

Freedom, Productivity, and Performance for Accelerated Computing

Next-Gen HPC + AI

Intel's Open, efficient, performant portfolio

Burnie Legette

AI/DataOps Solutions Architect

Intel Government Technologies



Data Center Requirements Are Evolving

Varying uses require unique optimization vectors

AI Everywhere

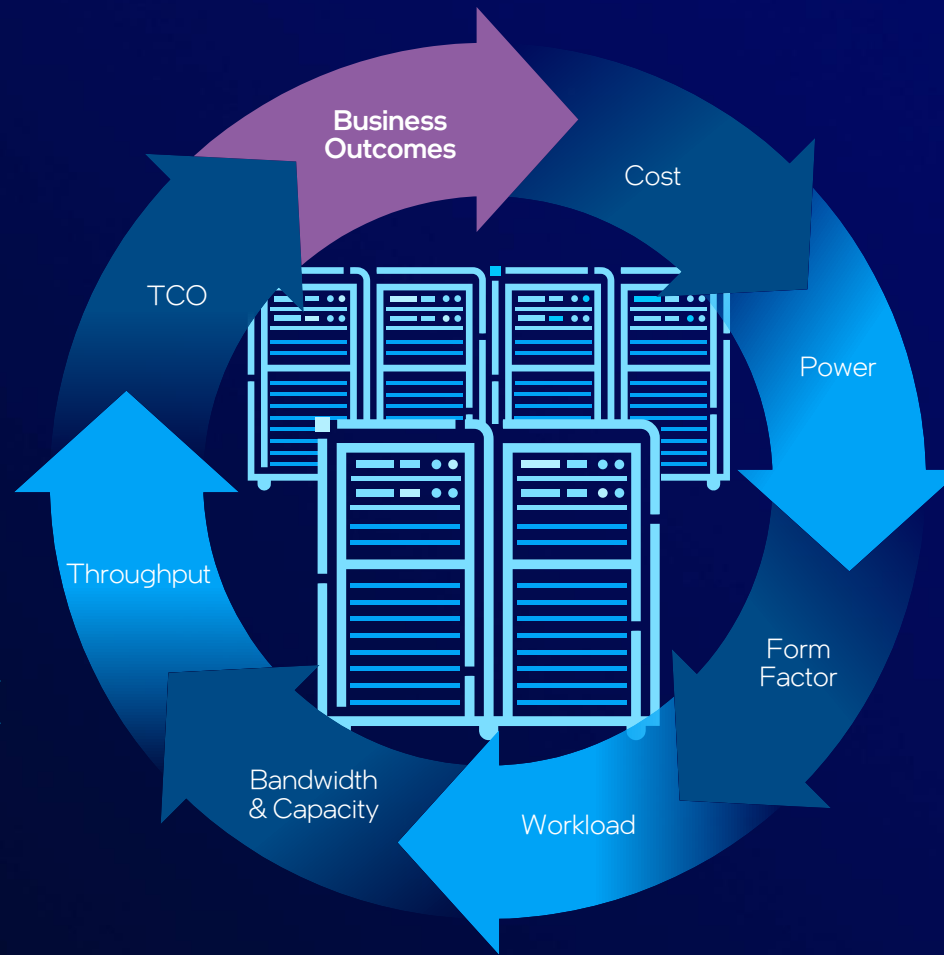
Enable performance at scale with accelerated hardware and open standards-based software

Throughput & Latency

Maximize system-level performance with best-in-class response time

Security, Quality & Reliability

Provide security, quality, and reliability for at-scale deployments



Efficiency & TCO

Increase rack density while meeting power efficiency requirements to improve TCO (total cost of ownership)

Sustainability

Minimizing carbon emissions through improved energy efficiency and circular product design

