Critical Considerations for AI Deployments

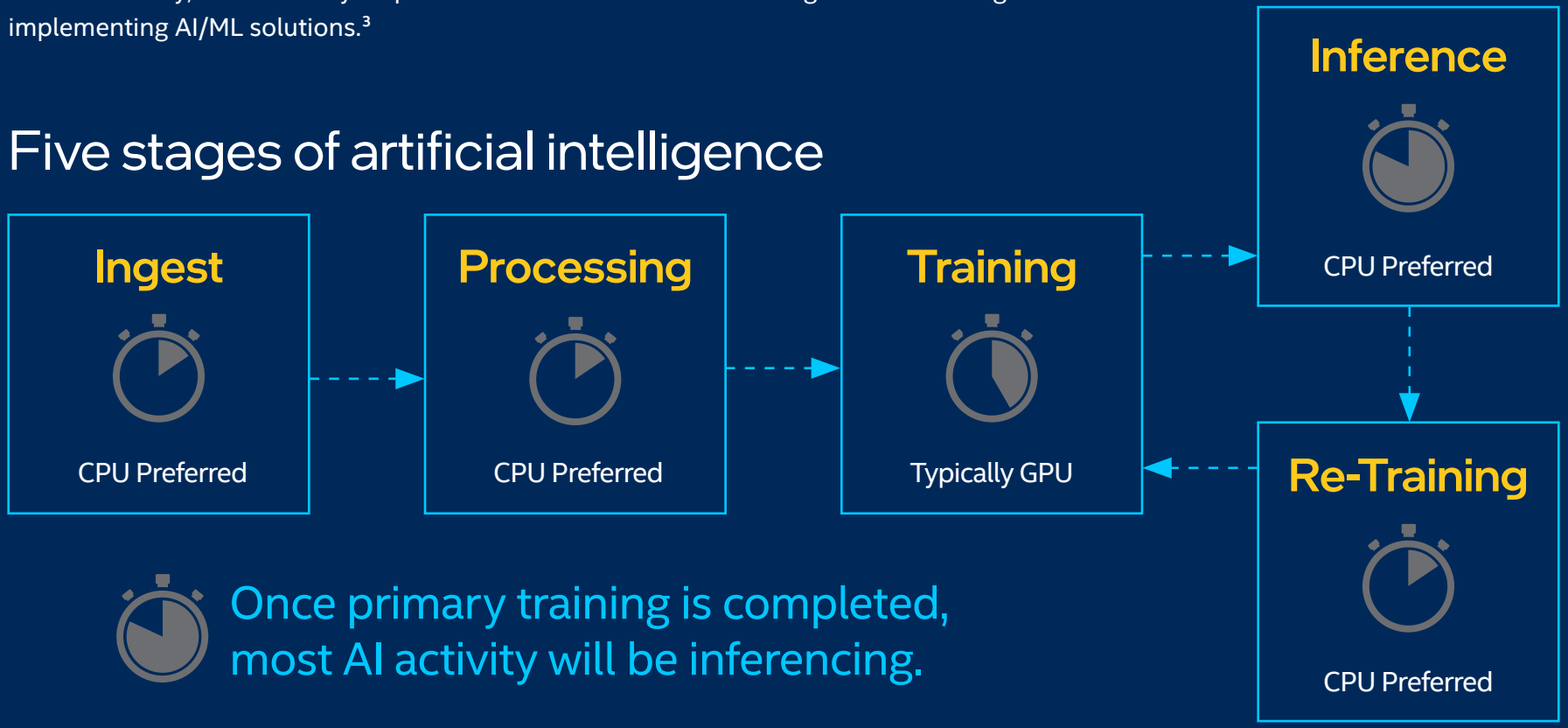
Look Beyond the GPU

While GPU solutions are impactful for training, AI deployment requires more.

It's anticipated that the AI industry will grow to tens of billions of dollars by the mid-2020s, with most of the growth in AI inference.[1] Intel Xeon Scalable processors represent approximately 70% of the processor units that are running AI inference workloads in the data center. [2]

GPUs are effective for training workloads but aren't required for all of the different stages of AI. In a 2021 study, 56% of study respondents cited cost as their most significant challenge to implementing AI/ML solutions.[3]

Five stages of artificial intelligence



Once primary training is completed, most AI activity will be inferencing.

