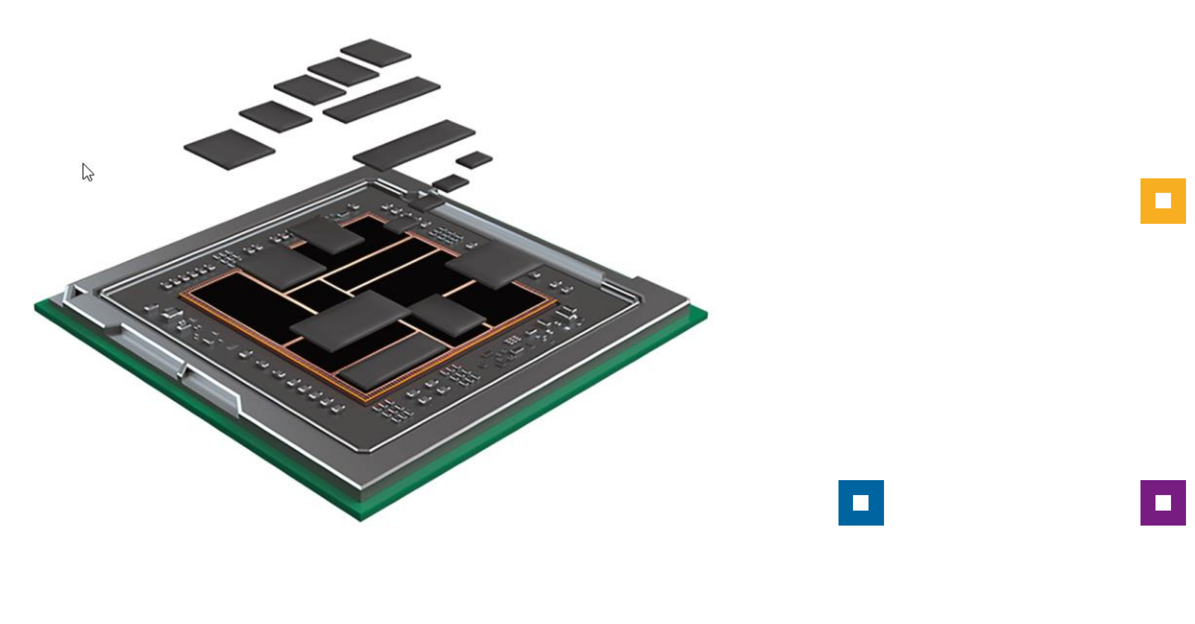
Beyond Moore's Law: Leveraging Advanced Packaging Technology

Cătălin Ciobanu, Transilvania University of Brașov (UnitBV)



Moore’s Law

Goordon Moore (1929 – 2023)

- Prof. Carver Mead popularized the phrase „Moore’s Law”
- „Moore's law is the observation that the number of transistors in an integrated circuit (IC) doubles about every two year”
- Several factors contributing to this exponential behavior (1975 IEEE International Electron Devices Meeting)
  - Metal-oxide semiconductor (MOS) technology
  - **The exponential rate of increase in die sizes** coupled with a decrease in defective densities
  - **Finer minimum dimension**

Source: https://en.wikipedia.org/wiki/Gordon_Moore

Transilvania University of Brasov

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Transistor count

50,000,000,000
10,000,000,000
5,000,000,000
1,000,000,000
500,000,000
100,000,000
50,000,000
10,000,000
5,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000

GC2 IPU
AMD Epyc Rome
72-core Xeon Phi
Centriq 2400
AWS Graviton2
SPARC M7
32-core AMD Epyc
IBM z13 Storage Controller
Apple A12X Bionic
18-core Xeon Haswell-E5
HiSilicon Kirin 990 5G
Apple A13 (iPhone 11 Pro)
Xbox One main SoC
AMD Ryzen 7 3700X
61-core Xeon Phi
HiSilicon Kirin 710
12-core POWER8
10-core Core i7 Broadwell-E
8-core Xeon Nehalem-EX
Qualcomm Snapdragon 835
Six-core Xeon 7400
Dual-core + GPU Iris Core i7 Broadwell-U
Dual-core Itanium 2
Quad-core + GPU GT2 Core i7 Skylake K
Pentium D Presler
POWER6
Quad-core + GPU Core i7 Haswell
Apple A7 (dual-core ARM64 "mobile SoC")
Itanium 2 with 9 MB cache
Core i7 (Quad)
AMD K10 quad-core 2M L3
Itanium 2 Madison 6M
Core 2 Duo Wolfdale
Pentium D Smithfield
Core 2 Duo Conroe
Itanium 2 McKinley
Cell
Core 2 Duo Wolfdale 3M
Pentium 4 Prescott-2M
Core 2 Duo Allendale
Pentium 4 Cedar Mill
AMD K8
Pentium 4 Prescott
Pentium 4 Northwood
Barton
Atom
Pentium 4 Willamette
Pentium III Tualatin
Pentium II Mobile Dixon
ARM Cortex-A9
AMD K7
Pentium III Coppermine
AMD K6-III
AMD K6
Pentium III Katmai
Pentium II Deschutes
Pentium Pro
Pentium II Klamath
Pentium
AMD K5
SA-110
Intel 80486
R4000
TI Explorer's 32-bit Lisp machine chip
ARM700
Intel 80386
Intel i960
ARM 3
Motorola 68020
DEC WRL MultiTitan
Intel 80286
ARM 9TDMI
Motorola 68000
Intel 80186
Intel 8086
Intel 8088
ARM 2
ARM 1
ARM 6
WDC 65C816
Novix NC4016
TMS 1000
Zilog Z80
Motorola 6809
WDC 65C02
RCA 1802
Intel 8085
Intel 8080
Intel 8008
Motorola 6800
MOS Technology 6502
Intel 4004

1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020

Year in which the microchip was first introduced

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Sursa: https://en.wikipedia.org/wiki/Moore%27s_law

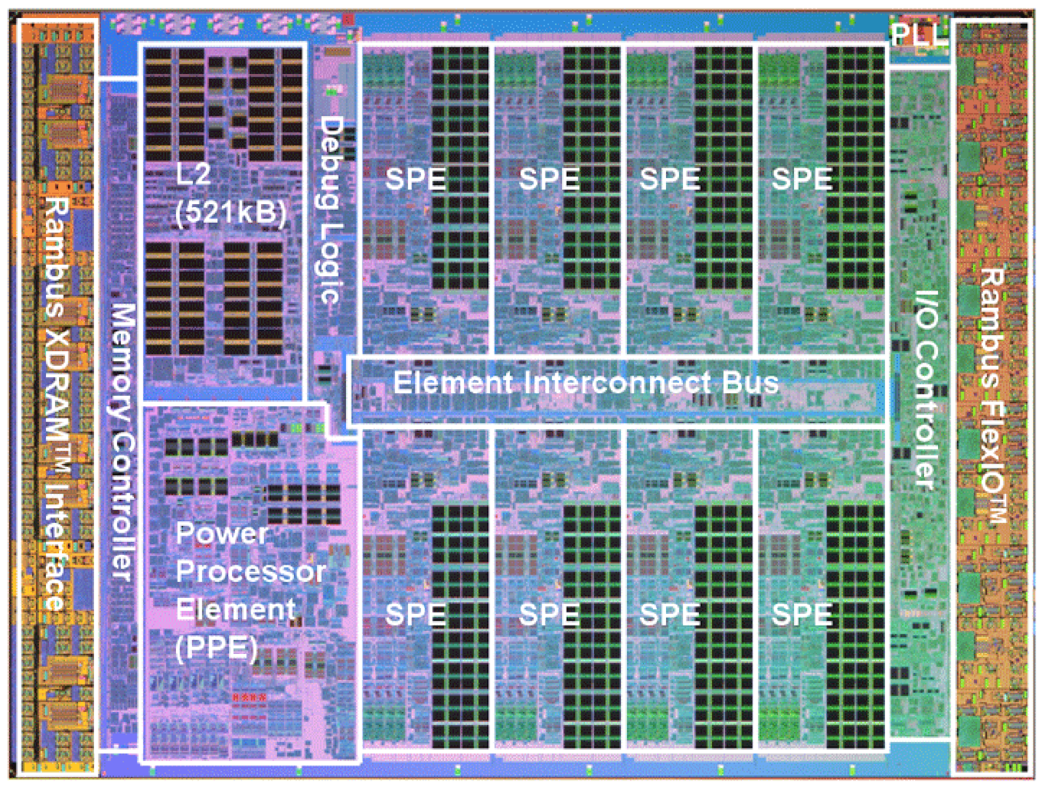
Beyond Moore's Law

- New **challenges** in semiconductor industry – multiple **constraints in advanced chip design**
  - **Power** limitations
  - **Thermal** limitations
  - **Limited** Instructions Per Clock (**IPC**) gains in new generations of processors
- Industry **trends**
  - **Multicore** designs
  - Wide use of **accelerators**
- Wafer costs for each new technology node are rising
  - **Monolithic** designs are **prohibitively expensive**
    - **Lower yields** inherent to large chips
- Reticle limit for future High-NA (Numerical Aperture) Extreme Ultraviolet (EUV) Lithography will be halved
    - **856mm²** -> **429mm²** due to the use of an amorphous lens array[1]

1: https://en.wikichip.org/wiki/mask

CELL Broadband Engine (2006) - Heterogeneous Multicore

- Power Processor Element (PPE) for control tasks
- Synergistic Processor Elements (SPE) for data-intensive processing



Source: https://commons.wikimedia.org/wiki/File:CELL_BE_processor_PS3_board.jpg
https://www.realworldtech.com/cell-shrink/3/

Solution – Advanced packaging - Chiplet technology

- ”Chiplet technology is a microelectronics design and manufacturing approach where multiple smaller dies or chiplets are combined into a single package, with each of chiplets performing a specific function.”[1]
- **Multiple chiplets** – connected to form a SoC-like solution
    - **Disaggregate** large designs
    - **Avoid** the silicon **reticle size limitation**
- Multiple **advantages** when using chiplets
    - **Flexibility**
    - **Scalability**
    - **Cost savings** – off-the-shelf chiplets
    - Mixing functions **from different process nodes**

Source: 1: Anachronix Chiplet Linked Webinar 2023

Chiplets Motivation

Chiplets Motivation

CHIPLET TECHNOLOGY

OUR MOTIVATION

Limited and Divergent Scale Factors

Silicon Area Scaling by Function

ANALOG I/O

MEMORY

LOGIC

28nm 16nm 10nm 7nm 5nm

Density gains diminishing

Increasing Costs[1]

Cost per yielded mm^2

6 4 2

45nm 32nm 28nm 20nm 14/16nm 7nm 5nm

Costs Increasing

9

[1] AMD Internally estimated cost per yielded mm2 for a 250mm2 die

AMD
together we advance_gaming

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

Moore’s Law Keeps Slowing

Barriers to Large Cache Size

MOORE’S LAW KEEPS SLOWING

90nm, 65nm, 45nm, 32nm, 22nm, 14nm, 10/7nm

2004 2007 2010 2013 2016 2019 [1]

Silicon Area Scaling by Function

Analog

SRAM

Logic

28nm 16nm 10nm 7nm 5nm

WHILE COSTS CONTINUE TO INCREASE

Normalized Cost/Yielded mm

6 4 2 0

45nm 32nm 28nm 20nm 14nm 7nm 5nm [2]

- **SRAM scaling & cost = barriers to large on-die caches**
  - Cost
  - Product flexibility
  - Latency

} Chiplets

© 2022 IEEE
International Solid-State Circuits Conference

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

26.4: 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU

6 of 36

Source: IEEE ISSCC 2022

Slowing of Moore’s Law + Increasing Cost

[AMD Official Use Only]

Technology Headwinds to Meeting the Demand

Slowing of Moore’s Law

90nm
65nm
45nm
32nm
22nm
14nm
10/7nm
2004 2007 2010 2013 2016 2019 [1]

Increasing Cost

Normalized Cost/yielded mm
6
4
2
-
45nm 32nm 28nm 20nm 14/16nm 7nm 5nm [2]

Reticle Limit

Die Size (MM²)
1000
100
2006 2009 2012 2014 2017 2020 [1]
Server CPU
GPU

Power of Top 100 Supercomputers

100 MW
10 MW
1 MW
100 KW
10 KW
2009 2019 [3]

4 DATA CENTER ENERGY EFFICIENCY PRESS SLIDES

[1] Naffziger, VLSI Short Course, 2020, [2] Cost per yielded mm2 for a 250mm2 die, [3] Data from top500.org

AMD

Source: www.techpowerup.com/287291/amd-announces-ambitious-goal-to-increase-energy-efficiency-of-processors-running-ai-training-and-high-performance-computing-applications-30x-by-2025 (Sep 29th 2021)

Application-Specific optimization, Modular design: chiplets + 3D Stacking

[AMD Official Use Only]

Efficiency through Domain Specific Architectures and Package Innovation

Application-Specific optimization provides better performance-per-Watt

Domain Specific

GPU

General Purpose CPU

Performance/Watt

Applications

AMD is leading the way in finding new approaches to reduce energy for compute

Modular design, chiplets and 3D stacking are the next frontier for efficiency gains

CHIPLETS

3D CHIPLETS

(Chiplet + Advanced 3D Stacking)

6 DATA CENTER ENERGY EFFICIENCY PRESS SLIDES

AMD

Source: www.techpowerup.com/287291/amd-announces-ambitious-goal-to-increase-energy-efficiency-of-processors-running-ai-training-and-high-performance-computing-applications-30x-by-2025 (Sep 29th 2021)

Chiplet Technology

Chiplet Technology and Heterogeneous Integration

- **Transition** from layout of laminate substrates to silicon substrate
- **Electrical and thermal** analysis **challenges**
- Multi-chip Modules (**MCM**) date back to the 1960s
- Sytem in Package (**SiP**) began to replace the term MCM in the late 1990s
- **2.5D IC packaging** - Silicon substrates – high density, using through-silicon vias (TSVs)

PCB
Board size/complexity reduction (SWaP)
System in Package (SiP)
Disaggregated SoC
SoC
MCM/SiP
Unpackaged Die
Bare Die
Bare Die
Chiplet
Chiplet
Chiplet
Chiplet
Heterogenous Integration
Chiplets
Laminate Substrate
Silicon Substrate

PCB to MCM/SiP Benefits

- Smaller footprint
- PCB simplification
- Higher bandwidth
- Lower power

SoC to HI Benefits

- Reduced NRE costs
- Shorter time to market
- Larger than reticle size designs
- More flexible IP use-model

Figure 1: Heterogeneous integration

Source: Cadence: Chiplets and Heterogeneous Packaging Are Changing System Design and Analysis

