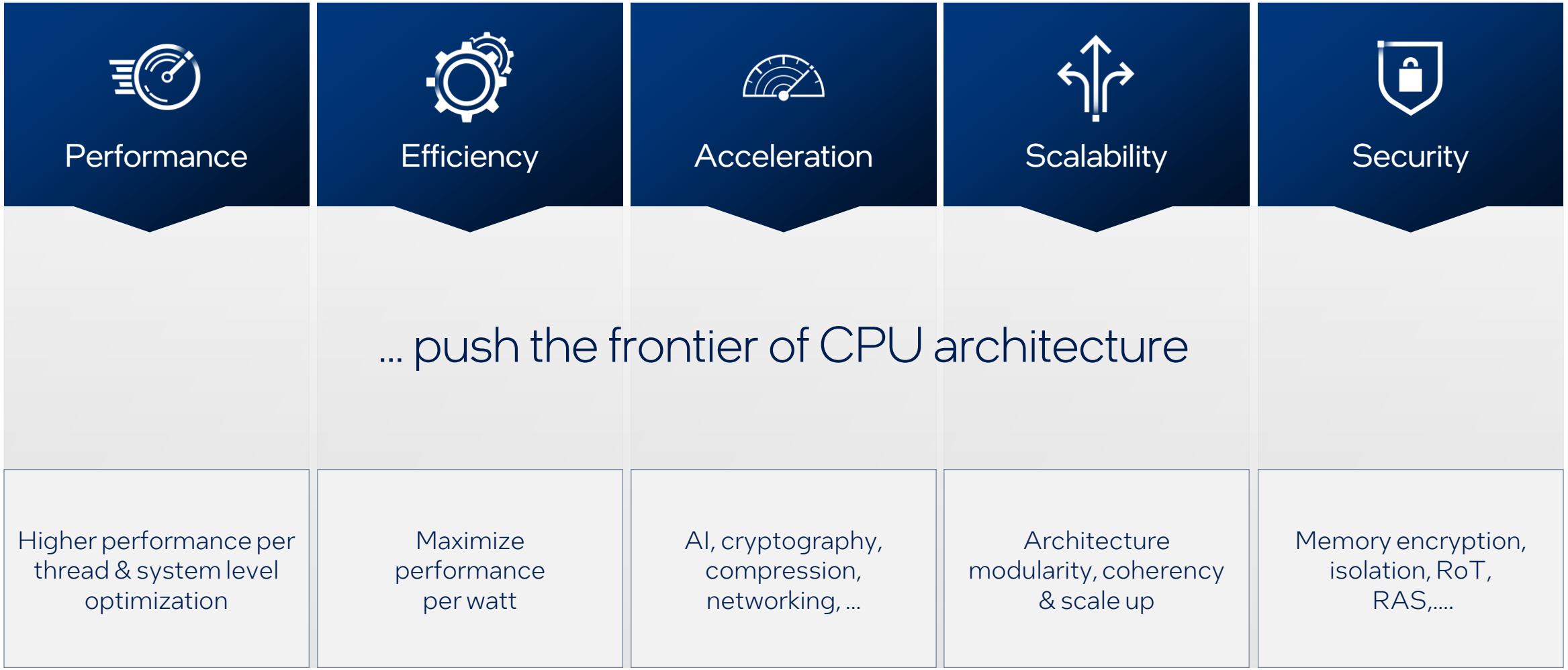
Diamond Rapids

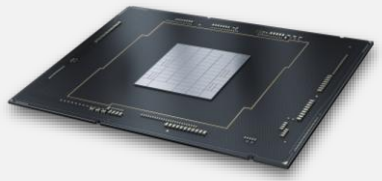
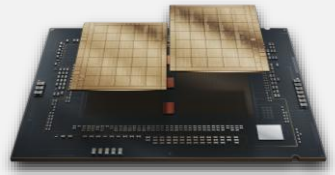
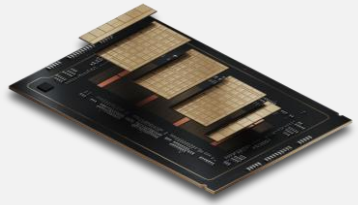
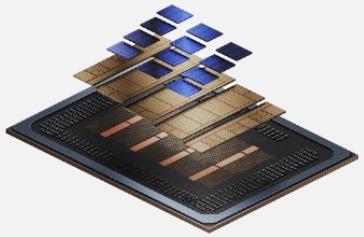
Intel's next generation Xeon CPU

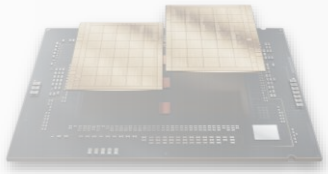
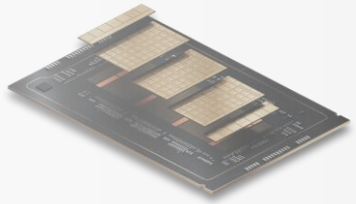
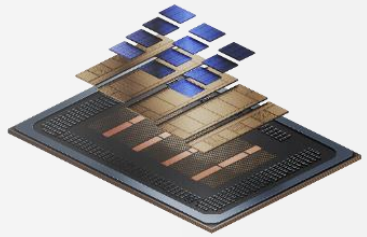
Krishnakanth Sistla - Intel Fellow | Chief Architect Xeon Performance

Evolving Data Center Workloads ….



Intel® Xeon® Architecture Journey

	Ring	Mesh	Tiled Mesh	Modular Mesh	3D Mesh
Architecture					
Packaging	Monolithic	Monolithic	Disaggregated 2.5D	Disaggregated 2.5D	2.5D & 3D
Partitioning	Centralized	Centralized	Tile Based Partition	Function Based Partition	Function Based Partition
Products	*Haswell, Broadwell, …*	*Skylake, Cascade Lake, Cooper Lake, Ice Lake*	*Sapphire Rapids, Emerald Rapids*	*Granite Rapids, Sierra Forest*	*Clearwater Forest*
Brand	Intel® Xeon® processors	1st, 2nd, 3rd Gen Intel® Xeon® Scalable processors	4th & 5th Gen Intel® Xeon® Scalable processors	Intel® Xeon® 6 with P-cores and E-cores	Intel® Xeon® 6+ processors