Software Compatibility

ISA consistency for software ecosystem compatibility

Intel Xeon 6 Processors for HPC+AI

World's Best
CPU for AI

The Most
Deployed Host CPU

Up to 128 P-cores

on 6900-series
up to 86 P-cores on 6700/6500-series

More bandwidth & cache

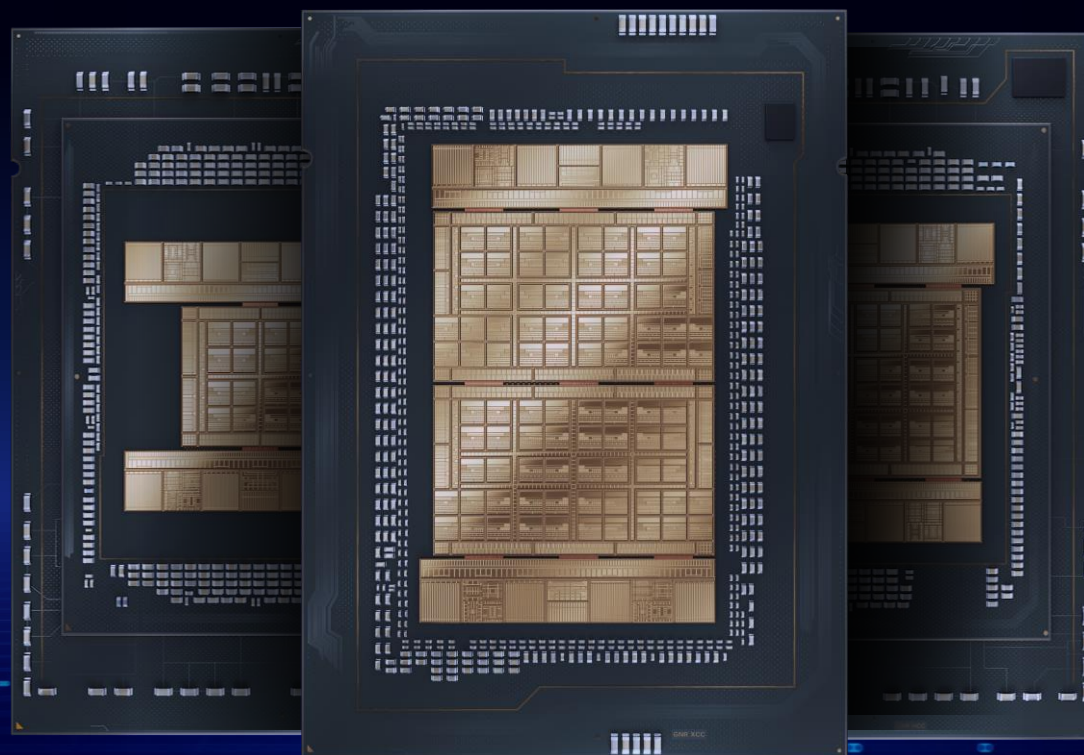
MRDIMM memory support
Up to 504MB low latency LLC