Chiplet Technology Challenges

Method of Moments
Finite Element Method
Boundary Element Method
Finite Difference Method
Solver Technologies

Electro-Magnetics
Circuit/Power Analysis
Heat Transfer
Fluid Dynamics
Solid Mechanics
Multi-Physics

Chip(let)
Packaging
PCB
Design/Layout Platforms

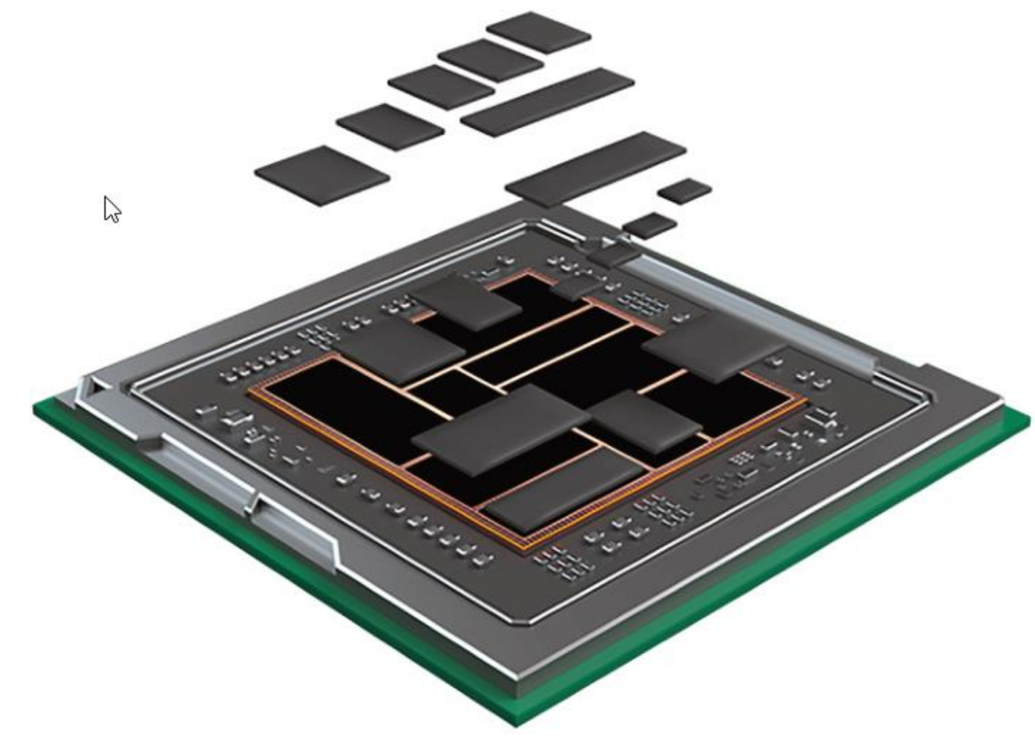
System Design, Optimization & Electrical Signoff

Figure 2: System-level electrothermal analysis

Source: 1: Cadence: Chiplets and Heterogeneous Packaging Are Changing System Design and Analysis

Chiplets Heterogeneous Integration (HI)

- Similar to designing a small PCB
- Each chiplet built with a common/known communication interface
  - PCIe
  - HBM
  - AIB
- Lower development cost – modular integration
- Lower manufacturing costs – purchasing known-good die (KGD)
- Cost advantage – volume manufacturing when reusing the same chiplets in many designs
- Many vendors exploring this space
  - Intel CO-EMIB – EMIB + Foveros in the same package
  - Intel Omni-Directional Interconnect (ODI) – horizontal communication (similar to EMIB) or vertically using TSVs (similar to Foveros)
  - TSMC's Chip-on-Wafer-on-Substrate (CoWoS)



Source: 1: Cadence: Chiplets and Heterogeneous Packaging Are Changing System Design and Analysis

Advanced Packaging Technology Evolution

System in Package (SiP) — 1998
RF Module — 2004
2.5D-IC (Silicon Interposer) — 2010
Ultra-High-Density RDL/FOWLP — 2016
Bump-Less 3D-IC — 2018
Heterogeneous Integration (Disaggregated SoC) Photonics — Now
Monolithic 3D-IC — Future

Figure 4: Evolution of advanced multi-chip(let) packaging technologies

Industry Shift Towards Multi-Die SoCs

- ❑Shift fueled by several converging trends
  - ❑ Some SoCs – too big for manufacturability
  - ❑ Some SoCs require different process nodes for optimal cost/perf
  - ❑ Desire for enhanced product scalability and composability
- ❑Lack of design ecosystem – pause/postpone multi-die projects
- ❑Early adopters developed proprietary die-to-die interfaces
  - ❑ Limits the ability to assemble dies from different verndors
- ❑Solution: standardized die-to-die interconnects
  - ❑ **Optical Interface Forum (OIF)** – The XSR and USR physical layer specifications optimized for die-to-die connectivity
  - ❑ **Chips Alliance** –The **AIB** specification which was originally introduced by Intel
  - ❑ **Open Compute Platform (OCP)** –The OpenHBI and Bunch-of-Wires (BOW) specifications optimized for different use cases
  - ❑ **Unified Chiplet Interconnect Express** (**UCIe**) –A comprehensive die-to-die interconnect specification covering multiple use cases and a complete protocol stack

Source: https://www.synopsys.com/designware-ip/technical-bulletin/ucie-multi-die-socs.html

Die-to-die interconnect standards

Alliance	OIF	OPEN Compute Project®		CHIPS ALLIANCE	UCIe Universal Chiplet Interconnect Express
Standard	XSR	BOW	OHBI	AIB	UCIe
Data Rate	112G / 224G	8G / 16G	8G / 16G	6G	16G / 32G
Protocol	Not Defined	Not Defined	Not Defined	Not Defined	Streaming, PCIe, CXL
Package Types	2D	2D, 2.5D	2D, 2.5D	Bridge	2D, 2.5D, Bridge
Target Applications	Optical Networking (CPO/NPO)	Cost sensitive aggregation cases	High density scale and split cases for data center	Mil-aero ecosystem	Scale & Split w/ streaming Aggregation w/ PCIe/CXL

Source: https://www.synopsys.com/designware-ip/technical-bulletin/ucie-multi-die-socs.html
Synopsys Multi-Die SoCs Gaining Strength with Introduction of UCIe

UCIe: Companies Join Forces – standardized die-to-die interconnect

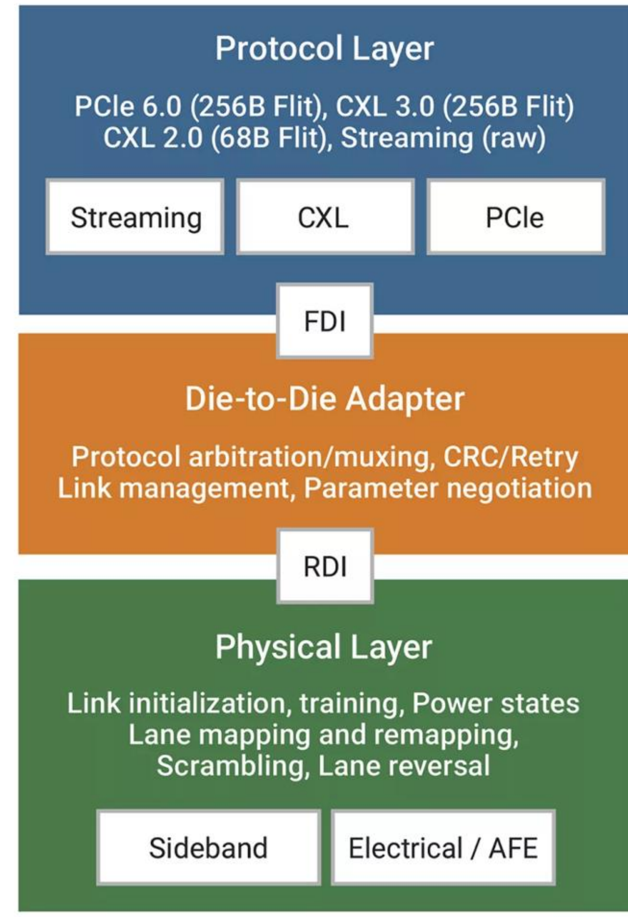
AMD
arm
ASE GROUP
Google Cloud
intel
Meta
Microsoft
Qualcomm
SAMSUNG
tsmc

Source: https://www.synopsys.com/designware-ip/technical-bulletin/ucie-multi-die-socs.html

Synopsys Multi-Die SoCs Gaining Strength with Introduction of UCIe

UCIe: a Complete Stack for Die-to-die interconnect

- ❑ Supports data rates from 8Gbps/pin 16Gbps/pin
  - ❑ Expected to support 32GBps/pin
- ❑ Supports all types of package technology
  - ❑ UCIe for advanced packages (silicon interposer, silicon bridge or RDL fanout)
  - ❑ UCIe for standard packages (organic substrate or laminate)
  - ❑ Both options share the same architecture and protocols
- ❑ Very competitive performance advantages to multi-die SoC designers
  - ❑ High energy efficiency (pJ/b)
  - ❑ High edge usage efficiency (Tbps/mm)
  - ❑ Low latency (ns)
- ❑ UCIe – compelling roadmap
  - ❑ Higher data rates
  - ❑ New protocols
  - ❑ 3d packaging
  - ❑ Security
  - ❑ Testability



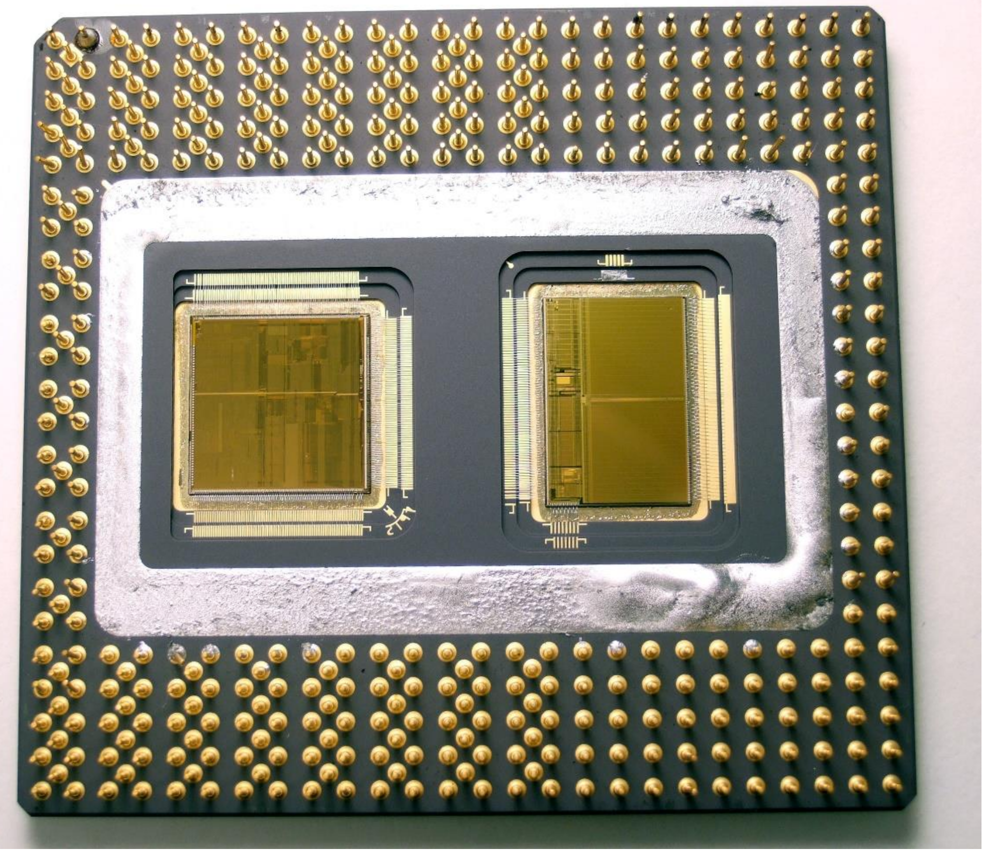
Source: https://www.synopsys.com/designware-ip/technical-bulletin/ucie-multi-die-socs.html
Synopsys Multi-Die SoCs Gaining Strength with Introduction of UCIe

Brief History of Multi Chip Modules

Intel Pentium Pro Multi-Chip Module (MCM) (1995)

Pentium Pro is packaged in Multi-Chip Module (MCM) – “on package cache”

- Separate L2 die in the package, same speed as CPU core
- The dies are connected to the package using conventional wire bonding
- Up to two 512KB cache dies
- 0.35 µm to 0.50 µm

