intel XEON Accelerate with Xeon

High Performance Computing & AI with Intel & Google Cloud

Chris Feltham

Intel-Google Alliance lead

EMEA South, UK&I

chris.feltham@intel.com

Julian Fischer

Intel-Google Alliance lead

EMEA North, META

julian.fischer@intel.com

intel

- Intel-Google Cloud Alliance
- Intel HPC & AI Portfolio

Intel-Google Cloud Alliance

Continual innovation, technology development, and customer success.

2016: Google Cloud CEO Diane Green announces strategic partnership with Intel

2016: 1st CSP enabled on Intel Xeon Scalable Processors

2017: TensorFlow software optimizations on Intel architecture

2017: GCE, GKE, Dataproc, Zync enabled on Intel Xeon Scalable Processors

2018: Google Cloud CEO Thomas Kurian announces Anthos partnership with Intel

2018: GCP instances for AI/ML enabled using Intel DL Boost

2018: Google Cloud awards Intel "Innovative Solution in Infrastructure"

2019: Anthos reference design enabled on Intel Select Solutions

2019: GCP on Intel helps Kinsta run Wordpress 200% faster

2019: C2, N2, M2, O2 enabled on 2nd Gen Intel Xeon Scalable Processors

2020: VMware Engine enabled on 2nd Gen Intel Xeon Scalable

2020: GCP N2 instances help goto reduce operating costs by 90%

2020: vRAN reference design enabled on 2nd Gen Intel Xeon Scalable

2021: HPC Simulation & Modelling enabled on Intel Select Solutions

2021: New N2 instances enabled on 2nd Gen Intel Xeon Scalable

2021: Google Cloud GM & VP NW Shailesh Shukia announces Telco partnership with Intel

2022: Intel OneDNN integrated into TensorFlow 2.9

2022: New M3 instances enabled on 3rd Gen Intel Xeon Scalable

2022: Google Cloud & Intel wins at StubHub and Asahi

2023: 1st CSP to market with 4th Gen Intel Xeon Scalable (Sapphire Rapids)

2023: 1st CSP to market with Intel-Google co-designed IPU

intel XEON Accelerate with Xeon

Intel-Google Cloud Alliance

A broad range of Intel-based compute instances targeting critical workload & solution categories.

Infrastructure Modernization	Application Modernization
Compute **Workload optimized solutions:** • General Purpose (N1, N2, **C3**) • Compute Optimized (C2) • Memory Optimized (M1, M2, M3) • AI DL & HPC Optimized	**Telco & Edge Solutions** **Telco strategic partnership to:** • accelerate 5G/LTE and Edge • drive monetization for telcos and the ecosystem **Google Distributed Cloud** • enabled by Anthos, a portfolio of HW and SW solutions extending Cloud to customer DC and Edge **Co-developed reference designs** • vRAN reference design for FlexRAN • Anthos Ready bare metal reference designs with Intel, to address vertical use cases
Enterprise **VMware, SAP & Oracle solutions optimized and certified on Intel:** • Google Cloud VMware Engine (VE1) • SAP-certified (N1, N2, M1, M2, M3, O2) • Bare Metal for Oracle (O2)	

Intel-Google Cloud Alliance

Accelerating our customers' digital transformation.

Technical

Technology previews and early access

Co-design of silicon powering new customer services (Intel IPU)

Code optimisation services: "Software Center of Excellence"

Commercial

Compute trial funding programs for Intel based Google instances – GCE, GKE, Dataproc, GCVE...

Cost optimisation solutions including real-time workload optimisation (Granulate) and instance analysis & recommendation (Densify)

Migration support programs

Software CoE

Engineering-led effort to optimise software performance and price/performance on Intel-based Google Cloud instances.

Discovery

▼

Performance Review

▼

Performance Report

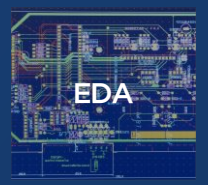
Successful engagements include:

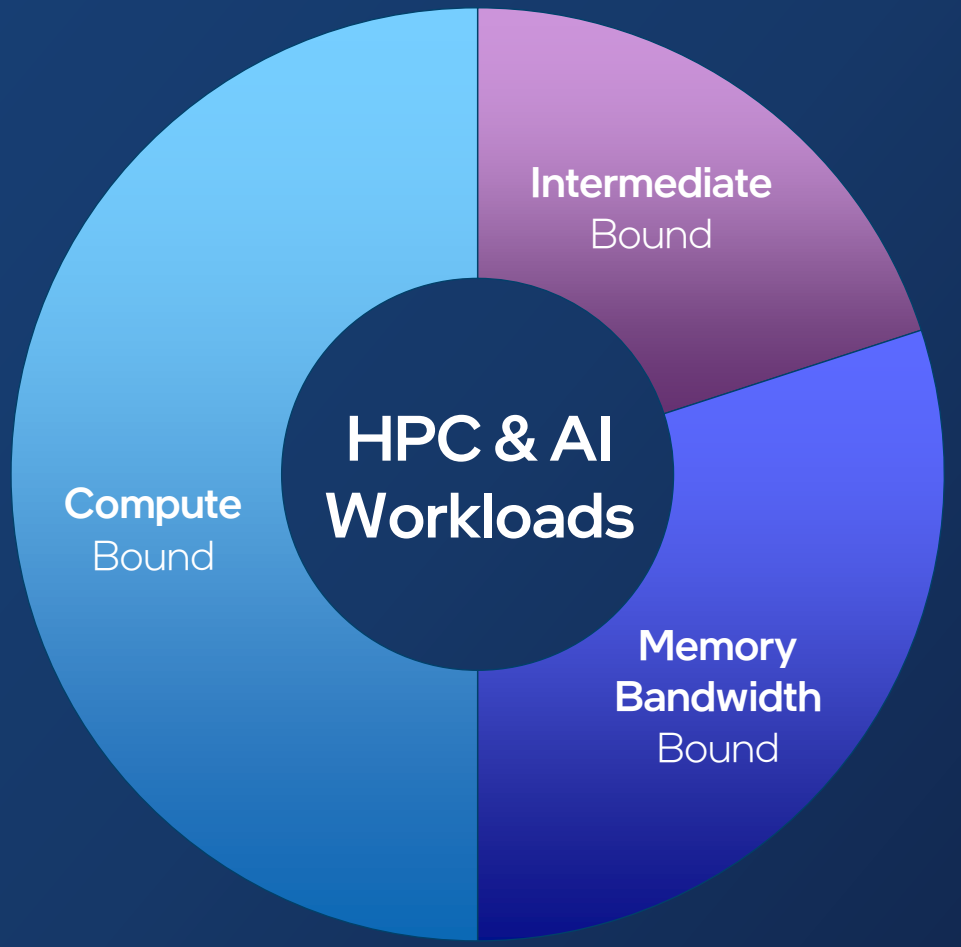
Major credit bureau
3X performance improvement
2X reduction in latency

