August 2024

Built for the Edge: The Intel® Xeon® 6 SoC

Praveen Mosur
Intel Fellow

Edge is the Next Frontier in Digital Transformation

- Secure, connected, managed
- Compute density for AI & scalar workloads
- Integrated connectivity
- Optimized for space & power constrained ruggedized environments

The Processor Built for the Edge

Consistent Architecture from Edge to Cloud

Compute Optimized

Compute
Scalar & data parallel workloads

Memory
Low latency, high bandwidth

IO
High bandwidth PCIe Gen 5

RAS
Server grade robustness

Edge Optimized

Security
Confidential AI enabled

Scalability
Multiple edge systems
based on one architecture

Integration
Ethernet and accelerators

Form Factor
Optimized for space and
power constrained environments

Intel® Xeon® 6 SoC

Scalable Architecture, Integrated Form Factor

Redwood Cove **P-Core**

Compatible **Packages** — **4Ch** | **8Ch**

Memory

up to **8Ch** DDR5

Multiplexed Combined Rank **MCR DIMMs**

Integrated **Ethernet**

8 ports

2x **100G**

4x **50G**

8x **25G, 10G, 1G, 100M**

PCI Express & Compute Express Link

32L PCIe 5.0

16L PCIe 4.0

16L CXL 2.0

Integrated **Accelerators**

Media Accelerator

Intel® **QuickAssist Technology**

Intel® **vRAN Boost**

Intel® **Dynamic Load Balancer**

Intel® **Data Streaming** Accelerator

Intel® Xeon® 6 SoC

Delivering Performance and Efficiency

up to
2X
increase in integrated Ethernet throughput[1]

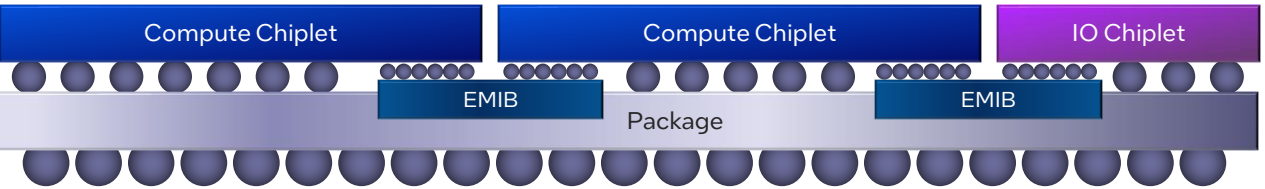
>3X
increase in core count and memory bandwidth[1]

up to
2.5X
increase in IO performance[1]

1. See appendix for platform configuration and workload details. Results may vary.

Modular Product Architecture

Flexible Design Enables Wider Range of Optimized Products



Support for a range of core counts and thermals, including rugged deployments

Separate compute and IO chiplets

Leading process technology delivers low power & efficient design

Embedded multi-die interconnect bridge (EMIB) packaging enables high bandwidth, low latency cache & memory

Performance-Optimized Core

Proven Intel® Xeon® Architecture

- Optimized for high performance per core
- Built on latest Intel 3 technology
- Improved power efficiency[1]
- Intel® Advanced Matrix Extensions

Enhanced µArch

- 8-wide decode, 6-wide allocate and 8-wide instruction retire
- 64KB I-cache & 48KB D-cache
- 512 OOO execution engine
- 2MB private L2 cache

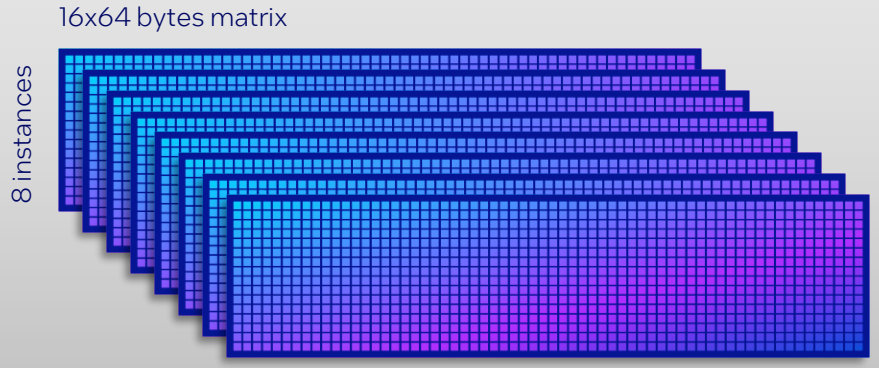
1. See appendix for platform configuration and workload details. Results may vary.

Integrated AI Acceleration

- Support for VNNI and Intel® AMX
- Support for FP16
- Increased inflight memory requests/BW[2]

	Intel® Xeon® D 2899NT processor (AVX512 VNNI)	Intel® Xeon® 6 SoC (Intel® AMX)
Resnet-50 (images/sec)[1]	Baseline	Over 8X
Visual Transformer[1]	Baseline	Over 6X

Intel® Advanced Matrix Extensions (Intel® AMX)



Store bigger chunks of data[2]

Instructions compute larger matrices in a single operation[2]

1. Results estimated based on internal Intel testing as of 5/18/24 using preproduction silicon. Results may vary.
2. In comparison to AVX512 VNNI.