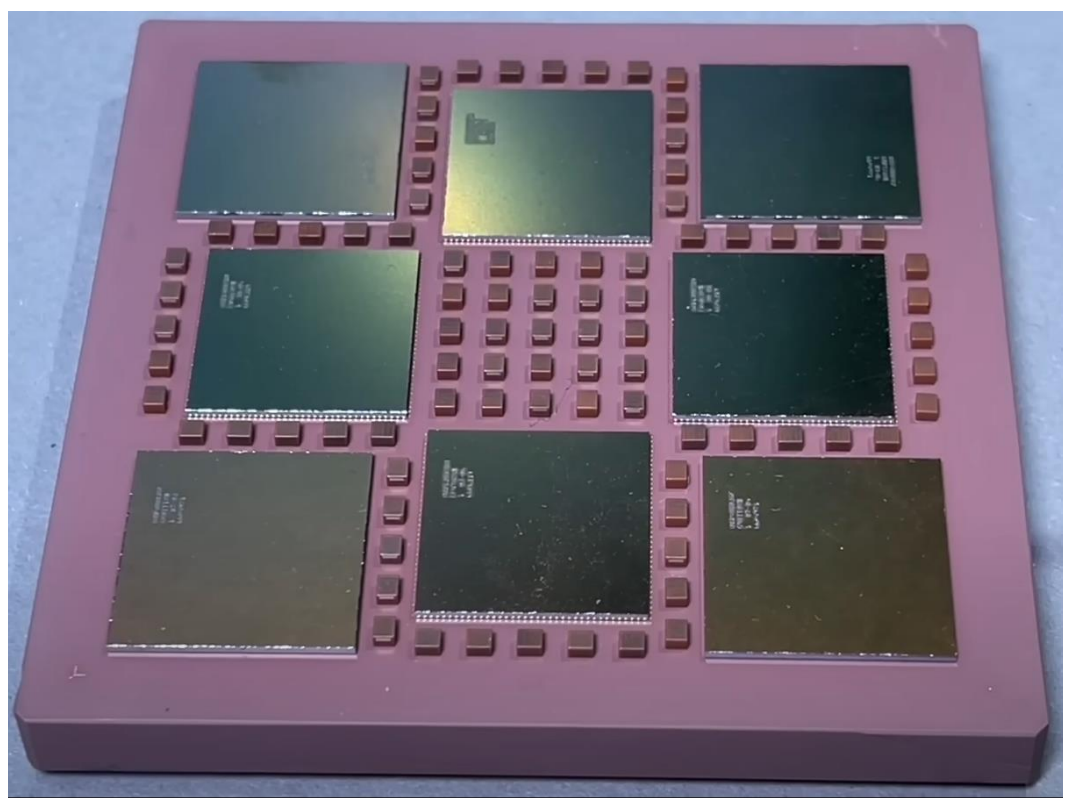


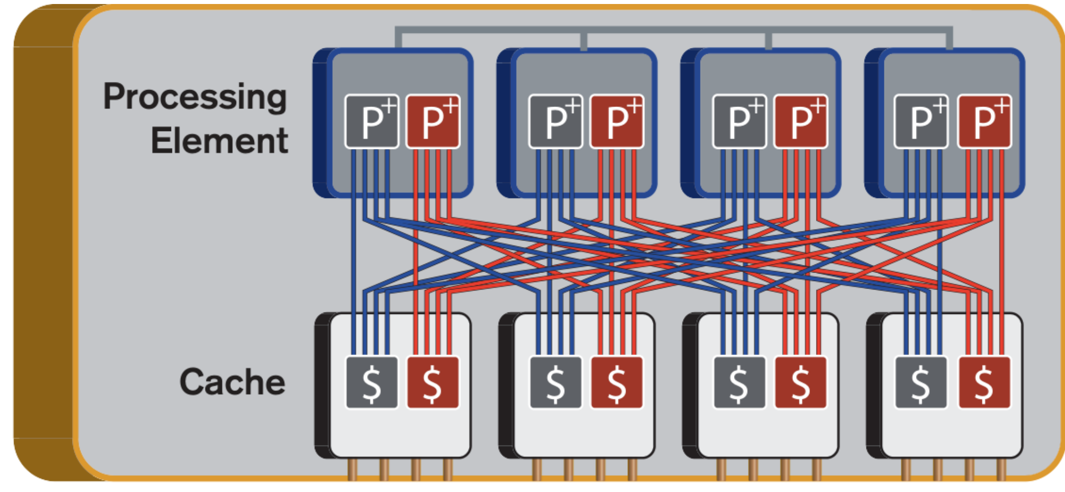
Source: https://en.wikipedia.org/wiki/Pentium_Pro#/media/File:Ppro512K.jpg
https://en.wikipedia.org/wiki/Pentium_Pro#/media/File:Pentiumpro_moshen.jpg

Cray X1 (2003)

8-chip Multi-Chip Module (MCM)

- ❑ 4 processor chips
- ❑ 4 custom streaming cache chips



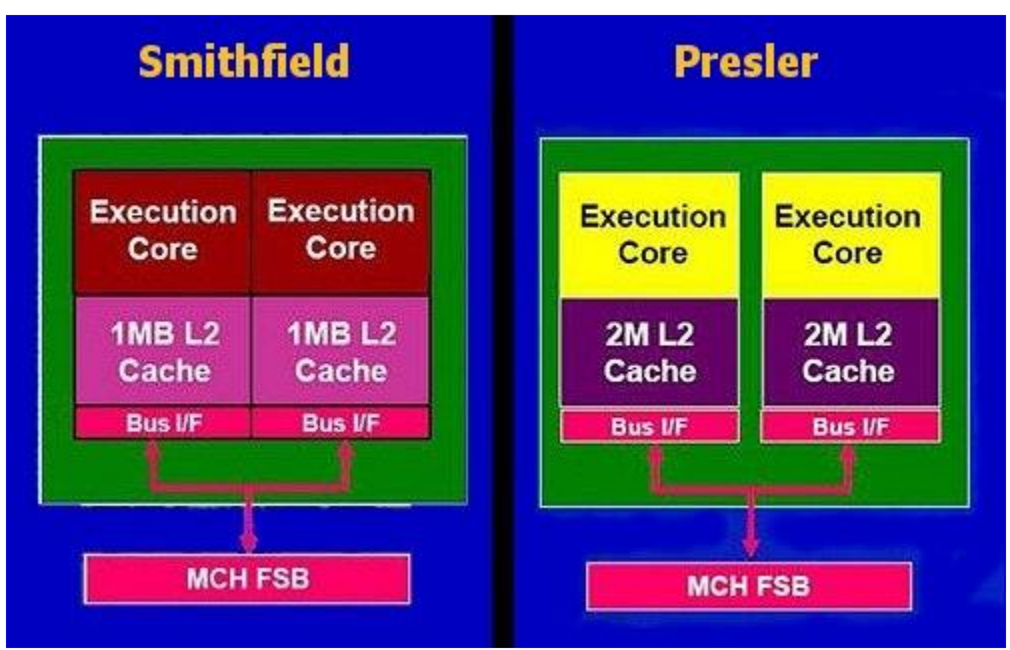


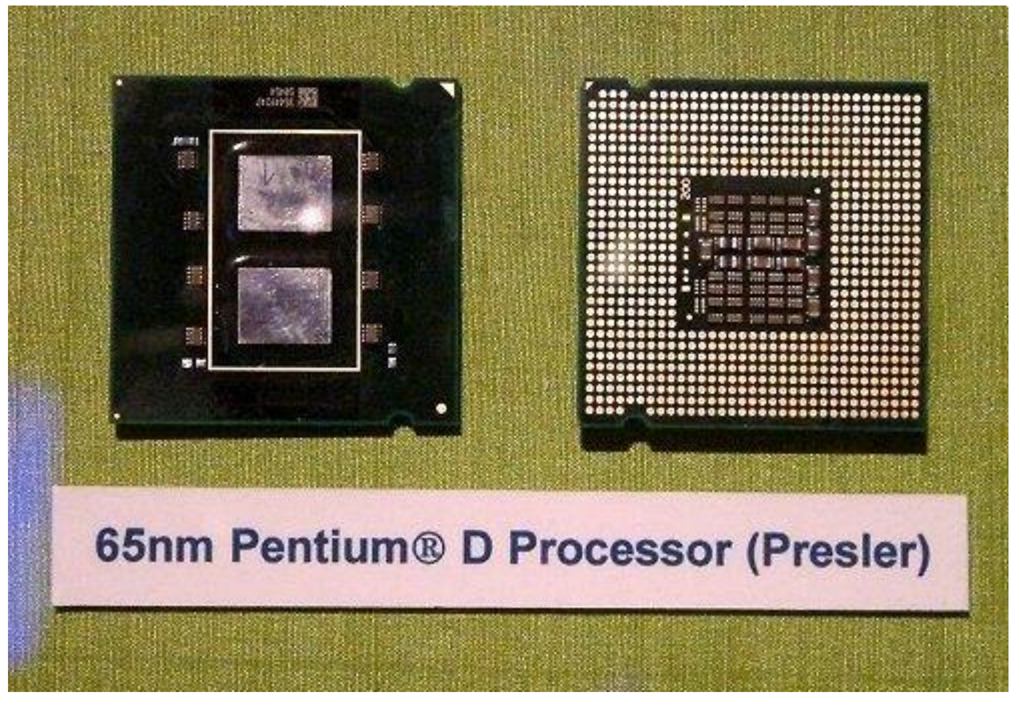
Source: CPU Galaxy - CRAY X1 Multi Chip Module - Monster CPU with 8 Cores – Teardown, Feb 18, 2023 https://www.youtube.com/watch?v=QdtcJTIcqDE
https://www.craysupercomputers.com/downloads/CrayX1/CrayX1E_Datasheet.pdf

Intel Pentium D Presler (2006)

Smithfield Pentium D 800 series had a single die with two cores
- Two single-core dies placed on the same substrate
- Reduced cache available to each core
- 90nm, 206 mm^2

Presler Pentium D 900 series used a Multi-Chip Module (MCM)
- Two single-core dies placed on the same substrate
- Each die could be sold as a Pentium 4 CPU
- Reduced cost due to higher yields
- 65nm, 162 mm^2 for both cores





Source: https://www.hardwarezone.com.sg/m/feature-intels-pentium-xe-965-and-955-processors/presler-core
https://en.wikipedia.org/wiki/Pentium_D

AMD Zen Embraces Chiplets, 3D Vcache

AMD EPYC Naples (2017)

- Four 14nm compute dies, each with 8 cores
- Each die connected directly to the others via Infinity Fabric

INFINITY FABRIC: DIE-TO-DIE INTERCONNECT

- Fully connected Coherent Infinity Fabric within socket
- Optimized low power, low latency MCM links between die
- 42GB/sec bi-dir BW per link, ~2pJ/bit TDP
- Single-ended, low power zero transmission
- Infinity Control Fabric connected between dies

Purpose-built MCM links optimized for power, bandwidth, and latency

72 | AMD EPYC EMBARGOED UNTIL JUNE 20TH AT 3:00 PM CENTRAL U.S. TIME AMD

Source: https://www.techpowerup.com/233381/amd-announces-high-performance-computing-platform-naples-is-epyc
https://en.wikipedia.org/wiki/Epyc
https://www.anandtech.com/show/11551/amds-future-in-servers-new-7000-series-cpus-launched-and-epyc-analysis/2

AMD 2nd Generation EPYC Rome (2019)

2nd gen. EPYC with Zen2 architecture

- 9 7m dies
- 8 7nm Complex Core Die (CCD) chiplets, 8 cores each
- A 14nm IO die
- Connected via second-gen infinity fabric

Source: https://www.hpcwire.com/2019/08/08/amd-launches-epyc-rome-first-7nm-cpu/

AMD Zen2 Rome Chiplets

Server IO Die
8.34 Billion FETs, 416 mm2

IFOP
4 x16 PCIe/IFIP
IFOP
4x DDR
4x DDR
IFOP
4 x16 PCIe/IFIP
IFOP

"Rome"

L3
IFOP
L3
Zen2 cores
Zen2 cores

7nm CCD: 8.34 Billion FETs, 74 mm2

IFOP
2x DDR
PCIe

Client IO Die
2.09 Billion FETs, 125 mm2

"Matisse"

"X570 Chipset"

Figure 2.2.1: Three heterogeneous technology chiplets leveraged to many products and markets.

Source: IEEE ISSCC 2020 AMD Chiplet Architecture for High-Performance Server and Desktop Products

AMD Zen2 Cost-performance scalability

Cost vs. Performance Scalability

2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0

Higher max performance than possible with a monolithic design

Cost scales down with performance through chiplet depopulation and harvesting

64 Cores | 48 Cores | 32 Cores | 24 Cores | 16 Cores

■ Chiplet 7nm + 14nm ■ Hypothetical Monolithic 7nm

Figure 2.2.2: Cost-performance scalability with chiplet design.

Source: IEEE ISSCC 2020 AMD Chiplet Architecture for High-Performance Server and Desktop Products

AMD Zen2 Infinity On-Package (IFOP) SerDes Architecture

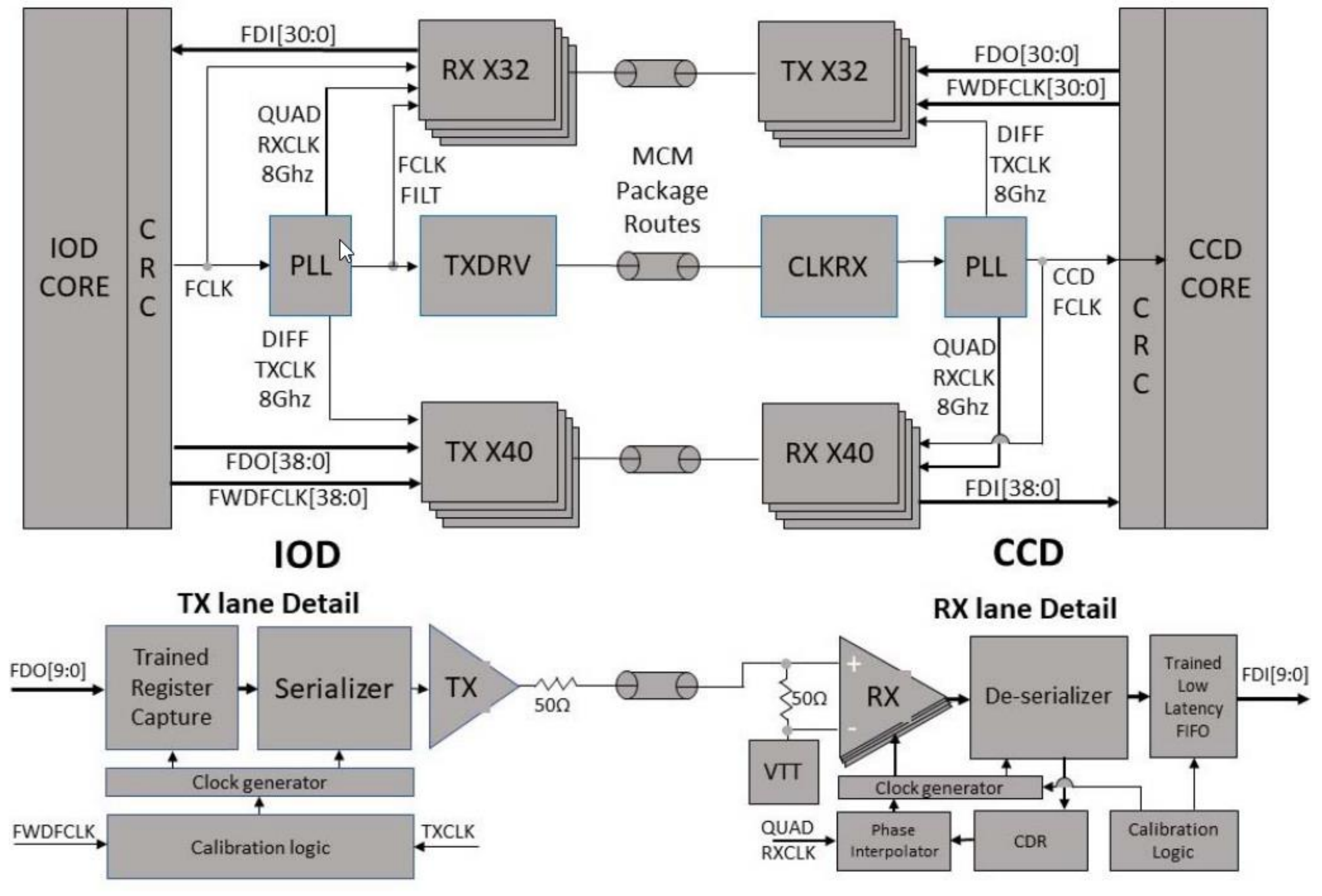


Figure 2.2.3: Infinity Fabric On-Package (IFOP) SerDes architecture.

