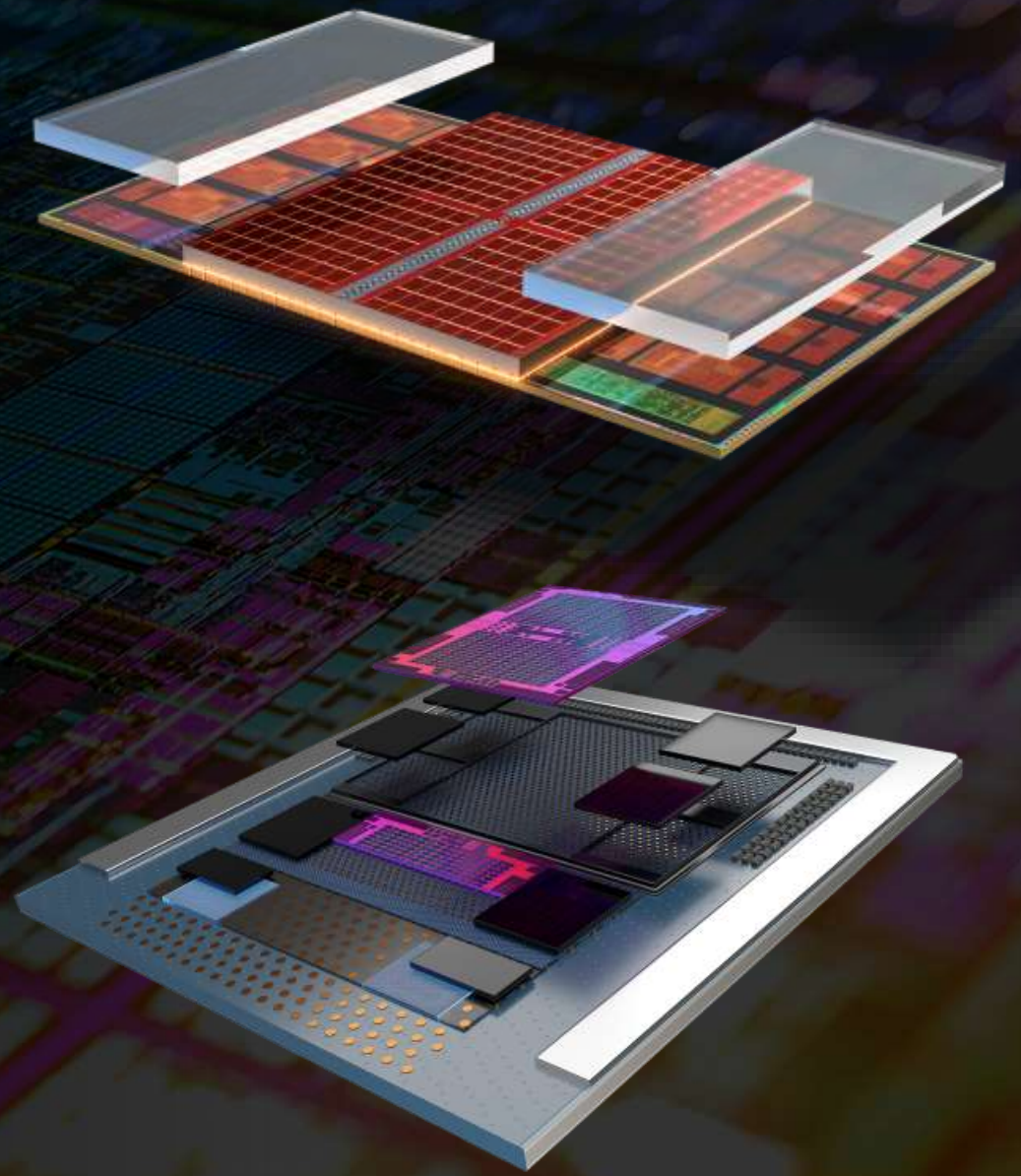
AMD

ADVANCED PACKAGING

ENABLING MOORE'S LAW'S NEXT FRONTIER THROUGH HETEROGENEOUS INTEGRATION



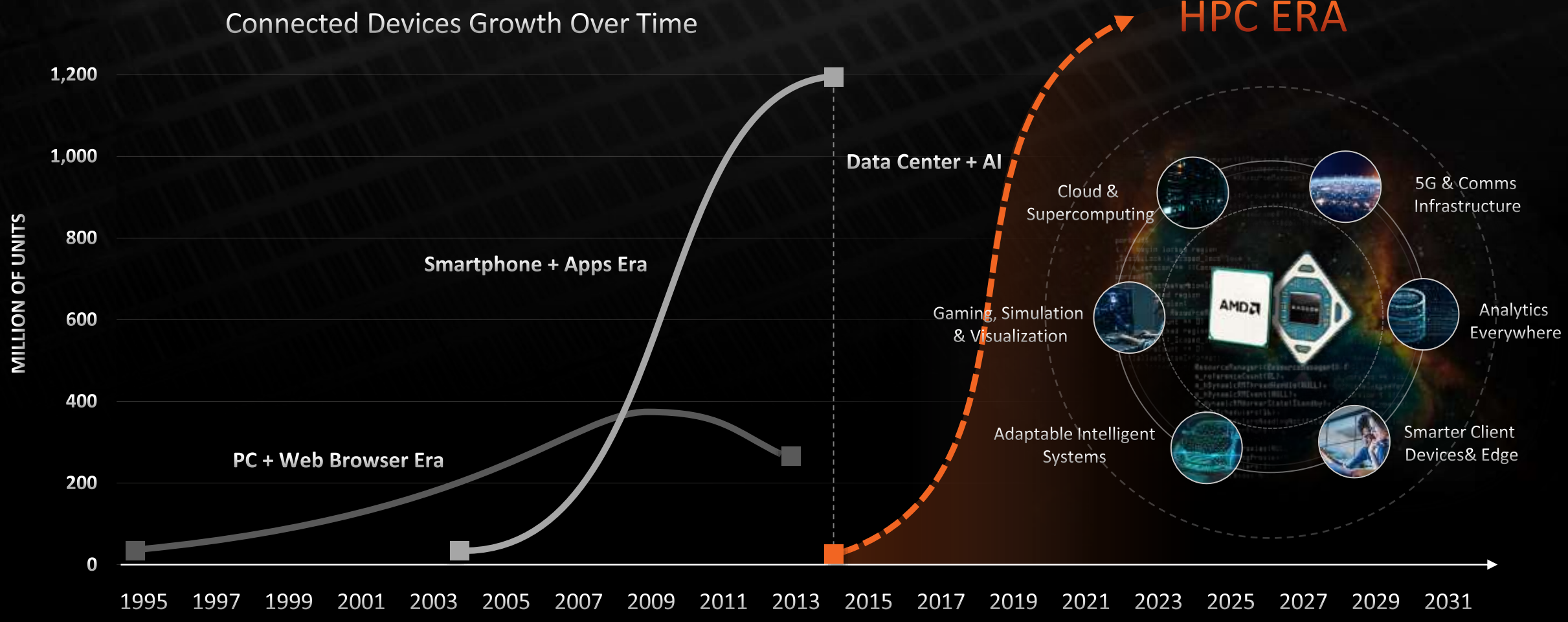
RAJA SWAMINATHAN
AMD SENIOR FELLOW &
ADVANCED PACKAGING LEADER

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 and technology as well as 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.

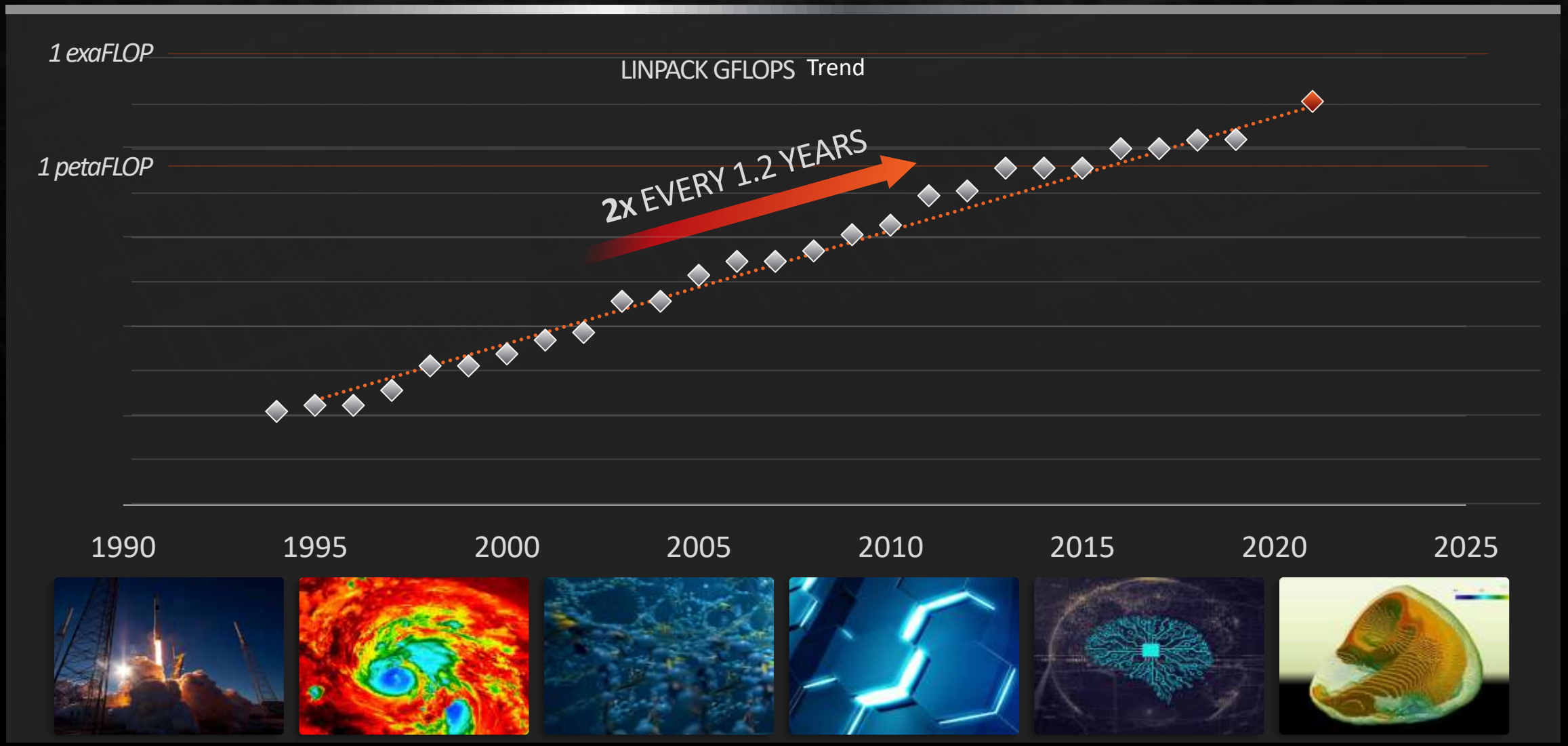
AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

EXPLOSION OF CONNECTED DEVICES



AMD

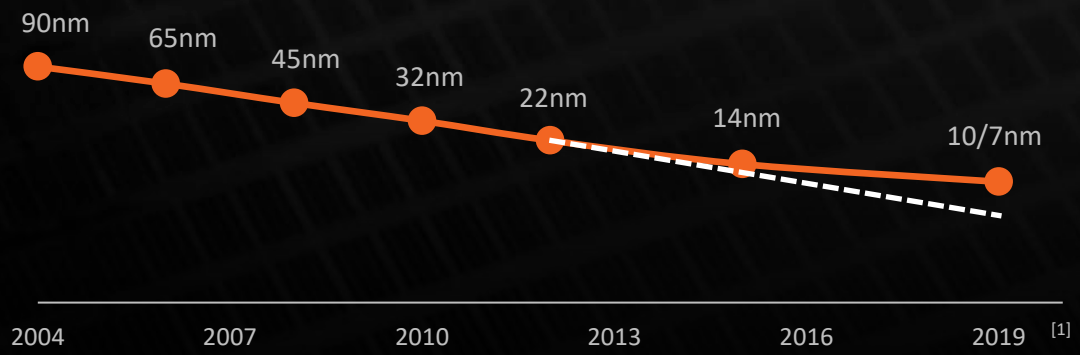
THE RELENTLESS DEMAND FOR MORE COMPUTE..