Unified Cache & Memory

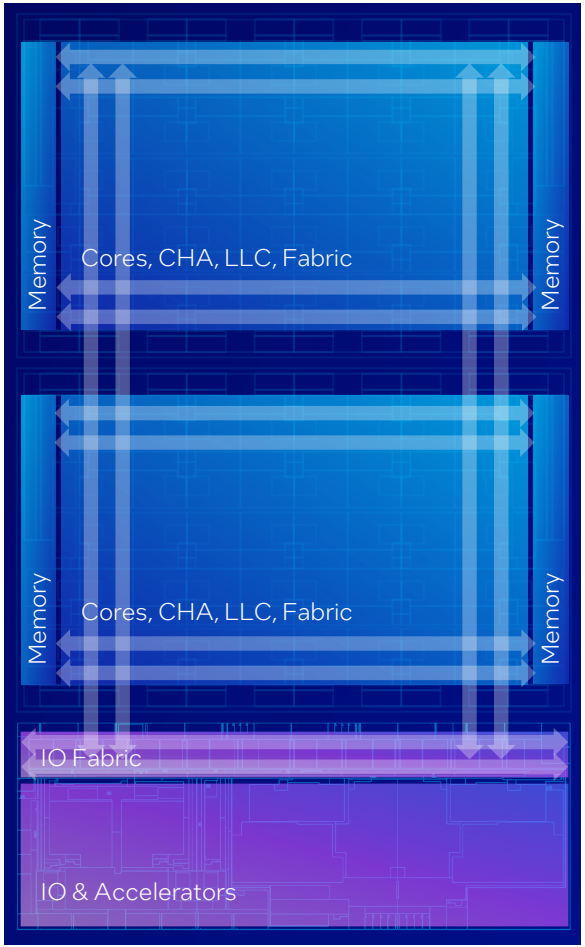
Mesh Superhighway Scales Naturally from Small to Large Configs

Logically monolithic mesh enables direct access between agents (reduced latency and jitter)

Fabric distributes IO traffic across multiple columns to ease congestion

Larger last level cache is shared by all cores and IO agents

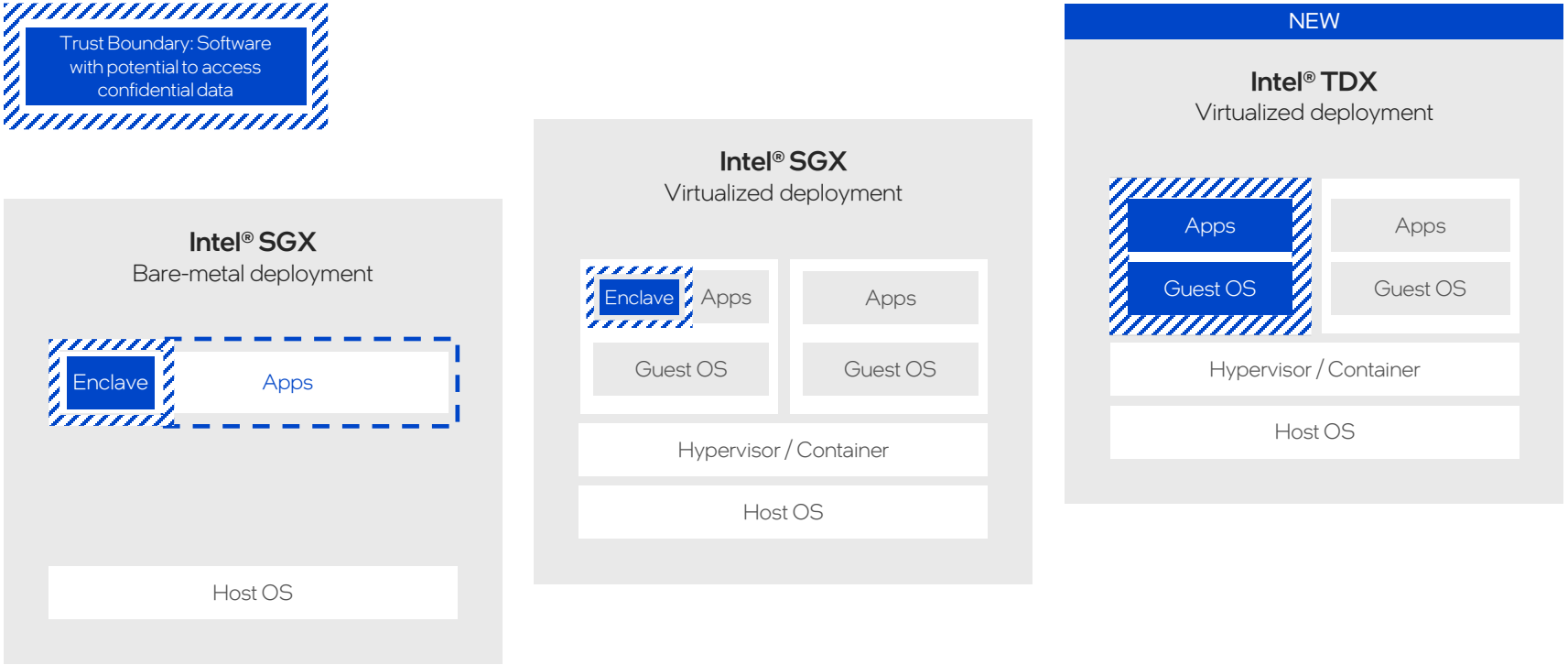
Data direct IO lowers latency and memory bandwidth for IO intensive workloads[1]