AI accelerators built-in

Intel® AMX, Intel® AVX-512,
and Intel® AVX-2

Comprehensive SW suite

AI development across classical ML and
small GenAI models



Superior I/O performance

up to 192 PCIe 5.0 lanes

High Single Threaded Performance

With Intel's latest generation P-core

Top Tier Memory Support

30% higher memory B/W with MRDIMMs
Expandability with CXL 2.0

Ready for Deployment

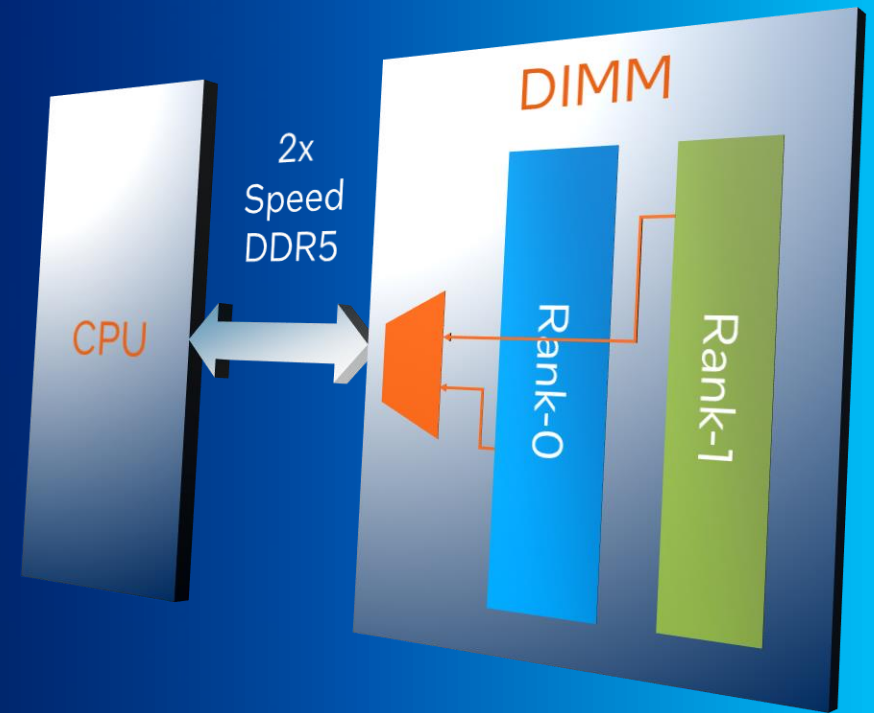
DC-MHS & NVIDIA MGX™
form factors supported

Multiplexed Rank DIMMs

First to market on Intel Xeon 6 processors with P-core

- MRDIMMs offer significant improvement for memory bound workloads like AI, HPC and other key workloads
- Improving latency and addressing TCO needs

MRDIMMs deliver up to **8800 MT/s** data rate on Intel Xeon 6 with P-cores



Developer Tools for Intel® Xeon® 6 Processors

Maximize Performance for AI & Accelerated Compute Workloads on Intel Xeon 6 Processors with P-cores

Accelerate AI Frameworks & Applications

Accelerate Generative AI/LLM, and other deep learning, data science pipelines using the [Intel® oneAPI Base Toolkit](#) & [Intel AI Tools](#).

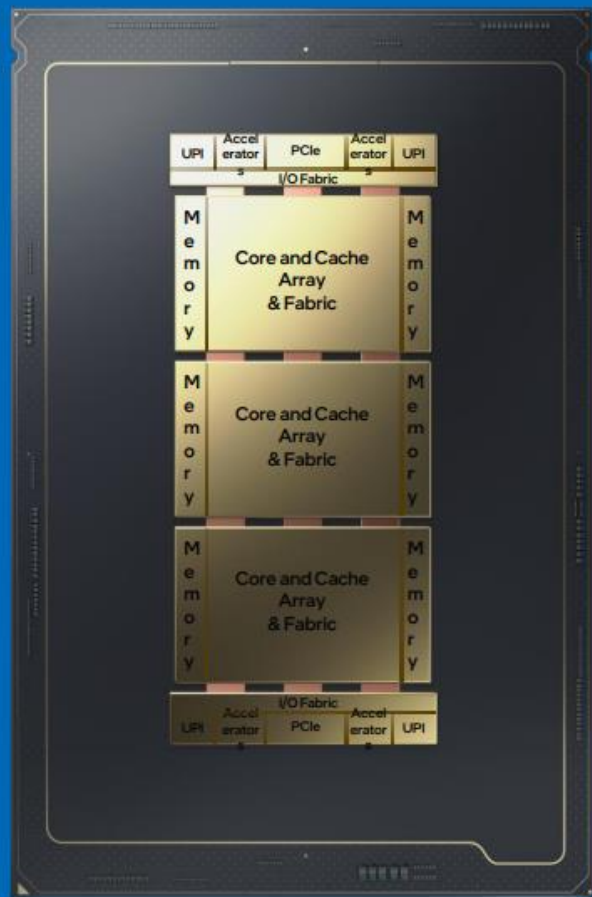
- Intel oneAPI library optimizations are regularly up streamed to the latest versions of [PyTorch](#),* [TensorFlow](#),* and other leading [deep learning frameworks](#), enabling developers to achieve orders of magnitude performance improvements on Intel hardware using their existing AI workflows.
- [Intel® oneAPI Deep Neural Network Library \(oneDNN\)](#) accelerates deep learning and generative AI models on Intel® Xeon 6 Processors with P-Cores, the first Intel CPU platform supporting AI acceleration with Intel® Advanced Matrix Extensions through FP16 and complex FP16 instructions (building on already existing int8 & BF16 support).
 - up to **3x better Llama2** performance vs. prior generation for large-language-model (LLM)¹
 - up to **1.86x gen-to-gen** performance improvement in AI inferencing²