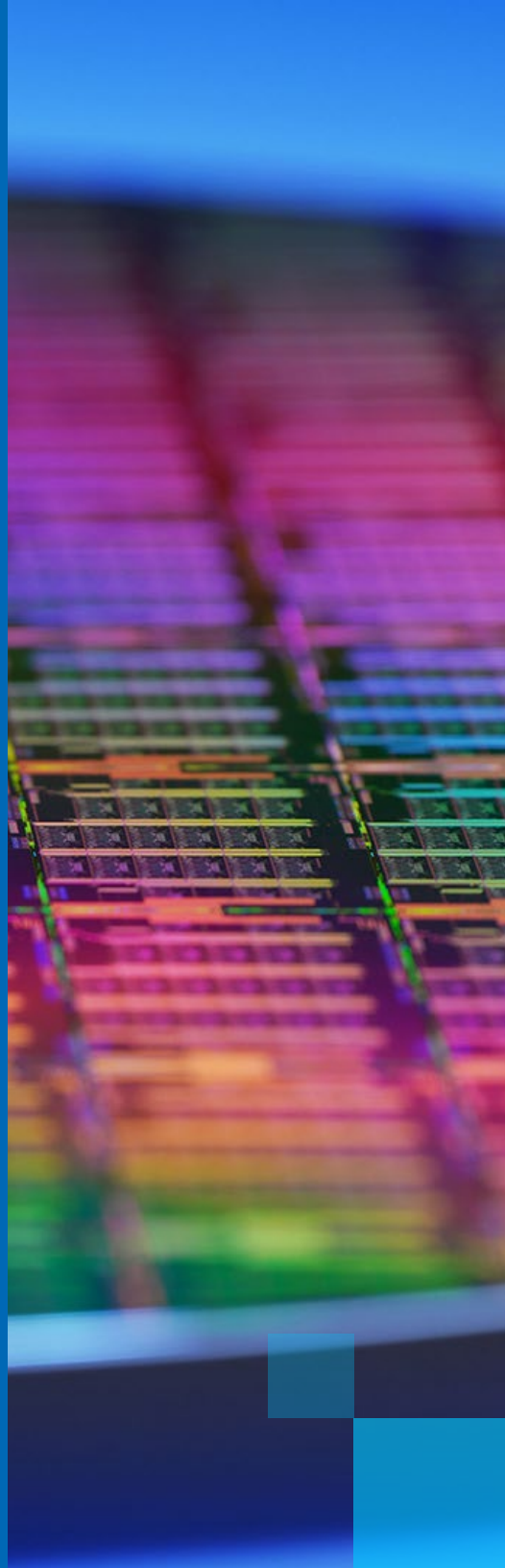
When evaluating AI solutions, make sure you have all the right details to inform your decision.

With proprietary platforms, your organization, enterprise architecture, data infrastructure, middleware and application stacks all have to be rearchitected with expert teams to maximize value from CAPEX and OPEX intensive resources.

Intel-based solutions provide the most workload flexibility for your organization, including AI.

These solutions adapt to your existing organization, deployment context, enterprise architec0tures, data infrastructures, middleware and application stacks so rearchitecting isn't required.

Intel offers the most robust toolkit to support your AI needs. Here are the most important things to keep in mind when considering the implementation of AI solutions across the five stages of AI execution.



01

Intel Accelerates Data Science

Data science workflows require highly interactive systems that handle massive volumes of data in memory, using algorithms and tools designed for single-node processing - current GPUs are generally a poor fit for many of these tasks.

- Data preprocessing today is done on a CPU and many practitioners spend a significant amount of their time using the highly popular Pandas library
- PMem can make it possible to load larger datasets into memory without falling back to disk - it can also act as a fast-cache configuration

Intel platforms with Intel® Optane™ persistent memory (PMem) offer large memory for data science workflows.

Intel's distribution of Modin is an open-source library which accelerates Pandas applications up to **90x**[4]

02

Intel® Xeon® Scalable Processor Enables Effective AI Data Preprocessing

Data infrastructure is already optimized for Intel and effective ingest. The result is a completely optimized pipeline scaling from PC and workstation to cloud to edge: customers can scale AI everywhere by leveraging the broad, open software ecosystem and unique Intel tools.

If you are accessing and processing data then storage and memory are critical - take advantage of a faster Intel storage subsystem that doesn't require use of a GPU.