1. Results estimated based on internal Intel testing as of 8/13/24 using preproduction silicon. Results may vary.

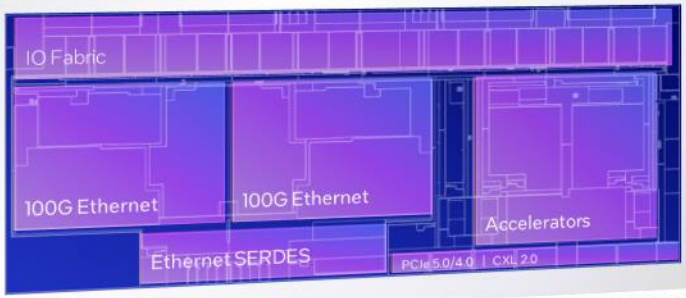
Intel® Software Guard Extensions (Intel® SGX) & Intel® Trust Domain Extensions (Intel® TDX)

Broad Array of Confidential Computing Options



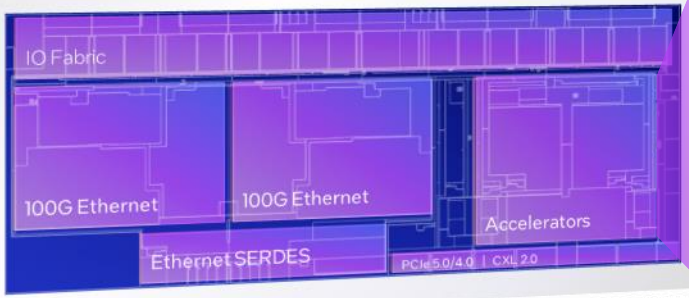
IO Chiplet

Power-Optimized on Intel 4 Process Node



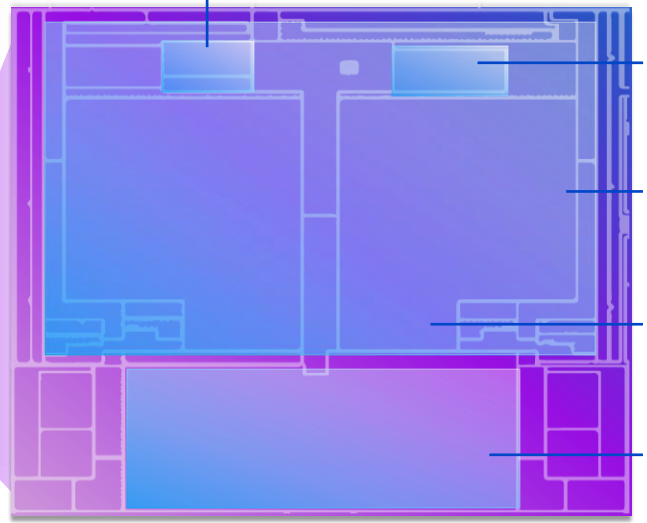
IO Chiplet

Power-Optimized on Intel 4 Process Node



Integrated Accelerators

Intel® Data Streaming Accelerator
(Intel® DSA)
for infrastructure processing & storage



Intel® Dynamic Load Balancer
(Intel® DLB)
for packet processing

Intel® Quick Assist Technology
(Intel® QAT)
for storage, CDN, security

Media Accelerator
for media processing

Intel® vRAN Boost
for telco radio access networks

Intel® Resource Director Technology (Intel® RDT)

End-to-End Resource Monitoring & Control

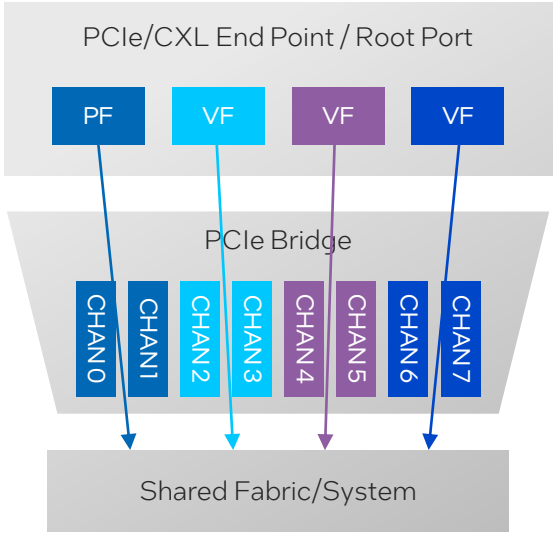
Extending Intel® RDT to IO Devices

End-to-end QoS management by extending QoS tagging to PCIe, CXL and integrated accelerators