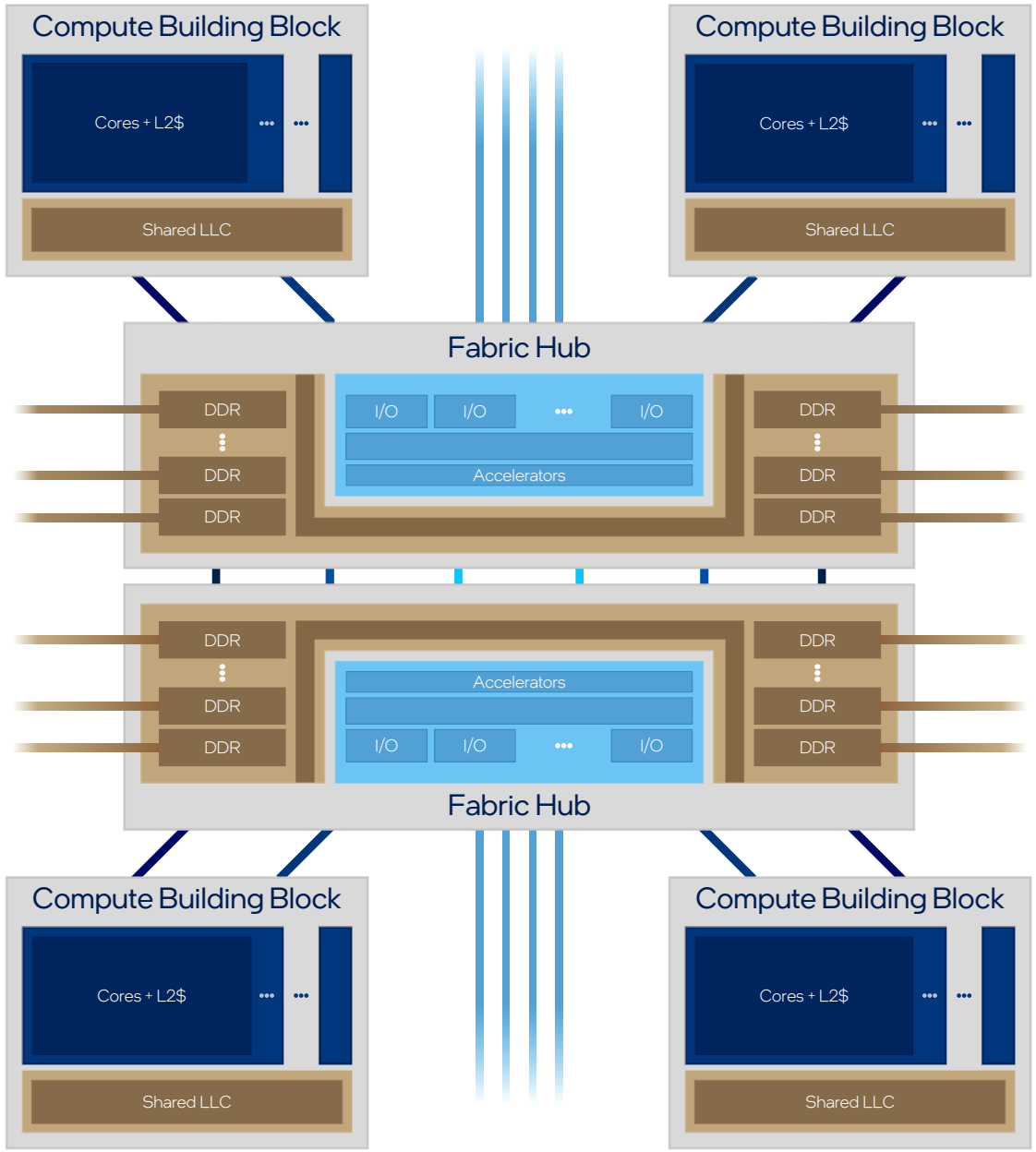
Tiled Mesh	Modular Mesh	3D Mesh
		
Disaggregated 2.5D	Disaggregated 2.5D	2.5D & 3D
Tile Based Partition	Function Based Partition	Function Based Partition
Sapphire Rapids, Emerald Rapids	*Granite Rapids, Sierra Forest*	*Clearwater Forest*
4th & 5th Gen Intel® Xeon® Scalable processors	Intel® Xeon® 6 with P-cores and E-cores	Intel® Xeon® 6+ processors

The Next Frontier of DC CPU Architecture

Defined by how effectively the processor moves, processes and protects data at hyperscale

Fan-out Fabric Architecture

- **Scalable Compute Building Blocks**
- **Centralized I/O & Memory in Fabric Hub**
- **Unified Memory Fabric**
- **Flexible I/O Fabric**