AMD

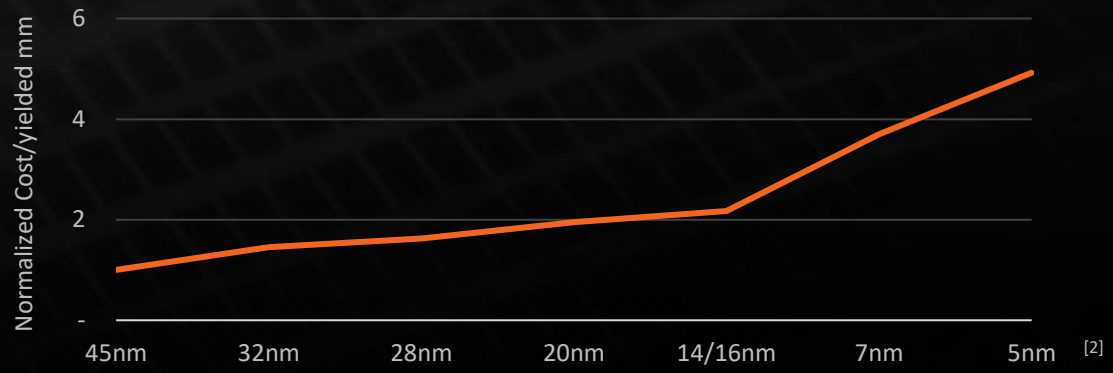
TECHNOLOGY HEADWINDS TO MEETING THE DEMAND

Slowing of Moore's Law



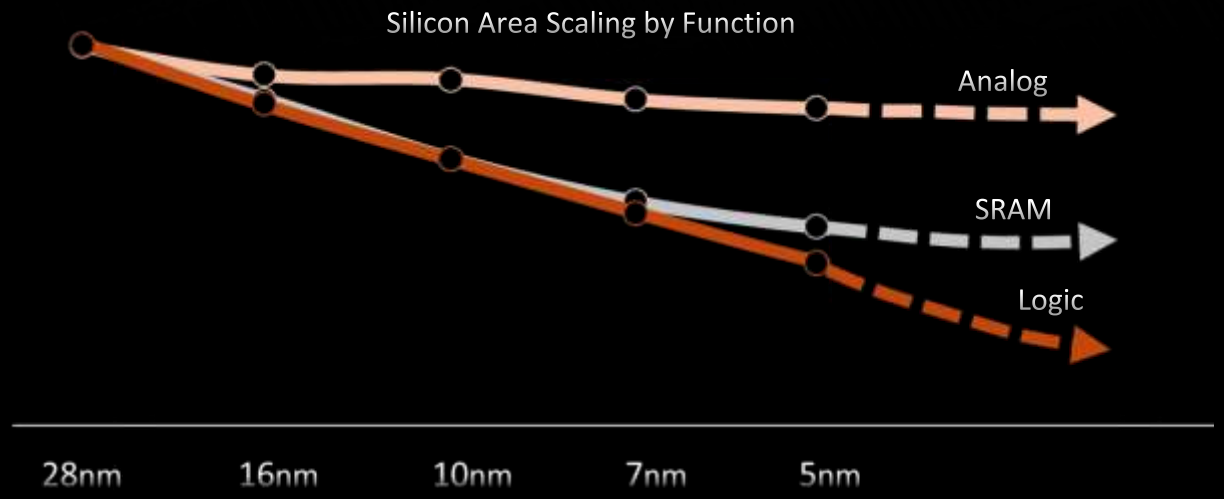
[1]

Increasing Cost

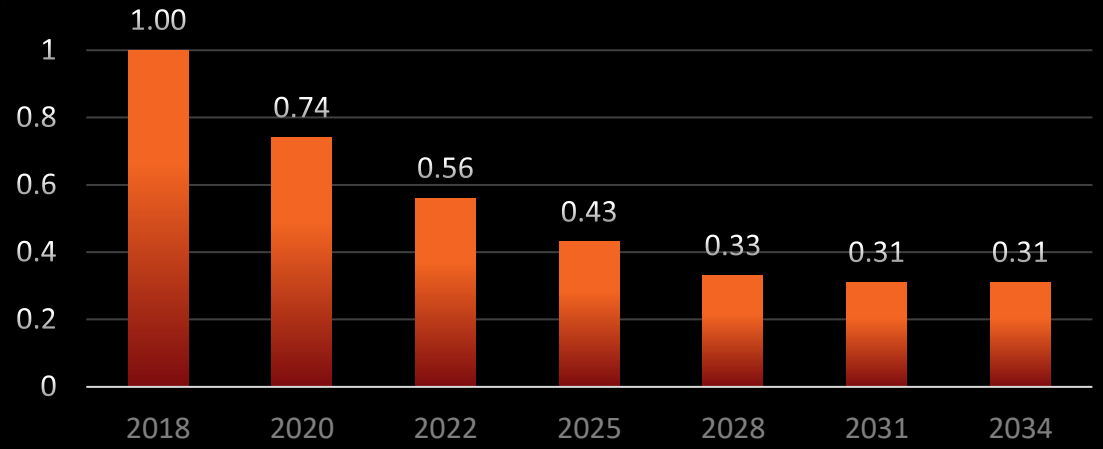


[2]

Limited and Divergent Scale Factors

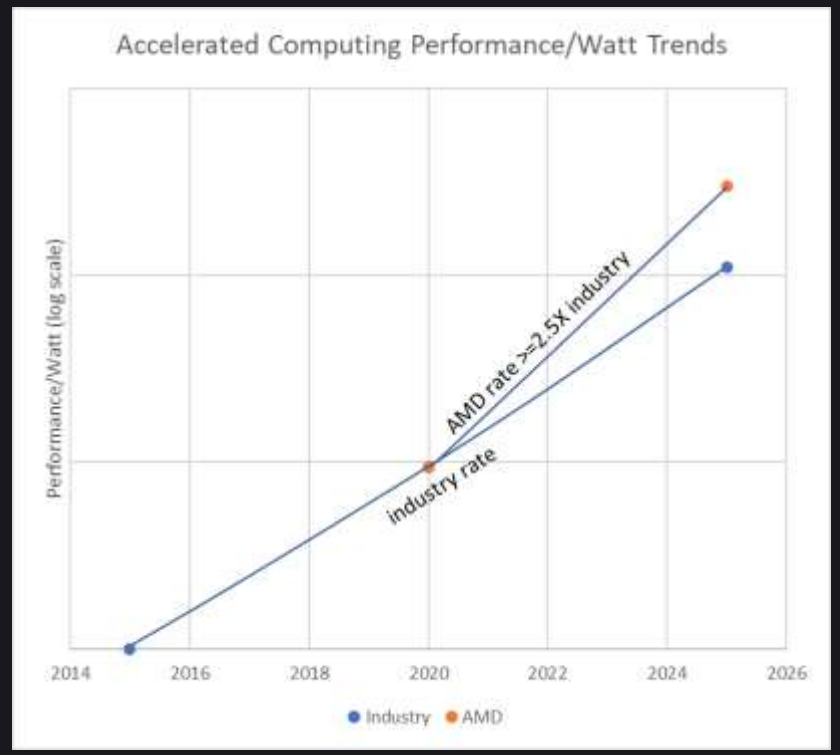


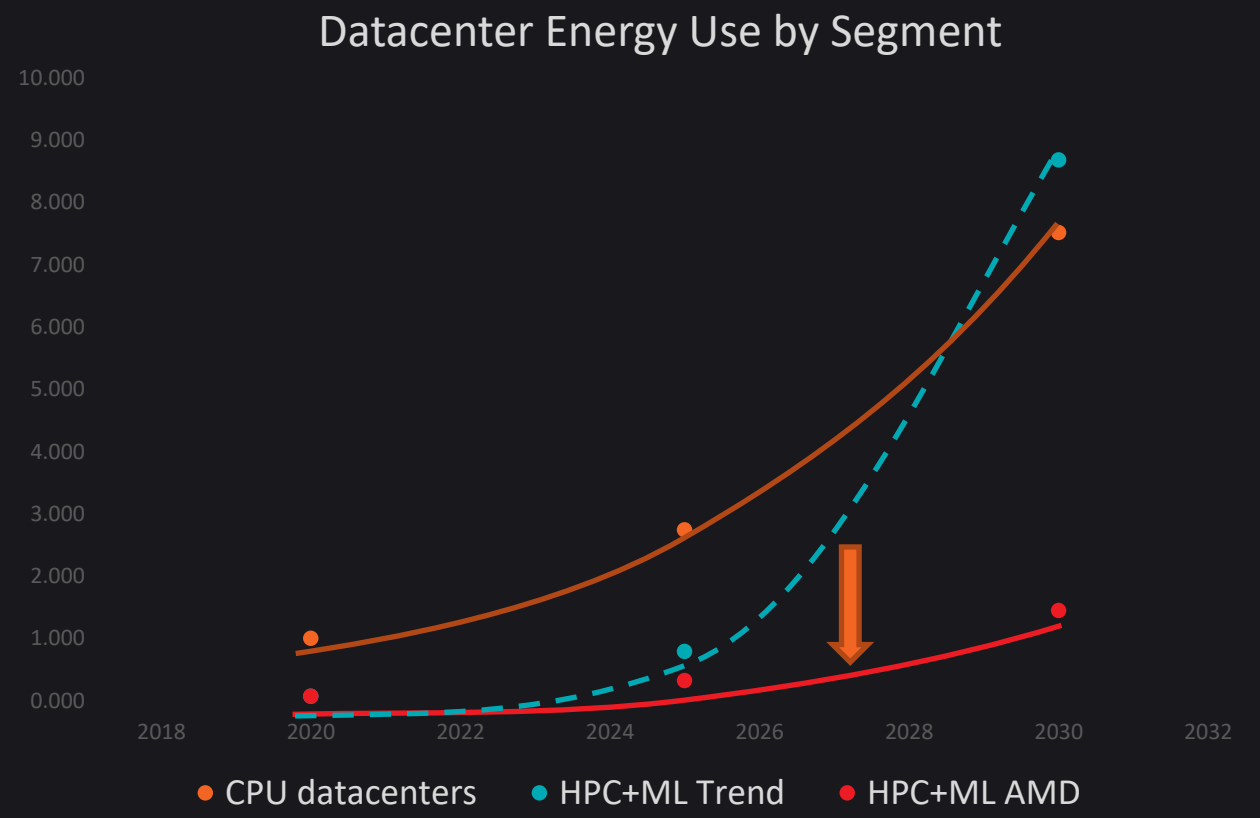
SoC Scaling Flattening



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

AMD'S EFFICIENCY GOAL FOR HPC/AI APPLICATIONS



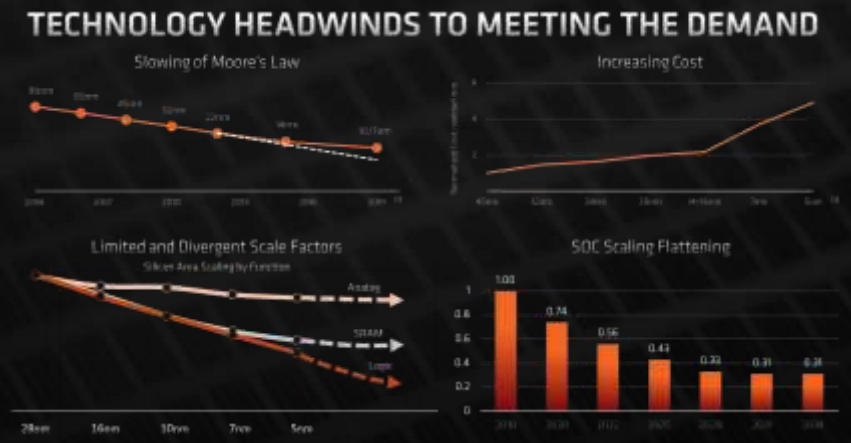


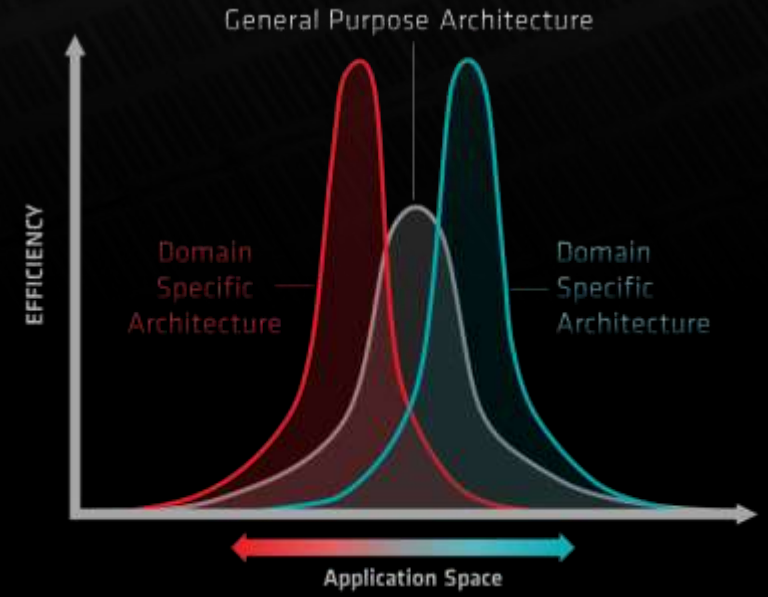
Exploiting architectural innovations, package and silicon technology advances, AMD's goal is to dramatically accelerate energy efficiency improvement rate to 30X by 2025

AMD

WHERE THIS ALL LEADS

- Compute demand increasing
- Significant barriers to traditional scaling
- Higher efficiency domain specific accelerators required
- Modular design supported by advanced packaging required





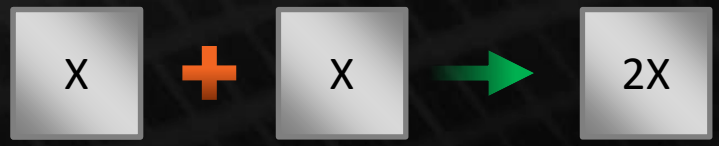
How to Architect, Design and Build Future Systems?