Reduces IO latency to memory through shared fabric & buffers

Architecture supports channel-granular based tagging – addressing head-of-line blocking

Platform QoS utility support



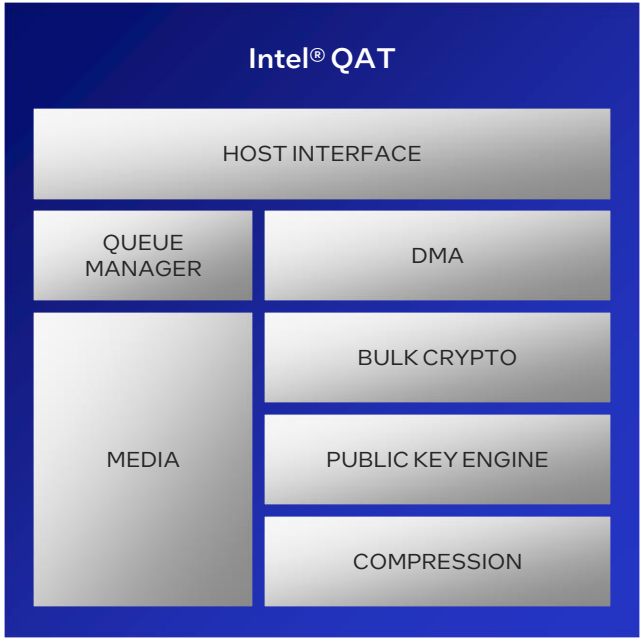
Channel to RMID/CLOS Tagging

CHANNEL	RMID	CLOS
CHANNEL	RMID	CLOS
CHANNEL	RMID	CLOS

RMID: Resource Monitoring ID
CLOS: Class of Service ID

5th Generation

Intel® QuickAssist Technology (Intel® QAT)



Acceleration Capabilities

- Bulk crypto for wired/wireless network security and encrypted storage
- Public key engine
- Lossless data compression for storage and network optimizations
- NEW: Media transcode for live OTT, VOD and broadcast media

Architected for Edge & Cloud native Application Integration

- Shared virtual memory
- 3 level rate limiting scheduler
- SR-IOV with support for 128 VFs

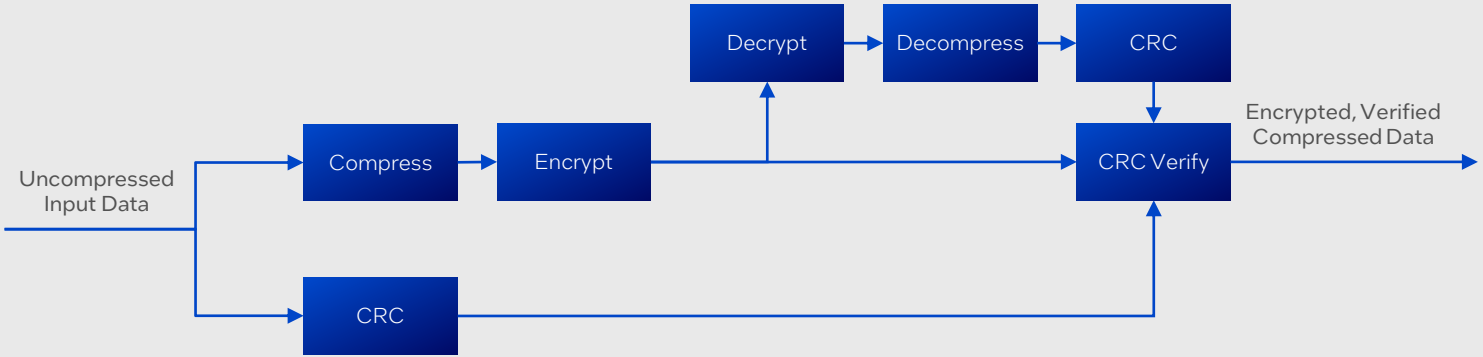
Intel® QuickAssist Technology

Lossless Compression

Single-pass fused operations enable encrypted/secure storage, advanced RAS, and lower latency and memory bandwidth

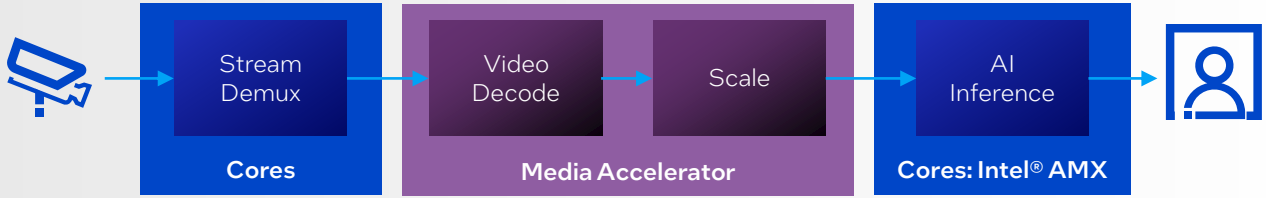
Multiple single-pass fused operations can be enabled based on usages