Source: IEEE ISSCC 2020 AMD Chiplet Architecture for High-Performance Server and Desktop Products

AMD Zen2 Infinity On-Package (IFOP) SerDes Architecture

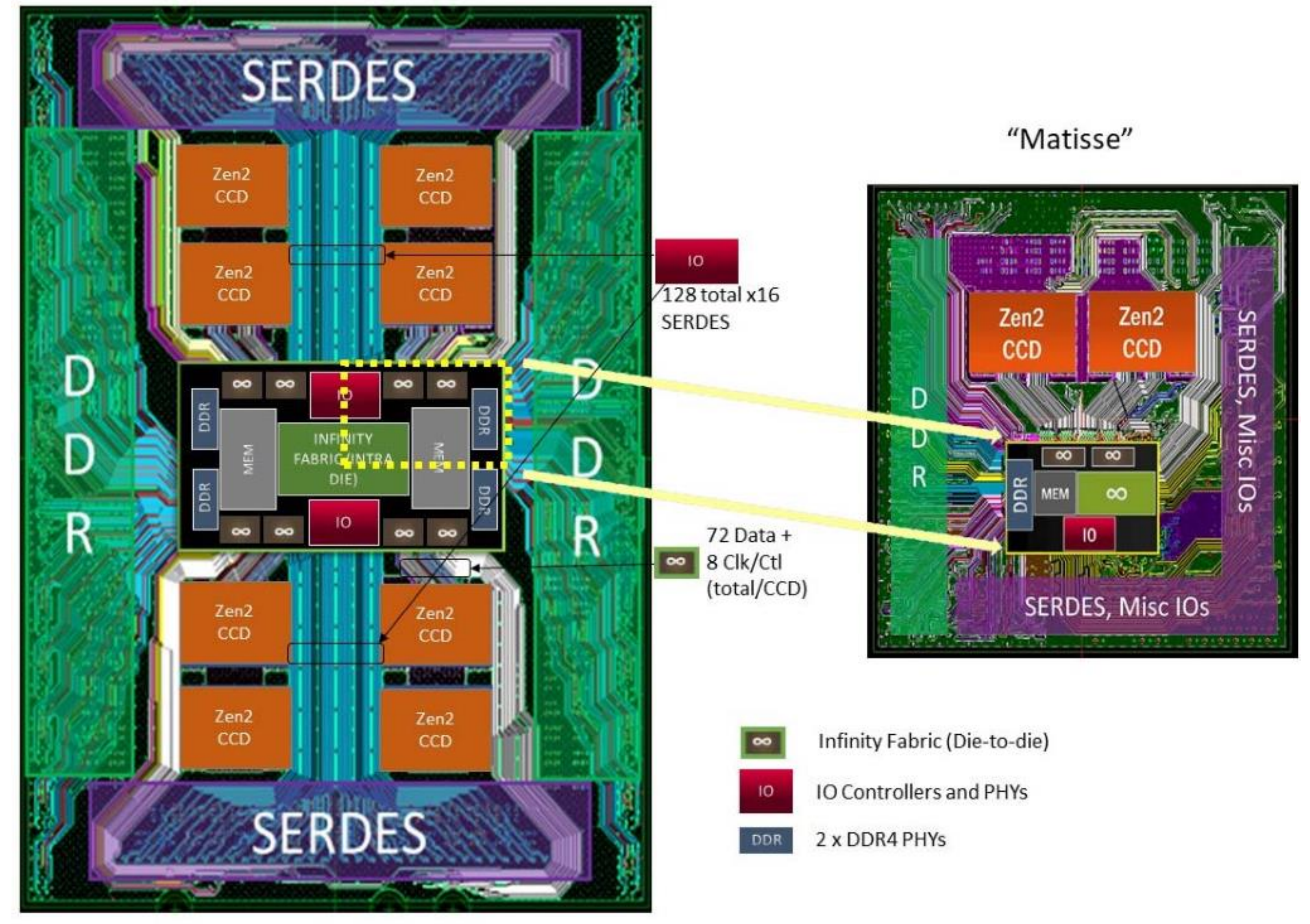
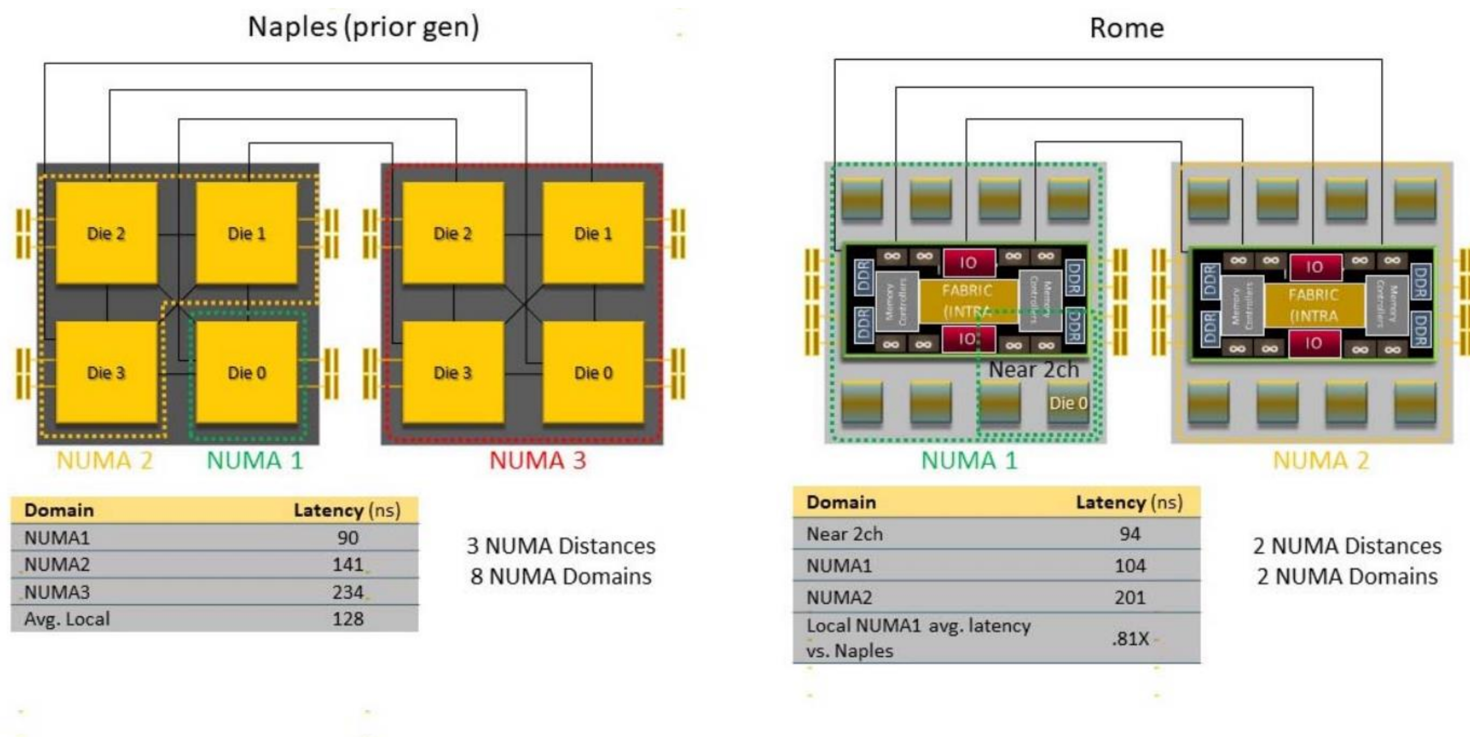


Figure 2.2.4: 'Rome' and 'Matisse' package design and IOD leverage.

Source: IEEE ISSCC 2020 AMD Chiplet Architecture for High-Performance Server and Desktop Products

AMD Naples vs. Rome



Domain	Latency (ns)
NUMA1	90
NUMA2	141
NUMA3	234
Avg. Local	128

Domain	Latency (ns)
Near 2ch	94
NUMA1	104
NUMA2	201
Local NUMA1 avg. latency vs. Naples	.81X

Figure 2.2.6: 'Rome' central IOD reduces the number of NUMA domains and distances for much improved memory latency attributes relative to it's predecessor.

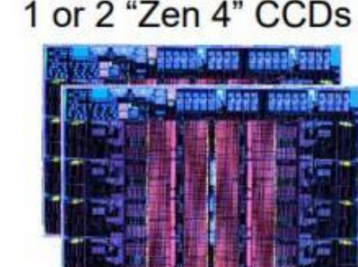
Source: IEEE ISSCC 2020 AMD Chiplet Architecture for High-Performance Server and Desktop Products

AMD Zen4 line-up

Product Configurations

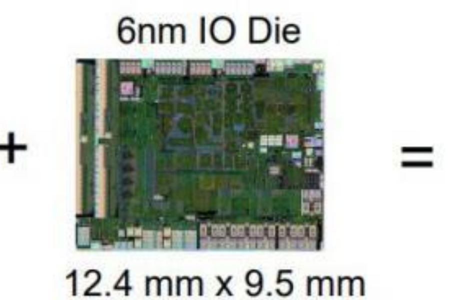
Desktop

1 or 2 "Zen 4" CCDs

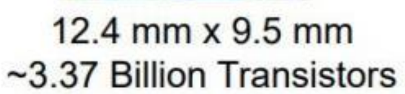


+

6nm IO Die

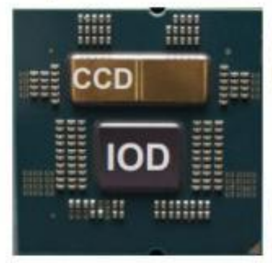


12.4 mm x 9.5 mm
~3.37 Billion Transistors



=

Ryzen™ 7000 Series

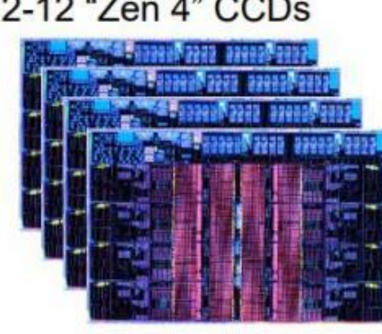


6-16 Cores
Boost up to 5.7 GHz

Up to 16.5 B transistors per socket

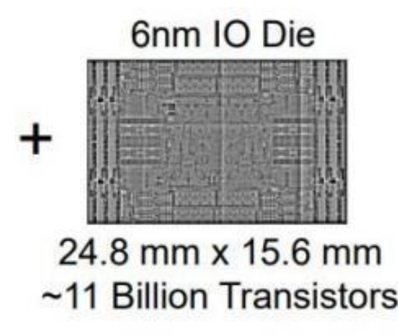
Server

2-12 "Zen 4" CCDs



+

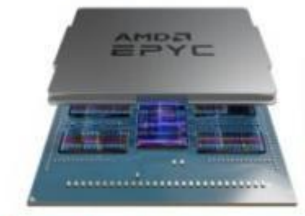
6nm IO Die



24.8 mm x 15.6 mm
~11 Billion Transistors

=

EPYC™ 9004 Series



Up to 90.2B transistors per socket

Product Positioning	Core Count	Default TDP	cTDP
Core Performance	16-48	320-360 W	400 W
Core Density	64-96	280-360 W	400 W
Total Cost of Ownership	16-32	200-280 W	400 W

© 2023 IEEE International Solid-State Circuits Conference

2.1: "Zen 4" - The AMD 5nm 5.7 GHz x86-64 Microprocessor Core

23 of 31

Source: IEEE ISSCC 2023

AMD Ryzen Zen4 IO die

GMI3 / IFOP3 PHY Port 0

GMI3 / IFOP3 PHY Port 1

2x 40-bit DDR5 PHY

2x 40-bit DDR5 PHY

Memory Controllers

GPU Complex

RDNA2 WGP (2x CU) 128 Shaders

DCN3.1.5 Block for Display Control?

VCN3.1.2 With AV1 Decode

Interconnect, PCIe Control, I/O MISC, Power Management, Security Processor, Audio DSP

Multi-Purpose I/O PHYs 4x USB3 or Display Output

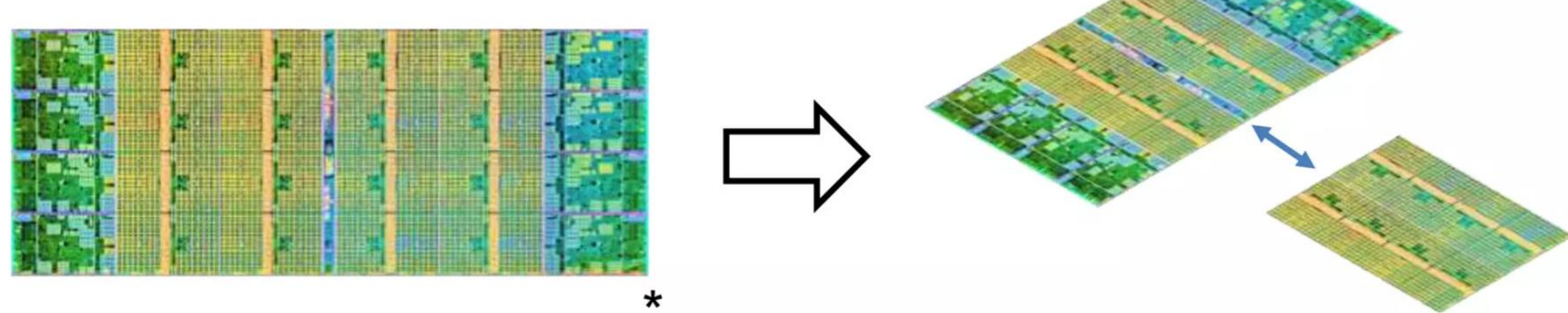
28x PCIe5 PHYs

6 nm client I/O die shot for Zen 4 Raphael by AMD (ISSCC2023)
Simple floorplan interpretation by Locuza, March 2023

Source: IEEE ISSCC 2023

2.5D chiplets, 3D chiplets

More than Moore



- 2.5D chiplets can provide product flexibility and reduce cost
- However, 3D can be even better!
  - Improves effective memory latency
  - Reduces long datapath and I/O's dynamic powers
  - Fits more transistors within a given package cavity size