Modular Design

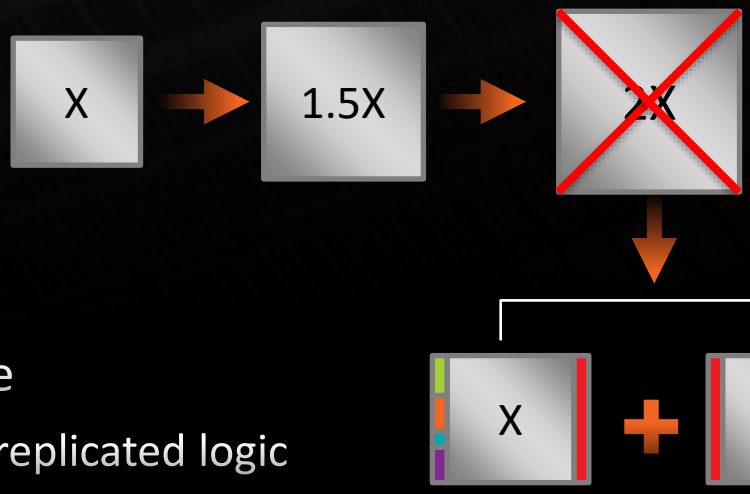
AMD

CHIPLETS BACKGROUND

Historically, except for the largest systems,
Moore's Law was sufficient to meet compute needs



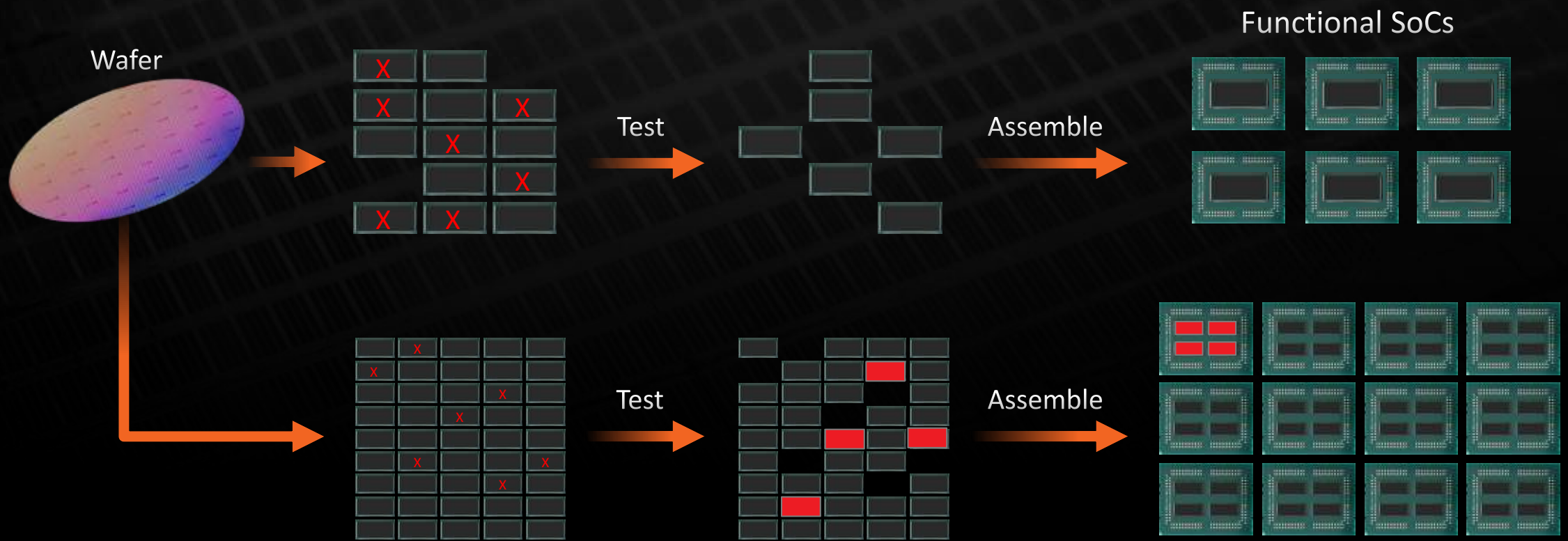
Current trends require a new approach



However, chiplets are not free

- Additional area for interfaces, replicated logic
- Additional design effort, complexity
- Past methodologies less suited for chiplets

HIGH-LEVEL APPROACH TO CHIPLETS

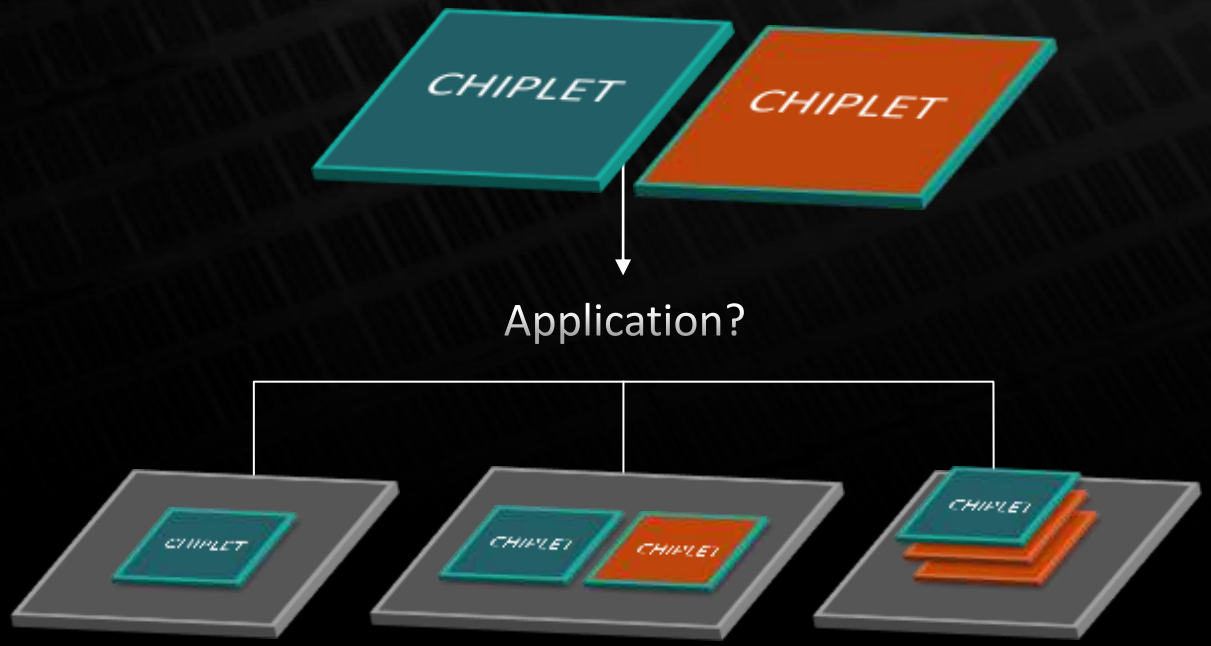


Many More Functional SoCs

Ability to mix and match at a finer grained level

MODULAR ARCHITECTURE GOALS

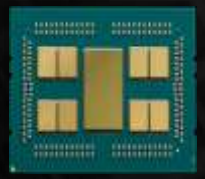
ENABLING A MORE FLEXIBLE APPROACH



- We want to build tailored products for specific markets by mixing and matching chiplet types
- We can now specialize a domain specific chiplet and include more or fewer of them for a given product
- More domain-specific products at higher yields … provided we can build low-overhead chiplets

PACKAGE ARCHITECTURES FOR CHIPLETS

MCM



INFO-R

INFO-L

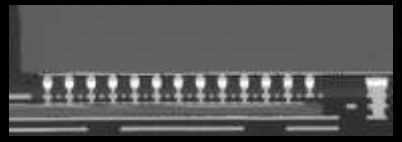
Foveros

Intel: Foveros

CoWoS-L

TSMC: INFO-R/-L, CoWoS-L

EMIB



Co-EMIB

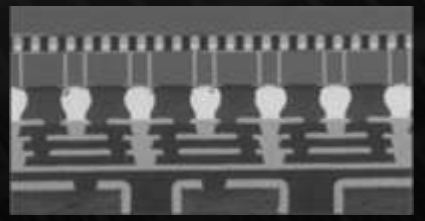
Intel: EMIB and Co-EMIB

FoCoS

ASE: FoCoS

Micro Bump 3D

Si Interposer + TSV



AMD Fiji GPU

Courtesy: TechSearch

Foveros-ODI

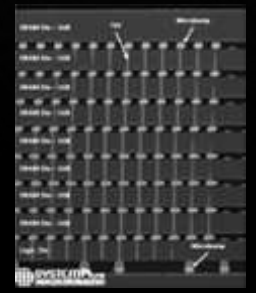
Intel: Omni-directional interconnect

INFO-POP

Apple A10 on FO+POP

Courtesy: SystemPlus Consulting

W2W stacking+ TSV+uBump



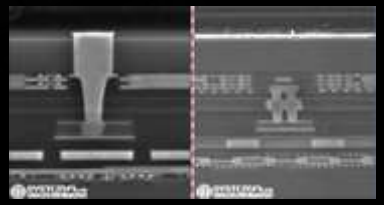
Samsung: HBM2 memory

Courtesy: SystemPlus Consulting

WoW

W2W F2F di-electric bonding+ TSV

W2W F2F hybrid bonding w/o TSV



Sony: CMOS Image Sensors

Courtesy: SystemPlus Consulting

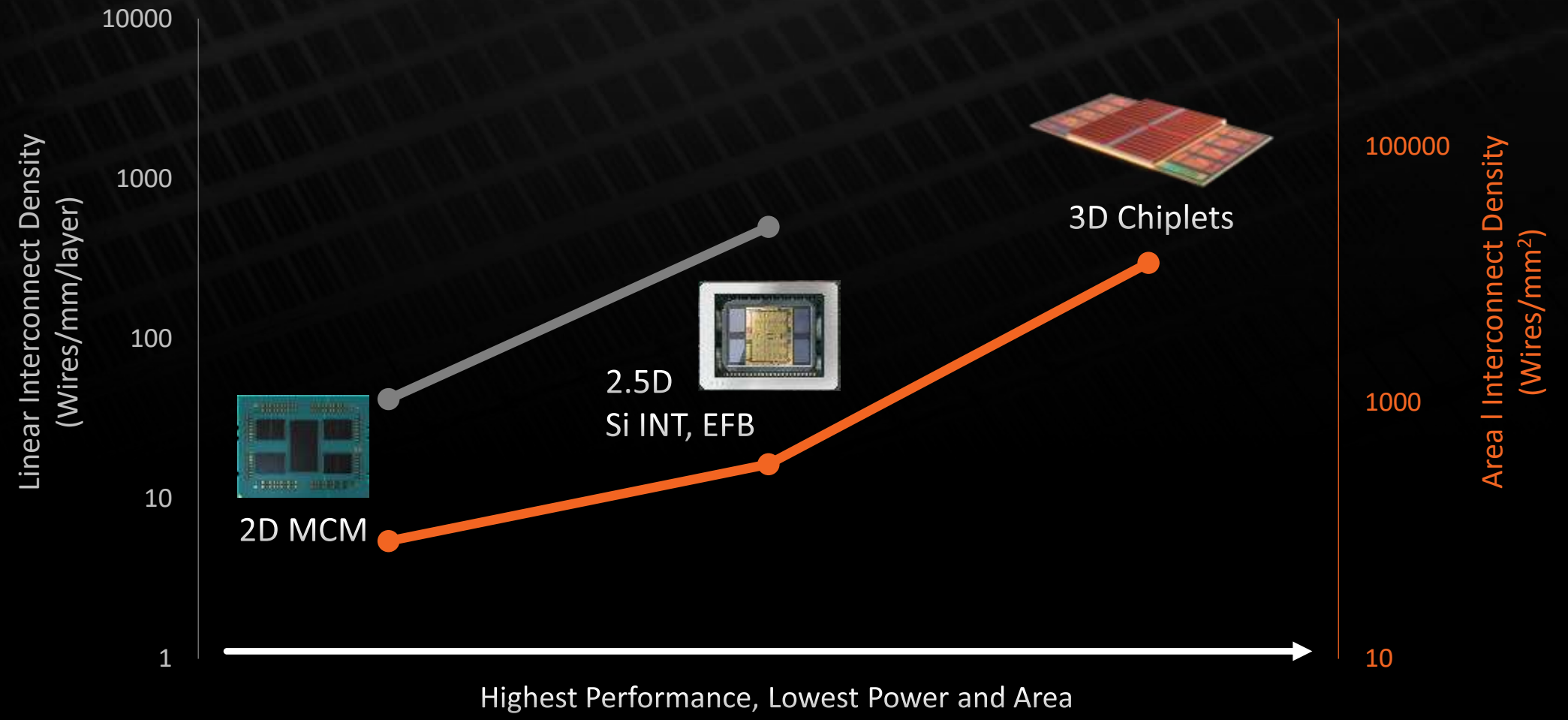
No single package architecture works for all products - choice based on product PPAC

Images courtesy: Intel, TSMC, Amkor, ASE from public material

AMD

IMPROVING KEY PARAMETERS

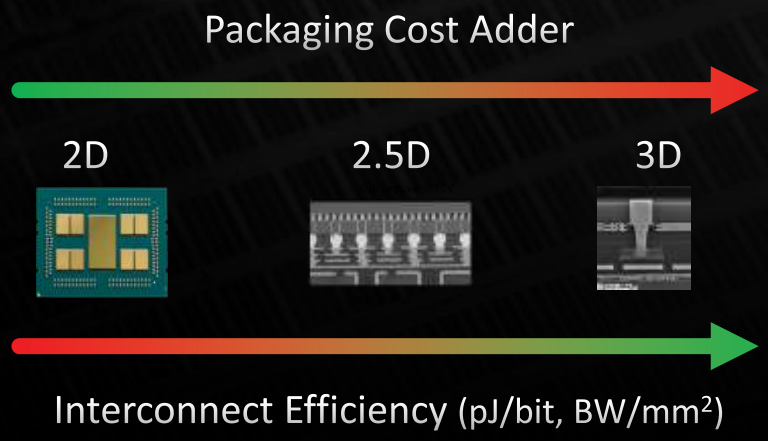
DRIVING HIGH-PERFORMANCE COMPUTING FORWARD



AMD

FINDING THE OPTIMAL SOLUTION

Chiplet package architecture selection requires balancing a complex equation…



[Value of modular product]
+
[Chiplet yield and technology cost benefit]