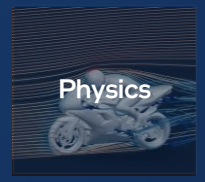
Video streaming platform
55% average improved performance
up to 7X speed up

E-Commerce solution provider
35% increase throughput
32% improved performance per $

Different Workloads Face Different Bottlenecks

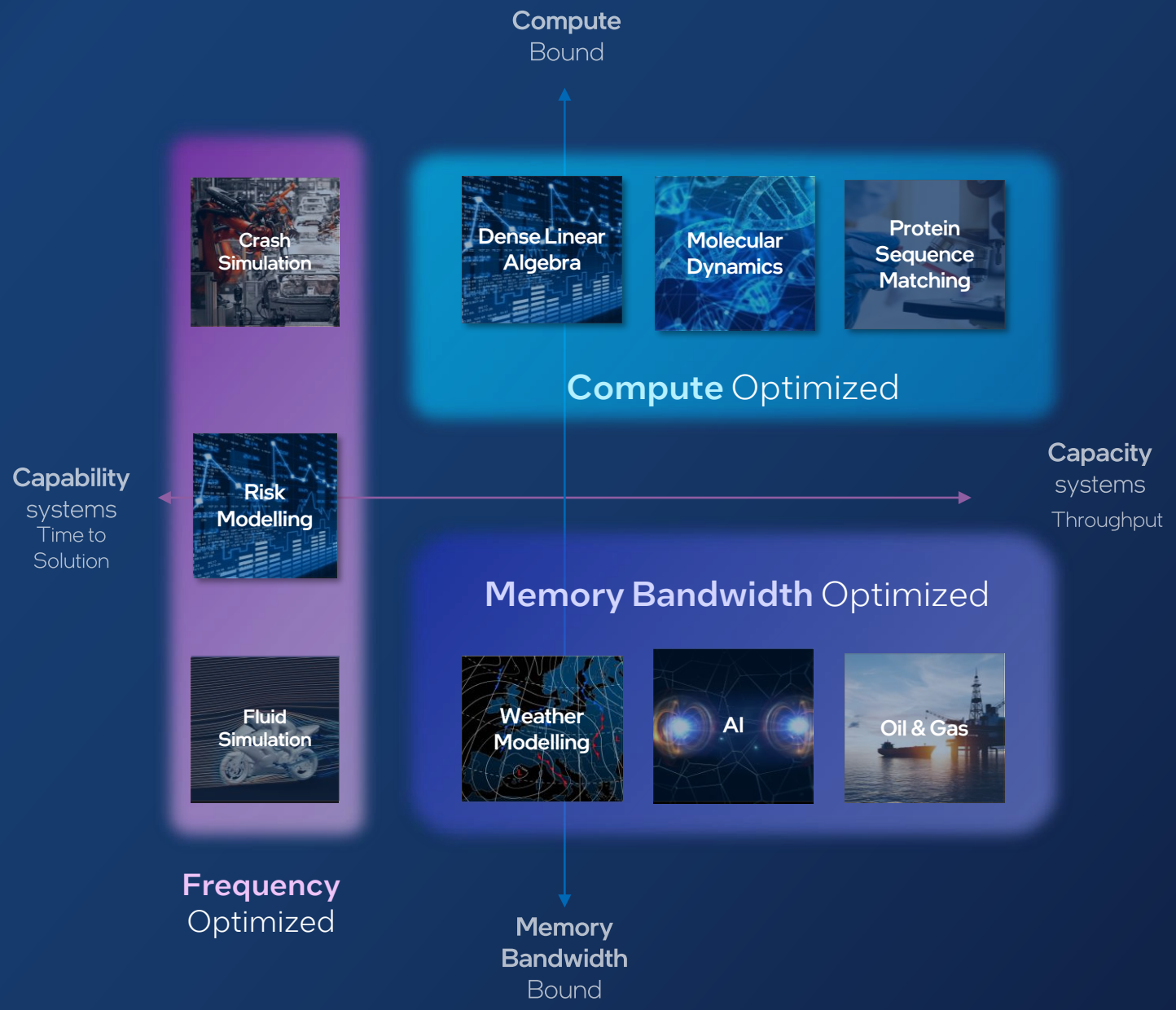






Optimizing for All HPC & AI Workloads

Compute Explosion
Emergence of AI
Need for High Bandwidth Memory
Density, Scalability & Sustainability



intel XEON Accelerate with Xeon

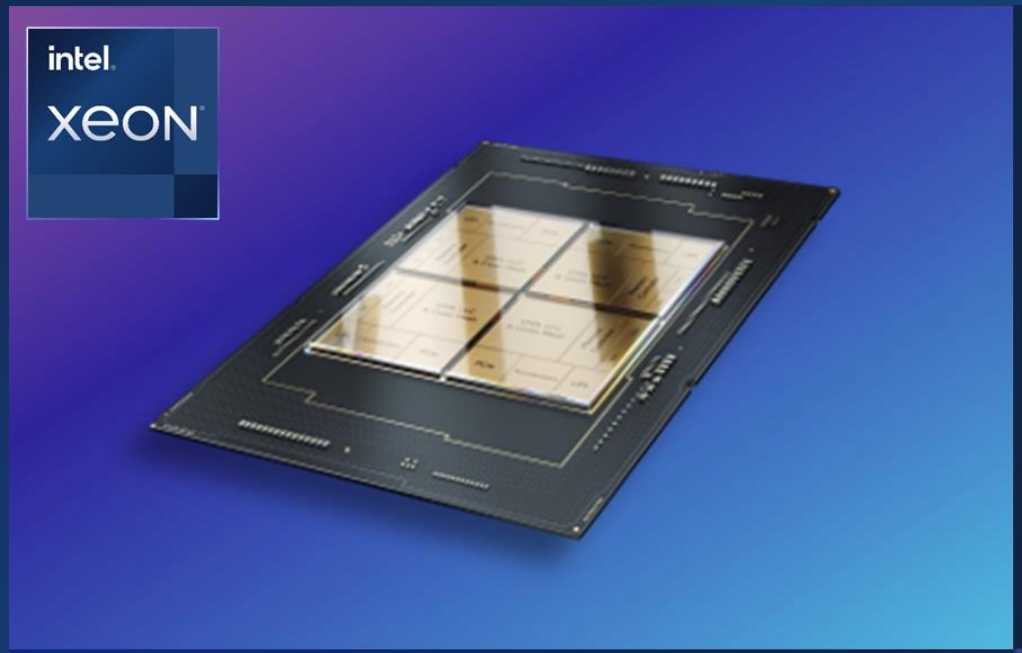
Broadest HPC Portfolio with Open Software Standards