Example: Compress + Encrypt + Verify

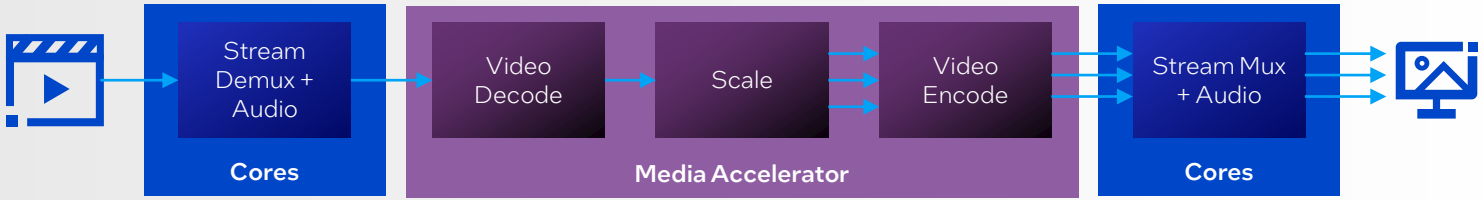


Media Accelerator:
Video Transcode and Visual Inference
for Live Over-the-Top, Video-on-Demand and Broadcast Media

Visual Inference



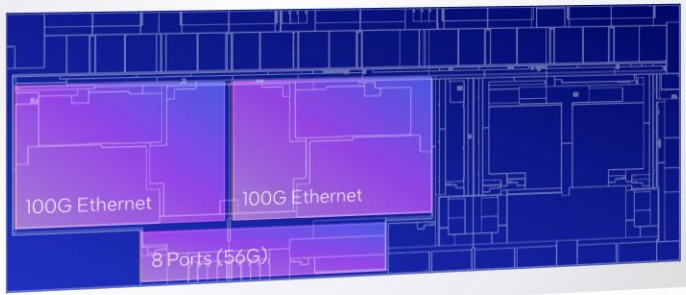
Full Offload Video Transcode



- Media encode/decode
- Scaling
- Crop
- 1080p30 AVC/HEVC/AV1

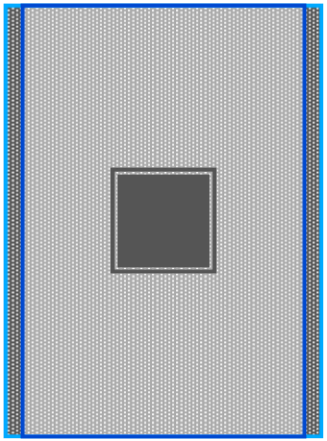
Integrated Intel® Ethernet Technology

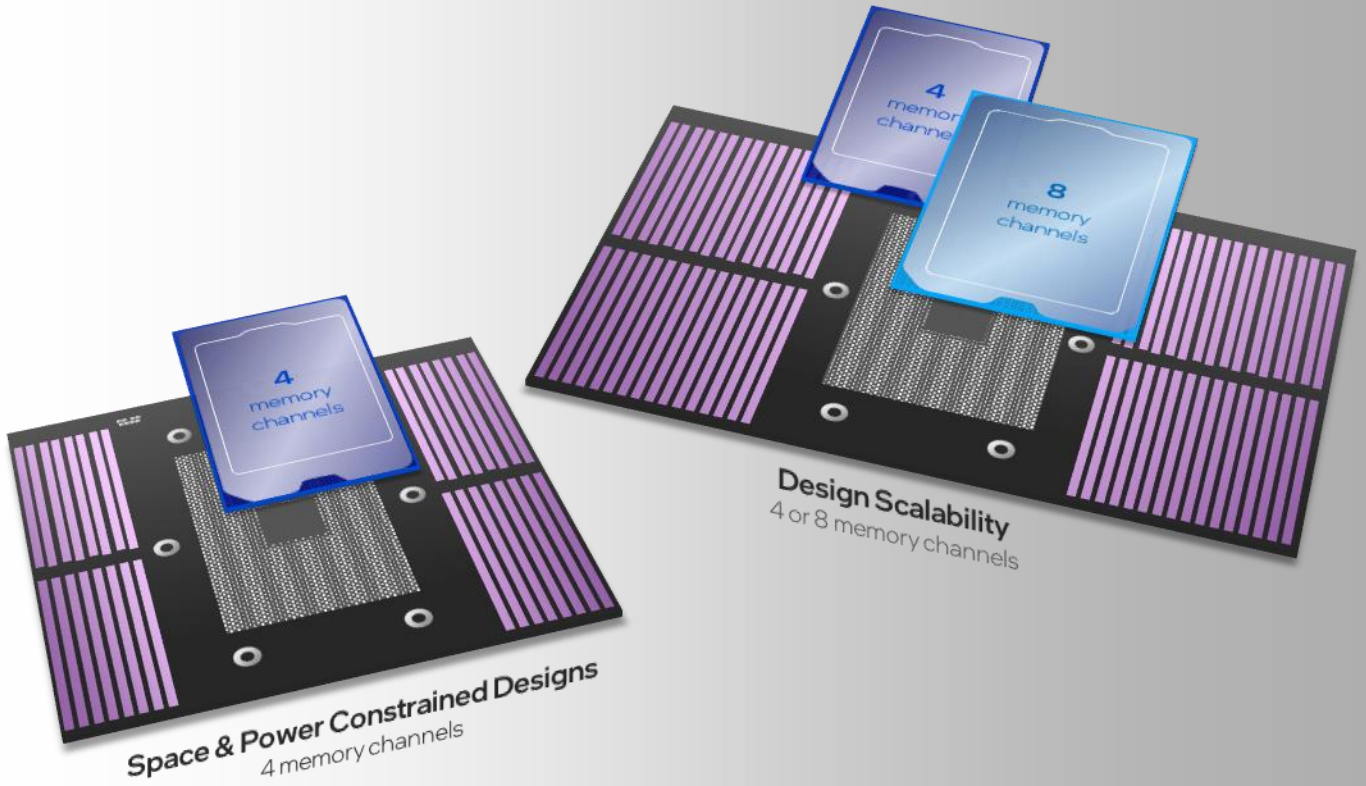
Latency Optimizations & New Features



- 2x100 Gbps programmable parsing
- Packet classification
- Integrated ACL processing
- Feature-rich RSS/flow director
- Advanced scheduling module: multiple layer hierarchical scheduler with dynamic updates, dual rate shaping, strict priority, WFQ or combination scheduling