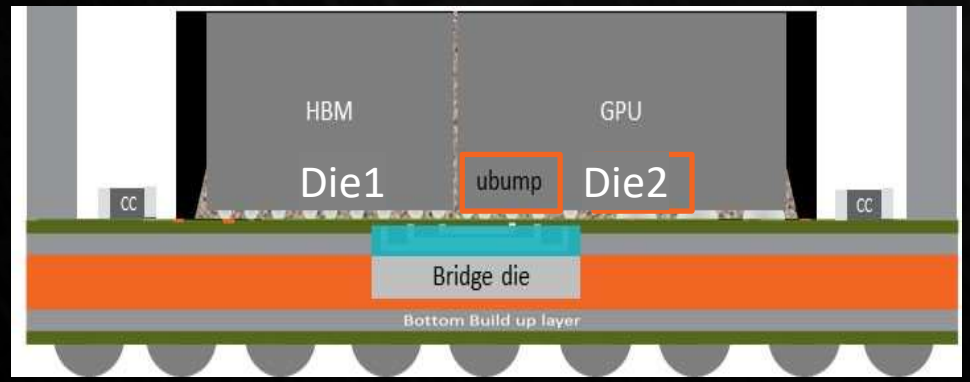
>

[Chiplet die overhead adder]
+
[Packaging cost adder]
+
[Power cost of interconnect]
+
[Engineering complexity of solution]

Architectural need for bandwidth, die partition options and package technology create a multi-disciplinary optimization equation

2.5D "BRIDGE" ARCHITECTURE LANDSCAPE

Substrate Embedded 2.5D

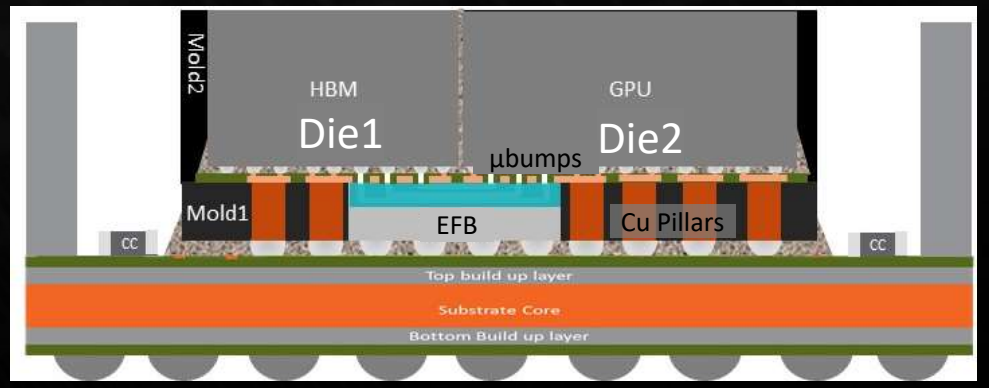


LOCALIZED INTERCONNECTS
Bridge Technologies

BETTER ELECTRICALS
Lower Parasitic Capacitance

Compared to Si Interposer Based 2.5D

Elevated Fanout Bridge 2.5D



SCALABLE SOLUTION
Lithographically Defined

STANDARD SUBSTRATES
Lower Cost

STANDARD FLIP CHIP PROCESS
Lower Complexity Bumping, Assembly Process

Compared to Substrate Embedded 2.5D

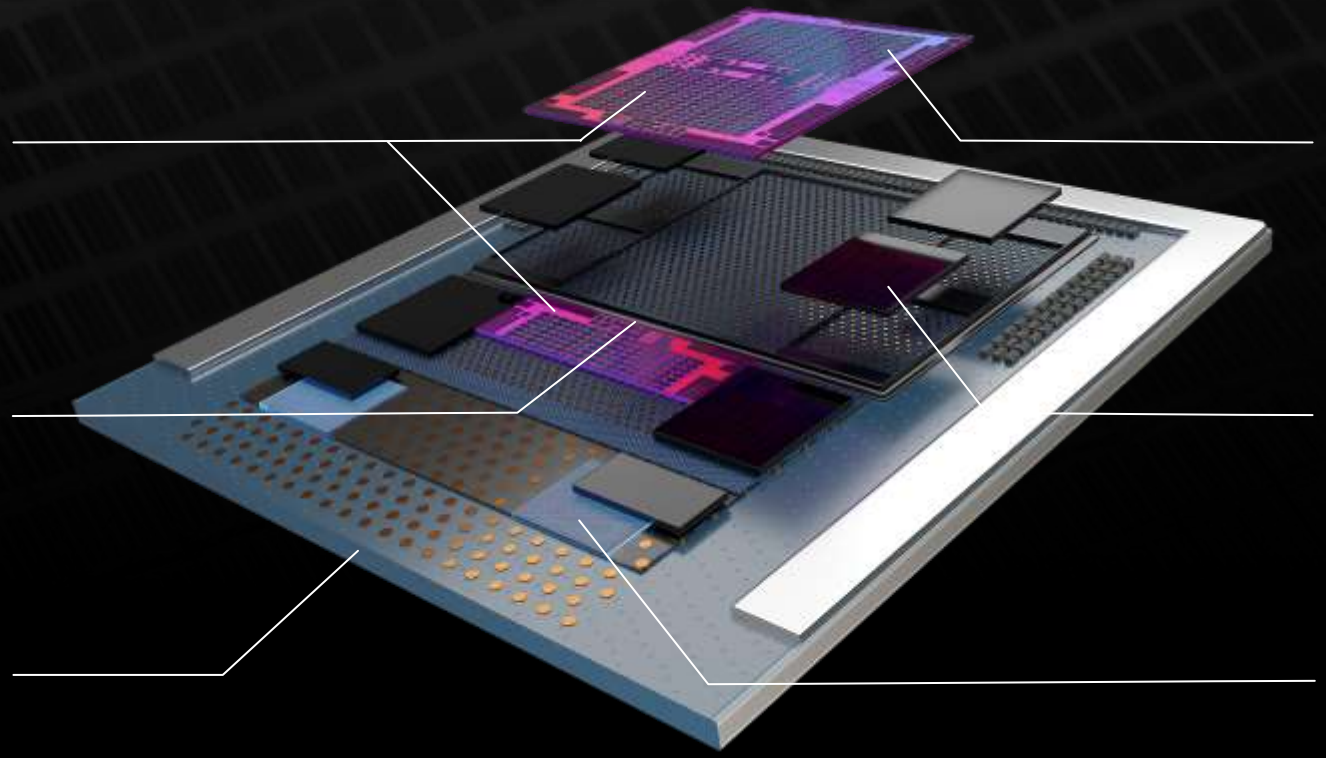
AMD INSTINCT™ MI200 SERIES

DEPLOYING EFB

TWO
AMD CDNA™2 DIES

ULTRA HIGH BANDWIDTH
DIE INTERCONNECT

COHERENT CPU-TO-GPU
INTERCONNECT



2ND GEN MATRIX
CORES FOR HPC & AI

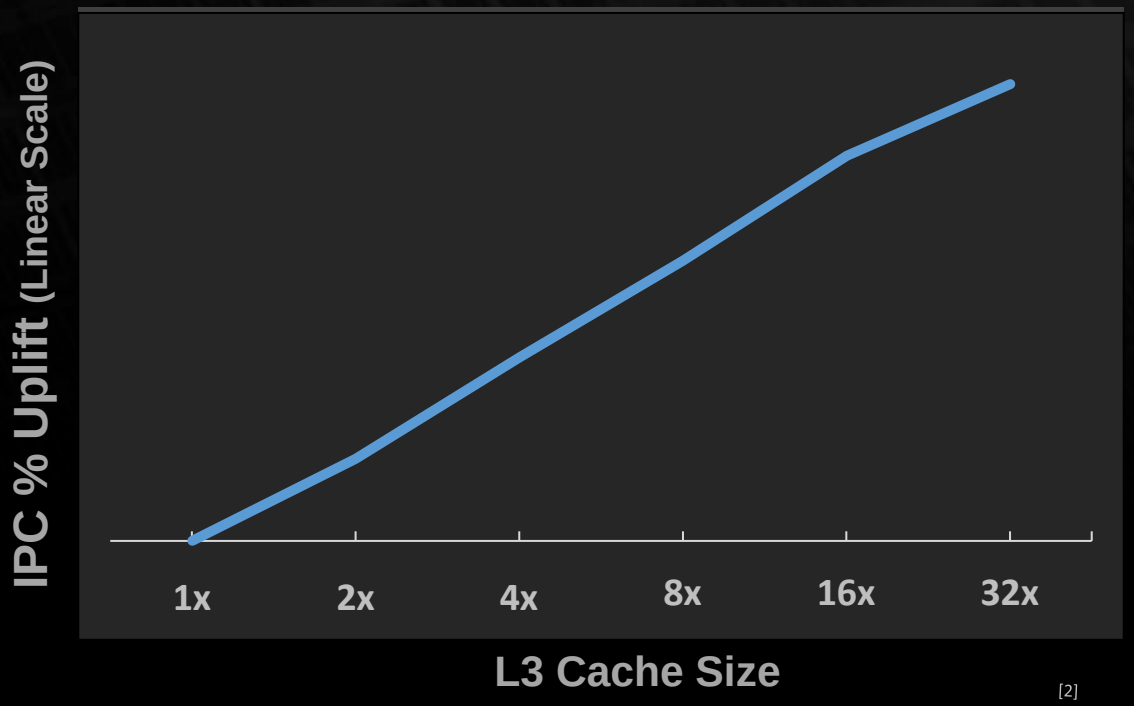
EIGHT STACKS
OF HBM2E

2.5D ELEVATED
FANOUT BRIDGE (EFB)

AMD INSTINCT™ MI200 OAM SERIES

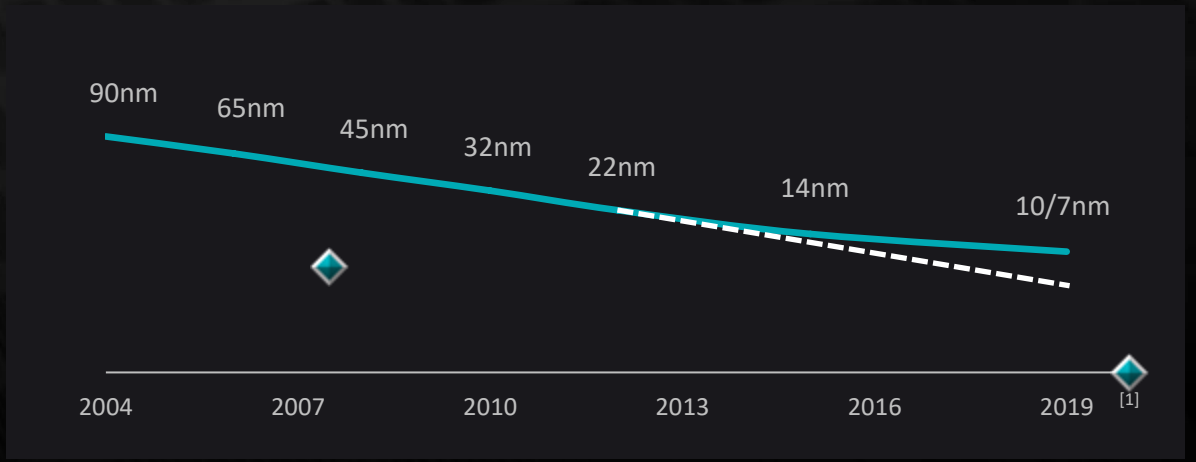
3D V-CACHE: MOTIVATION

Large L3 Caches Provide IPC Uplift

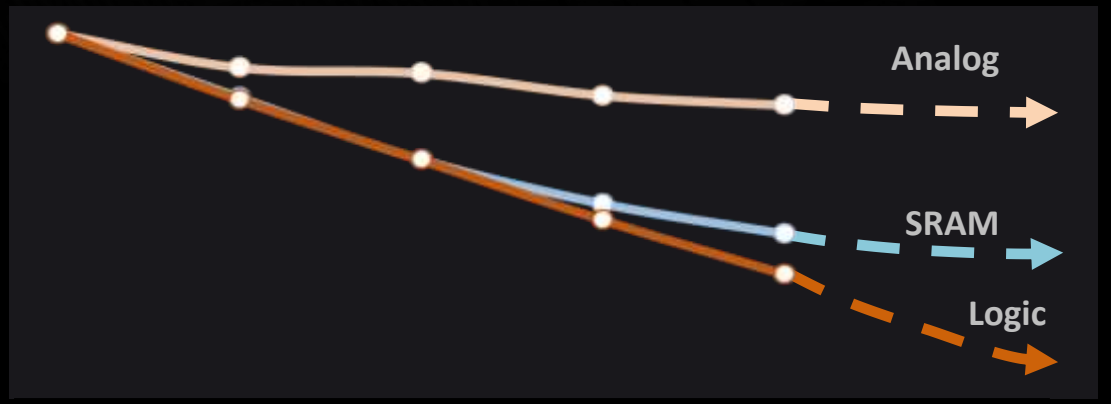


[2]

Moore's Law slowing down



SRAM doesn't scale effectively



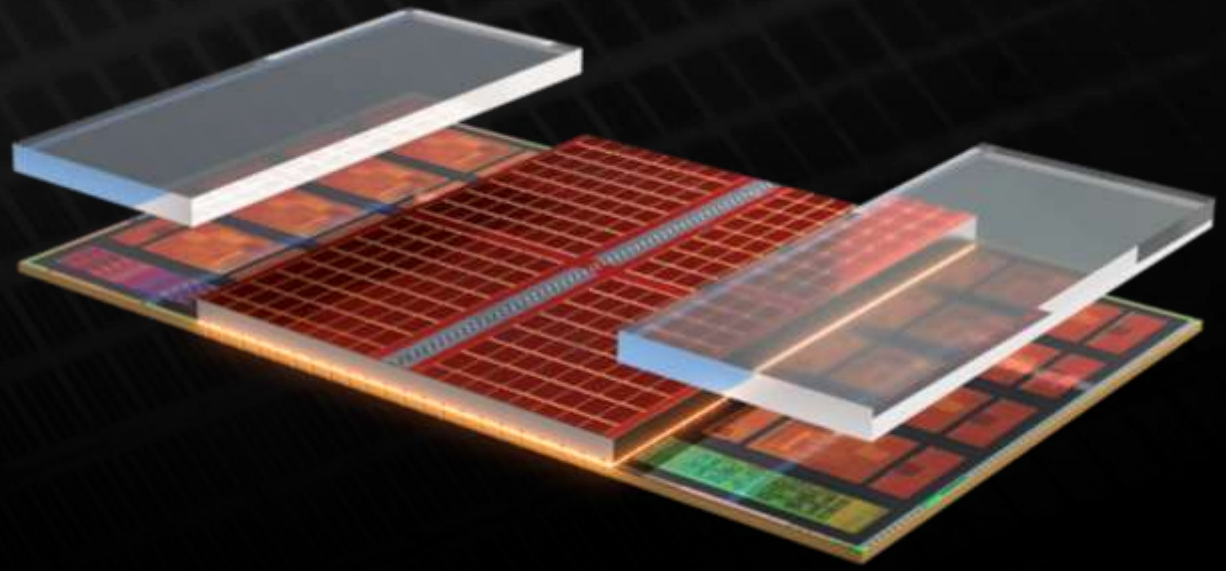
[1] Naffziger, VLSI Short Course, 2020, [2] Cost per yielded mm² for a 250mm² die