*Hypothetical processor with large cache

© 2022 IEEE
International Solid-State Circuits Conference

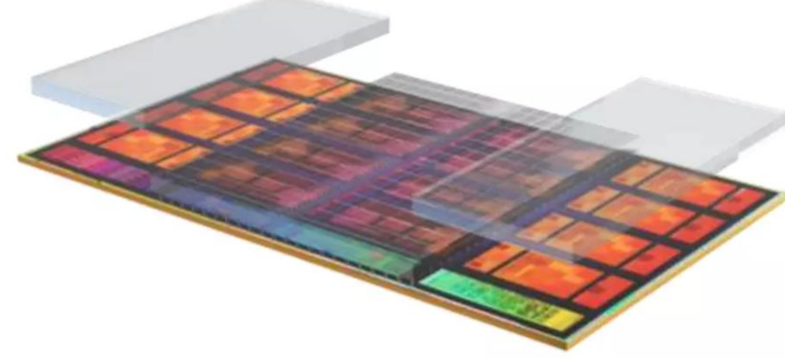
26.4: 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU

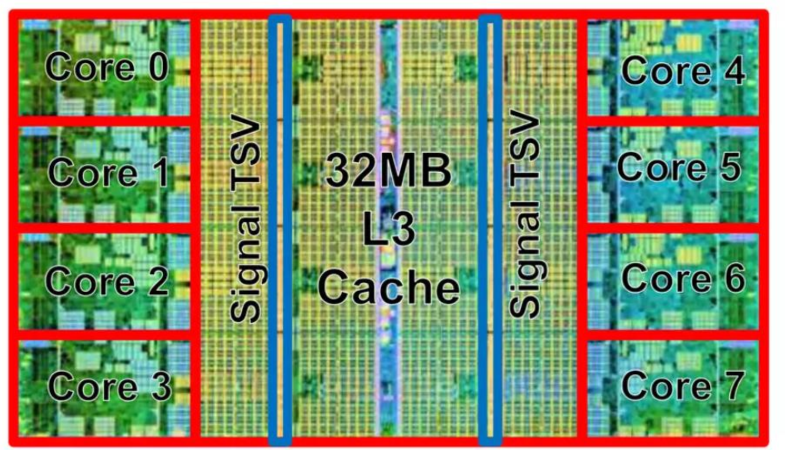
7 of 36

Source: IEEE ISSCC 2022

AMD Ryzen Zen3 3D V-cache

AMD 3D V-Cache™ Components: CCD





- "Zen 3" x86-64 CPU Core Complex Die (CCD)
- TSMC 7nm technology
- 8 cores per Core Complex (CCX)
- 32MB shared L3 Cache
- +19%[1] IPC (Ave) vs. "Zen 2"
- 81mm^2
- <u>AMD 3D V-Cache™ support integrated from Day 1</u>

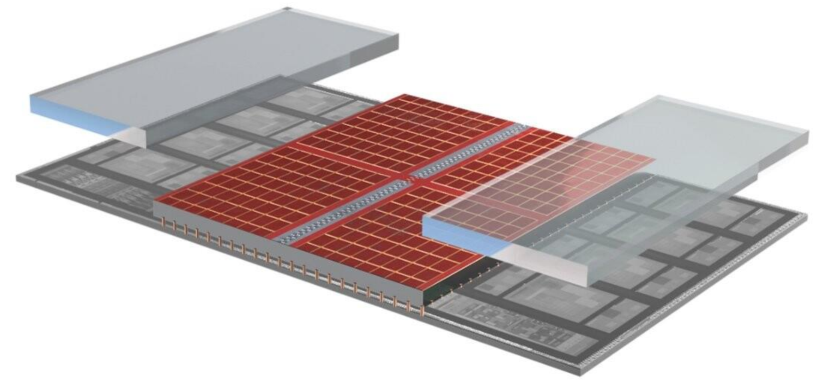
[1]SEE ENDNOTES: R5K-003

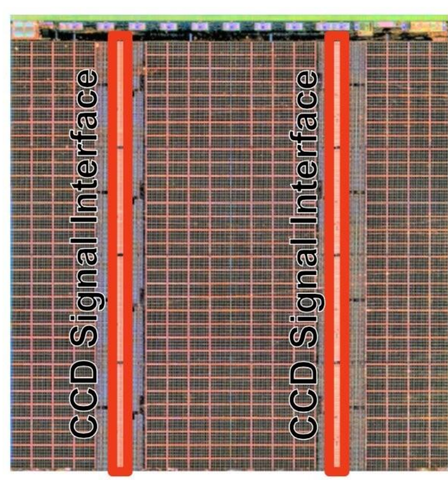
© 2022 IEEE
International Solid-State Circuits Conference

Source: IEEE ISSCC 2022

AMD Ryzen Zen3 3D V-cache

AMD 3D V-Cache™ Components: L3D





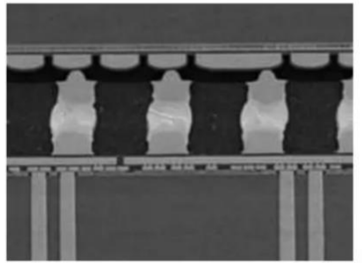
- AMD 3D V-Cache™ extended L3 Die (L3D)
- TSMC 7nm FinFET Technology
- 13 layers Cu + 1 layer Al metal stack
- 64MB L3 Cache Extension
- 41mm^2

© 2022 IEEE
International Solid-State Circuits Conference

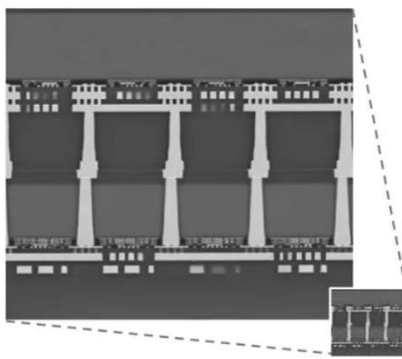
Source: IEEE ISSCC 2022

AMD 3D V-cache Interface

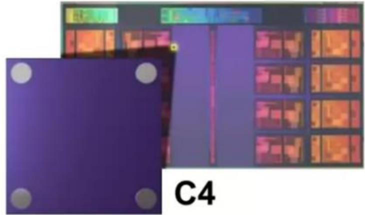
Micro Bump vs. Hybrid Bond



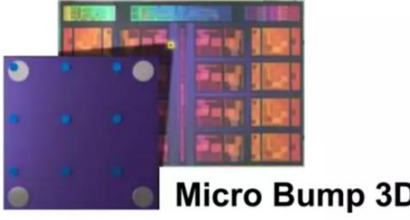
Micro Bump 3D



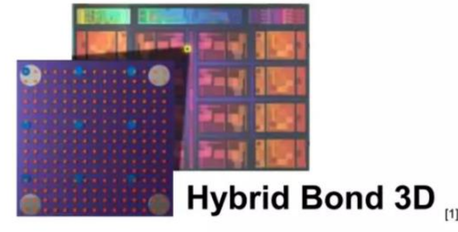
Hybrid Bond 3D



C4



Micro Bump 3D



Hybrid Bond 3D [1]

- **Compared to Micro Bump 3D solutions, Hybrid Bond offers**
  - >15x interconnect density
  - >3x interconnect energy efficiency
  - Superior thermal conductance

SEE ENDNOTES: EPYC-027

C4 and Micro Bump 3D illustrations are hypothetical

[1] Swaminathan, Hot Chips Tutorial, 2021

© 2022 IEEE
International Solid-State Circuits Conference

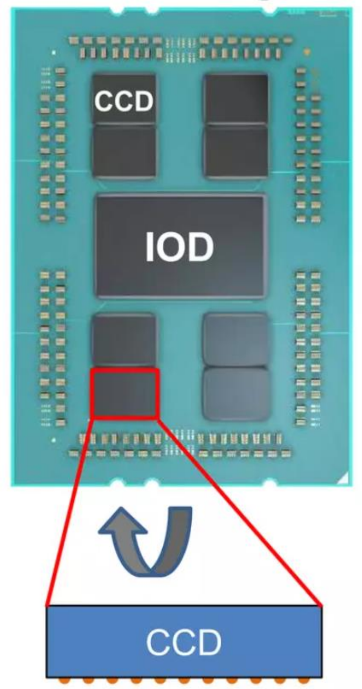
26.4: 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU

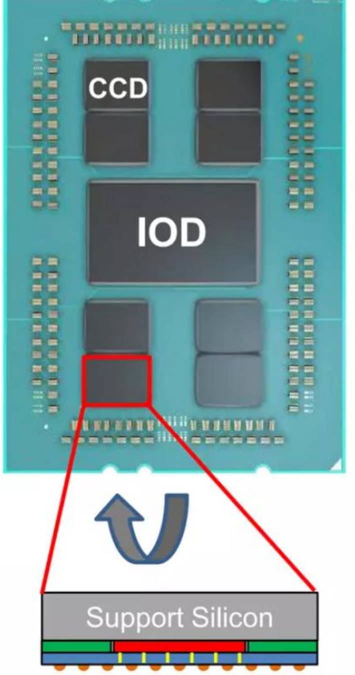
15 of 36

Source: IEEE ISSCC 2022

AMD 3D V-cache Interface

Server Configurations





- **AMD 3rd Gen EPYC™ Server CPU**
  - Up to 8 "Zen 3" CCDs
  - 1 I/O Die (IOD)
- **AMD 3rd Gen EPYC™ Server CPU with AMD 3D V-Cache™**
  - Up to 8 thinned CCDs + L3Ds
  - Support silicon added to match 2D CCD Z-height
- **Both designs compatible with the same package**

© 2022 IEEE
International Solid-State Circuits Conference

Source: IEEE ISSCC 2022

AMD 3D V-cache Interface

Cache Interface Illustration

64 MB AMD 3D V-Cache™

32 B/Cycle Bi-Directional Bus

32 MB L3 Cache

Core
Core
Core
Core

Core
Core
Core
Core

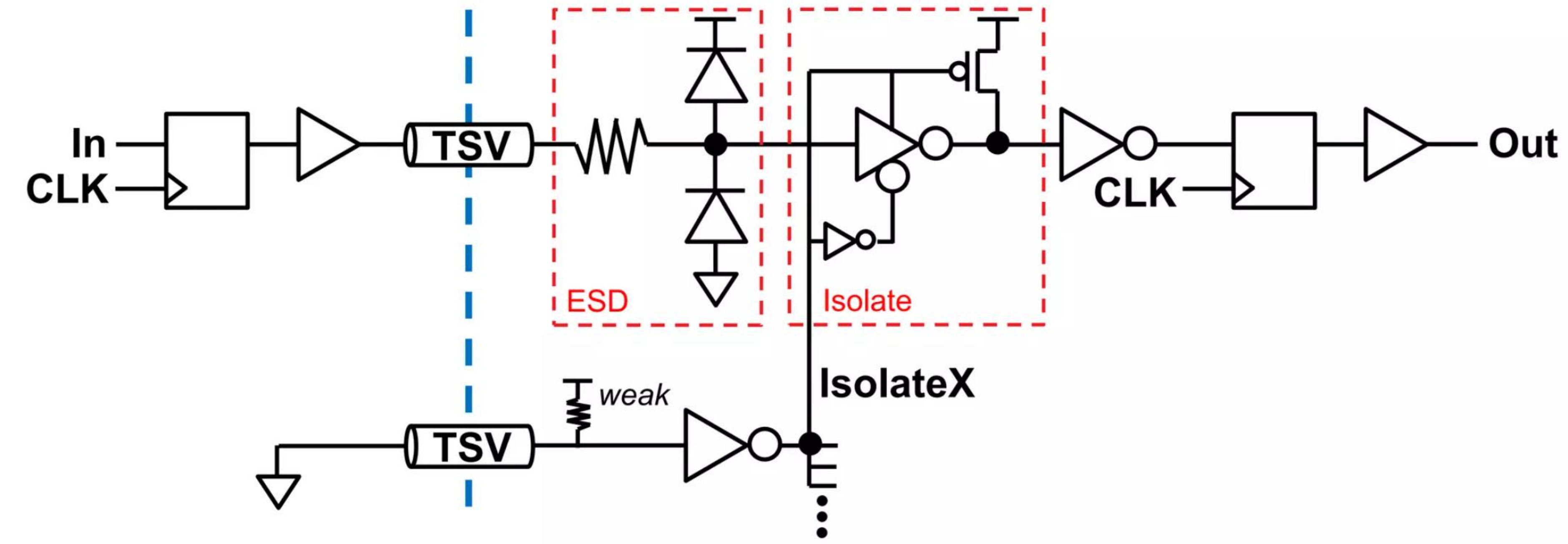
TSV Columns

[5]

[5] Burd, ISSCC, 2022

© 2022 IEEE

AMD Hybrid-Bonded 64MB Stacked Cache



Source: IEEE ISSCC 2022

AMD 64MB Stacked Cache Performance

Server Performance

24.4
JOBS/HOUR

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

40.6
JOBS/HOUR

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

RESULTS MAY VARY. SEE ENDNOTES: MLNX-001R

© 2022 IEEE
International Solid-State Circuits Conference

26.4: 3D V-Cache: The Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU

31 of 36

Source: IEEE ISSCC 2022

Intel Tiles with EMIB and Foveros

Intel Lakefield (2020): Foveros Die Interface (FDI) die stacking

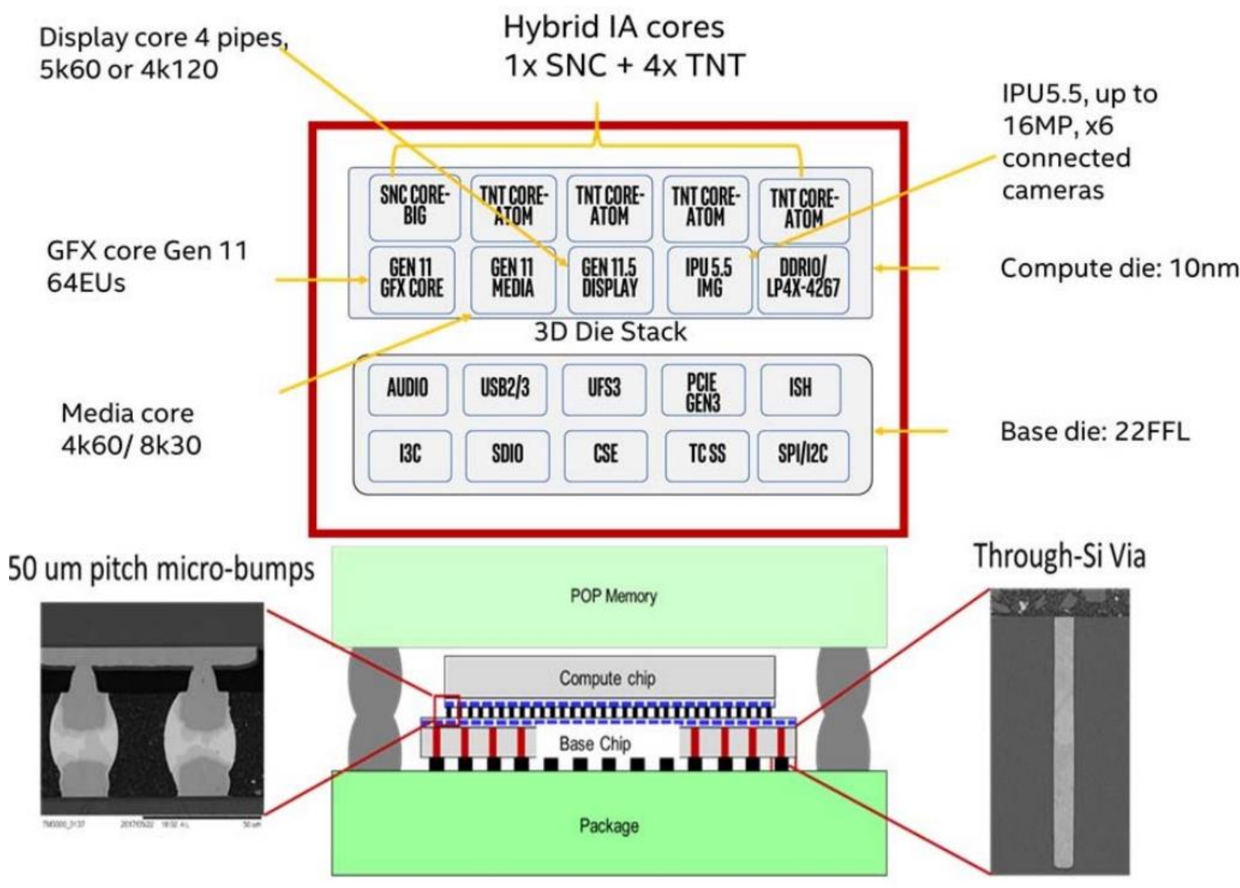


Figure 8.1.1: System partitioning with 3D stacking across compute and base die.