Scalable Compute

Maximum Memory Bandwidth

High Compute Density

4th Gen Intel® Xeon® Scalable Processors for HPC Accelerated Performance



- Built-in acceleration to boost application performance – Intel AVX 512, Intel DSA, Intel AMX
- New µarch, built on Intel 7, with up to 60 performance cores for added compute
- Increased memory bandwidth and higher speeds with DDR5
- Higher IO bandwidth with PCIe 5. and support for coherent interface with CXL 1.1

Intel Advanced Vector Extensions 512 (AVX-512)

Intel Advanced Matrix Extensions (AMX)

Intel Data Streaming Accelerator (DSA)

Google Cloud is currently the only hyperscale cloud service provider with general availability of instances ("C3") based on 4th Gen Intel Xeon Scalable Processor

intel XEON Accelerate with Xeon

https://cloud.google.com/compute/docs/general-purpose-machines#c3_machine_types

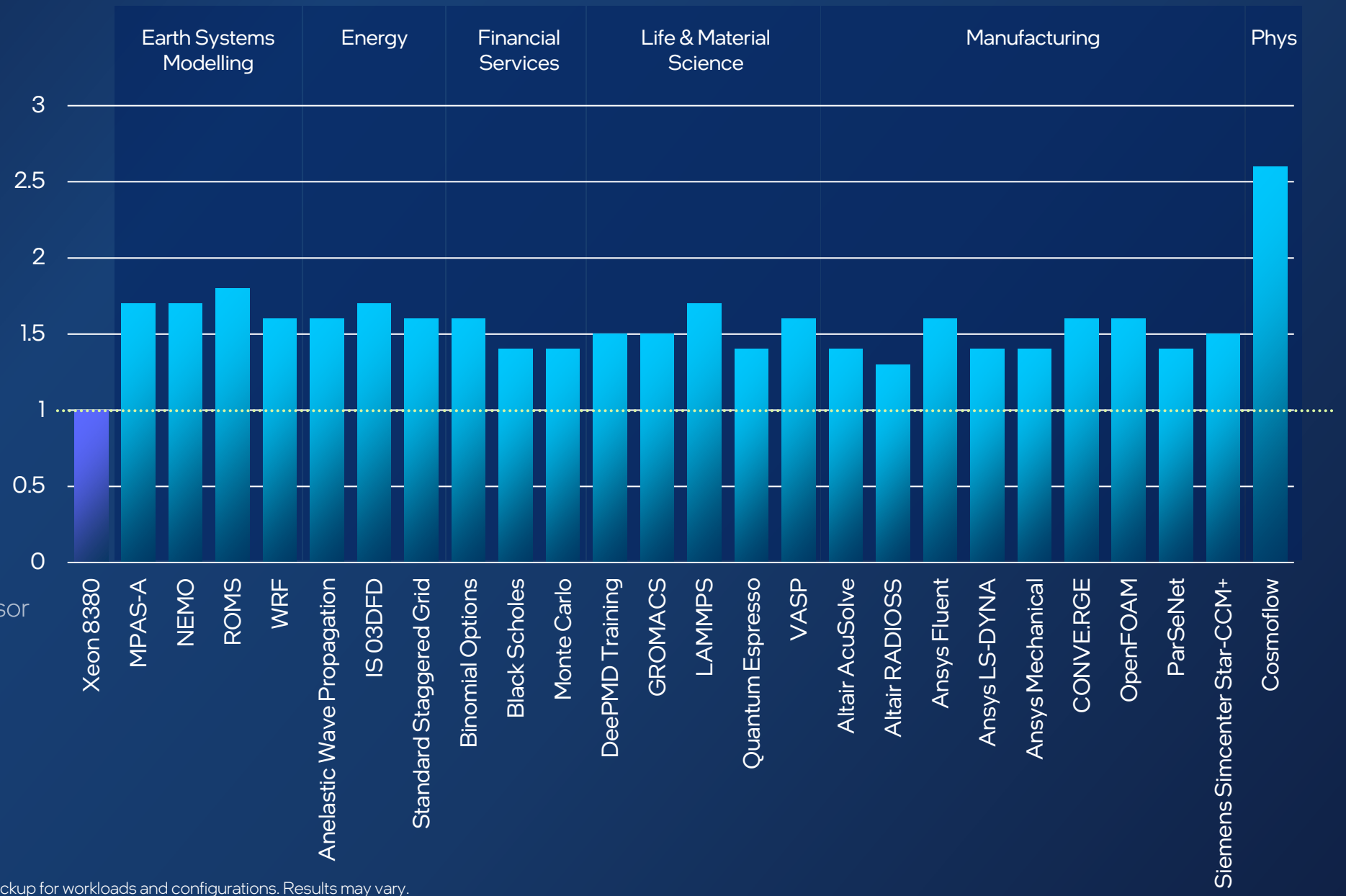
intel XEON

Up to 2.6x Performance On Real Workloads

2S 4th Gen Intel® Xeon® processor vs.
2S 3rd Gen Intel® Xeon® 8380 processor

Relative performance. Higher is better

Intel Xeon 8380 Baseline



intel XEON Accelerate with Xeon

See backup for workloads and configurations. Results may vary.