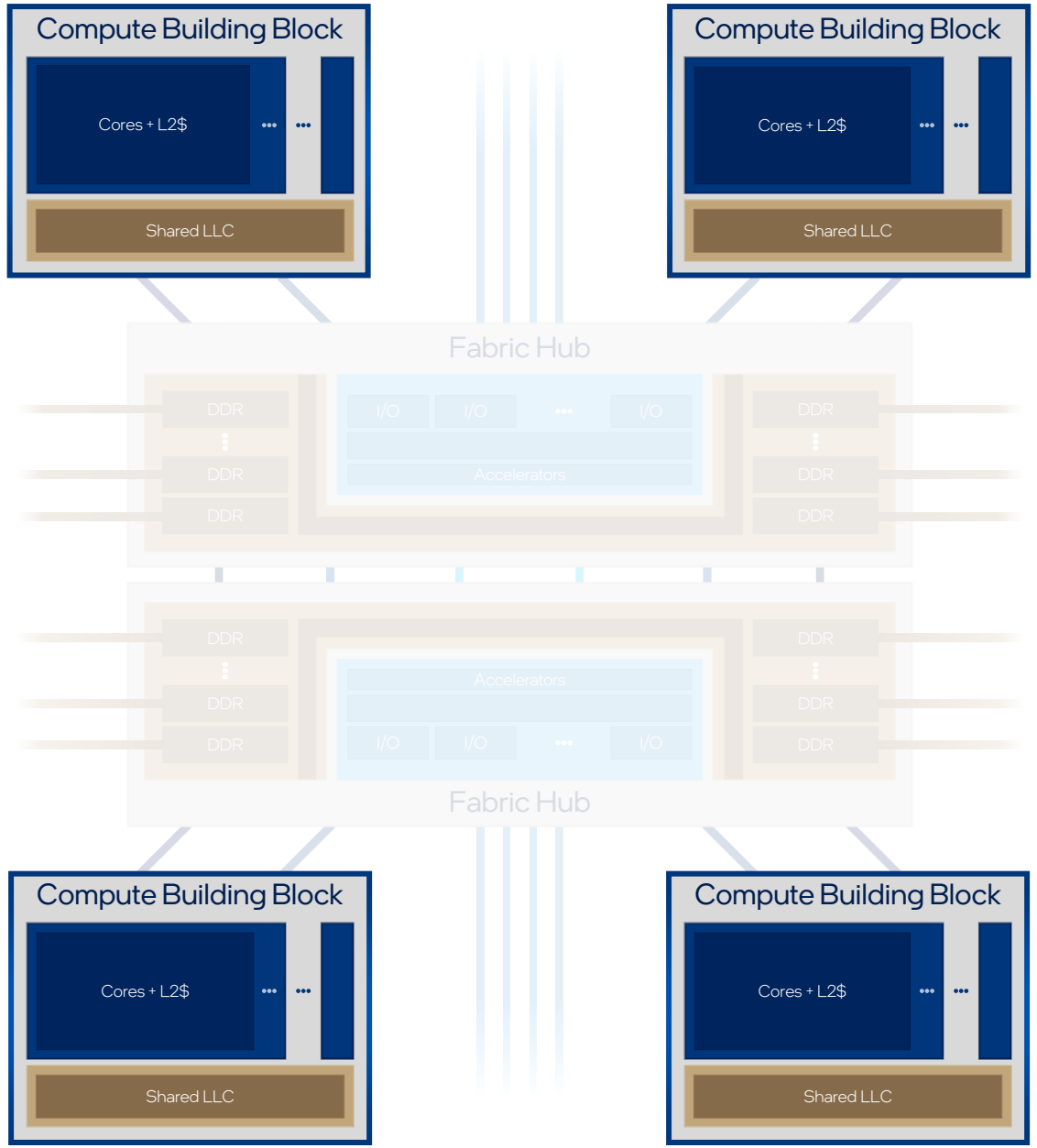
Fan-out Fabric Architecture

Scalable Compute Building Blocks

Centralized I/O & Memory in Fabric Hub

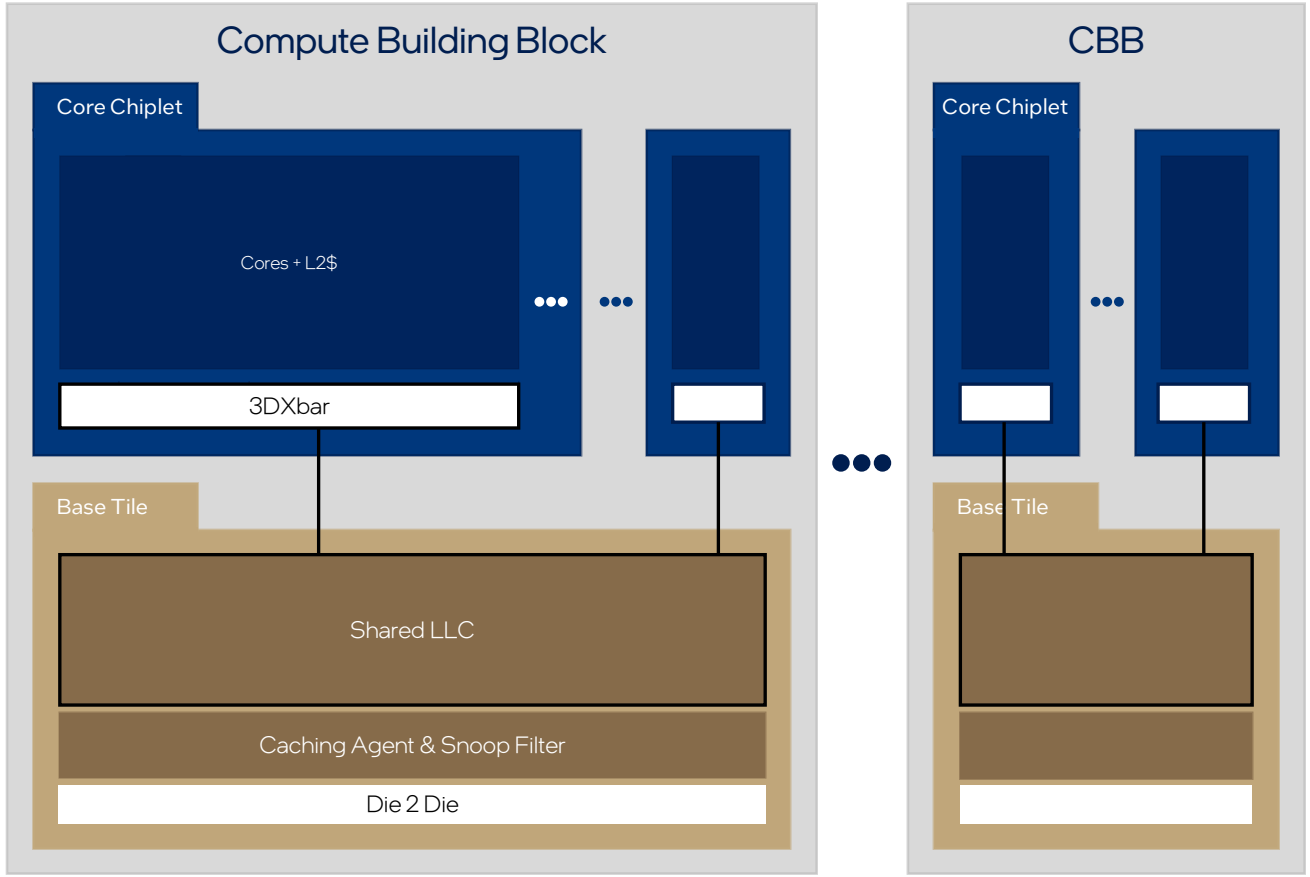
Unified Memory Fabric

Flexible I/O Fabric



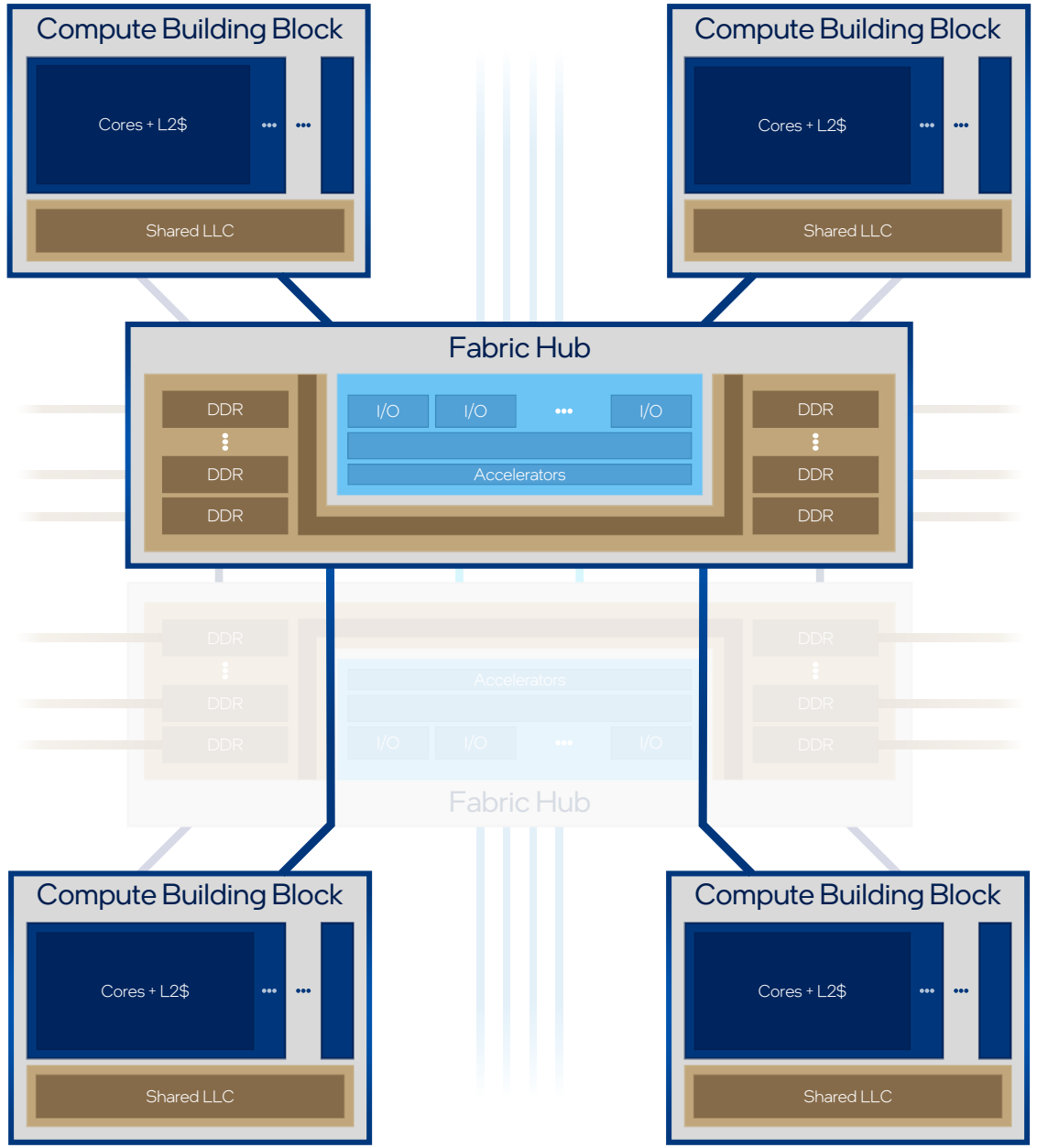
Scalable Compute Building Blocks

- Up to 16 cores per **Core Chiplet**
- **3D Xbar** connects cores to LLC **on Base Tile**
- LLC in **Base Tile** shared across **Compute Building Block**
- Scalable number of **core chiplets** per **base Tile**
- Scalable number of **Compute Building Blocks** / Chip