[2] Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

AMD

3D V-CACHE: IMPLEMENTATION

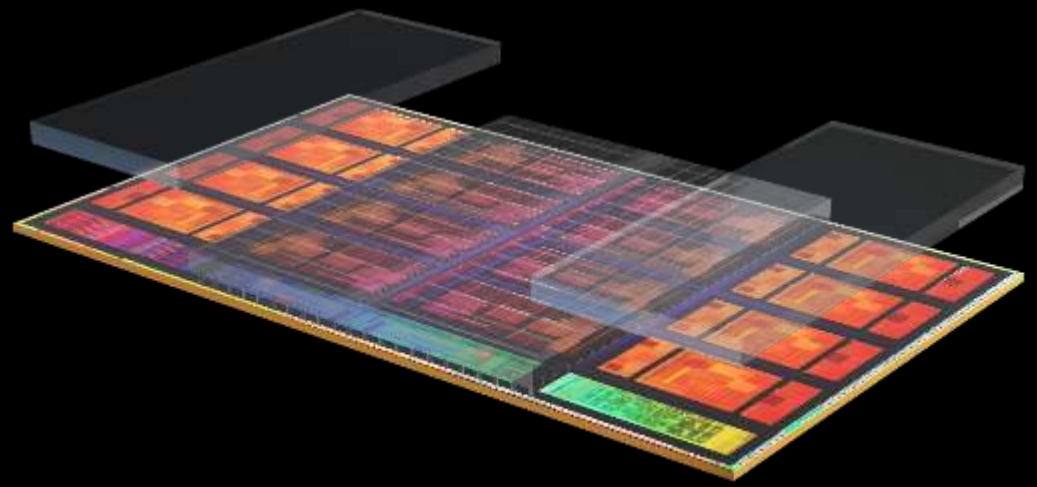
Industry’s first high-performance processor product with Hybrid Bonded 3D cache die

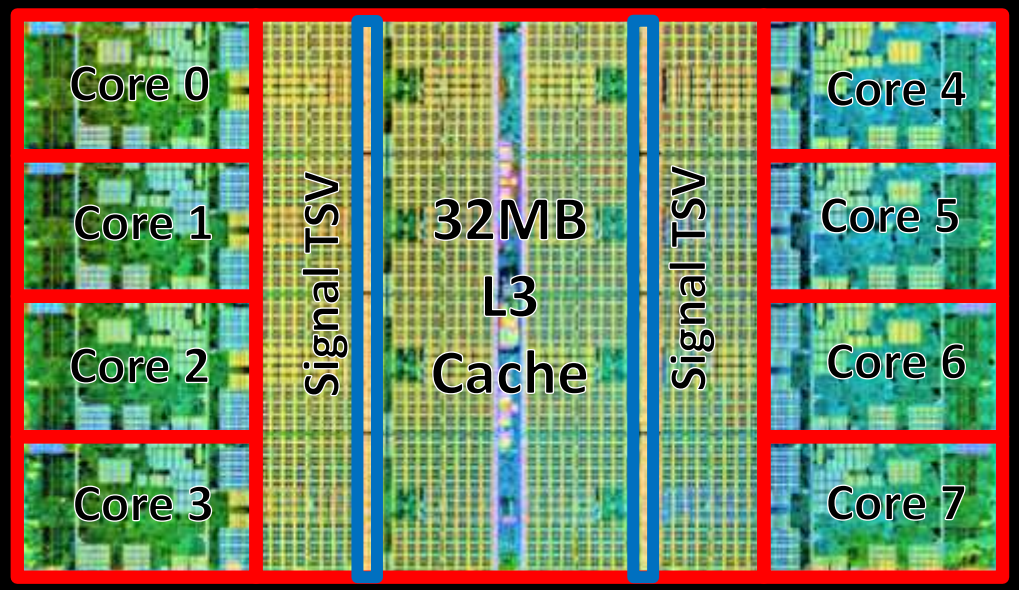


- ✓ Improves effective memory latency
- ✓ Reduces long data path and I/O’s dynamic powers
- ✓ Fits more transistors within a given package cavity size

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

AMD 3D V-CACHE™ COMPONENTS: CCD





"Zen 3" x86-64 CPU Core Complex Die (CCD, N7)

8 cores per Core Complex (CCX)

32MB shared L3 Cache

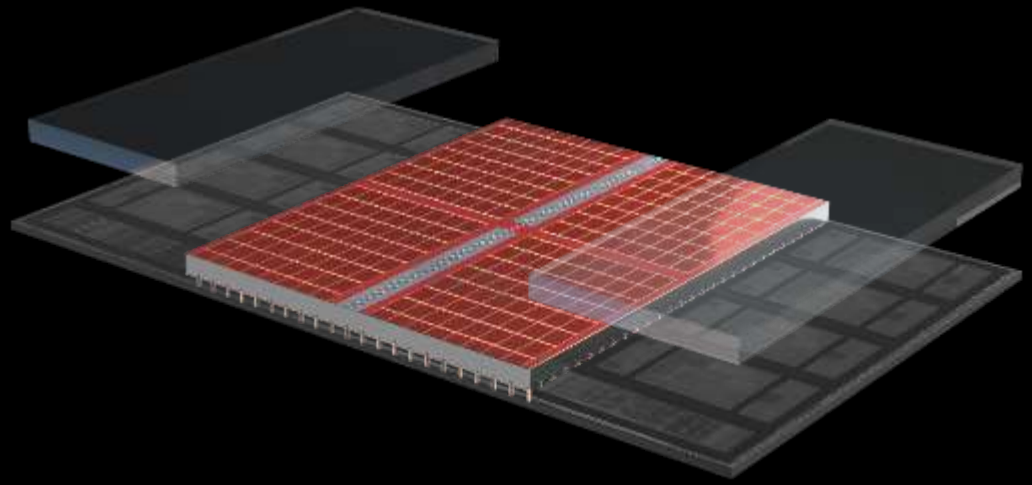
+19%[1] IPC (Ave) vs. "Zen 2"

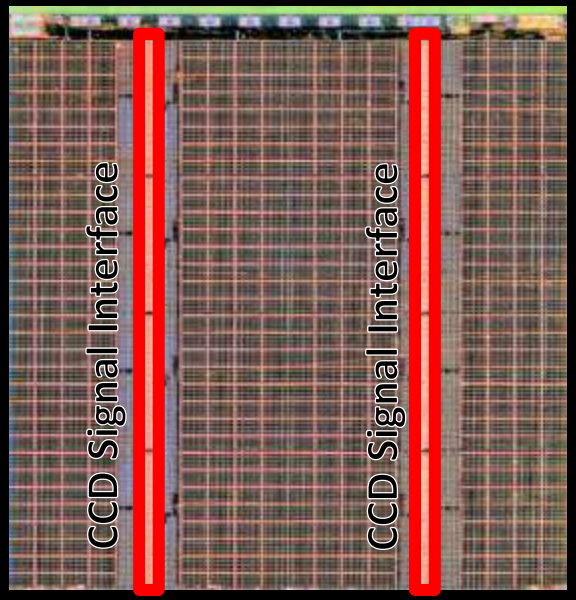
AMD 3D V-Cache™ support integrated from Day 1

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

AMD

AMD 3D V-CACHE™ COMPONENTS: L3D





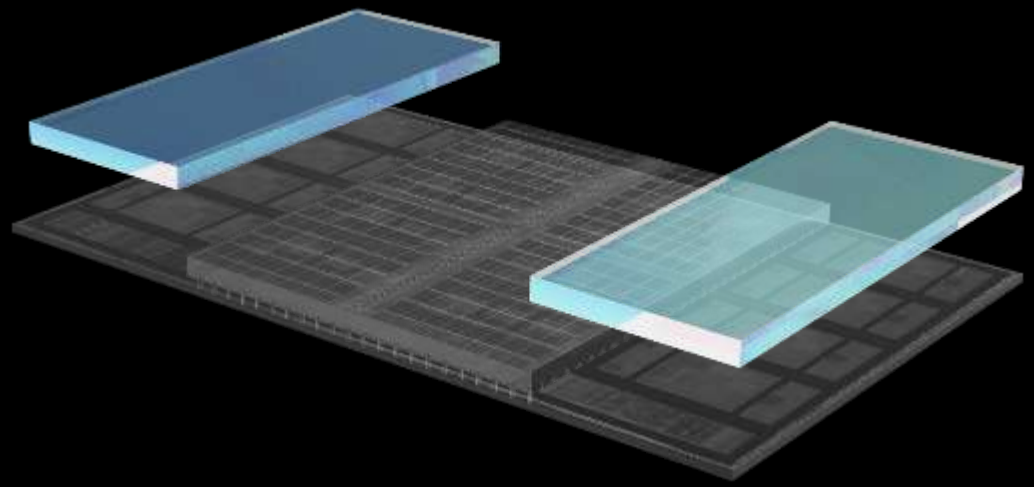
AMD 3D V-Cache™ extended L3 Die (L3D, N7)

64MB L3 Cache Extension

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

AMD

AMD 3D V-CACHE™ COMPONENTS: STRUCTURAL DIE



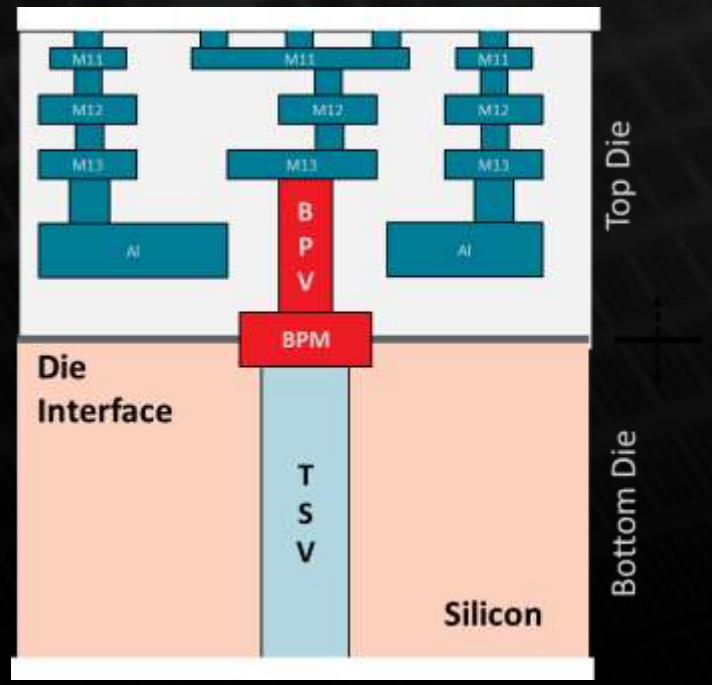
AMD 3D V-Cache™ Structural Dies

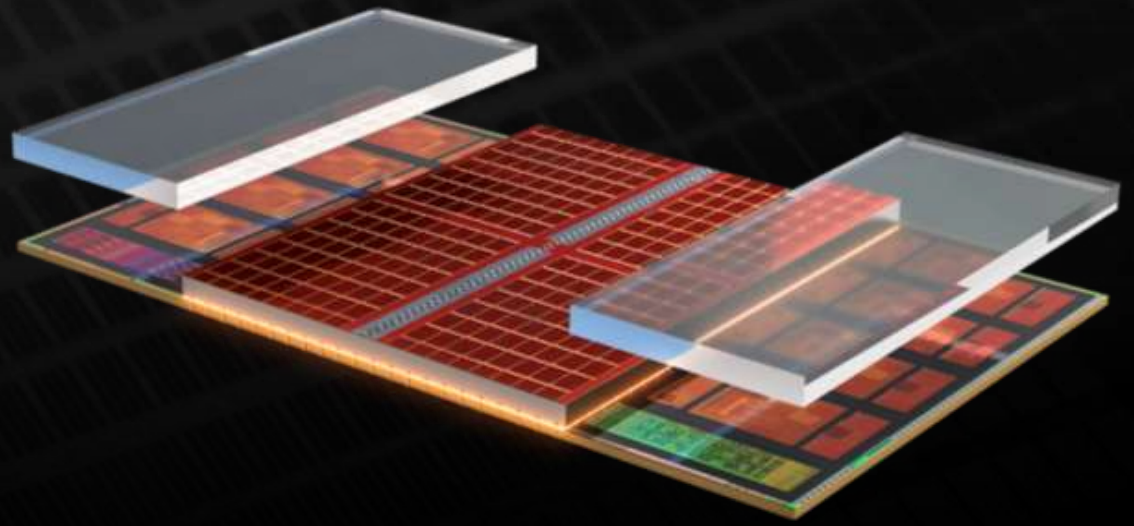
Structural support for thinned CCD

Thermal dissipation for CPU cores

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

3D V-CACHE: BRINGING IT TOGETHER





- TSMC SoIC™ process: Cu-Cu Hybrid bonded using Bond Pad Metal (BPM) pads
- BPM interfaces with TSV; Bond Pad Via (BPV) connects BPM to M13
- 9u minimum TSV/Hybrid Bond pitch