Accelerate AI & General Compute Workloads

Build, analyze, optimize, and scale applications with the latest techniques in vectorization, multithreading, multi-node parallelization, and memory, using the [Intel® oneAPI Base Toolkit](#), [Intel® Distribution for Python](#), [Intel® oneAPI HPC Toolkit](#).

- Accelerate math functions across multiple domains such as BLAS, LAPACK and FFT with [Intel® oneAPI Math Kernel Library \(oneMKL\)](#) performance tuning for up to **2.5x better HPCG** performance vs. prior generation with MRDIMM³
- Push your application's efficiency further with [Intel® oneAPI DPC++/C++ Compiler's](#) improved data access through preloading cache reducing latency & Intel® Advanced Matrix Extensions-FP16 instruction support leveraged by [Intel® oneAPI Deep Neural Network Library \(oneDNN\)](#).
- [Intel® Fortran Compiler](#) supports backend code generation and enriched performance tuning for latest Intel Xeon 6 processors.
- [Intel® MPI Library](#) now supports 128-core tuning and optimizations for scale out and scale up.
- [Intel® VTune™ Profiler's](#) new features such as hotspots, microarchitecture and memory access, I/O and platform diagram makes identifying performance bottlenecks and memory issues easier.
- [Intel® Threading Building Blocks \(oneTBB\)](#) is enhanced to scale parallel execution performance on Intel Xeon 6 processor's higher CPU core count to accelerate multi-threaded applications.



1) See [9A2] at [intel.com/processorclaims](https://www.intel.com/processorclaims): Intel® Xeon® 6

2) See [9A3] at [intel.com/processorclaims](https://www.intel.com/processorclaims): Intel® Xeon® 6

3) See [9H10] at [intel.com/processorclaims](https://www.intel.com/processorclaims): Intel® Xeon® 6