"At Cadence, we're excited that initial testing indicates a 15% reduction in simulation runtime per core for Clarity workloads running on C3 vs. C2. These significantly faster runtimes on Google's Cloud Platform will help to increase engineering productivity for our joint customers"

-- Ben Gu, Vice President of R&D, Cadence

"At Palo Alto Networks, we develop and deploy deep learning models for inline threat detection in our customers' network traffic. Inference latency is critical for our AI workloads. By adopting C3 VMs with Intel Sapphire Rapids and the new AMX instruction set for AI, we are seeing 2x performance for some of our inline models, compared to the previous generation N2 Ice Lake VMs"

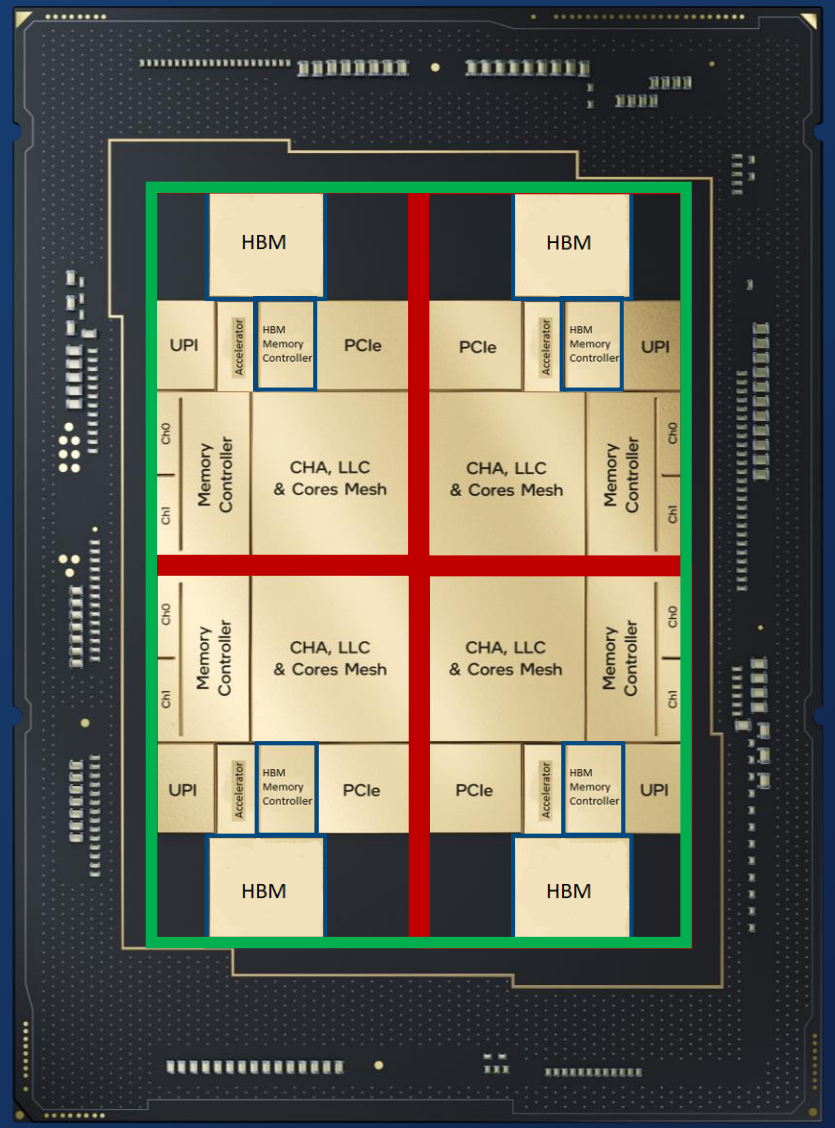
-- Suiqiang Deng, Senior Distinguished Engineer, Palo Alto Networks

https://cloud.google.com/blog/products/compute/c3-machine-series-on-intel-sapphire-rapids-now-ga

Clustering Support

Optimizing Latency and Bandwidth

Sub-NUMA Clustering (SNC4)	Each Compute Tile a NUMA domain with associated Local Memory
UMA Clustering (Quadrant)	CHA and MC Affinity CHA and Cores no Affinity Socket is a single NUMA domain



"C3 VM shapes are optimized for the underlying NUMA architecture to deliver consistent performance" - Google

Intel Advanced Matrix Extensions (AMX)

Intel AVX-512

85 x int8 ops/cycle/core with 2 x FMA
vpmaddubsw → vpmaddwd → vpaddd

Intel AVX-512 VNNI

256 x int8 ops/cycle/core with 2 x FMA
vpdpbusd

Intel AVX-512

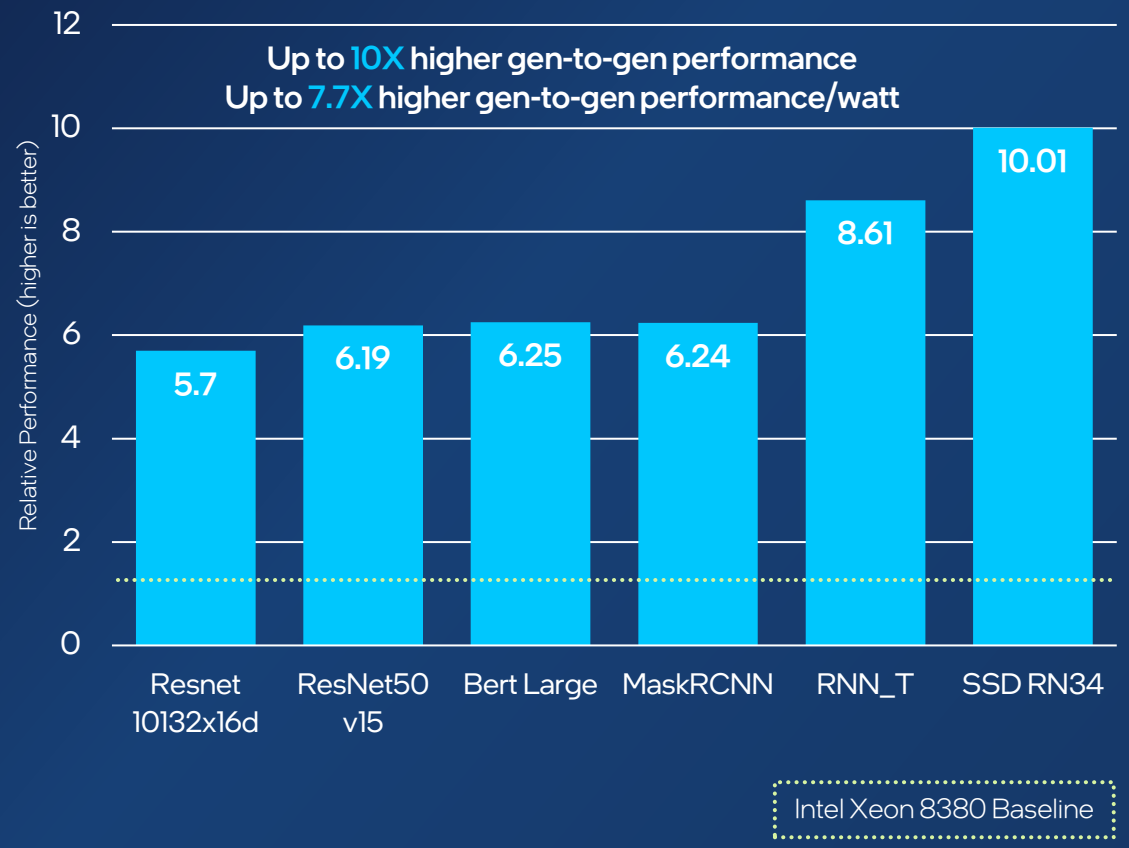
2048 x int8 ops/cycle/core
Multi-fold MACs in one instruction
tdpbusd