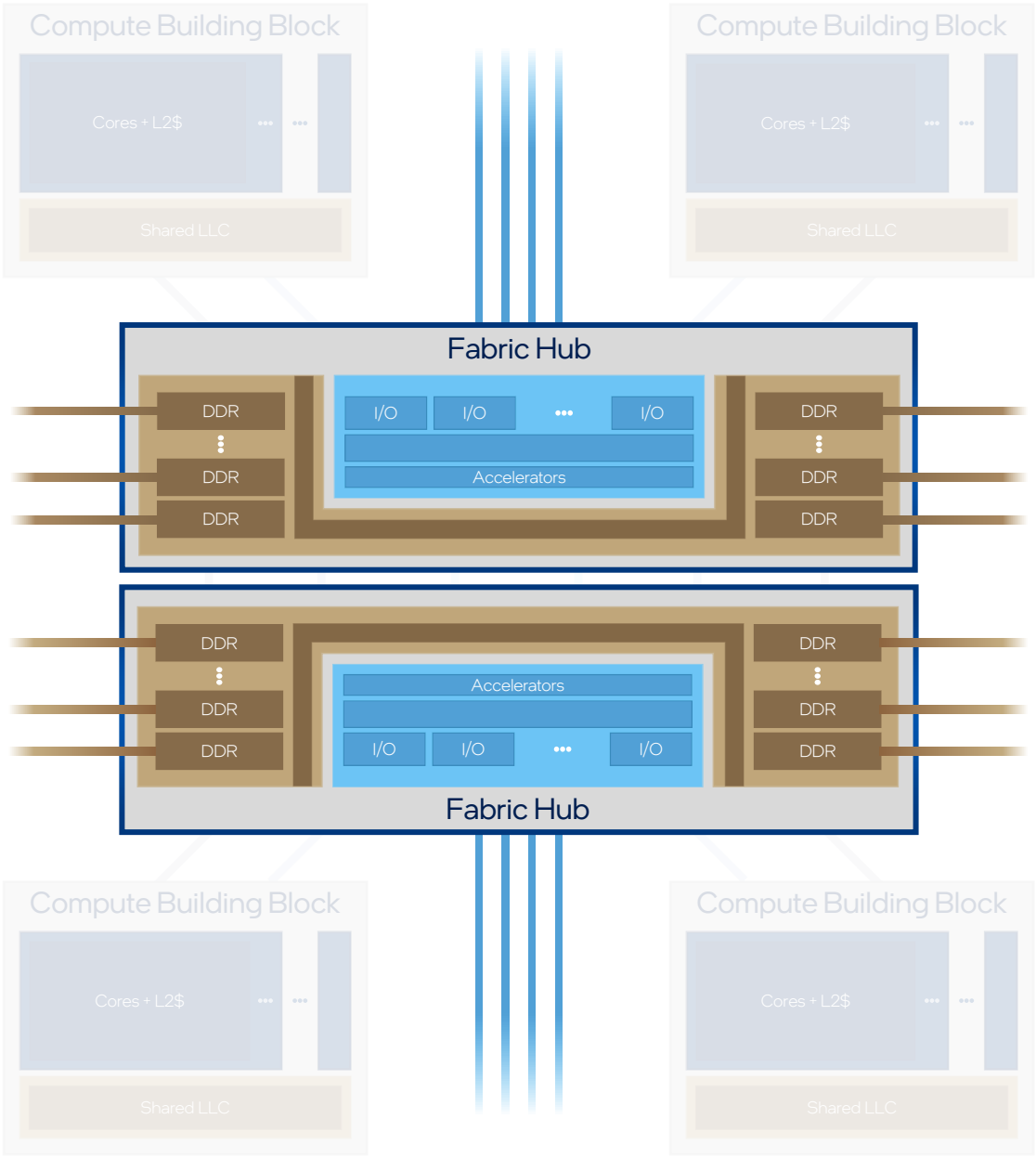
Fan-out Fabric Architecture

- **Scalable Compute Building Blocks**
- **Centralized I/O & Memory in Fabric Hub**
- **Unified Memory Fabric**
- **Flexible I/O Fabric**



Fan-out Fabric Architecture

- **Scalable Compute Building Blocks**
- **Centralized I/O & Memory in Fabric Hub**
- **Unified Memory Fabric**
- **Flexible I/O Fabric**



Scalable Fabric Hub

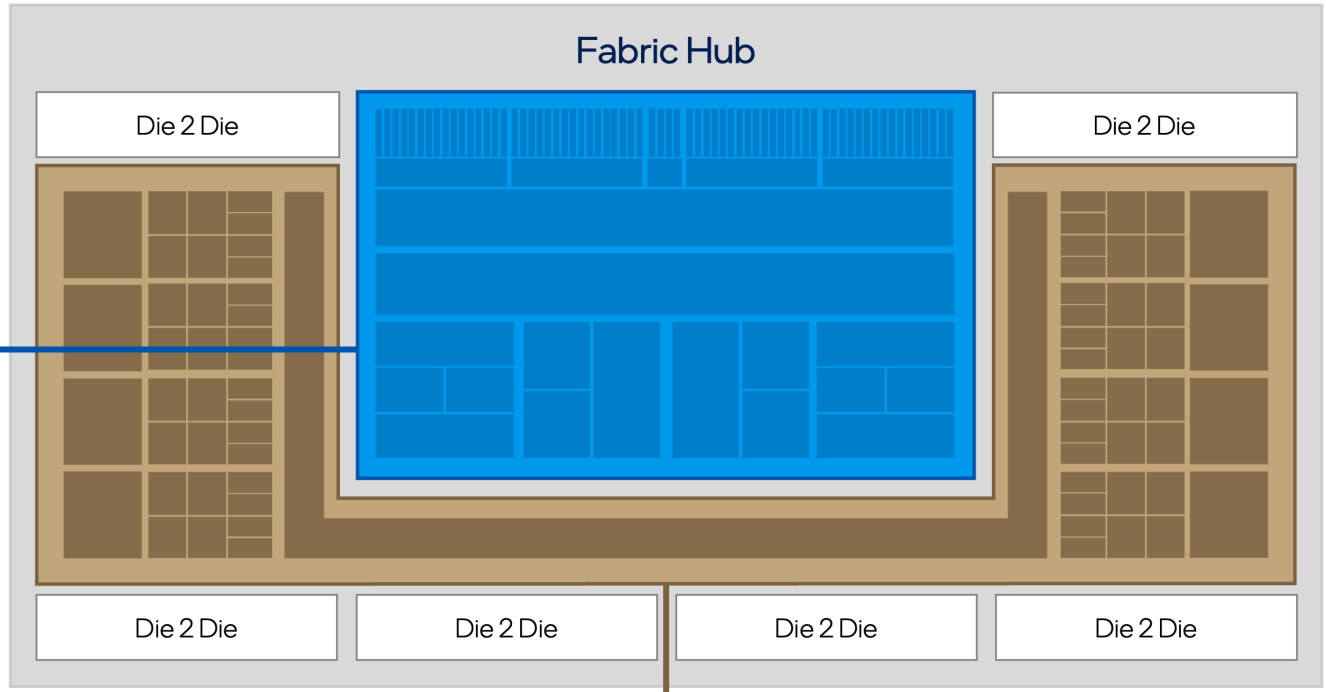
Flexbus I/O Fabric

Flexible I/O Subsystem	I/O PHY	I/O Cache	Accelerators
I/O Interconnect	Snoop Filter		

Unified Memory Fabric

Memory Subsystem with up to 1.6 TB/s bandwidth

Memory Value Function	DDR PHY	Home Agent	Snoop Filter
Mem Controller	Mem Encryption	Mem Interconnect	



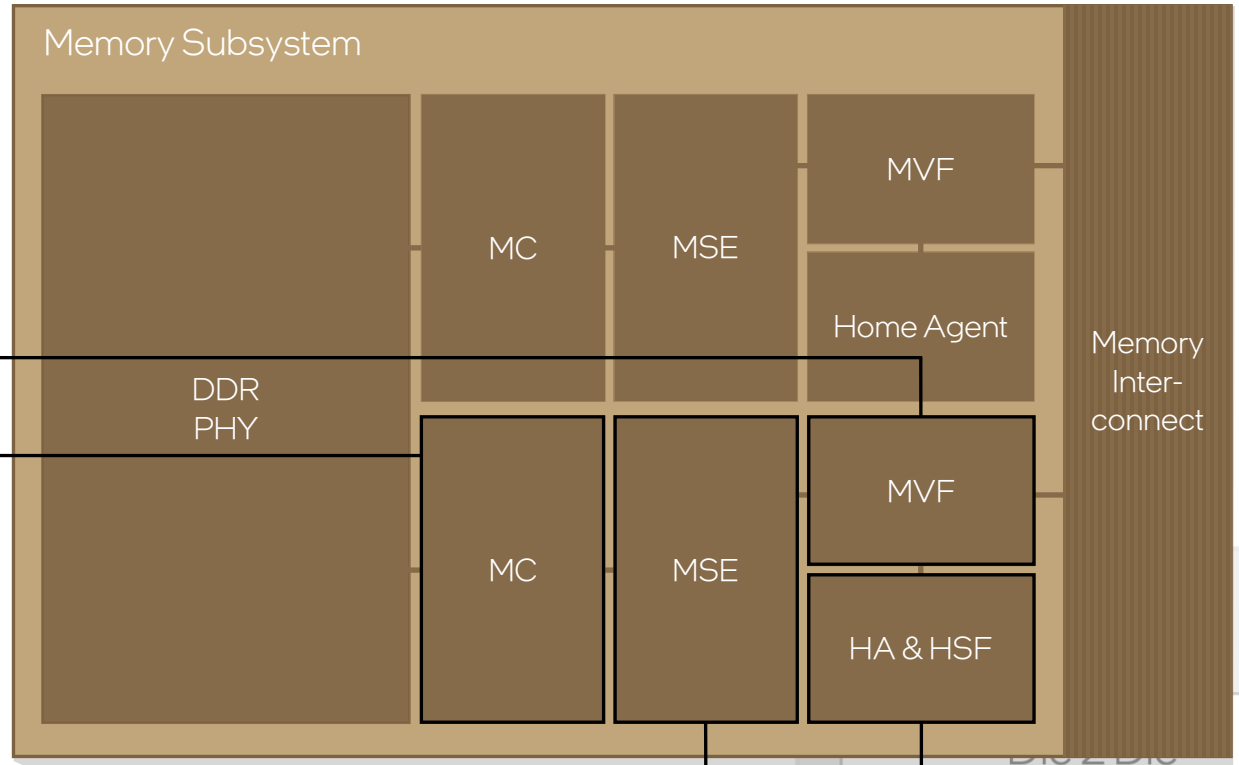
Unified Memory Fabric

Memory Controller (MC)
schedules DDR accesses with robust ECC, power management, and row-hammer protection