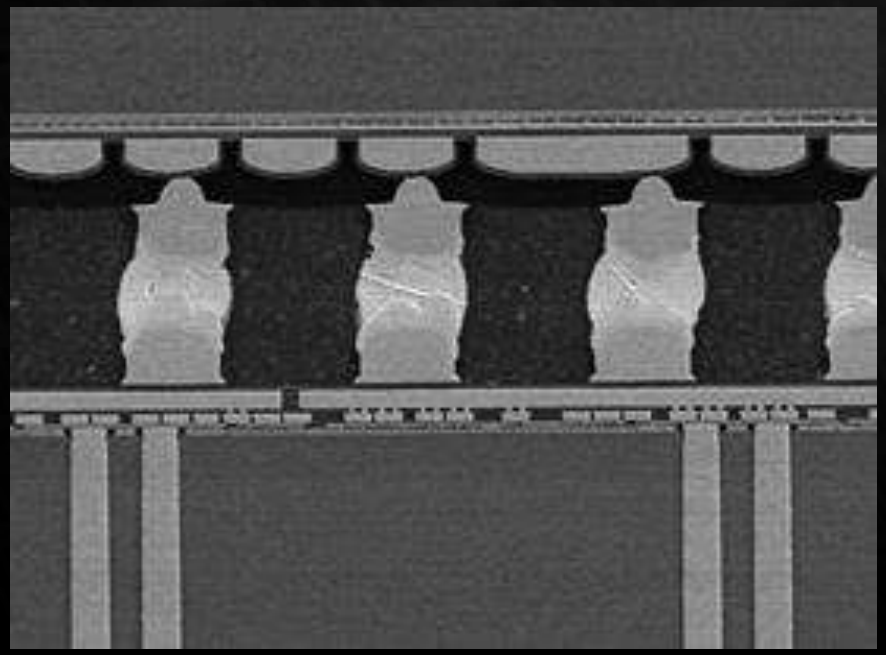
CCD face-down
- C4 interface to substrate
- TSV interface to L3D

L3D face-down
- Hybrid Bonded (HB) to CCD

Structural Dies
- Oxide bonded to CCD

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

OTHER 3D ARCHITECTURE



Micro Bump (50u→36u pitch)

~50u tall TSV

AMD 3D CHIPLETS

Hybrid Bond (9u pitch)

Back End Like TSV

>3X

Interconnect Energy Efficiency

Compared to Micro Bump 3D

>15X

Interconnect Density

Compared to Micro Bump 3D

BETTER PARASITICS

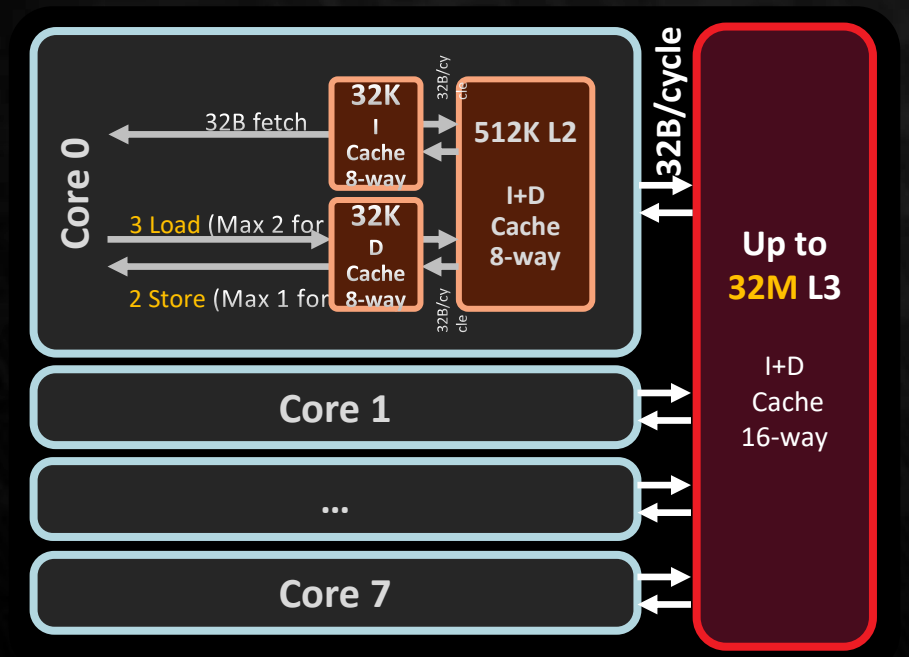
Lower TSV capacitance, inductance

Compared to Micro Bump 3D

Based on AMD engineering internal analysis, May 2021 See endnotes.

3D AMD V-CACHE: ARCHITECTURE

Zen 3 Cache Hierarchy



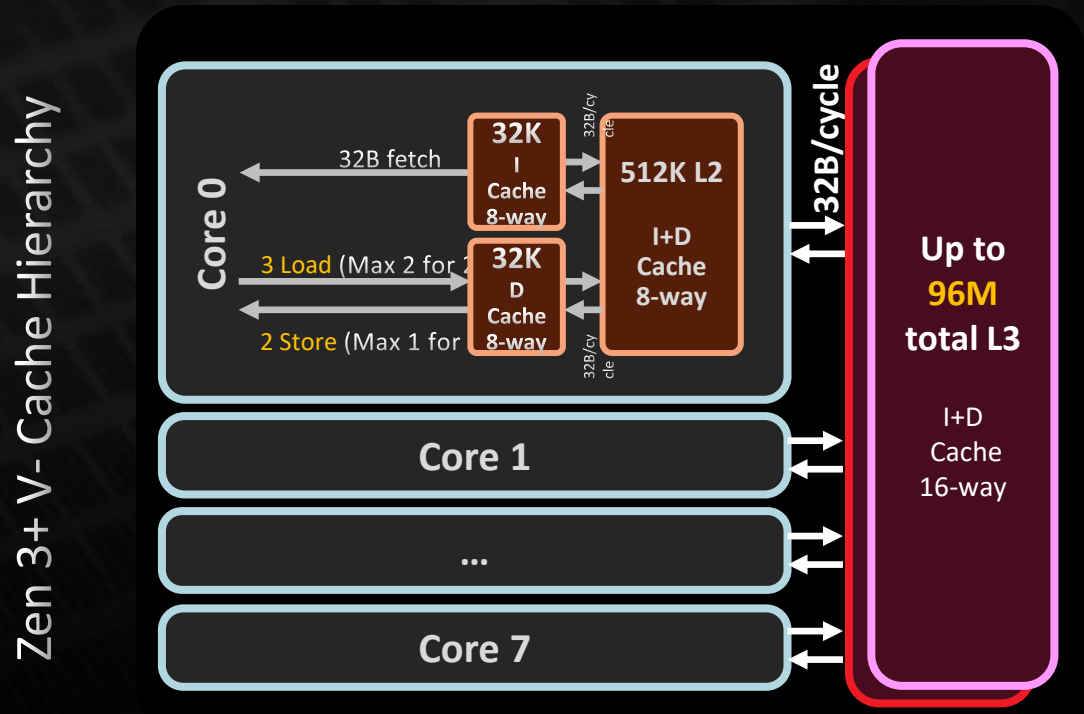
8 Cores per CCD

- 32K I-Cache + 32K D-Cache
- Private 512K L2 per core

Shared 32MB L3 between 8 cores

- 16-way set associative; 32B/cycle interface to each core
- DECTED ECC for enhanced data reliability

Zen 3+ V- Cache Hierarchy



96MB shared L3 Cache between 8 cores

- 16-way set associative
- 32B/cycle interface to each core

>2 TB/s L3 bandwidth; +4 cycles latency

Each die's L3 includes its own

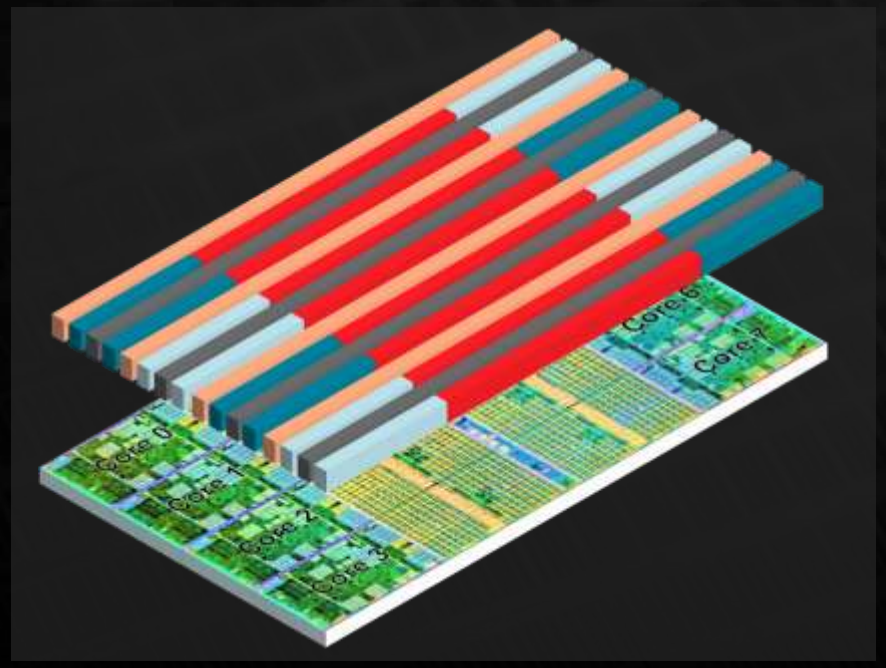
- Data arrays, Tag arrays, LRU arrays

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

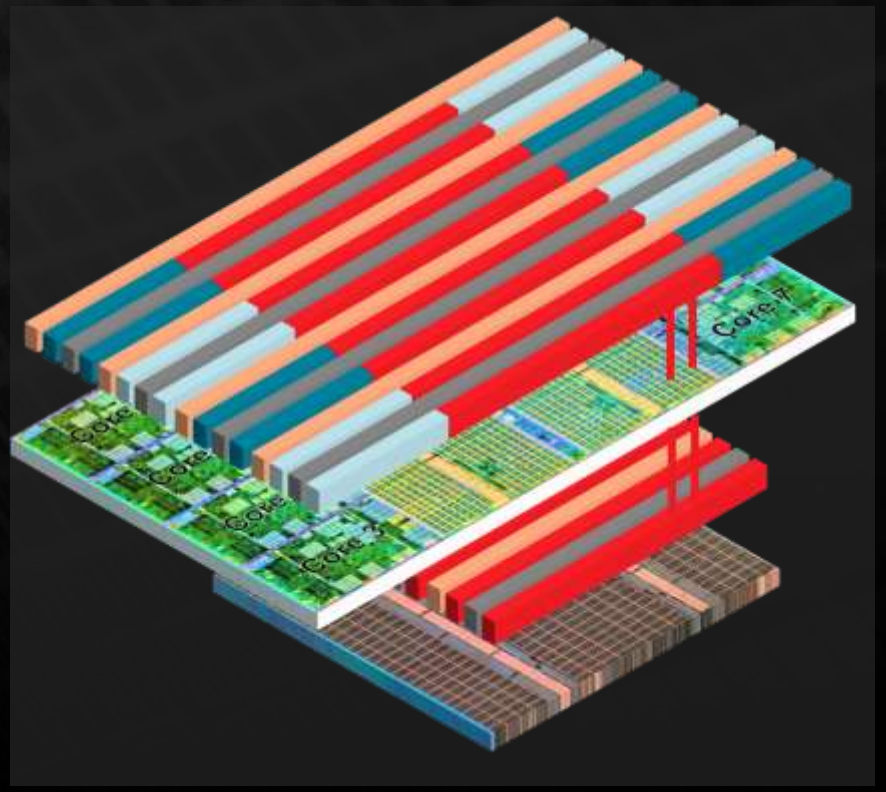
AMD

3D AMD V-CACHE: POWER DELIVERY

Zen 3 PDN



Zen 3+ V- Cache PDN



CCD primary supplies

- RVDD: Ungated supply
- VDD: Per-core gated supply
- VDDM: Gated L2/L3 SRAM supply

L3D implementation

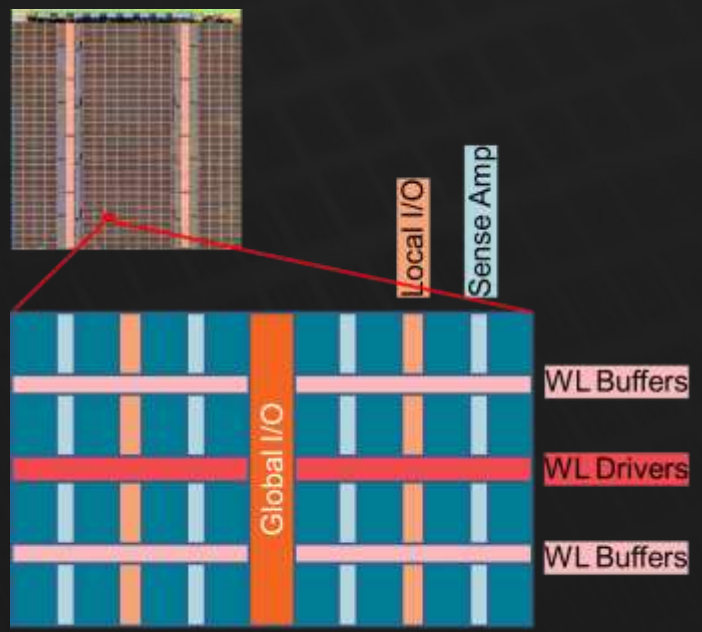
- RVDD and VDDM delivered to L3D through power TSVs
- Power TSVs in channels between CCD array macros