Intel AMX is fully supported in Google Cloud C3 instances, and open source frameworks including TensorFlow and PyTorch

DL Inference

Real-Time Inference Performance

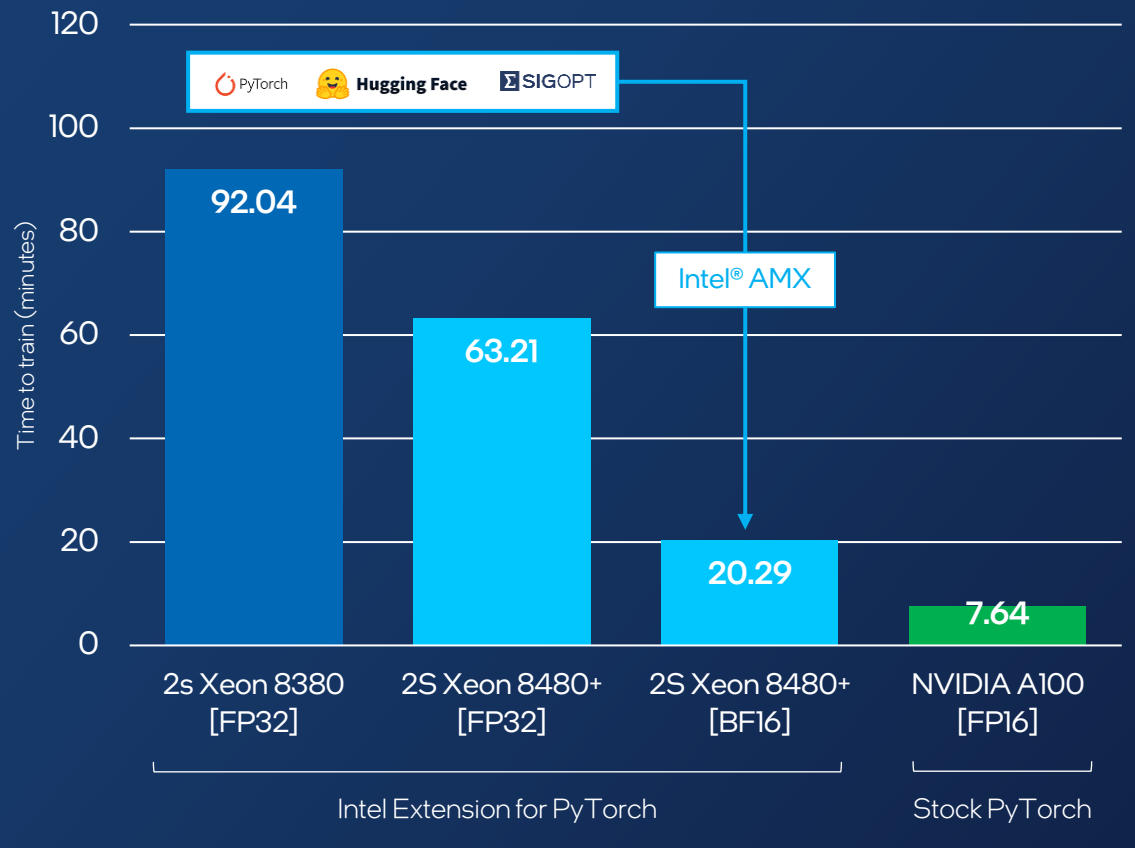
2S Intel Xeon 8480+ [AMX BF16] vs. 2X Intel Xeon 8380 [FP32]



E2E Inference Pipeline

DLSA HuggingFace Bert-large [IMDB]

Fine tuning time-to-train 2S Xeon vs. NVIDIA A100



intel XEON Accelerate with Xeon

See backup for workloads and configurations. Results may vary.

Intel® Xeon® CPU Max Series

oneAPI

Scalable Compute

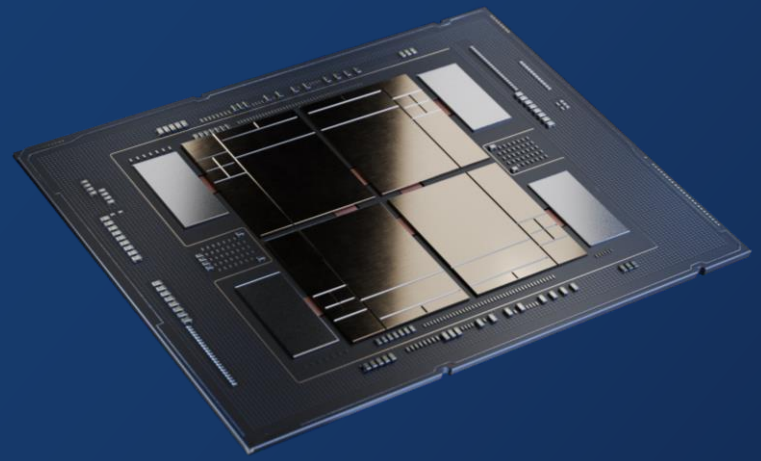
Maximum Memory Bandwidth

High Compute Density

intel XEON Accelerate with Xeon

intel XEON MAX SERIES

The only x86 CPU with High Bandwidth Memory (HBM)



HBM

64GB HBM2e

up to **112.5MB** shared LLC

DDR5
8ch/CPU @ 4800 MTS (1DPC)
16 DIMMs/CPU

~1TB/s memory bandwidth

1GB/core HBM capacity

HBM Only Mode

Workloads <64GB capacity

HBM

No code change.
No DDR.

