/13/23. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

Adv Pkg Performance and Efficiency (Slide #12/13/14): Based on AMD engineering internal analysis

3D Hybrid Bonding (Slide #17): Based on AMD engineering internal analysis , May 2021

RYZEN™ 9 7950X3D(Slide 21): Based on AMD Internal Analysis

SP5-050: EDA RTL Simulation comparison based on AMD internal testing completed on 4/13/2023 measuring the average time to complete a graphics card test case simulation. comparing: 1x 16C EPYC™ 9384X with AMD 3D V-Cache Technology versus 1x 16C AMD EPYC™ 9174F on the same AMD “Titanite” reference platform. Results may vary based on factors including silicon version, hardware and software configuration and driver versions.

MI300-03: Measurements by AMD Performance Labs June 4, 2022, on current specification and/or estimation for estimated delivered FP8 floating point performance with structure sparsity supported for AMD Instinct™ MI300 vs. MI250X FP16 (306.4 estimated delivered TFLOPS based on 80% of peak theoretical floating-point performance). MI300 performance based on preliminary estimates and expectations. Final performance may vary.

Disclaimer and Attributions

DISCLAIMER

The information contained herein is for informational purposes only, and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD’s products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18