Memory Value Function (MVF)
supports CXL memory (1LM or Flat2LM) and mirroring

Memory Encryption Engine (MSE)
supports DDR or CXL memory encryption

Home Agent (HA) & Home Snoop Filter (HSF)
On-Die Home Snoop Filter for coherency



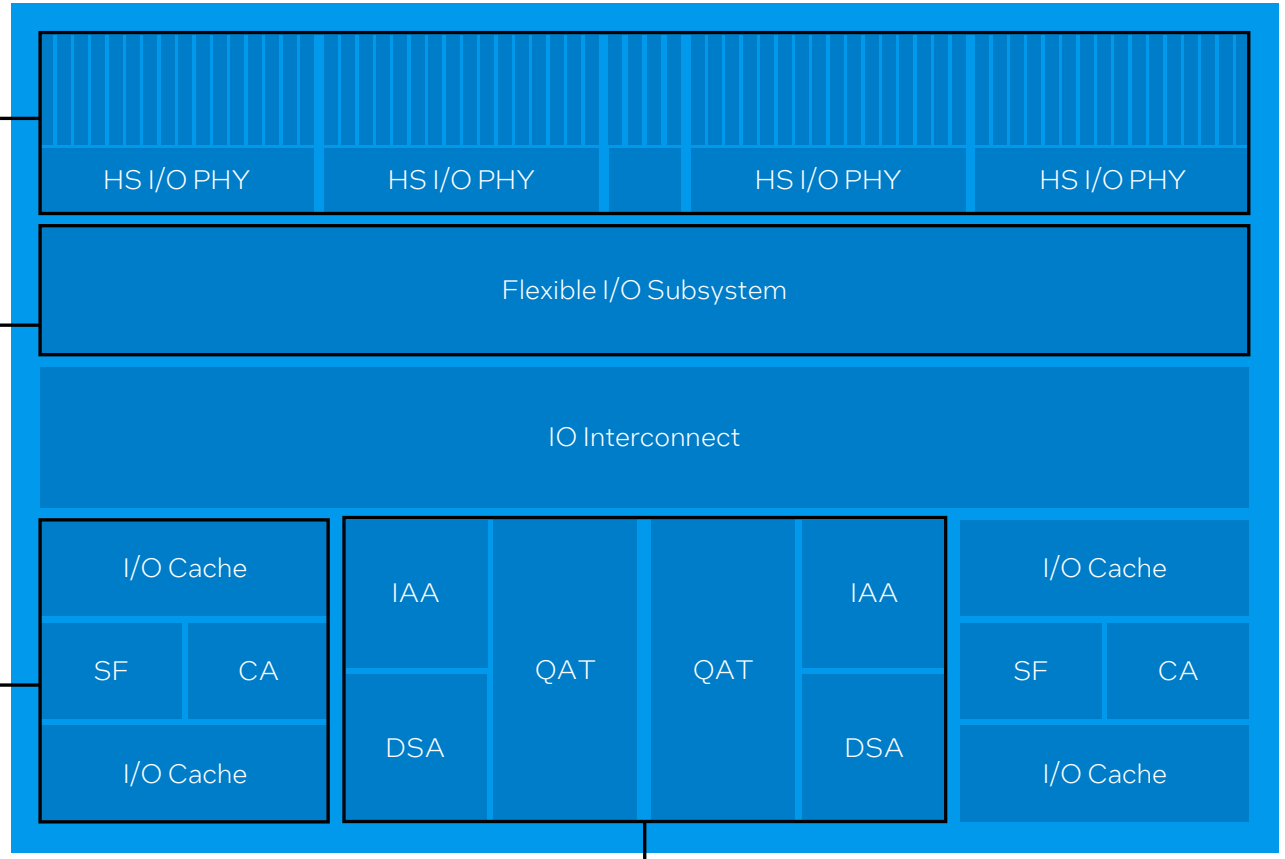
Flexbus I/O Fabric

I/O PHY
4 x 16 high speed I/O lanes per Fabric Hub
Additional 4 lanes of PCIe Gen4 for Platform Infra

Flexible I/O Subsystem
Each x16 can be used as PCIe Gen6, CXL 3, UPI 3 ports, or some combinations

I/O Caching & Filtering
Up to 16MB of IO Cache, with on-die Snoop filtering and caching agent for coherency

Accelerator Complexes
Each Fabric Hub houses 2 Accelerator complexes with Intel® QAT, Intel® DSA, Intel® IAA



Power & Thermal Management

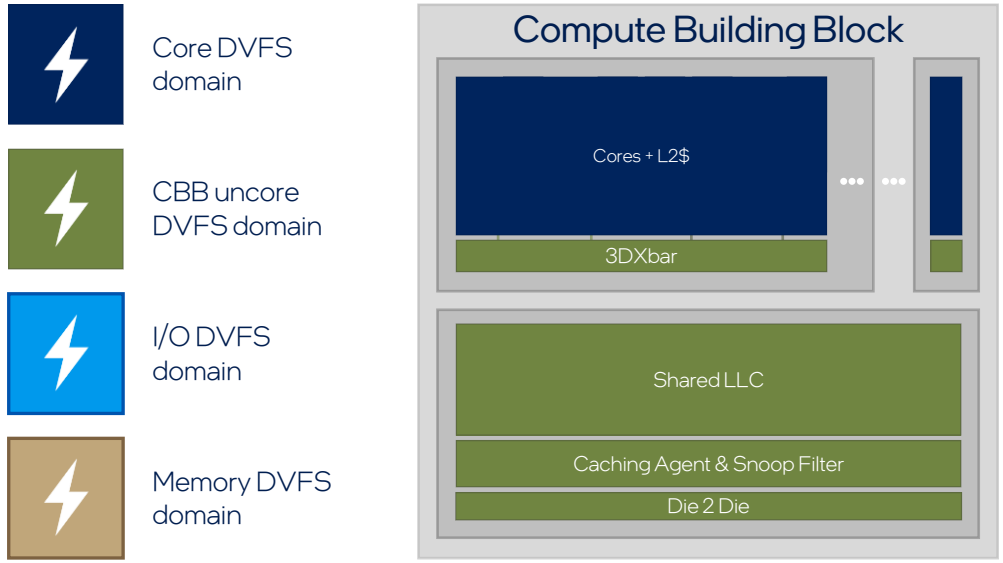
Distinct DVFS domains

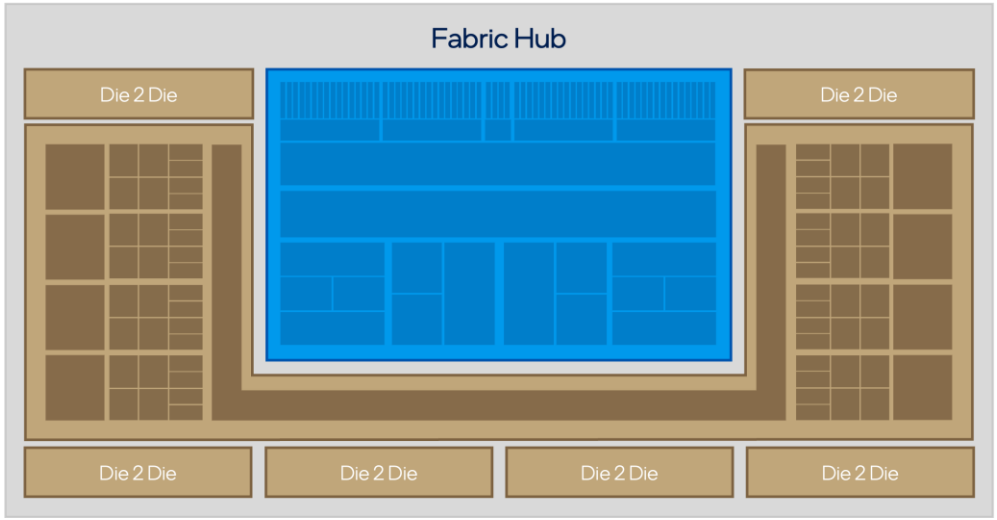
Core and CBB power management

- New core idle power state with L2 cache retention
- Priority Core Turbo support