System boots and operates with HBM only

HBM Flat Mode

Flat memory regions with HBM and DDR

HBM
DDR

Code change may be needed to optimize performance

Flexibility for apps that require large memory capacity

HBM Caching Mode

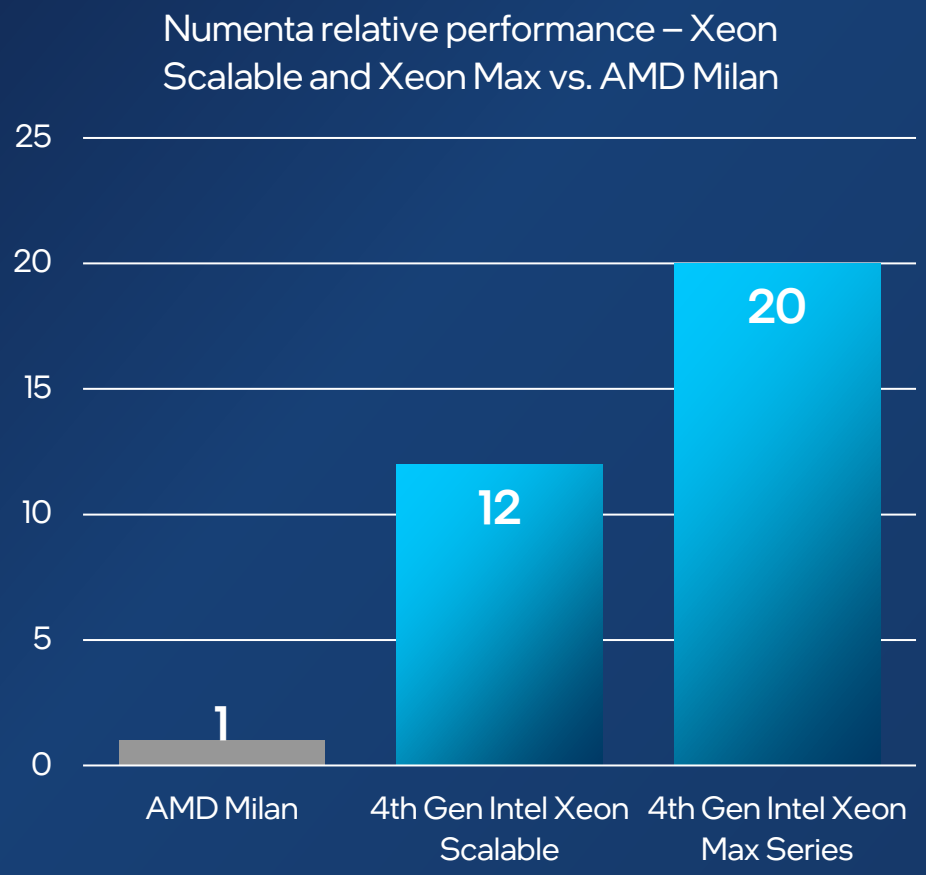
DRAM backed cache

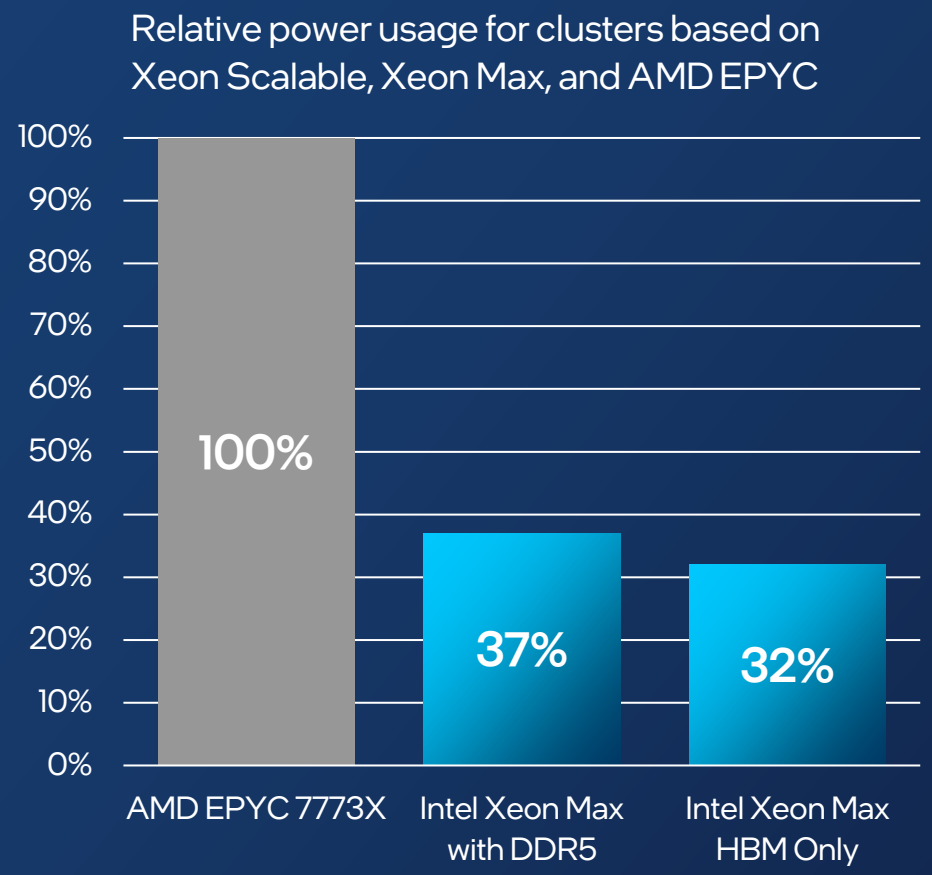
HBM DDR

No code change.
HBM caches DDR.