Source: ISSCC 2020 Lakefield and Mobility Compute: A 3D Stacked 10nm and 22FFL Hybrid Processor System in 12×12mm2, 1mm Package-on-Package

Transilvania University of Brasov

Intel Lakefield (2020): Foveros Die Interface (FDI) die stacking

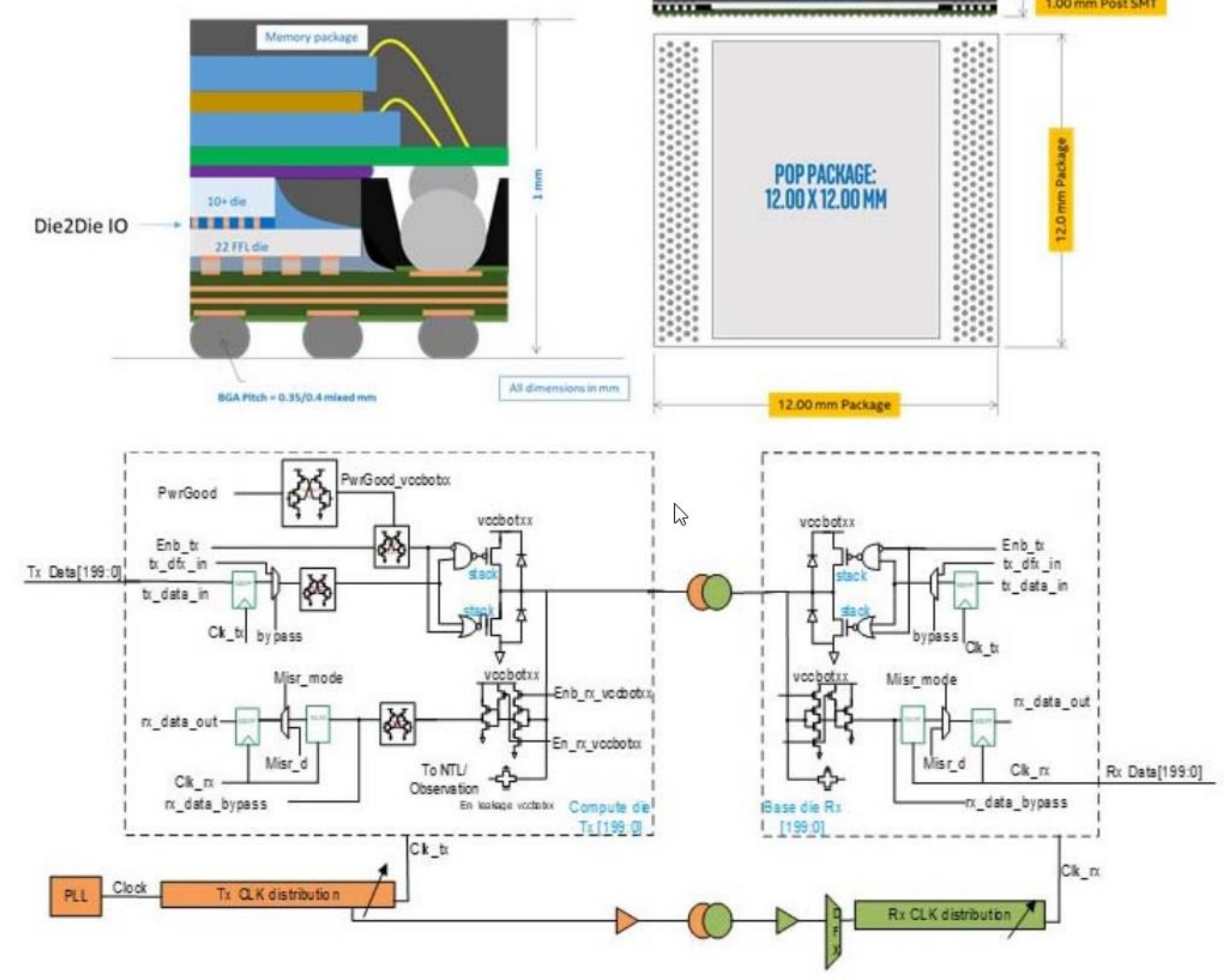


Figure 8.1.3: Simplified Die2Die IO circuits.

Source: ISSCC 2020 Lakefield and Mobility Compute: A 3D Stacked 10nm and 22FFL Hybrid Processor System in 12×12mm2, 1mm Package-on-Package

Intel Sapphire Rapids – Multi-Tile Design

Ice Lake

Single Monolithic Die

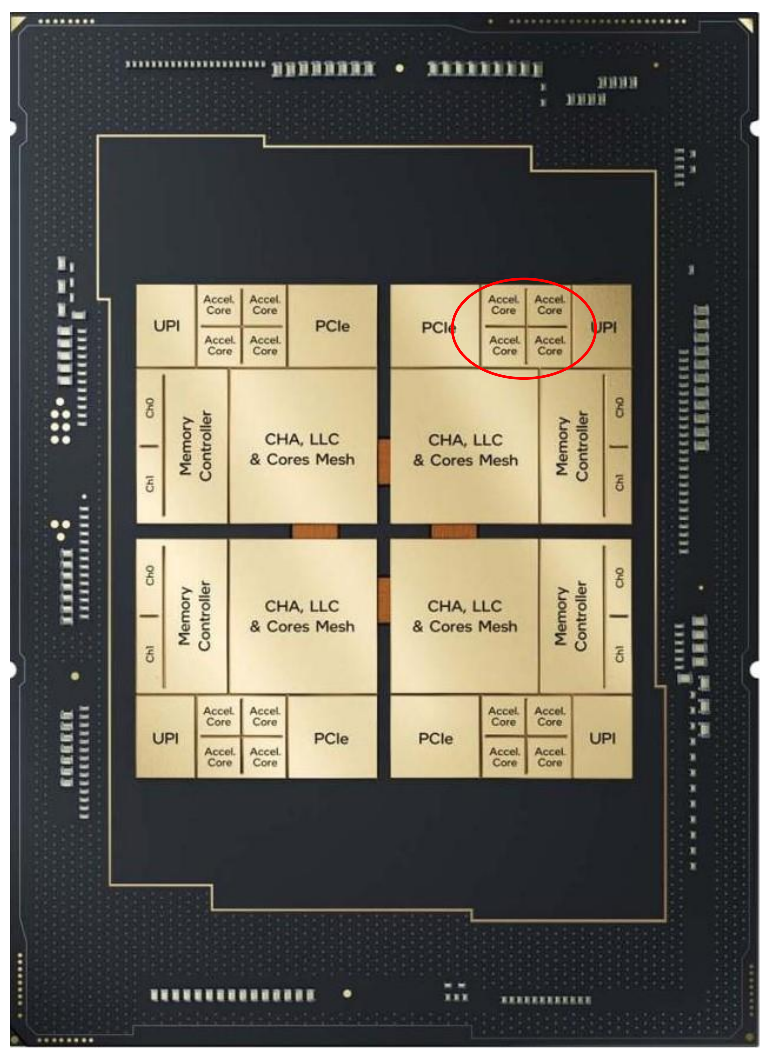
Sapphire Rapids

Multi-Tile Design for Increased Scalability

Delivers a scalable, balanced architecture leveraging existing software paradigms for monolithic CPUs via a modular architecture

Source: https://hc33.hotchips.org/assets/program/conference/day1/HC2021.C1.4%20Intel%20Arijit.pdf

Intel Sapphire Rapids (2023)



Sapphire Rapids Accelerators

- ❑ Intel QuickAssist Technology (QAT)
  - ❑ Faster compression and encryption
- ❑ Intel Dynamic Load Balancer (DLB)
  - ❑ Load balancing
  - ❑ Queue management
  - ❑ Packer prioritization
- ❑ In-Memory Analytics Accelerator (IAA)
  - ❑ In-memory databases
  - ❑ Big data analytics
- ❑ Data Streaming Accelerator (DSA)
  - ❑ High performance data copy and transformation

Source: https://www.intel.com/content/www/us/en/newsroom/news/innovation-artificial-intelligence-news.html

Intel Sapphire Rapids (2023)

Enabling the Intel® Accelerator Engines

Tools for developers to take advantage and deploy today

IAA, DSA, QAT, DLB

UPI | Accelerators | PCIe | PCIe | Accelerators | UPI

Memory Controller | CHA, LLC & Cores Mesh | CHA, LLC & Cores Mesh | Memory Controller

Memory Controller | CHA, LLC & Cores Mesh | CHA, LLC & Cores Mesh | Memory Controller

UPI | Accelerators | PCIe | PCIe | Accelerators | UPI

IAA, DSA, QAT, DLB

Intel® Advanced Matrix Extensions (Intel® AMX)
- TensorFlow
- PyTorch
- ONNX Runtime
- OpenVINO
- oneDNN (Intel oneAPI)

Intel® Advanced Vector Extensions (Intel® AVX) for vRAN
- FlexRAN
- Data Plane dev Kit (DPDK)*

Intel® In-memory Analytics Accelerator (Intel® IAA)
- Intel Query Processing Library

Intel® Data Streaming Accelerator (Intel® DSA)
- Storage Perf Dev Kit (SPDK)*
- Data Plane Dev Kit (DPDK)*

Intel® QuickAssist Technology (Intel® QAT)
- QATzip* (Intel lib)
- OpenSSL**
- Boring SSL

Intel® Dynamic Load Balancer (Intel® DLB)
- VPP IPsec
- Data Plane Dev Kit (DPDK)*

intel XEON Accelerate with Xeon

*Intel open-source library (not part of stock SW).
**Difference between Intel version and stock version.
***Intel® QPL and *Intel® DML in open-source beta, v1.0.0 coming shortly.

Source: www.servethehome.com/4th-gen-intel-xeon-scalable-sapphire-rapids-leaps-forward/4th-gen-intel-xeon-scalable-sapphire-rapids-acceleration-engines-enablement/

Intel Sapphire Rapids (2023)

A Higher Performance Server Architecture

Benefits of Intel® Accelerator Engines

Intel® Advanced Matrix Extensions (Intel® AMX)	Intel® Dynamic Load Balancer (Intel® DLB)	Intel® Data Streaming Accelerator (Intel® DSA)	Intel® In-Memory Analytics Accelerator (Intel® IAA)	Intel® QuickAssist Technology (Intel® QAT)
Up to	Up to	Up to	Up to	Up to
8.6x	96%	1.7x	2.1x	84%
higher speech recognition inference performance with built-in AMX BF16 vs. FP32	lower latency at the same throughput for Istio-Envoy Ingress with Intel® DLB vs. software for Istio Ingress gateway	higher IOPs for SPDK-NVMe with built-in Intel® DSA vs. ISA-L software	higher RocksDB performance with Intel® IAA vs Ztsd software	fewer cores to achieve same connections/s on NGINX with built-in QAT vs. out-of-box software

Accelerators Enable Step Function Performance Beyond Base Architecture

See backup for workloads and configurations. Results may vary.

Source: www.servethehome.com/4th-gen-intel-xeon-scalable-sapphire-rapids-leaps-forward/4th-gen-intel-xeon-scalable-sapphire-rapids-acceleration-engines-enablement/

Intel Sapphire Rapids – EMIB

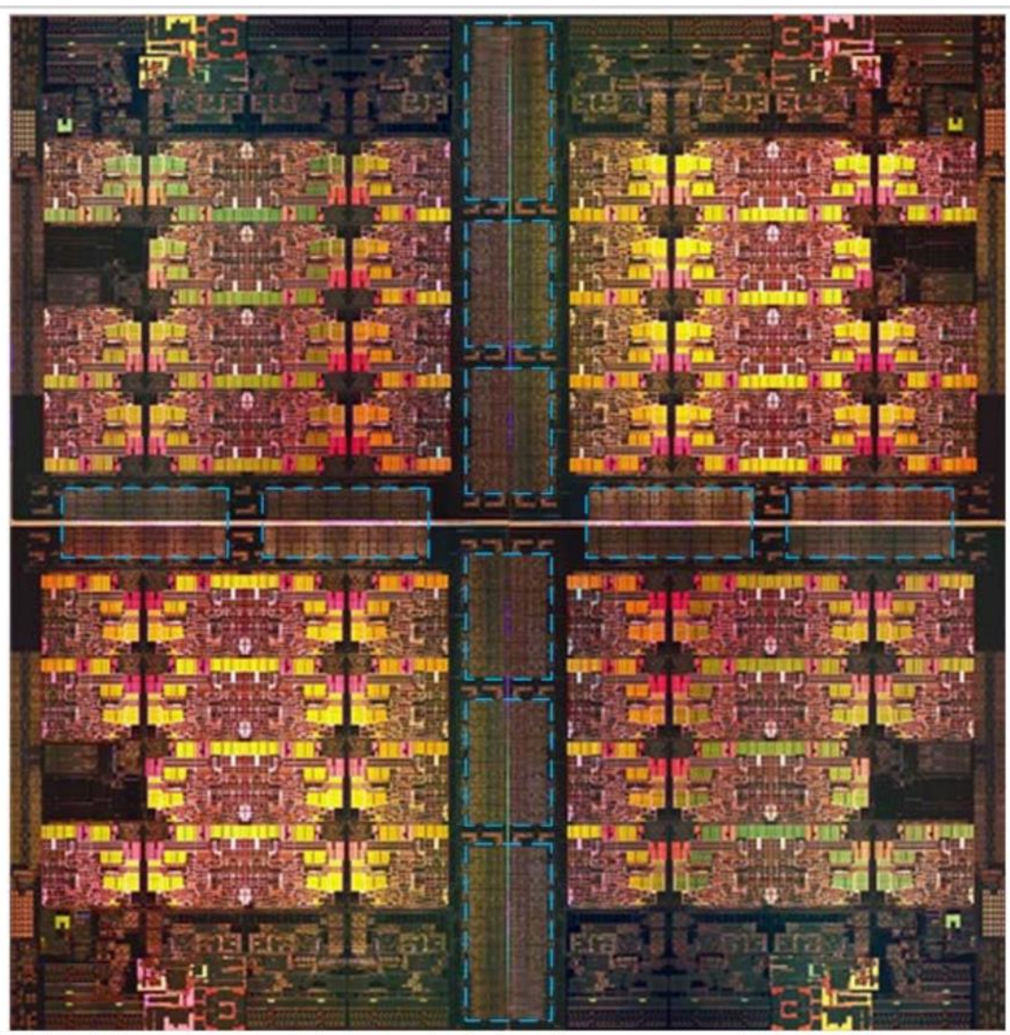


Figure 2.2.7: Die photo of left and right die arranged in 2×2 quasi-monolithic configuration. EMIB placement highlighted in blue.