Common Platform Design for Flexible Scalability



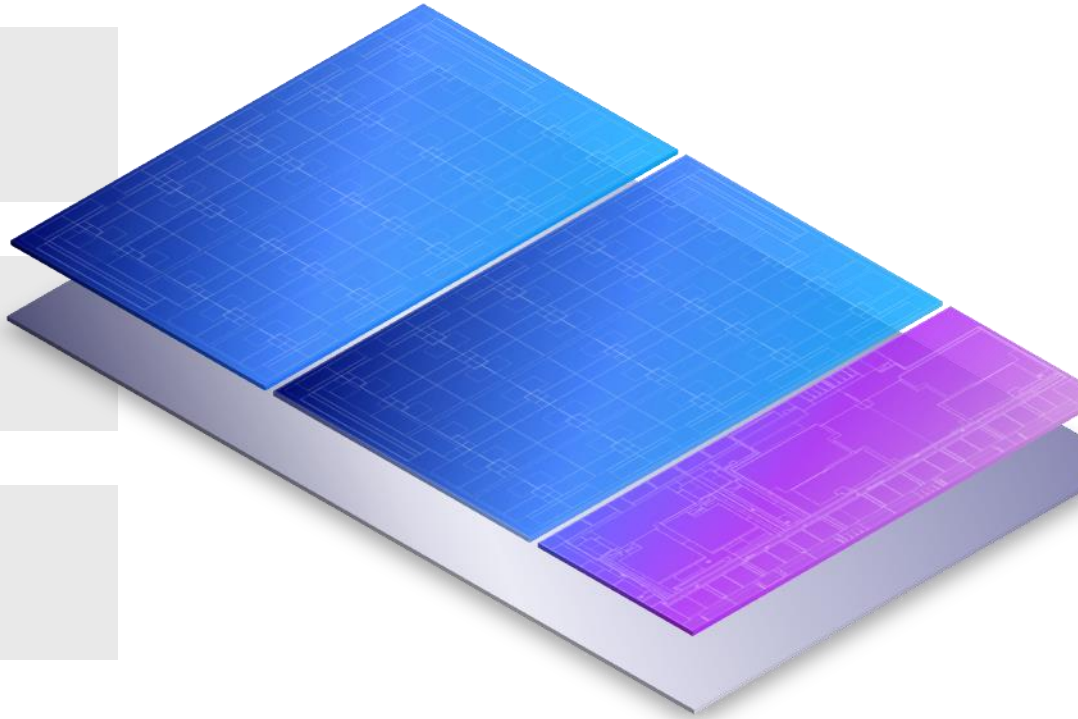


Intel® Xeon® 6 SoC: Built for the Edge

Multiple edge systems based on one architecture

Integrated acceleration enabling Confidential AI

Optimized for space and power constrained ruggedized environments



Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

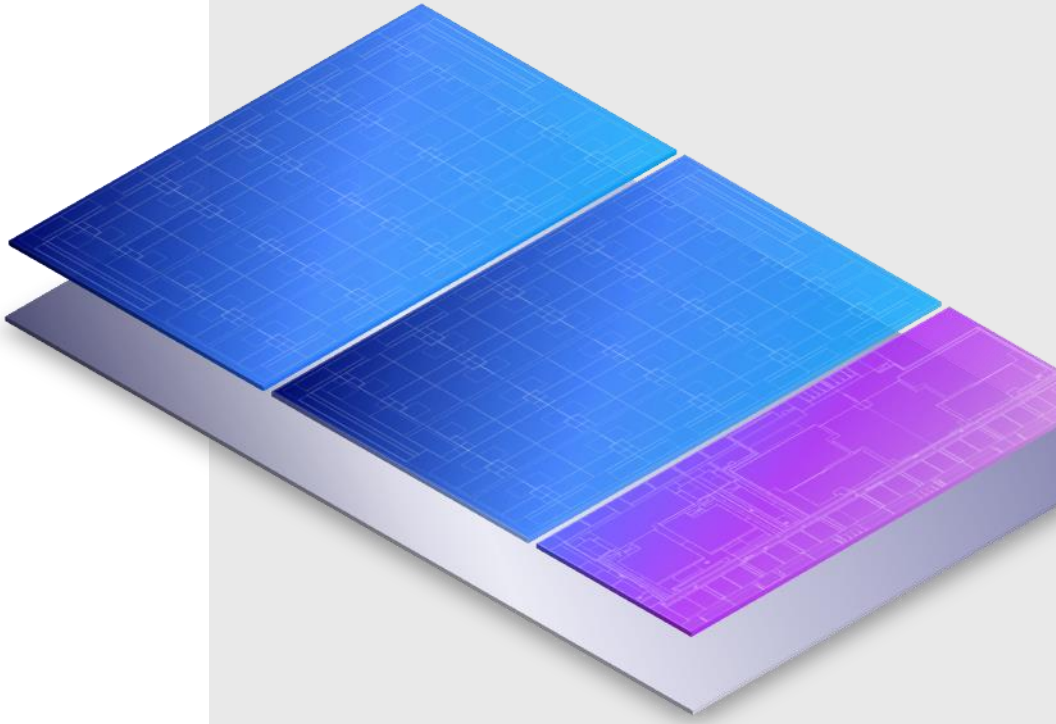
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Availability of accelerators varies depending on SKU. Visit https://ark.intel.com/content/www/us/en/ark.html for additional product details.

Your costs and results may vary.

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

Thank You.




intel

Appendix

APPENDIX – Configuration and System Details for AI-Related Performance Claims

Resnet50-v1-5:
EDL-D: Test by INTEL as of 05/22/24. 1-node, 1x Intel(R) Xeon(R) D-2899NT CPU @ 2.20GHz, 22 cores, HT On, Turbo Off, NUMA 1, Total Memory 128GB (4x32GB DDR4 3200 MT/s [3200 MT/s]), BIOS IDVLCRB1.86B.0027.D11.2306160450, microcode 0x1000260, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 223.6G KINGSTON SUV400S37240G, 1x 240M Disk, Ubuntu 22.04.3 LTS, 5.15.0-27-generic, GCC 11.3, Resnet50-v1-5 with AVX512 VNNI, score =.511; Perf/Watt=4.5

GNR-D: Test by INTEL as of 05/18/24. 1-node, 1x Genuine Intel(R) CPU $0000%@, 42 cores, HT On, Turbo On, NUMA 1, Total Memory 128GB (4x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 2x Intel Corporation, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 240M Disk, 1x 240M UDisk, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 11.3, ResNet50-v1-5 with Intel® AMX, score =4198; Perf/Watt=22.3

SW config: Topology – ResNet50-v1-5; Framework – Pytorch using IPEX (Throughput tests); Batch Size –128; Accuracy – 0.781; Function – Inference; Data Type – int8; No. of instances – 1 (All cores utilized in a single instance); Source Code - https://github.com/intel-innersource/frameworks.ai.infrastructure.dlboost.pytorch/blob/2024_ww03; IPEX: 2.2.0+gitf1112eb; PT : 2.2.0a0+git2b96221