Whole apps may fit in HBM cache. Blurs line between cache and memory.

intel XEON Accelerate with Xeon





Up to 20X NLP speed up compared to AMD Milan, using Numenta

68% lower power usage than a Milan-X cluster for the same HPCG performance

Intel® Data Center GPU Max Series

oneAPI

Scalable Compute

Maximum Memory Bandwidth

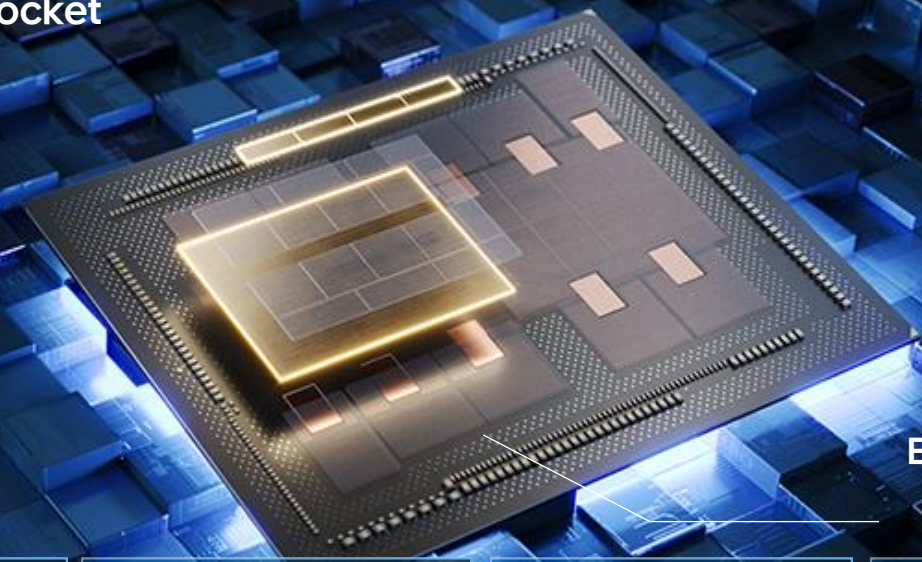
High Compute Density

intel DATA CENTER GPU MAX SERIES

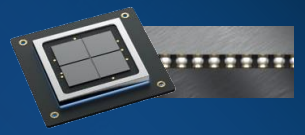
Intel® Data Center GPU Max Series

Highest
Compute Density in a Socket

Rambo Cache
(Random Access Memory, Bandwidth Optimized)

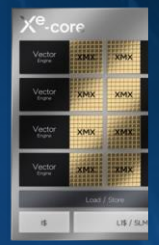


Base Tile

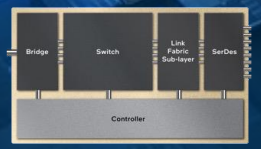


Constructed with **EMIB and Foveros**

Up to **128** **Xᵉ HPC Cores**



52TF Peak FP64 Throughput

16 

Xᵉ Links for GPU-to-GPU communication

Up to **128GB** HBM2e

Up to **408MB** Rambo L2 Cache

intel XEON Accelerate with Xeon

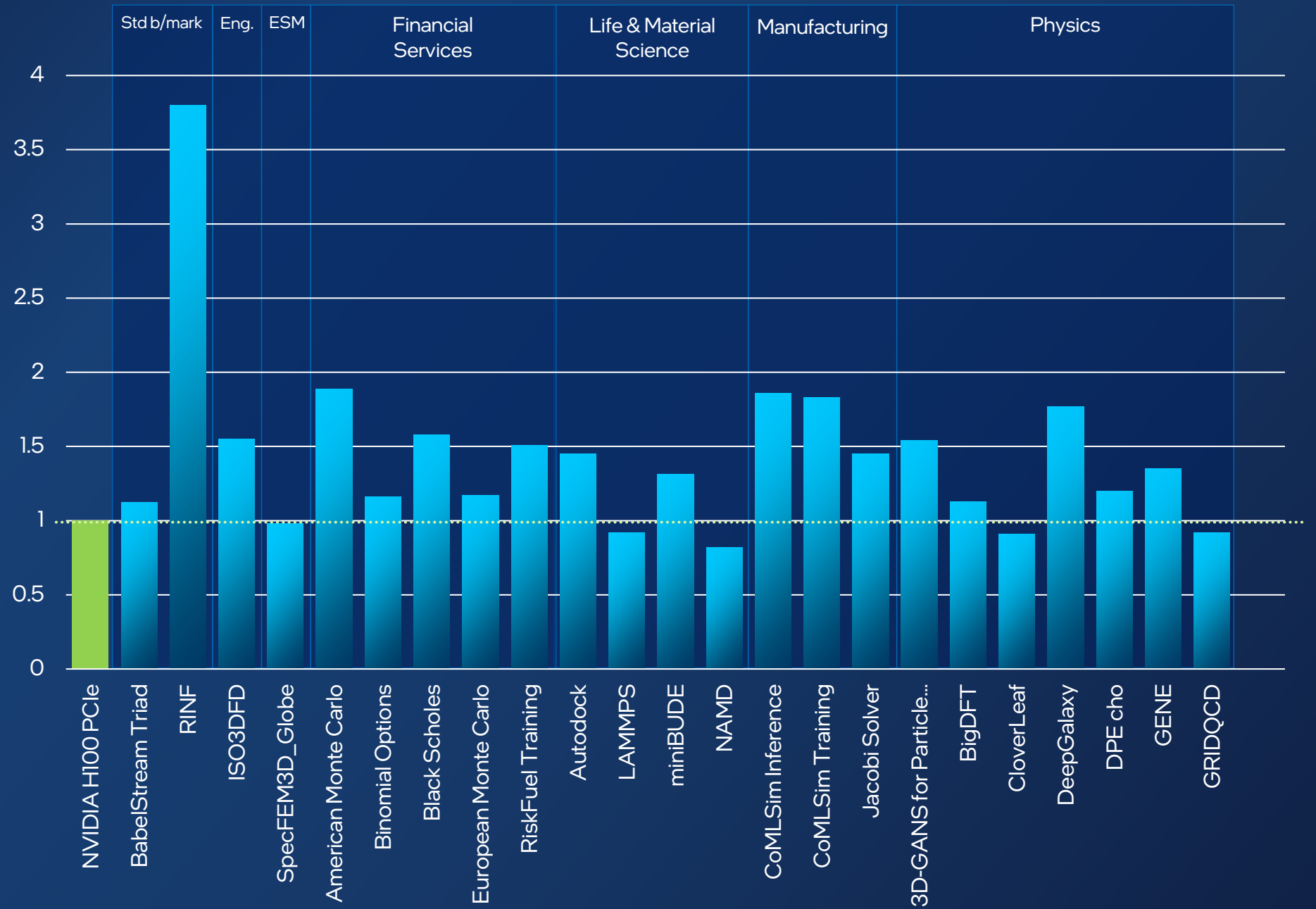
intel DATA CENTER GPU MAX SERIES

Average of 1.3X Performance Advantage

Intel Data Center GPU Max 1550 vs. NVIDIA H100 PCIe

Relative performance. Higher is better

NVIDIA H100 baseline



intel XEON Accelerate with Xeon

oneAPI

Scalable Compute

Maximum Memory Bandwidth

High Compute Density

Open & Full Stack Solution

.....accelerate multiarchitecture, multivendor programming

Workload Apps

Middleware & Frameworks

oneAPI

Virtualization

Operating Systems

Level Zero

intel XEON Accelerate with Xeon

oneAPI

New 2023 Release

Open, Standards-Based unified programming model

- Optimized to deliver multi-architectural performance
- Extends support for SYCL and latest Fortran standards
- Compiler and tools support for OpenMP 5.1
- Optimizations for TensorFlow & PyTorch
- Enhanced CUDA-to-SYCL code migration capabilities