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

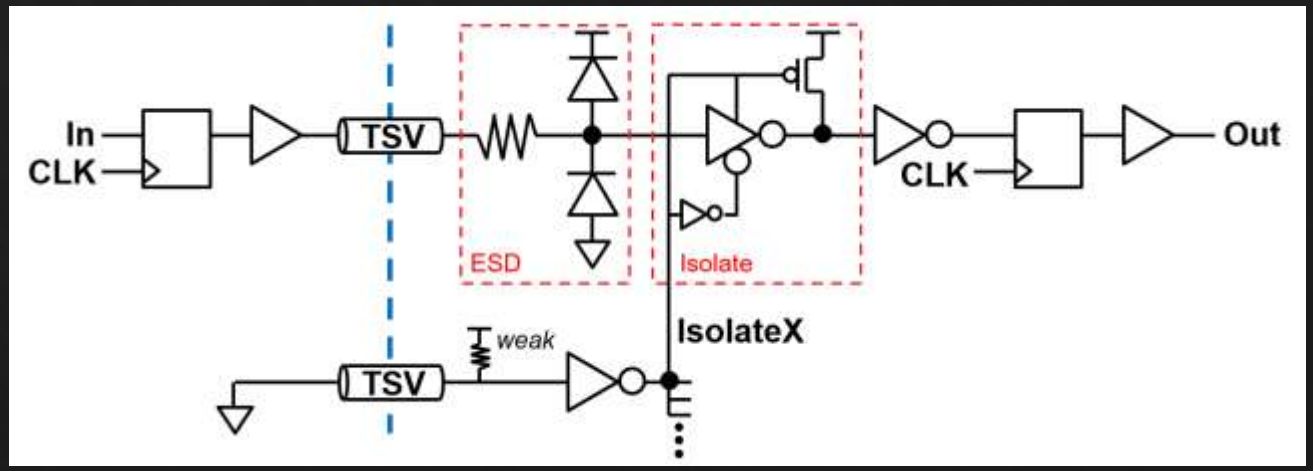
AMD

3D AMD V-CACHE: SIGNAL INTERFACE

L3D SRAM Arrays



Interface electricals



L3D consists of

512 128kB data macros + 1088 6kB tag macros

Dual-rail array design

VDDM powers bitcells, RVDD powers peripheral circuits

Simple Digital interface between dies

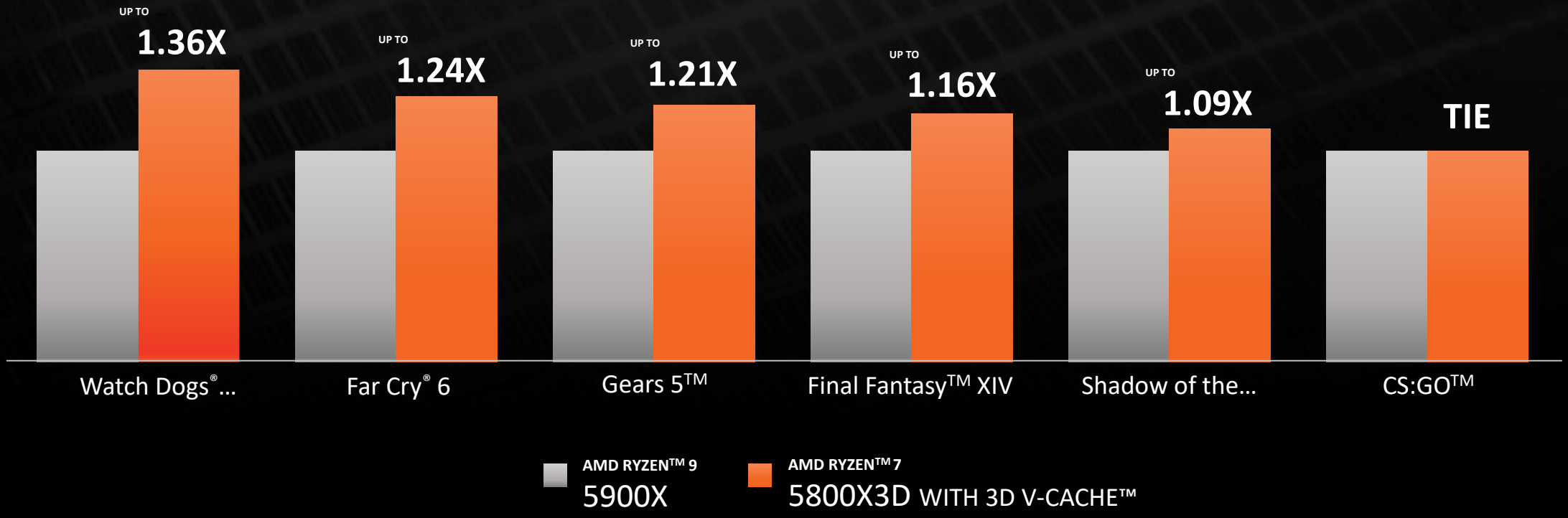
Enabled by Hybrid Bond technology's low parasitics

Wuu et.al, 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU, 2022 IEEE International Solid-state circuits conference

AMD

DESKTOP GAMING PERFORMANCE

AMD RYZEN™ 7 5800X3D WITH AMD 3D V-CACHE™



ADVANCED PACKAGING ENABLING GENERATIONAL PERFORMANCE GAINS

Based on AMD engineering internal analysis, May 2021

AMD

SERVER PERFORMANCE

3X	768MB	UP TO 64	SP3
L3 CACHE COMPARED TO 3RD GEN AMD EPYC CPUS	L3 CACHE PER SOCKET	"ZEN 3" CORES	SOCKET COMPATIBLE

50% AVERAGE UPLIFT ACROSS TARGETED WORKLOADS

ELECTRONIC DESIGN AUTOMATION

24.4 JOBS/HOUR

3RD GEN AMD EPYC™ 16-CORE WITHOUT AMD 3D V-CACHE

~66% FASTER RTL VERIFICATION

SYNOPSYS® VCS®

40.6 JOBS/HOUR

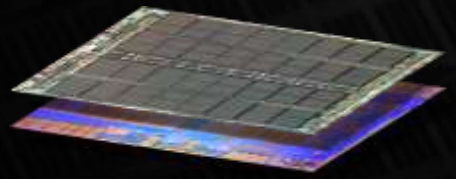
3RD GEN AMD EPYC™ 16-CORE WITH AMD 3D V-CACHE

SEE ENDNOTES: MLNX-021R

FUTURE OF 3D STACKING

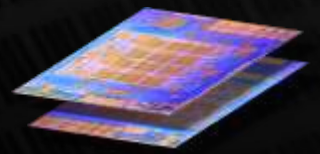
TSV Pitch

DRAM on CPU



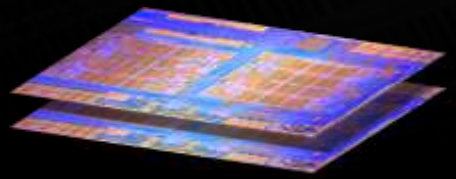
Full die-to-die

Cores on Cores

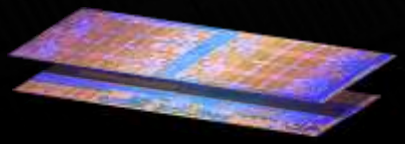


IP on IP

CPU on CPU

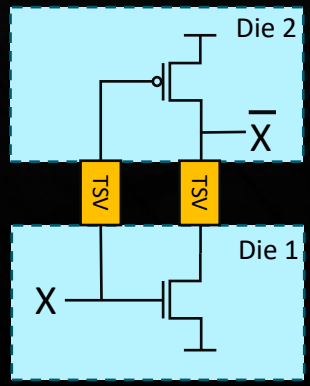


Cores on Uncore



Macro on Macro

IP Folding/Splitting



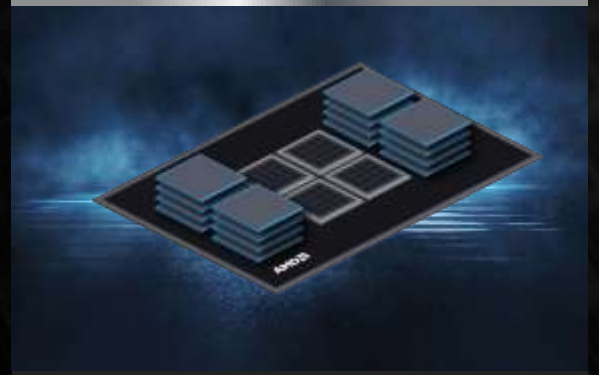
Circuit Slicing

ADVANCED PACKAGING CAN ENABLE INTEGRATION SCHEMES
NOT POSSIBLE WITH MONOLITHIC DESIGNS

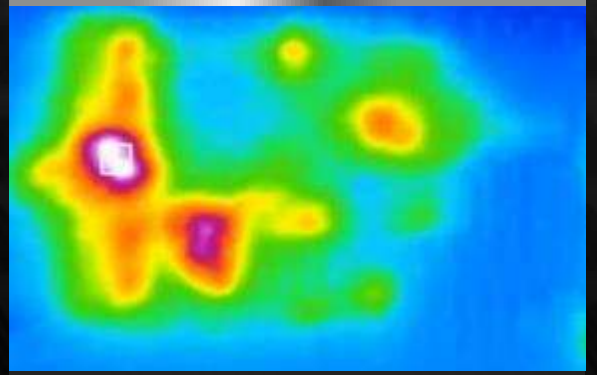
Based on AMD engineering internal analysis, May 2021

AMD

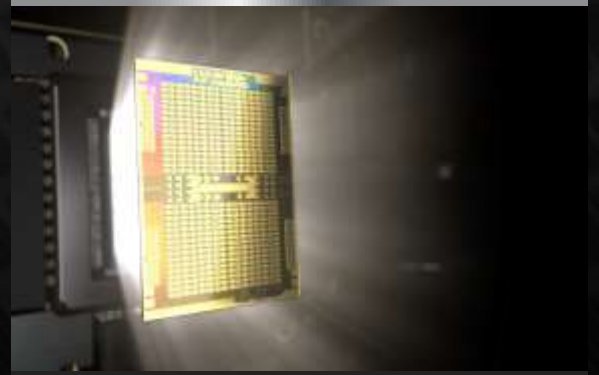
MANY AREAS NEED FURTHER INNOVATION



TEST AND KGD



THERMALS



POWER DELIVERY

SYSTEM-LEVEL INTEGRATION

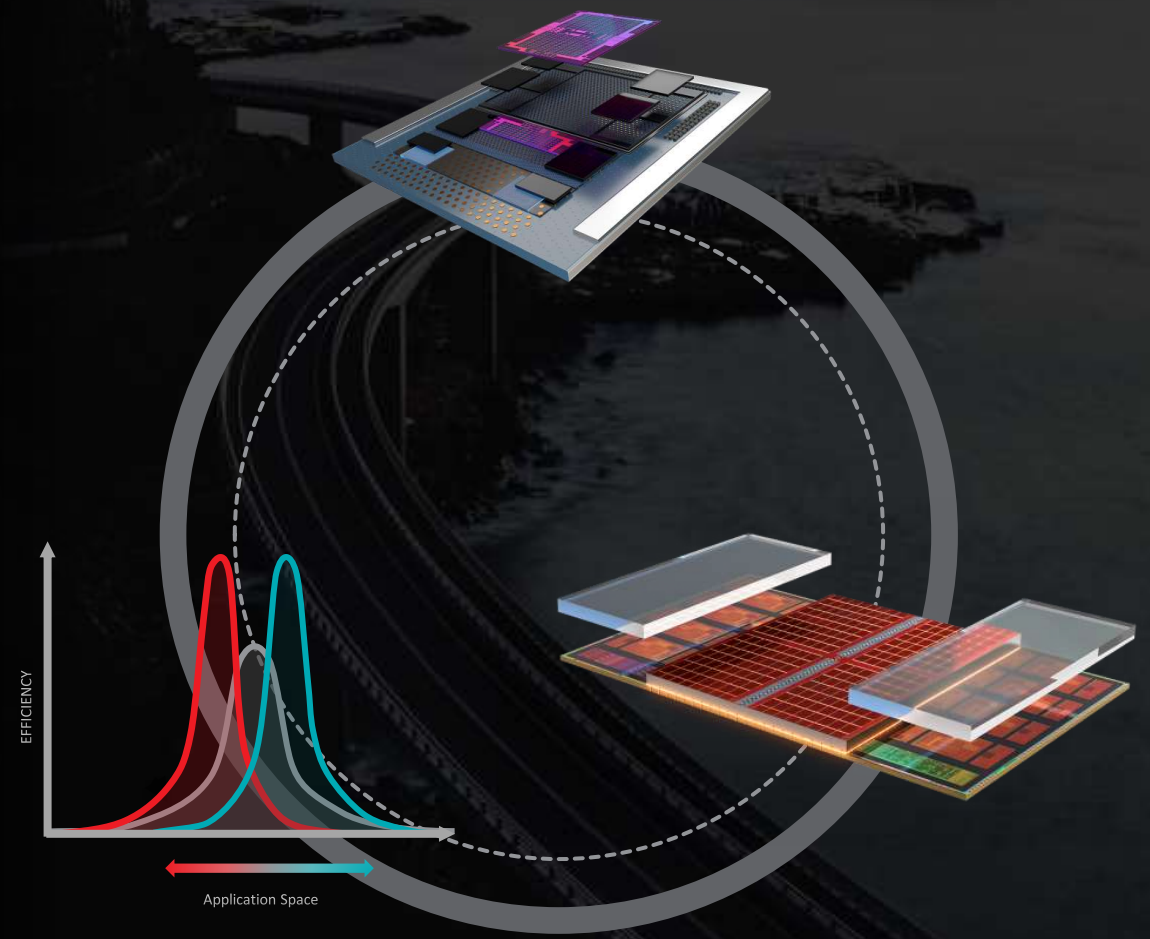
3DIC DTCO

AMD

THE ROAD AHEAD

TAILORED COMPUTING SOLUTIONS

- Technology trends are pushing the industry to heterogeneous domain specific computing
- Economics require small die and chiplet architectures
- Advanced packaging is the frontier for technology-architecture synergy enabling cost-effective, efficient designs



AMD

ENDNOTES

AMD 3D Chiplet Technology

Competition 3D architecture picture from SystemPlus. Intel Core i5-L16G7: the first utilization of Intel's Foveros Technology with Package-on-Package configuration in a consumer product.. https://www.systemplus.fr/reverse-costing-reports/intel-foveros-3d-packaging-technology/

3D Chiplet Gaming Demo And Performance Chart

Testing by AMD performance labs as of April 28, 2021 based on the average FPS of 32 PC games at 1920x1080 with the High image quality preset using an AMD Ryzen™ 9 5900X processor vs. 12-Core 3D Chiplet Prototype. Results may vary. R5K-078.