Power Efficiency Claim - Perf/Watt on EDL-D: 4.5; Perf/Watt on GNR-D: 22.3

Vision Transformer Base:
EDL-D: Test by INTEL as of 05/22/24. 1-node, 1x Intel(R) Xeon(R) D-2899NT CPU @ 2.20GHz, 22 cores, HT On, Turbo Off, NUMA 1, Total Memory 128GB (4x32GB DDR4 3200 MT/s [3200 MT/s]), BIOS IDVLCRB1.86B.0027.D11.2306160450, microcode 0x1000260, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 223.6G KINGSTON SUV400S37240G, 1x 240M Disk, Ubuntu 22.04.3 LTS, 5.15.0-27-generic, GCC 11.3, VisionTransformer Base with AVX512 VNNI, score =.119

GNR-D: Test by INTEL as of 05/18/24. 1-node, 1x Genuine Intel(R) CPU $0000%@, 42 cores, HT On, Turbo On, NUMA 1, Total Memory 128GB (4x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 2x Intel Corporation, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 240M Disk, 1x 240M UDisk, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 11.3, Vision Transformer Base with Intel® AMX, score =794

SW config: Topology – Vision-Transformer Base; Framework – Pytorch using IPEX (Throughput tests); Batch Size –128; Accuracy – 0.80; Function – Inference; No. of instances – 1 (All cores utilized in a single instance); Source Code - https://github.com/intel-innersource/frameworks.ai.infrastructure.dlboost.pytorch/blob/2024_ww03; IPEX: 2.2.0+gitf1112eb; PT : 2.2.0a0+git2b96221
Dateset - imagenet-1k; ViT Base: 12 encoders with 12 bidirectional self-attention heads totaling 86 million parameters, 768 Hidden Size, Input 224x224

APPENDIX – Configuration and System Details for DDIO-Related Performance Claims

GNR-D with Packets in Memory: 1-node, 1x GENUINE INTEL(R) XEON(R), 42 cores, HT On, Turbo On, NUMA 1, Integrated Accelerators Available [used]: DLB 1 [0], DSA 1 [0], IAA 0 [0], QAT 0 [0], Total Memory 128GB (4x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic. Software: VPP 23.02-release, 64B packets, 300000 routes,. BIOS: Non Allocating Write Enabled; Cores Active: 1C1T, Throughput: 10.314Gbps, Memory Bandwidth: 2.52GB/s, L2 miss latency: 107ns
Test by Intel as of 08/13/24.