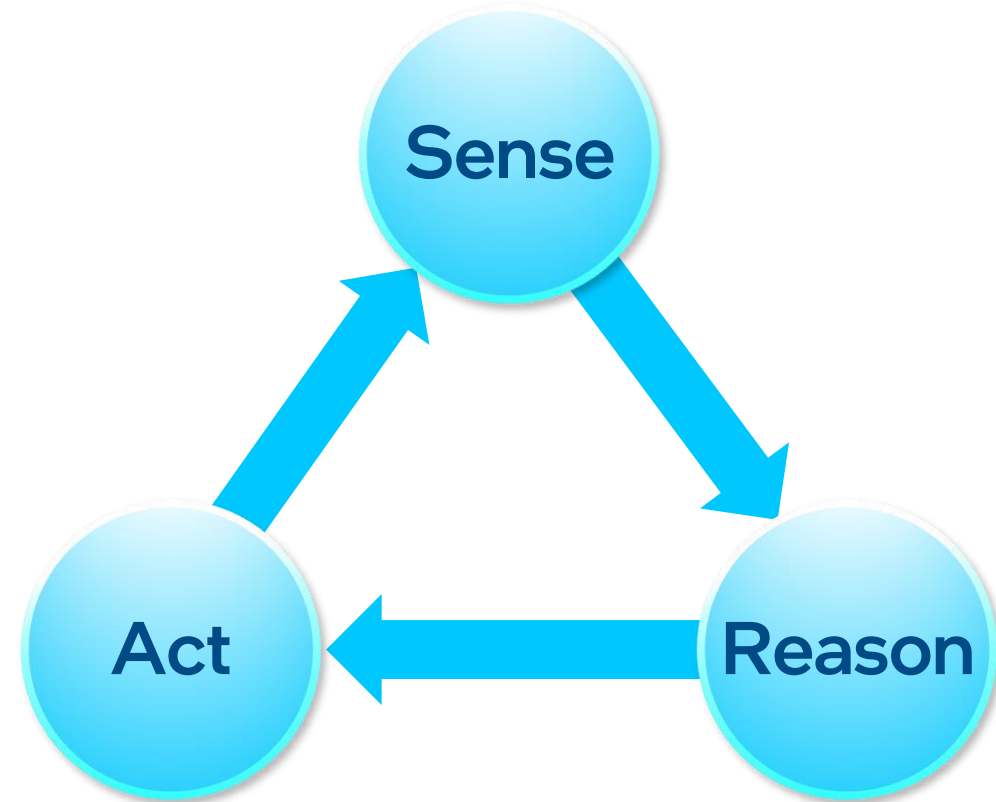
Agentic-Physical Era AI

From Perception To Reasoning/Action

- Promises:
 - Kahneman's (S2 → S1) loop
 - Learned object features → action motifs
 - Inference time scaling → reasoning
 - Human-in-the-loop → Humanoids*
- Challenges:
 - LLM models running out of training data
 - Current algorithms need revisit **
 - Physical world is much harder ***
 - Safe Superintelligence?