Fabric Hub power and thermal management

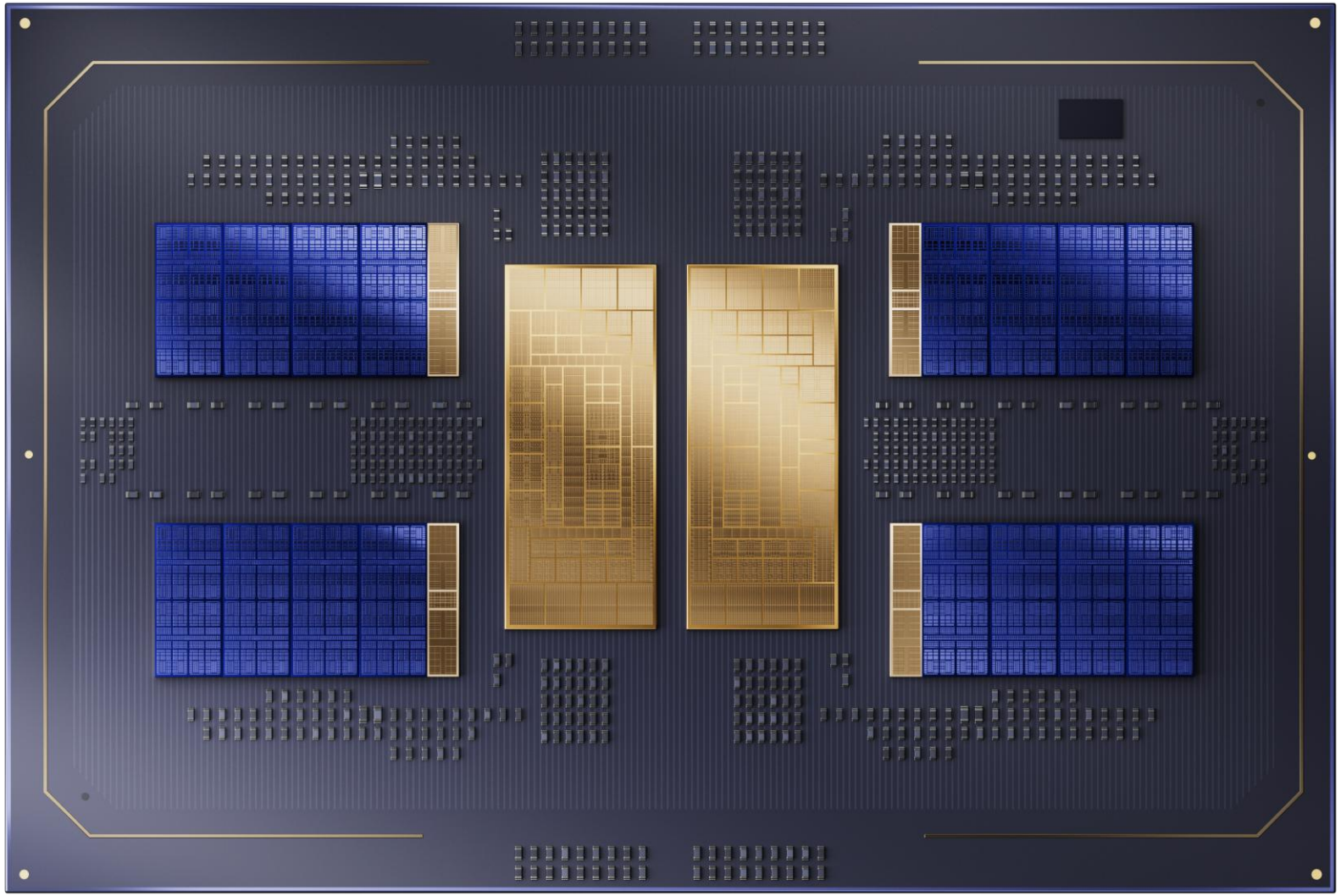
- MR4 based memory thermal management
- L0p support on both UXI and PCIe
- Use-case based power direction to I/O & Mem





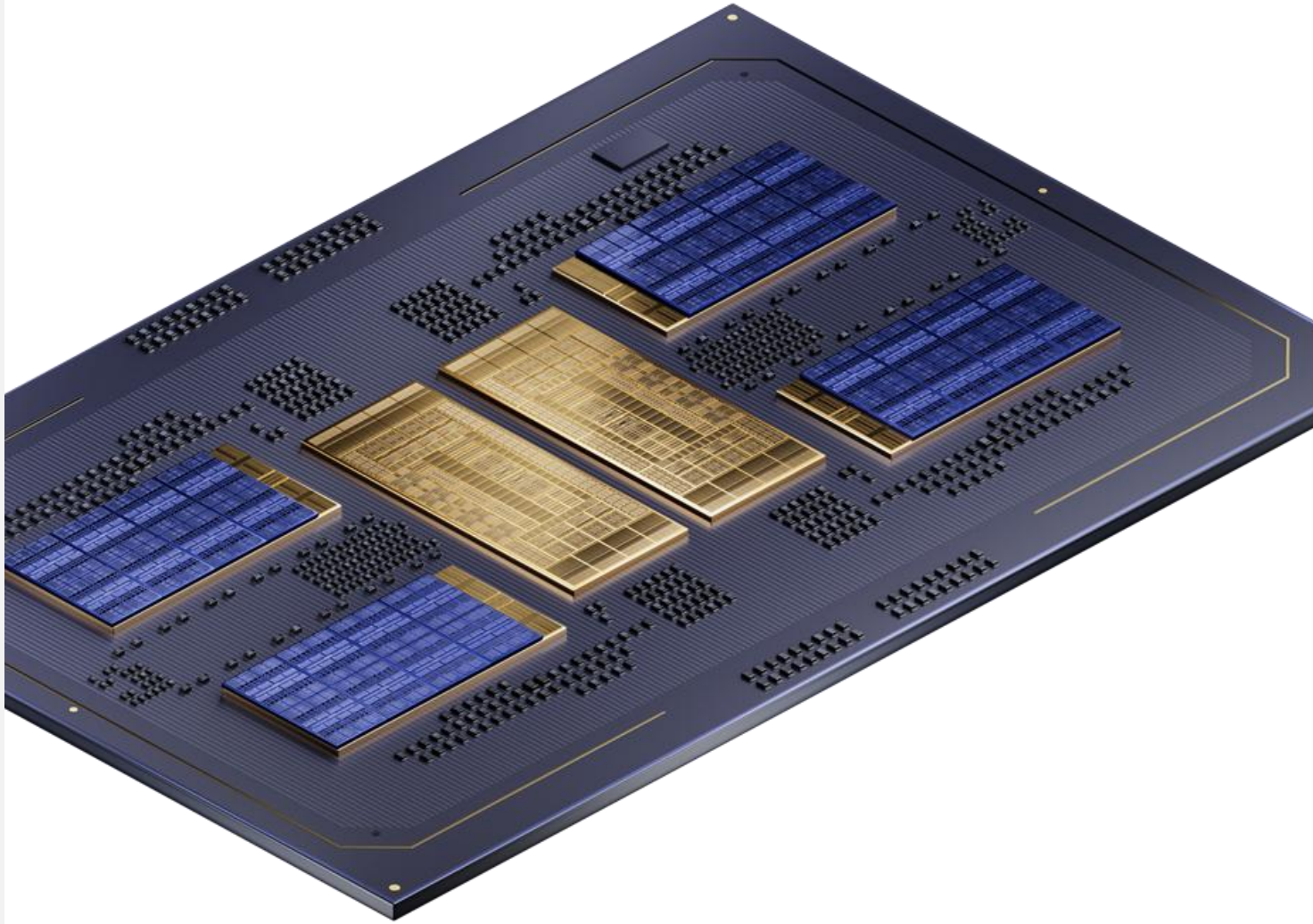
Diamond Rapids

First Intel® Xeon® built with Fan-out Fabric Architecture



Diamond Rapids SoC Overview

Compute	4 Compute Building Blocks
	up to 1.28 GB of LLC across all CBBs
	Up to 256 cores total
Memory	16 channels
	Up to 8000MT/s DDR5
	Up to 12800MT/s MRDIMM
I/O	128 Lanes
	PCIe Gen 6, CXL 3.0, UPI 3
	8 lanes PCIe Gen 4
Security & Acceleration	Intel® TDX, Intel® TDX connect, Intel® SGX
	Intel® QAT, Intel® DSA, Intel® IAA
	Intel® AMX, Intel®AVX-10