intel XEON Accelerate with Xeon

Storage

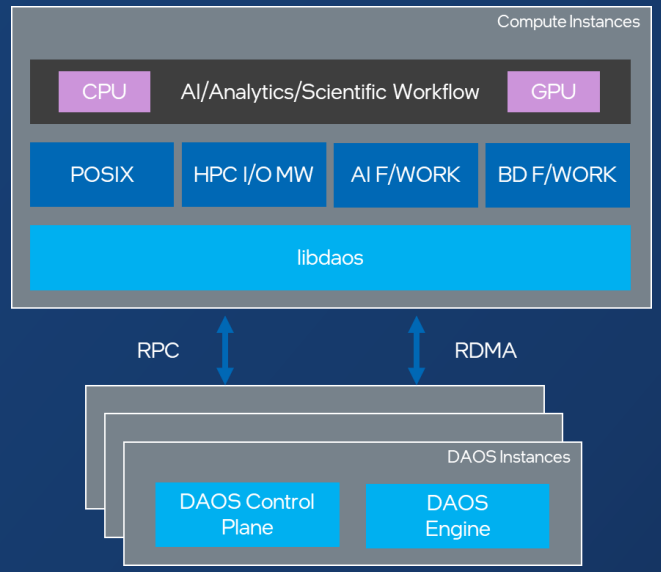
Distributed Asynchronous Object Storage (DAOS)

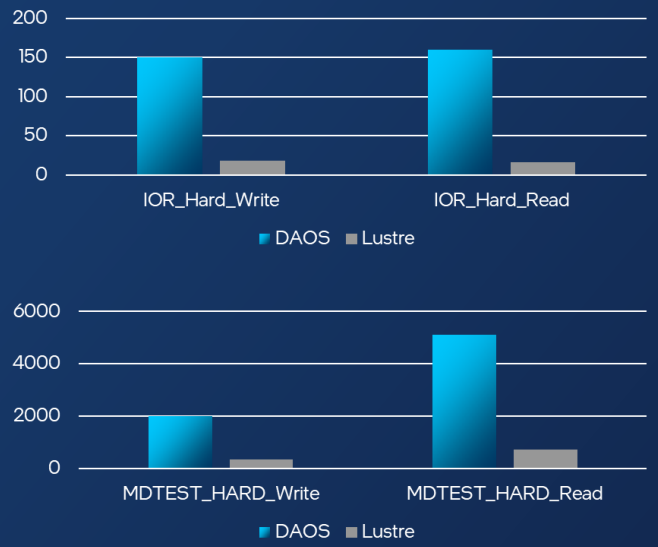
Intel Infrastructure Processing Unit (IPU), and Google Hyperdisk

Extreme storage performance

Intel Distributed Asynchronous Object Storage (DAOS) now available in Google Cloud.

Architected from ground up for NVM technologies – the foundation of the Intel Exascale storage stack.





Source: University of Cambridge

96 GiB/s read

60 GiB/s write

0.28ms random read IO latency

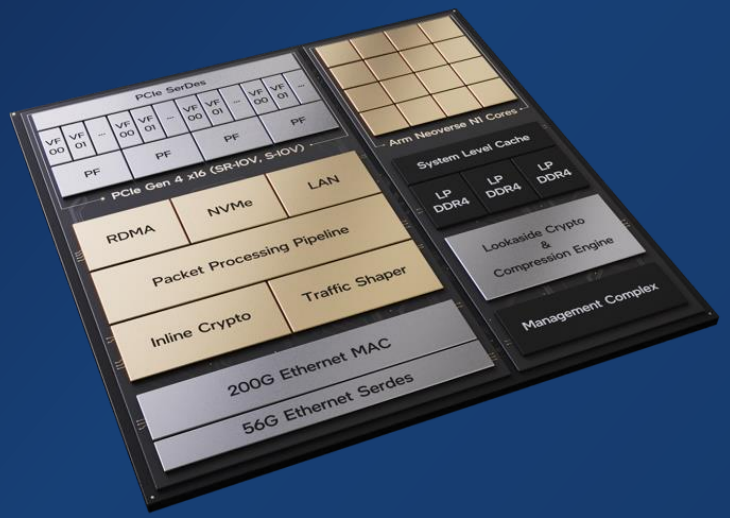
0.36ms random write IO latency

intel XEON Accelerate with Xeon

https://www.intel.com/content/www/us/en/high-performance-computing/resources/daos-google-cloud-performant-hpc-white-paper.html

Intel IPU

Intel-Google co-designed Infrastructure Processing Unit (IPU) ES2000, codename Mount Evans launched with C3 instances.



"We are pleased to have co-designed the first ASIC Infrastructure Processing Unit with Google Cloud, which has now launched in the new C3 machine series. A first of its kind in any public cloud, C3 VMs will run workloads on 4th Gen Intel Xeon Scalable processors while they free up programmable packet processing to the IPUs securely at line rates of 200Gb/s. This Intel and Google collaboration enables customers through infrastructure that is more secure, flexible, and performant"

-- Nick McKeown, SVP, Intel Fellow and GM Intel Network and Edge Group

---

Intel IPU ES2000 powers Google Hyperdisk – the next generation of network block storage

80% higher IOPS per vCPU[1]	Decouples storage performance from compute instance size	Hyperdisk Extreme: For performance critical applications including high end DB
	Scale performance and capacity independently from each other	Hyperdisk Throughput: For scale out workloads including Kafka and Hadoop

intel XEON Accelerate with Xeon

1. https://cloud.google.com/blog/products/compute/introducing-c3-machines-with-googles-custom-intel-ipu

Summary

Different HPC & AI workloads have different requirements – we're tackling them all with a broad and open portfolio of compute and storage.

Google Cloud is the only hyperscaler with 4th Generation Intel Xeon Scalable Processor in general availability.

Take advantage of Intel HPC and software expertise in region, as well as trial programs to evaluate compute options in Google Cloud.

intel XEON Accelerate with Xeon

Backup: performance claim configurations