Mostly Transformer-LLM Based

Text/Image/... In → Text/Image/... Out



Embodiment Policy → Action

Neural → Neural + Symbolic

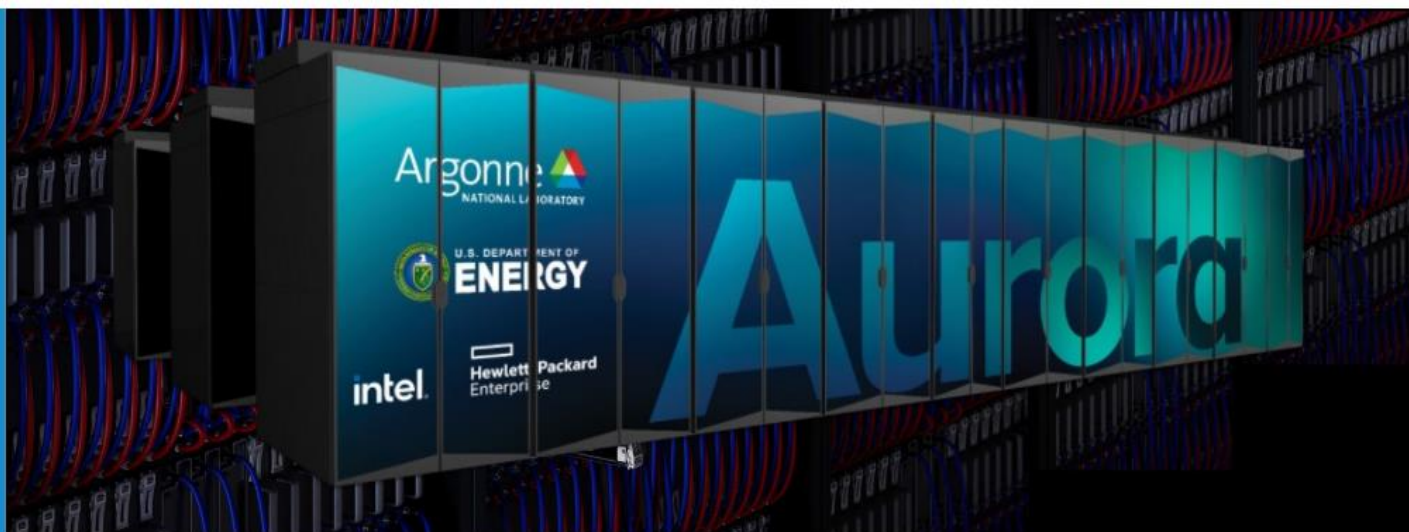
(Transformer + RL/MPC) Loop

Explainable AI

• Generally Capable Agents in Open-Ended Worlds, Jim Fan, NVIDIA GTC'24 <link>

** Objective-Driven AI, Prof. Yann Lecun <link>

*** ASAP: Aligning Simulation and Real-World Physics... <link>



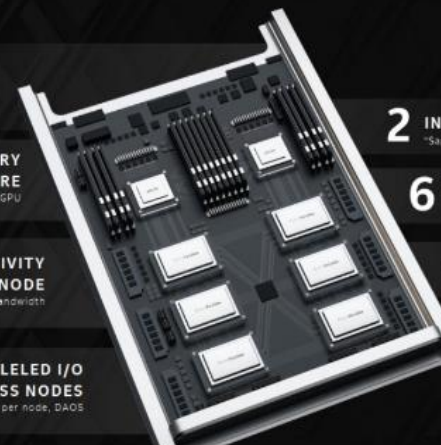
AURORA – BRINGING IT ALL TOGETHER

LEADERSHIP
PERFORMANCE
For HPC, Data Analytics, AI

UNIFIED MEMORY
ARCHITECTURE
Across CPU & GPU

ALL-TO-ALL CONNECTIVITY
WITHIN NODE
Low latency, high bandwidth

UNPARALLELED I/O
SCALABILITY ACROSS NODES
8 fabric endpoints per node, DAOS



2 INTEL XEON™ SCALABLE PROCESSORS
"Sapphire Rapids"

6 X^e ARCHITECTURE BASED GPUs
"Ponte Vecchio"

oneAPI
Unified programming model

Peak Performance
≥ 2 Exaflops DP

Intel GPU
Ponte Vecchio (PVC)

Intel Xeon Processor
**Sapphire Rapids with
High Bandwidth Memory**

Platform

Compute Node
2 Xeon SPR+HBM processors
6 Ponte Vecchio GPUs
Node Unified Memory
Architecture
8 fabric endpoints

GPU Architecture
Intel XeHPC architecture
High Bandwidth Memory Stacks

Node Performance
>130 TF

System Size
>9,000 nodes

Aggregate System Memory
>10 PB aggregate System Memory


















































System Interconnect
HPE Slingshot 11
Dragonfly topology with adaptive routing

Network Switch
25.6 Tb/s per switch (64 200 Gb/s ports)
Links with 25 GB/s per direction

Programming Environment

- C/C++
- Fortran
- SYCL/DPC++
- OpenMP offload
- Kokkos
- RAJA

Intel® HPC + AI Portfolio

Open Software Environment													
Deep Learning Acceleration	 Dedicated Deep Learning Training and Inference												
General Acceleration	 AI Inference, VDI, Media Analytics, Real-Time Dense Video  Parallel Compute, HPC, AI for HPC												
General Purpose	<table border="0"><tr><td data-bbox="596 893 861 1019"> </td><td data-bbox="871 893 1411 1019">Real-Time, Medium Throughput, Low Latency, and Sparse Inference</td><td data-bbox="1533 893 1921 1019">  </td><td data-bbox="1931 893 2425 1019">Medium to Small Scale Training and Fine Tuning</td></tr><tr><td data-bbox="596 1031 1131 1156">   </td><td data-bbox="1141 1031 1411 1156">Edge and Network AI Inference</td><td data-bbox="1635 1031 1819 1128"></td><td></td></tr><tr><td data-bbox="596 1168 1090 1288"> </td><td></td><td data-bbox="1533 1168 1921 1288">  </td><td data-bbox="1931 1168 2425 1288">Inference on Client</td></tr></table>	 	Real-Time, Medium Throughput, Low Latency, and Sparse Inference	  	Medium to Small Scale Training and Fine Tuning	   	Edge and Network AI Inference			 		  	Inference on Client
 	Real-Time, Medium Throughput, Low Latency, and Sparse Inference	  	Medium to Small Scale Training and Fine Tuning										
   	Edge and Network AI Inference												
 		  	Inference on Client										