GNR-D with Packets in L3 Cache due to DDIO: 1-node, 1x GENUINE INTEL(R) XEON(R), 42 cores, HT On, Turbo On, NUMA 1, Integrated Accelerators Available [used]: DLB 1 [0], DSA 1 [0], IAA 0 [0], QAT 0 [0], Total Memory 128GB (4x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic. Software: VPP 23.02-release, 64B packets, 300000 routes,. BIOS: Allocating Write Enabled; Cores Active: 1C1T, Throughput: 12.63Gbps, Memory Bandwidth: 0.109GB/s, L2 miss latency: 33ns
Test by Intel as of 08/13/24.

Claim: For VPP workload on GNR-D Placing packets into L3 cache with DDIO results in up to 23% better performance compared to placing packets in memory in performance with 1C1T and 64B

APPENDIX - Configuration and System Details for Media-Related Performance Claims

GNR-D: 1-node, 1x GENUINE INTEL(R) XEON(R), 32 cores, HT On, Turbo On, NUMA 1, Integrated Accelerators Available [used]: DLB 1 [0], DSA 1 [0], IAA 0 [0], QAT 0 [0], Total Memory 128GB (4x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS KVLDCRB1.IPC.0020.D62.2402220845, microcode 0x80000873, 1x I210 Gigabit Network Connection, 1x 223.6G KINGSTON SA400S37240G, 1x 1.8T INTEL SSDPEDME020T4F, CentOS Stream 9, 6.6.0-gnr.bkc.6.6.15.6.20.x86_64, ffmpeg 6.1.1, GNR-D_MediaIP Driver Package Rel69, score=832FPS.
Test by Intel as of 05/15/24.

APPENDIX - Configuration and System Details for IO-Related Performance Claims (MultiTunnel IPsec)

ICX-D: 1-node, 1x Intel(R) Xeon(R) D-2899NT CPU @ 2.20GHz, 22 cores, HT On, Turbo On, NUMA 1, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAA 0 [0], QAT 0 [0], Total Memory 64GB (4x16GB DDR4 3200 MT/s [3200 MT/s]), BIOS IDVLCRB1.86B.0021.D41.2112031014, microcode 0x1000150, 1xE810 2CQDA2, 1x 223.6G INTEL SSDSC2KB240G8, 1x 240M UDisk, Ubuntu 22.04 LTS, 5.15.0-27-generic. Software: VPP IPSec 24.02-release, 32000 SAs, 512B packets Score=55Gb/s.
Test by Intel as of 07/16/24.

GNR-D: 1-node, 1x Genuine Intel(R) CPU $0000%@, 42 cores, HT On, Turbo On, NUMA 1, Integrated Accelerators Available [used]: DLB 1 [0], DSA 1 [0], IAA 0 [0], QAT 0 [0], Total Memory 128GB (4x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 2xE810 2CQDA2, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic, VPP IPSec 24.02-release, 32000 SAs, 512B packets Score=180Gb/s.
Test by Intel as of 04/22/24.

APPENDIX - Configuration and System Details for IO-Related Performance Claims (NGFW)

EDL-D: Test by COMPANY as of 08/22/23. 1-node, 1x Intel(R) Xeon(R) D-2899NT CPU @ 2.20GHz, 22 cores, HT On, Turbo Off, NUMA 1, Total Memory 128GB (4x32GB DDR4 3200 MT/s [3200 MT/s]), BIOS IDVLCRB1.86B.0027.D11.2306160450, microcode 0x1000260, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 223.6G KINGSTON SUV400S37240G, 1x 240M Disk, Ubuntu 22.04.3 LTS, 5.15.0-27-generic, GCC 11.3, NGFW 24.01-RC2, score =35Gb/s.

GNR-D: Test by COMPANY as of 12/18/23. 1-node, 1x Genuine Intel(R) CPU $0000%@, 42 cores, HT On, Turbo On, NUMA 1, Total Memory 128GB (4x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS KVLDCRB1.SYS.0017.D23.2310101159, microcode 0xf0970506, 2x Intel Corporation, 1x I210 Gigabit Network Connection, 1x Ethernet interface, 1x 240M Disk, 1x 240M UDisk, 1x 476.9G SAMSUNG MZVL2512HCJQ-00A00, Ubuntu 22.04 LTS, 5.15.0-27-generic, GCC 11.3, NGFW 24.01-RC2, score =96Gb/s.