Source: ISSCC 2022 Sapphire Rapids: The Next-Generation Intel Xeon Scalable Processor

Intel Sapphire Rapids Multi-Die Fabric (MDF) IO

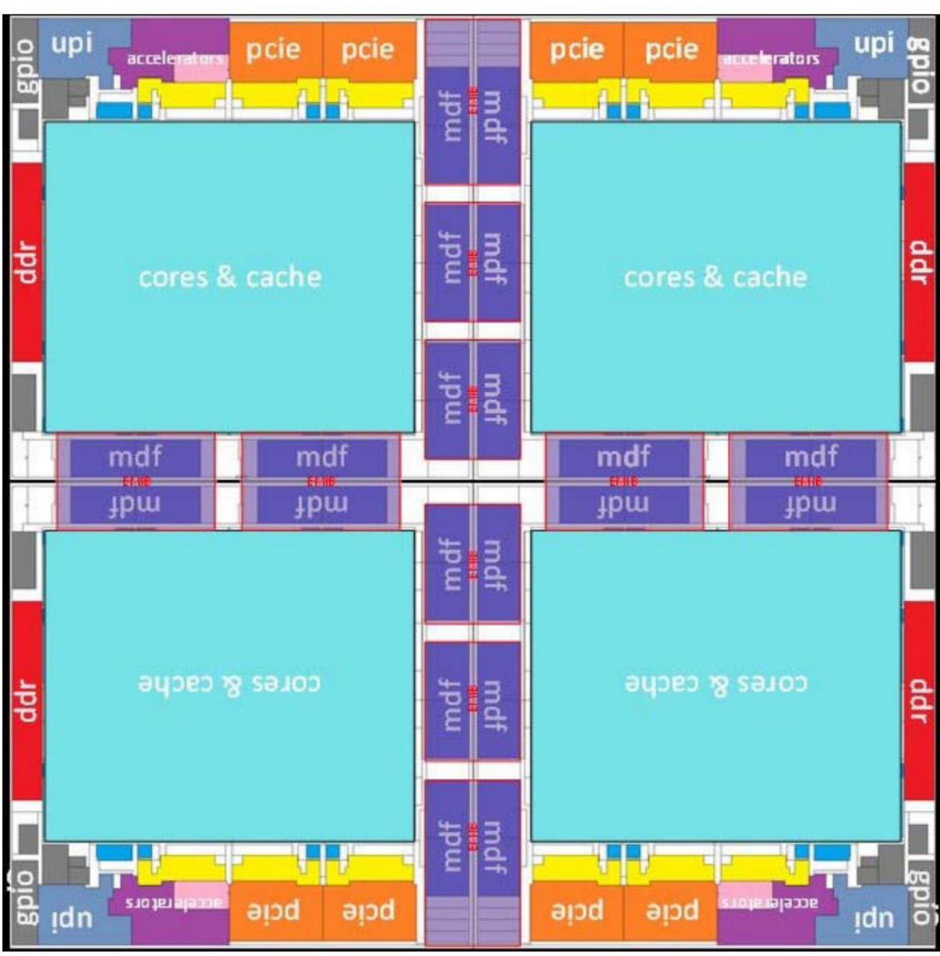


Figure 2.2.1: Die floorplan in 2×2 quasi-monolithic configuration. EMIB highlighted at die-to-die interfaces.

Source: ISSCC 2022 Sapphire Rapids: The Next-Generation Intel Xeon Scalable Processor

Intel Sapphire Rapids Cross-die timing model

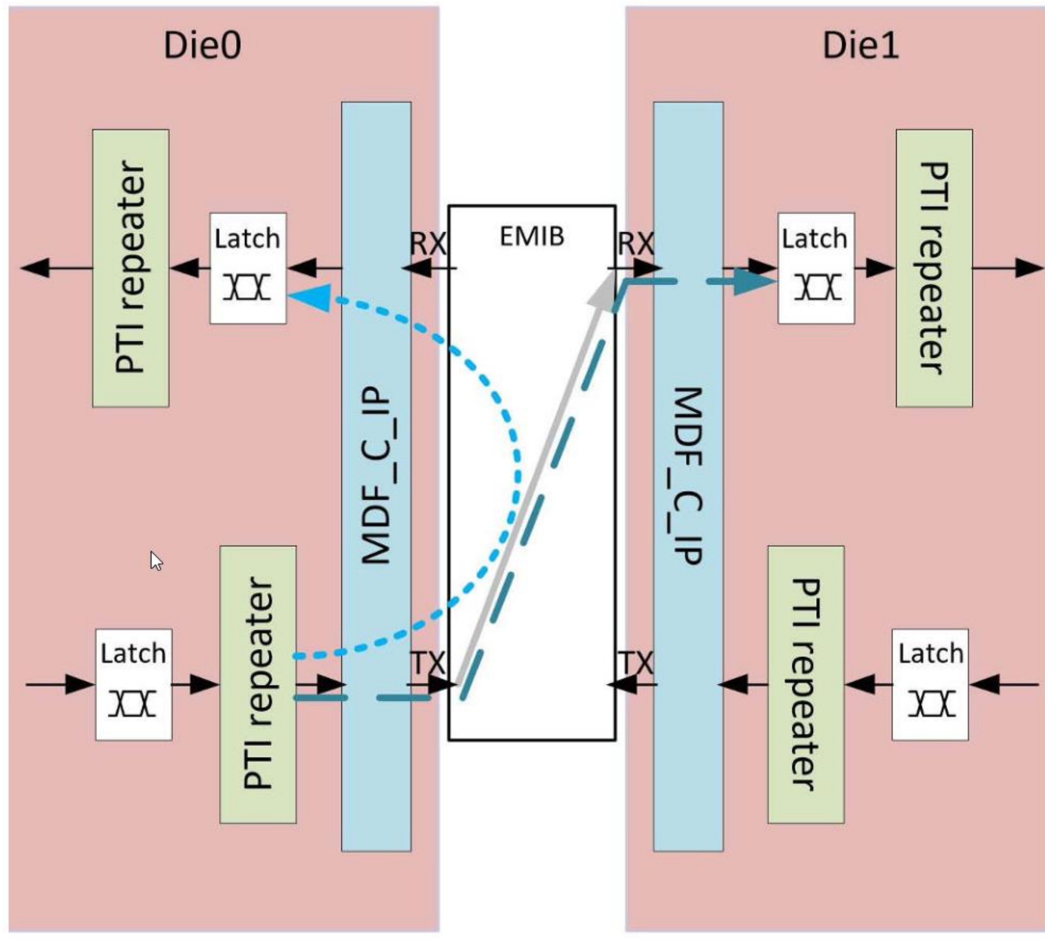


Figure 2.2.2: A cross-die timing model, or a single-die loopback model.

Source: ISSCC 2022 Sapphire Rapids: The Next-Generation Intel Xeon Scalable Processor

Chiplets-augmented GPUs

AMD RDNA3 – Chiplets for GPUs

CHIPLET TECHNOLOGY
CAN IT WORK FOR GRAPHICS?

Traditional Monolithic

I/O I/O I/O I/O
CCX CCX
CCX CCX
CCX CCX
CCX CCX
DDR DDR DDR DDR
I/O I/O I/O I/O

EPYC CPU Server

7nm CCD (×8)
14nm I/O Die
I/O
DDR
100's of signals

"Navi21" GPU

Shader Engine
INFINITY CACHE
INFINITY FABRIC
10's of 1000's of signals

- Chiplets enabled use of advanced nodes where they benefit CPU performance but mature nodes for IO and interfaces
- High speed organic package links meet CPU Bandwidth requirements
- GPU shader engines require massive amounts of connectivity compared to CPUs
- A different approach is required

11

AMD
together we advance_gaming

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

AMD RDNA3 – Graphics Compute Die and Memory Cache Die

- Each MCD has 16MB of cache

MCD MCD MCD
GCD
MCD MCD MCD

Chiplets

The right process technology for the right job

300mm² 5nm GCD | 6x37mm² 6nm MCD

Source: https://www.anandtech.com/show/17638/amd-reveals-radeon-rx-7900-xtx-and-7900-xt-first-rdna-3-parts-to-hit-shelves-in-december
https://www.tomshardware.com/news/amd-rdna-3-gpus

AMD RDNA3

CHIPLET TECHNOLOGY

A BETTER WAY TO PARTITION

The graphics engine is what benefits from advanced N5 technology

- AMD Infinity Cache™ critical to performance but barely shrinks into N5
- GDDR6 interfaces are also large and won't shrink at all

Split those poorly scaling components off as a chiplet and shrink the GFx core into N5

Full N5 performance, better yield for perf/$ and configurability

"Navi21"

AMD Infinity Cache™

Memory Interfaces

Graphics Engine

"Navi31"

MCD

MCD

MCD

GCD

MCD

MCD

MCD

12

AMD
together we advance_gaming

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

AMD RDNA3 Chiplets

CHIPLET TECHNOLOGY

HOW TO CONNECT THE CHIPLETS?

- GCD-MCD partitioning is great, but the bandwidth requirements are still extremely high
- Over 10X what a CPU CCD requires in EPYC
- Breakthrough Advanced packaging and a new interface is required:
  - **High Performance Fanout and Infinity links**

Total Bandwidth

14.0x
12.0x
10.0x
8.0x
6.0x
4.0x
2.0x
0.0x

CPU Chiplet BW
MCD required BW

"Navi31"

MCD
MCD
MCD
GCD
MCD
MCD
MCD

13

AMD
together we advance_gaming

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

RDNA3 Infinity Links

CHIPLET TECHNOLOGY

INFINITY FANOUT LINKS

BANDWIDTH DENSITY

- Infinity Links, operating at 9.2Gb/s with High Performance Fanout provide almost 10X the BW density of the IFOP links used in Ryzen and EPYC
- Enables industry-leading peak bandwidth of 5.3TB/s

MCD

Die-to-die
Fan out routing

GCD

Bandwidth Density

12.0x
10.0x
8.0x
6.0x
4.0x
2.0x
0.0x

Organic Package Links

GPU Infinity Links

AMD
together we advance_gaming

14

SEE ENDNOTE RX-817

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

RDNA3 Infinity Links

CHIPLET TECHNOLOGY

HIGH PERFORMANCE FANOUT INTERCONNECT

Organic substrate

High Perf Fanout

25 wires on organic substrate compared to 50 wires on High Performance Fanout

Images approximately to scale

15

AMD
together we advance_gaming

Source: https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus

AMD Instinct MI300

The world's first integrated data center CPU + GPU

AMD INSTINCT™

MI300

Breakthrough architecture to power the exascale AI era

Source: CES 2023 AMD Keynote

AMD Instinct MI300

AMD INSTINCT™ MI300

The world's first data center integrated CPU + GPU

AMD CDNA 3 — Next-Gen Accelerator Architecture

ZEN 4 — 24 Leadership Data Center CPU cores

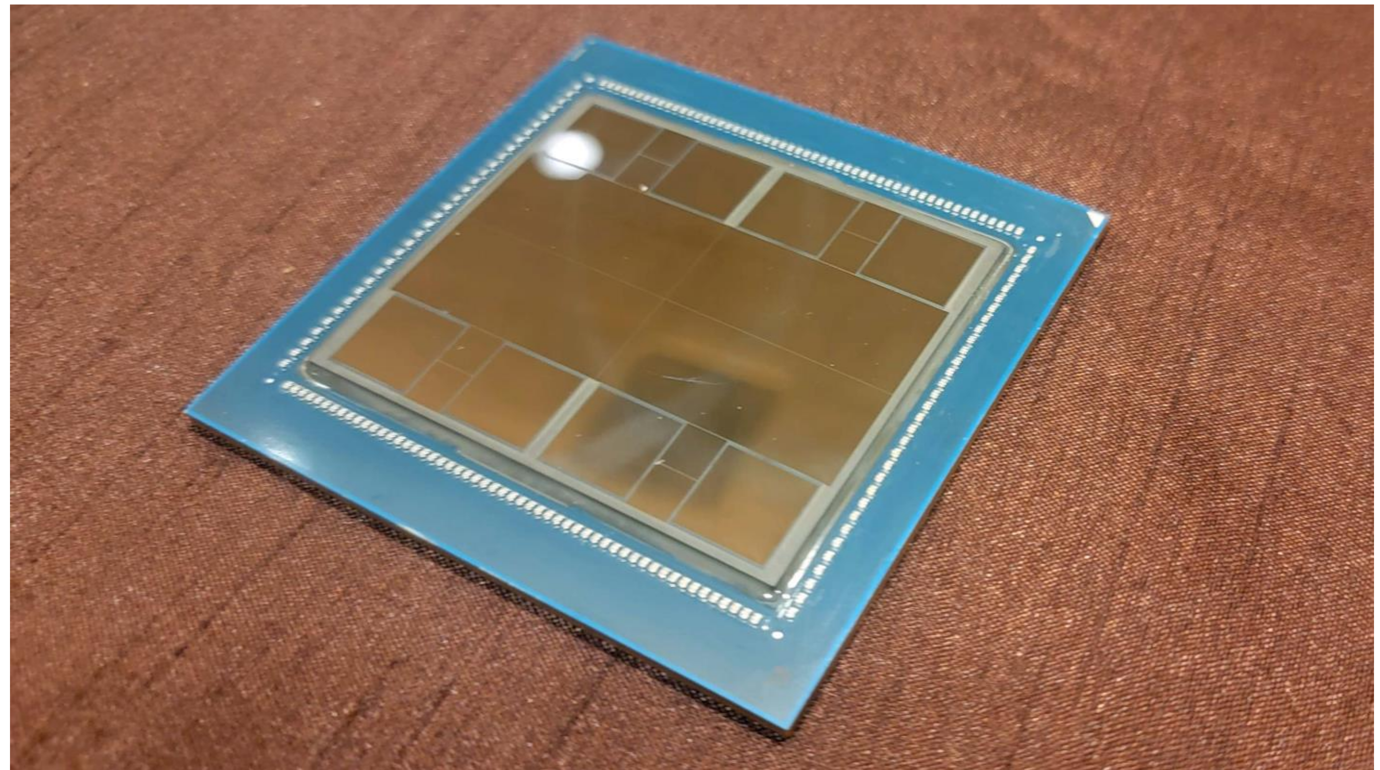
146B Transistors

128GB HBM3

3D Advanced Chiplet Packaging

Source: CES 2023 AMD Keynote

AMD Instinct MI300



MI300 – a disaggregated design

- ❑ Multiple TSMC 5nm chiplets
- ❑ 3D Stacking to place them over a base die
- ❑ On-package High Bandwidth Memory (HBM)