ENDNOTES

MLNX-001R: EDA RTL Simulation comparison based on AMD internal testing completed on 9/20/2021 measuring the average time to complete a test case simulation. comparing: 1x 16C 3rd Gen EPYC CPU with AMD 3D V-Cache Technology versus 1x 16C AMD EPYC™ 73F3 on the same AMD "Daytona" reference platform. Results may vary based on factors including silicon version, hardware and software configuration and driver versions.

MLNX-021R: AMD internal testing as of 09/27/2021 on 2x 64C 3rd Gen EPYC with AMD 3D V-Cache (Milan-X) compared to 2x 64C AMD 3rd Gen EPYC 7763 CPUs using cumulative average of each of the following benchmark's maximum test result score: ANSYS® Fluent® 2021.1, ANSYS® CFX® 2021.R2, and Altair Radioss 2021. Results may vary.

MLN-075A: Altair™ Radioss™ comparison based on AMD internal testing as of 09/27/2021 measuring the time to run the neon, t10m, and venbatt test case simulations using a server with 2x AMD EPYC 75F3 versus 2x Intel Xeon Platinum 8362. Neon crash impact is the max result test case. Results may vary.

MLN-080B: ANSYS® CFX® 2021.1 comparison based on AMD internal testing as of 09/27/2021 measuring the average time to run the Release 14.0 test case simulations (converted to jobs/day - higher is better) using a server with 2x AMD EPYC 75F3 utilizing 1TB (16x 64 GB DDR4-3200) versus 2x Intel Xeon Platinum 8380 utilizing 1TB (16x 64 GB DDR4-3200). Results may vary.

MLN-130A: ANSYS® Mechanical® 2021 R2 comparison based on AMD internal testing as of 09/27/2021 measuring the average of all Release 2019 R2 test case simulations using a server with 2x AMD EPYC 75F3 versus 2x Intel Xeon Platinum 8380. Steady state thermal analysis of a power supply module 5.3M (cg1) is max result. Results may vary.

MI200-01 - World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.

MI200-02 - Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X accelerator (128GB HBM2e OAM module) at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak double precision matrix (FP64 Matrix) theoretical, floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator resulted in 19.5 TFLOPS peak double precision (FP64 Tensor Core) theoretical, floating-point performance. Results found at:
https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.

MI200-07 - Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). The highest published results on the NVidia Ampere A100 (80GB) SXM GPU accelerator resulted in 2.039 TB/s GPU memory bandwidth performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf

MI200-15A - Testing Conducted by AMD performance lab as of 10/7/2021, on a single socket Optimized AMD EPYC™ CPU server, with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology, using LAMMPS ReaxFF/C, patch_2Jul2021 plus AMD optimizations to LAMMPS and Kokkos that are not yet available upstream resulted in a median score of 4x MI250X = 19,482,180.48 ATOM-Time Steps/s Vs. Dual AMD EPYC 7742@2.25GHz CPUs with 4x NVIDIA A100 SXM 80GB (400W) using LAMMPS classical molecular dynamics package ReaxFF/C, patch_10Feb2021 resulted in a published score of 8,850,000 (8.85E+06) ATOM-Time Steps/s. https://developer.nvidia.com/hpc-application-performance 19,482,180.48/8,850,000=2.20x (220%) the/1.2x (120%) faster. Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:lammps Information on LAMMPS: https://www.lammps.org/index.html Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

ENDNOTES:

MI200-16A - Testing Conducted by AMD performance lab as of 10/18/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU powered server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology, using HACC, plus AMD optimizations to HACC that are not yet available upstream resulted in a median score of 1x MI250X = 4,400,000 (4.40E+06) Particles/s Vs. Testing Conducted by AMD performance lab as of 10/18/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W), using HACC resulted in a median score of 1x A100 = 2,290,000 (2.29E+06) Particles/s. Information on HACC: https://asc.llnl.gov/coral-2-benchmarks/gpu-versions-and-other-supplementary-material https://asc.llnl.gov/sites/asc/files/2020-09/coral-hacc-benchmark-summary-v1.7.pdf Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-17A - Testing conducted by AMD performance lab as of 10/13/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology, using LSMS, plus AMD optimizations to LSMS that are yet available upstream resulted in a median score of 1x MI250X = 3,950,000,000 (3.95E+09) Atom Interactions/s Vs. Testing conducted by AMD performance lab as of 9/27/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W), using LSMS resulted in a median score of 2,440,000,000 (2.44E+09) Atom Interactions/s. Information on LSMS: https://github.com/mstsuite/lsms, Information on GFortran: https://gcc.gnu.org/fortran/, Information on GCC Compiler: https://gcc.gnu.org/ Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-19A - Testing Conducted by AMD performance lab as of 10/1/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology running AMG (Set up) FOM, resulting in a median score of 4x MI250X = 16,773,660,000 FOM_Setup / Sec (Setup Phase Time) Vs. Testing Conducted by AMD performance lab as of 10/1/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 4x NVIDIA A100 SXM 80GB (400W) running AMG (Set up) FOM, resulting in a median score of 4x A100 = 5,507,144,000 FOM_Setup / Sec (Setup Phase Time). Information on AMG_Setup: https://asc.llnl.gov/coral-2-benchmarks , https://asc.llnl.gov/sites/asc/files/2020-09/AMG_Summary_v1_7.pdf , Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-20A - Testing Conducted by AMD performance lab as of 10/1/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server, with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology using AMG (Solve) FOM resulting in a median score of 4x MI250X = 73,318,380,000 FOM_Solve / Sec (Solve Phase Time) Vs. Testing Conducted by AMD performance lab as of 10/1/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 4x NVIDIA A100 SXM 80GB (400W), using AMG (Solve) FOM resulting in a median score of 4x A100 = 31,476,470,000 FOM_Solve / Sec (Solve Phase Time). Information on AMG_Solve: https://asc.llnl.gov/coral-2-benchmarks, https://asc.llnl.gov/sites/asc/files/2020-09/AMG_Summary_v1_7.pdf Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-21A - Testing Conducted by AMD performance lab as of 9/22/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using Nvidia Nbody 32 CUDA sample version 11.2.152 converted to HIP plus AMD optimizations to Nbody 32 that are not yet available upstream resulting in a median score of 2.3x MI250X = 31.72 Particles (Body-to-Body) Interactions/s Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using Nbody 32 sample code version 11.2.152 resulting in a median score of 14.12 Particles (Body-to-Body) Interactions/s. Information on Nbody 32: https://developer.download.nvidia.com/compute/DevZone/C/html_x64/Physically-Based_Simulation.html, https://github.com/AMD-HPC/nbody-nvidia. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-22A - Testing Conducted by AMD performance lab as of 9/22/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250 X OAM GPU (128GB HBM2e) with AMD Infinity Fabric™ technology, using Nbody 64 CUDA Sample version 11.2.152 converted to HIP. Nvidia Nbody 64 samples code version 11.2.152, plus AMD optimizations to Nbody 64 that are not yet available upstream resulted in a median score of 19.245 Particles (Body-to-Body) Interactions/s. Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark Nvidia Nbody 64 sample code version 11.2.152 resulting in a median score of 7.631 Particles (Body-to-Body) Interactions/s. Information on Nbody 64: https://developer.download.nvidia.com/compute/DevZone/C/html_x64/Physically-Based_Simulation.html, https://github.com/AMD-HPC/nbody-nvidia. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI200-23A - Testing Conducted by AMD performance lab as of 10/6/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using Quicksilver - LLNL-CODE-684037 converted to HIP, plus AMD optimizations to Quicksilver that are on AMD Github branch resulted in a median score of 214,000,000 Segments/s Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using Quicksilver - LLNL-CODE-684037 run with CUDA code version 11.2.152 resulted in a median score of 85,500,000 Segments/s. Information on Quicksilver: AMD branch based on LLNL version for this testing: https://github.com/moes1/Quicksilver/tree/AMD-HIP , LLNL version: https://github.com/LLNL/Quicksilver & Quicksilver info sheet: https://hpc.llnl.gov/sites/default/files/Quicksilver_CTS.pdf. Note: A proxy app for the Monte Carlo Transport Code, Mercury. LLNL-CODE-684037. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

ENDNOTES:

MI200-24A - Testing Conducted by AMD performance lab as of 10/12/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using benchmark OpenMM_amoebagk v7.6.0, (converted to HIP) and run at double precision (8 simulations*10,000 steps) plus AMD optimizations to OpenMM_amoebagk that are not yet upstream resulted in a median score of 387.0 seconds or 223.2558 NS/Day Vs. Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark OpenMM_amoebagk v7.6.0, run at double precision (8 simulations*10,000 steps) with CUDA code version 11.4 resulted in a median score of 921.0 seconds or 93.8111 NS/Day. Information on OpenMM: https://openmm.org/ Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-25A - Testing Conducted by AMD performance lab as of 9/30/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology using MILC benchmark version 7.8.1 developer version MILC_QCD on Github, Apex Medium test module, plus AMD optimizations to MILC that are not yet available upstream resulted in a median score 1,604.567 Total Time (Seconds). Vs. Dual AMD EPYC 7742@2.25GHz CPUs with 1x NVIDIA A100 SXM 80GB (400W) using MILC benchmark version develop_c30ed15e (quda0.8-patch4Oct2017), Apex Medium test module, resulted in a published score of 2,262 Total Time (Seconds). https://developer.nvidia.com/hpc-application-performance Nvidia MILC Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:milc Information on MILC: https://web.physics.utah.edu/~detar/milc/ MILC Manual Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-26A - Testing Conducted by AMD performance lab as of 10/14/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server, with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using benchmark HPL v2.3, plus AMD optimizations to HPL that are not yet upstream resulted in a median score of 42.26 TFLOPS Vs. Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark HPL Nvidia container image 21.4-HPL resulting in a median score of 15.33 TFLOPS. Information on HPL: https://www.netlib.org/benchmark/hpl/ Nvidia HPL Container Detail: https://ngc.nvidia.com/catalog/containers/nvidia:hpc-benchmarks Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI200-27 - The AMD Instinct™ MI250X accelerator has 220 compute units (CUs) and 14,080 stream cores. The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores. MI200-27

MI200-31 - As of October 20th, 2021, the AMD Instinct™ MI200 series accelerators are the “Most advanced server accelerators (GPUs) for data center,” defined as the only server accelerators to use the advanced 6nm manufacturing technology on a server. AMD on 6nm for AMD Instinct MI200 series server accelerators. Nvidia on 7nm for Nvidia Ampere A100 GPU. https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/ MI200-31

MI200-33 - Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Calculations by AMD Performance Labs as of OCT 5th, 2020 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI100 memory bus interface is 4,096 bits and memory data rate is 2.40 Gbps for total memory bandwidth of 1.2288 TB/s ((2.40 Gbps*4,096 bits)/8) MI200-33

MI100-03 - Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak double precision (FP64), 46.1 TFLOPS peak single precision matrix (FP32), 23.1 TFLOPS peak single precision (FP32), 184.6 TFLOPS peak half precision (FP16) peak theoretical, floating-point performance. Published results on the NVidia Ampere A100 (40GB) GPU accelerator resulted in 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16) theoretical, floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-03

MI100-04 - Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 184.57 TFLOPS peak theoretical half precision (FP16) and 46.14 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 26.5 TFLOPS peak theoretical half precision (FP16) and 13.25 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-04

ENERGY EFFICIENCY METHODOLOGY ENDNOTES

- The goal is calculated by taking the performance of these compute nodes as measured by standard performance metrics
  - HPC: Linpack DGEMM kernel with 4k matrix size
  - AI training: lower precision training-focused floating-point math GEMM kernels such as FP16 or BF16 operating on 4k matrices.
- Divide these FLOP rates by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.
  - Segment-specific datacenter power utilization effectiveness (PUE) and with equipment utilization taken into account.
- AMD energy consumption is calculated using internal targets of improvement in energy efficiency across each computing segment - reflecting expected growth rates in each GPU for HPC and AI computing segments relative to 2020 worldwide shipments
  - Applied to the same computation requirement as projected assuming baseline trends.
  - The energy consumption baseline uses the same industry energy per operation improvement rates as from 2015-2020, extrapolated to 2025.
- The measure of energy per operation improvement in each segment from 2020-2025 is weighted by the projected worldwide volumes multiplied by the Typical Energy Consumption (TEC) of each computing segment to arrive at a meaningful metric of actual energy usage improvement worldwide.

COPYRIGHT AND DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

This information is provided ‘as is.” AMD makes no representations or warranties with respect to the contents hereof and assumes no responsibility for any inaccuracies, errors, or omissions that may appear in this information. AMD specifically disclaims any implied warranties of non-infringement, merchantability, or fitness for any particular purpose. In no event will AMD be liable to any person for any reliance, direct, indirect, special, or other consequential damages arising from the use of any information contained herein, even if AMD is expressly advised of the possibility of such damages.

AMD


AMD