Diamond Rapids Construction

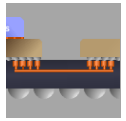
Process

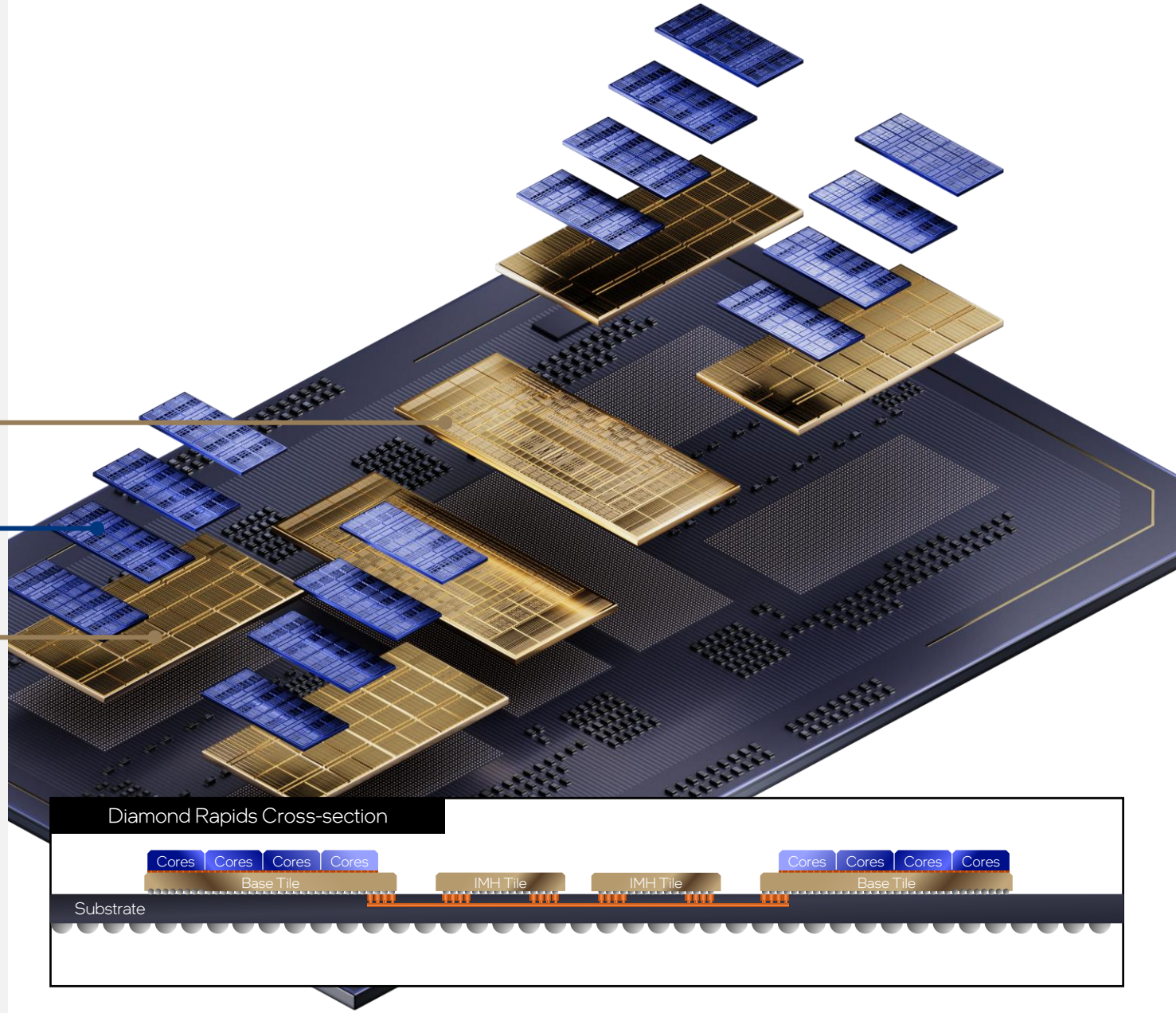
2x Fabric Hub Tile	intel 3
16x Core Chiplets	intel 18A-P
4x Base Tile	intel 3-T

Packaging

Hybrid Bonding Interface (HBI)
Foveros 3D Direct

Die to Die Interface
Substrate Copper Links (UCIe-S – 16G)

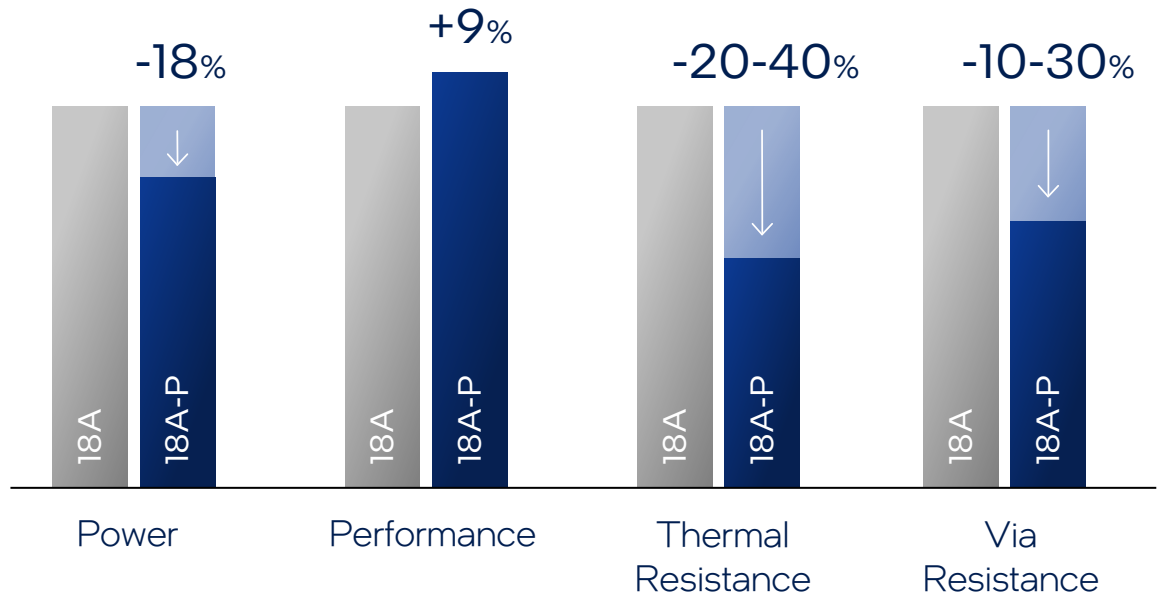




intel 18A-P

Intel 18A-P

- Expanded low-power, high-performance device offerings
- PowerBoost dual-contact low-resistance option
- New 5th logic Vt option between ULVT and LVT
- 33% tighter skew corners
- Improved process variation