Source: https://www.anandtech.com/show/18721/ces-2023-amd-instinct-mi300-data-center-apu-silicon-in-hand-146b-transistors-shipping-h223

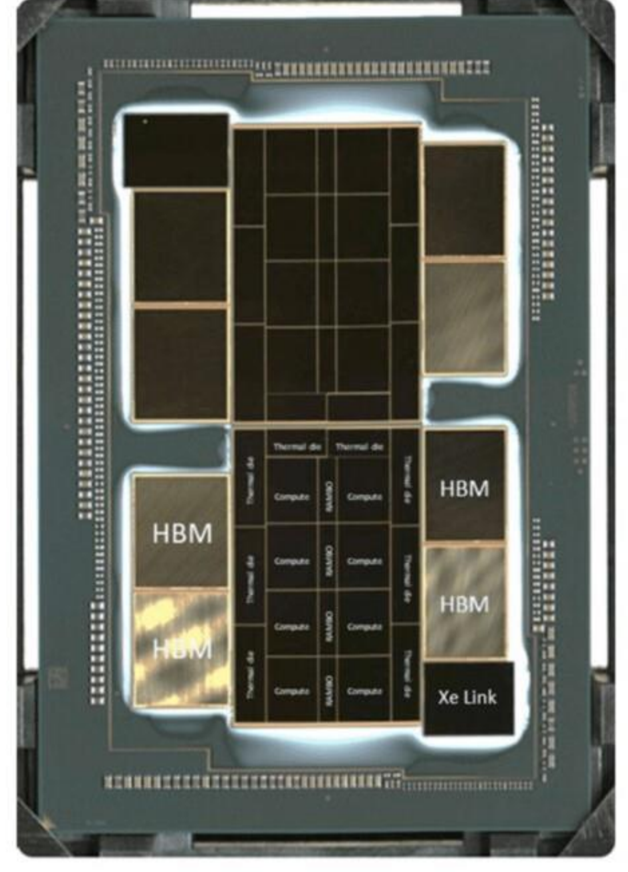
Intel Ponte Vecchio

Intel Ponte Vecchio

- ❑ Intel 10nm process
- ❑ Die Size of 1280mm2

Source: https://www.techpowerup.com/gpu-specs/intel-ponte-vecchio.g1046#:~:text=Intel's%20Ponte%20Vecchio%20GPU%20uses,12%20(Feature%20Level%2012_1).

Raja Koduri / Twitter



Integration	Foveros + EMIB
Power Envelope	600W
Transistor count	> 100B
Total Tiles	63 (**47 functional** + 16 thermal Tiles)
HBM count	8
Package Form factor	77.5 x 62.5 mm (4844 mm^2)
Platforms	3 platforms
IO	4x16 90G SERDES, 1x16 PCIe Gen5
Total Silicon	3100 mm^2 Si
Silicon footprint	2330 mm^2 Si footprint
Package layers	11-2-11 (24 layers)
2.5D Count	11 2.5D connections
Resistance	0.15 mΩ R_{path}/tile
Package pins	4468 pins
Package Cavity	186 mm^2 x4 cavities

S **Figure 2.1.7: Ponte Vecchio chip photographs and key attributes.**

IEEE ISSCC 2022

-lots-of-bandwidth

Intel Ponte Vecchio

RAMBO = Random Access Memory, Bandwidth Optimized

Figure 2.1.1: 3D and 2D system partitioning with Foveros and EMIB on PVC.

- 47 Tiles + 16 thermal tiles
  - 16 compute tiles, TSMC N5, 2.6TB/s speeds to the chip fabric
  - 8 tiles for RAMBO cache, Intel 7, 15MB per tile, 1.3 TB/s connection
  - 2 Foveros base tiles, Intel 7
  - 2 Xe-Link tiles, TSMC N7
  - 8 HBM2e tile
  - 11 Intel's embedded multi-die interconnect bridge (EMIB) tiles
- The package has 4844mm^2 with 4468 pins
  - 2330mm^2 of silicon for the 47 tiles
- Fully Integrated Voltage Regulators (FIVR)
- Compute Express Link (CXL) interface

Source: https://www.techpowerup.com/292250/intel-details-ponte-vecchio-accelerator-63-tiles-600-watt-tdp-and-lots-of-bandwidth
IEEE ISSCC 2022

Intel Ponte Vecchio Foveros and EMIB

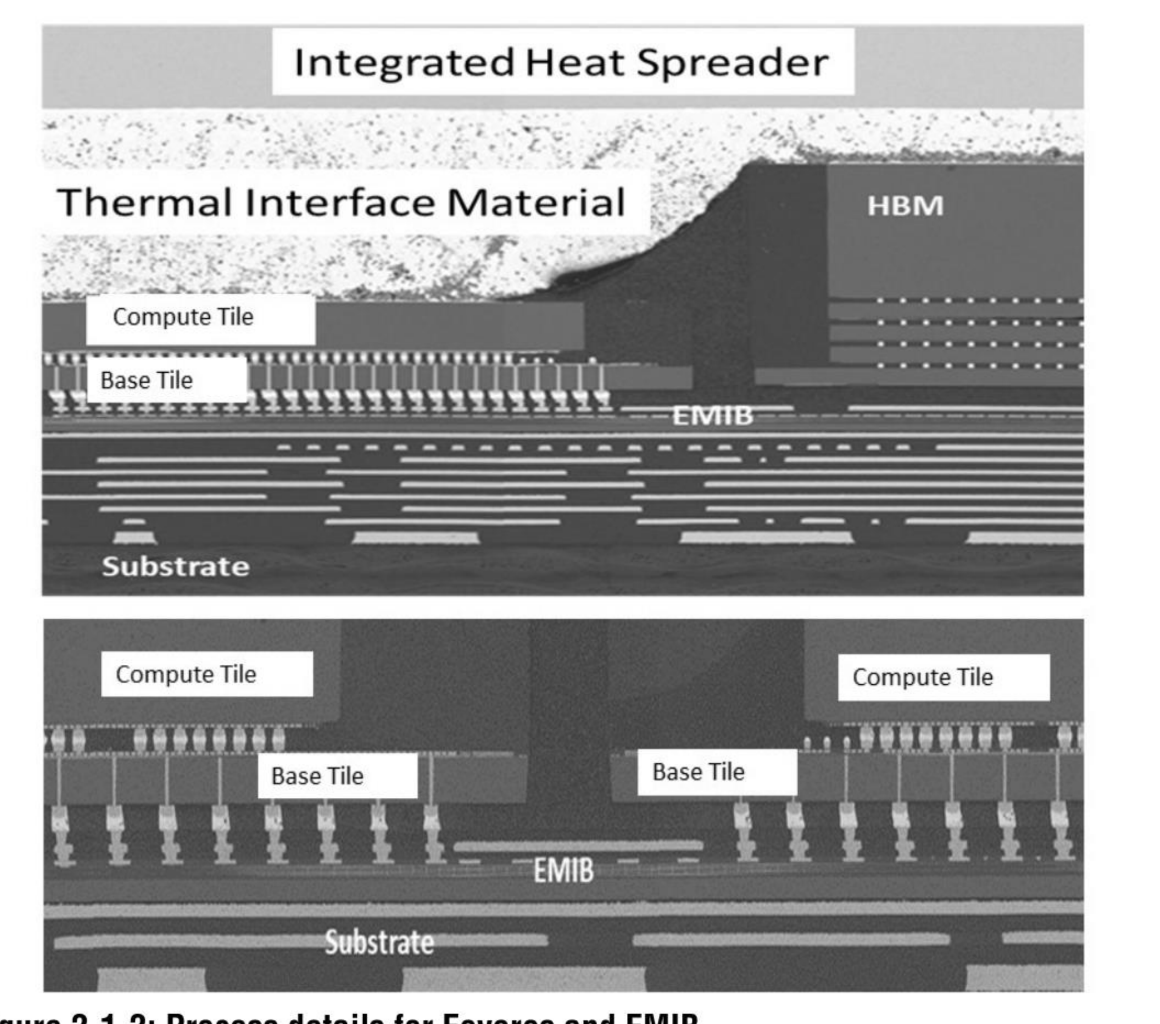
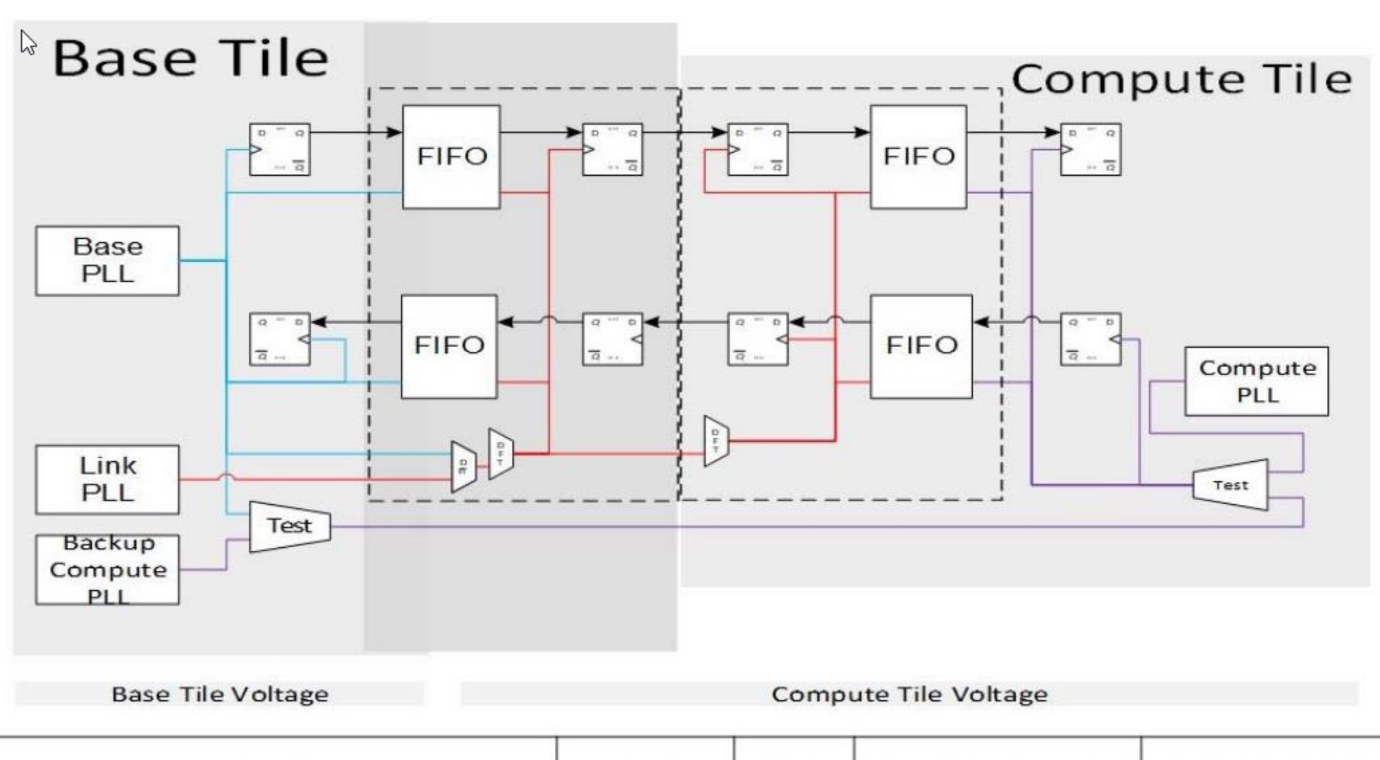


Figure 2.1.2: Process details for Foveros and EMIB.

SourceIEEE ISSCC 2022 Ponte Vecchio: A Multi-Tile 3D Stacked Processor for Exascale Computing

Intel Ponte Vecchio Die-to-die IO



Project	Clock Area	Skew	Freq ratio @ 0.65V	Cdyn/mm² ratio
Tiger lake Graphics (Intel 7)	1X	1X	1X	1X
PVC 2XCLK Graphics (Intel 7 Foveros)	16X	1X	2.6X	1.3X

Figure 2.1.4: PVC clocking, die-to-die IO and comparisons.

Intel GPU Max 1550 (2023)

Intel GPU Max 1550

- ❑ 600W TDP
- ❑ 128GB of HBM2e, 1024 bit interface
- ❑ Memory bandwidth: 3.27 TB/s

Source: https://www.techpowerup.com/gpu-specs/intel-ponte-vecchio.g1046#:~:text=Intel's%20Ponte%20Vecchio%20GPU%20uses,12%20(Feature%20Level%2012_1). https://ark.intel.com/content/www/us/en/ark/products/232873/intel-data-center-gpu-max-1550.html

Conclusions

- Industry has shifted towards chiplets
- UCIe creates a solid foundation for chiplet-based designs
- Future chips may be monolithic 3D ICs
- Chiplets solve several issues of current chips, beyond Moore's Law
  - Reticle limit
  - Better yields due to smaller silicon area of a single silicon die
  - Multi-vendor integration
- Universities need to introduce chiplet-based design in the curricula

Source: 1: Anachronix Chiplet Linked Webinar 2023

Thank you!
Questions?