*based on an industry standard core sub block

Intel® Advanced Performance Extensions

Intel® APX

Recompile without source code changes

Performance increase on recompiled SW

Legacy x86 code compatible

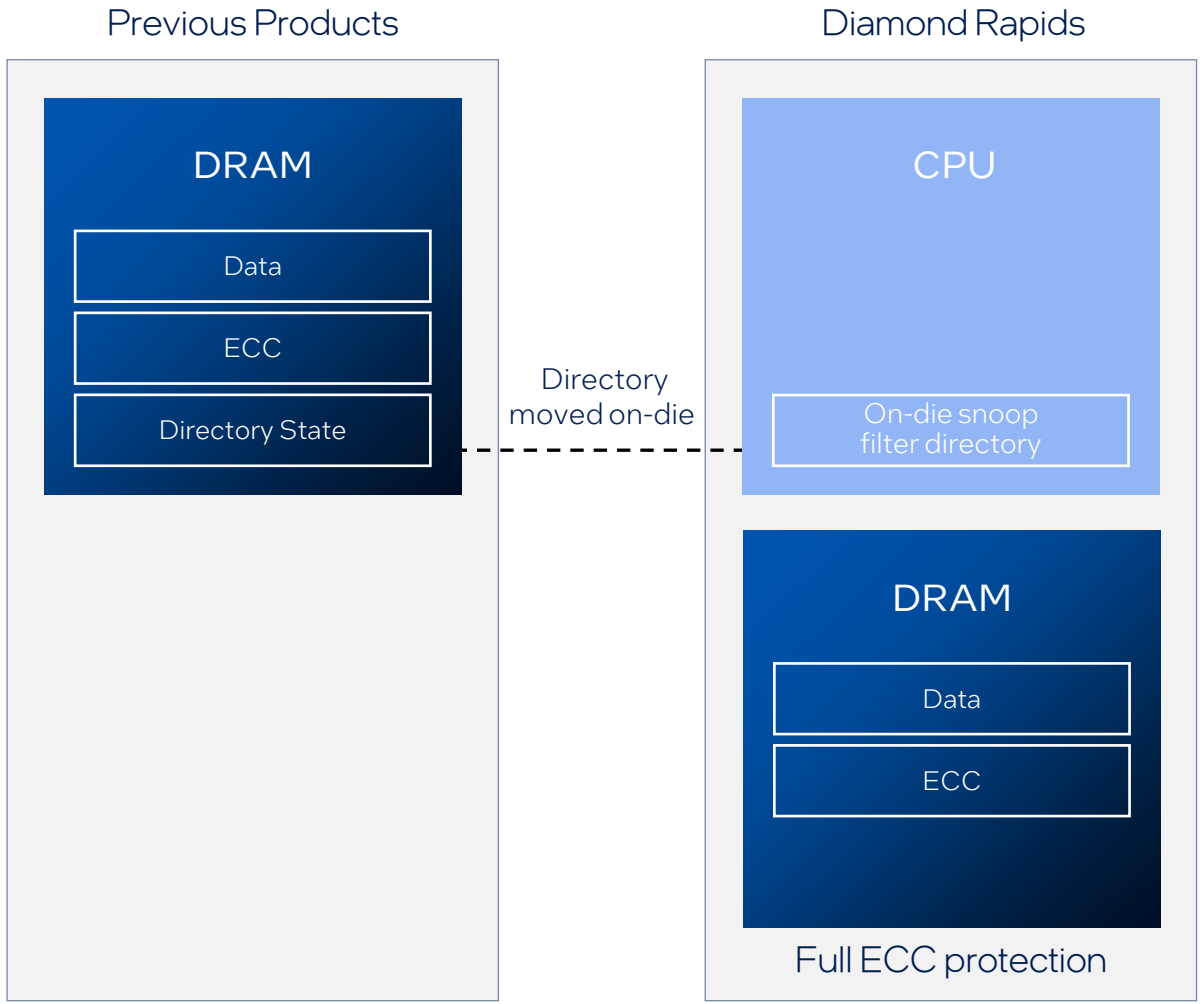
Security / RAS

On-die snoop filter
replaces DRAM-resident directory state

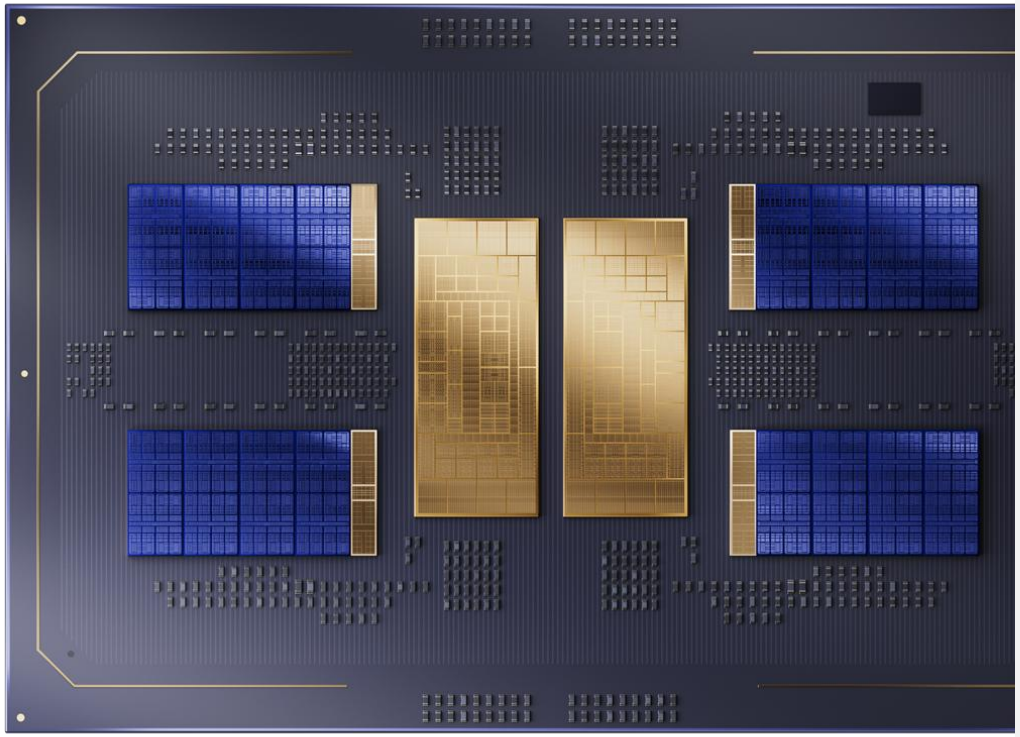
Preserves full ECC protection
by eliminating directory storage in memory

Lower latency & reduced coherence traffic
for cache lookups

Simpler memory subsystem
with improved reliability



Designed for Next Gen Agentic AI Infrastructure

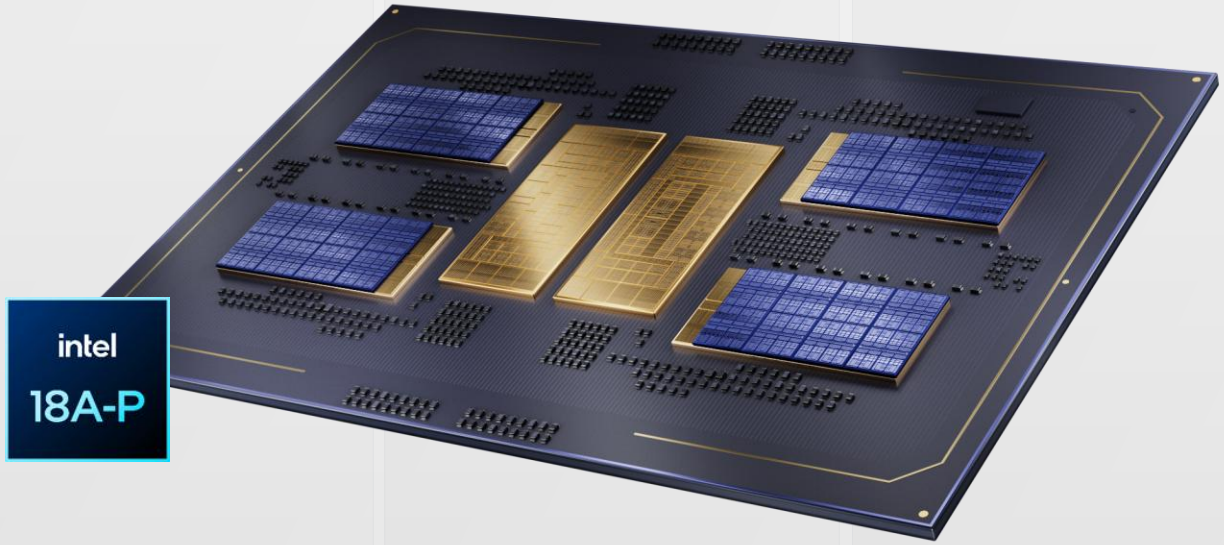


High-performance core count	Up to **256** cores
Matrix multiplication acceleration & Enhanced vector instruction sets	**AMX & AVX 10.2** With FP8, FP16 & BF16 support
Large, fast, coherent cache	**1.28GB** Last Level Cache
High bandwidth memory support	**1.6TB/s** 12800MT/s MRDIMM
High-Bandwidth, low latency I/O Fabric	**128 Lanes** up to 2TB/s Bi-directional bandwidth

Performance | Efficiency | Acceleration | Scalability | Security

Diamond Rapids

Intel's next generation Xeon CPU with performance cores



High-throughput Compute	Efficient Performance	Infrastructure & AI Acceleration	Modular Architecture	Multi Level Security
Up to 256 cores	Unified Memory Fabric	Intel AMX	Fan-out Fabric Architecture	Intel TDX, Intel SGX
Built-with 18A-P	Independent DVFS	Intel AVX 10.2	Modular CBBs	Improved RAS
Massive LLC	Lower Power I/O Fabric	Intel QAT/ Intel DSA/Intel IAA	Foveros 3D Direct (HBI)	Enhanced QoS & Telemetry

Notices & Disclaimers

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available.

These are not "commercial" names and not intended to function as trademarks.

Performance varies by use, configuration and other factors. Learn more at intel.com/performanceindex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.


intel
HOT
CHIPS
Thank You