03

Intel provides scalability for AI Training

Habana® Gaudi® provides customers with cost-efficient AI training, ease of use and system scalability – integration of the Gaudi platform eliminates storage bottlenecks and optimizes utilization of AI compute capacity.[5]

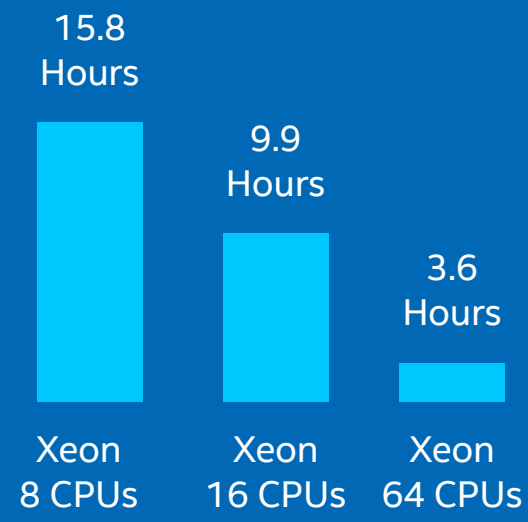
The Habana® Gaudi® AI Training Processor powers Amazon EC2 DL1 instances delivering up to **40%** **better price performance than comparable Nvidia GPU-based training instances** according to AWS testing[6] - this processor is also available for on-premises implementation with the Supermicro X12 Habana Gaudi AI Training Server.

Existing Intel® Xeon® Scalable processors scale well with intermittent training sets during off-peak cycles overnight or on weekends

The upcoming launch of Next Gen Intel® Xeon® Scalable processors (code named Sapphire Rapids) with AMX and BrainFloat16 will deliver even higher performance and scalability.

Time-To-Train on Intel Xeon[7]

MLPerf Results for ResNet-50
Lower is Better



04

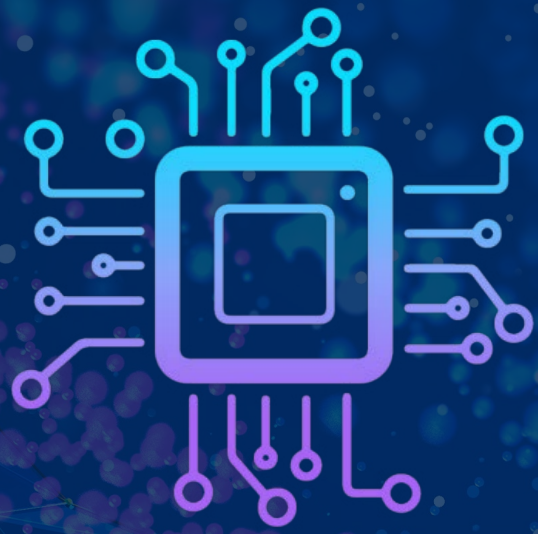
Intel® Xeon® Scalable boosts Machine Learning performance

Elevate effectiveness of machine learning workloads through the performance of Intel hardware.

New built-in acceleration capabilities in 3rd Generation Intel® Xeon® Scalable processors deliver

1.5x greater AI performance

than other CPUs across 20 key customer workloads, the majority of which are machine learning workloads.[8]



The AI accelerators built into Intel® Xeon® Scalable processors provide

10-to-100x
performance improvements[9]

for AI frameworks and libraries like Spark for data processing, TensorFlow, PyTorch, Scikit-learn, NumPy and XGBoost.

You don't need to break the bank for effective graph analytics - a single Intel® Xeon® Scalable processor-based server with sufficient memory is a much better choice for large-scale, general-purpose graph analytics.

Get faster analytics insights, up to 2x faster graph analytics computations (Katana Graph) for recommender systems and fraud detection

2x faster

on average when using 3rd Gen Intel® Xeon® Scalable Processors with Intel® Optane™ persistent memory 200 series.[10]

05

For Inference, Intel® Xeon® Scalable Processors are the Go To Solution

AI deployment is about inference, and Intel is the most globally trusted hardware for inference!
The performance capabilities of Intel hardware can drive the inferencing success your business operation relies on.

Intel® Xeon® Scalable is the only x86 data center CPU with built-in AI acceleration. Utilize Intel® Xeon® Scalable processors for more cost-effective inferencing rather than leveraging new Nvidia hardware that will add deployment and recurring cost.

30% higher average AI Performance across 20 workloads with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs Nvidia A100 GPU[11] (Geomean of 20 workloads) without adding the cost and complexity of a GPU

Dual socket servers with Next Gen Intel® Xeon® Scalable processors (code-named Sapphire Rapids) can infer over 24k images/second compared with 16k on a Nvidia A30 GPU[12]

This means Intel can deliver better than 1.5x the performance of Nvidia's mainstream inferencing GPU for 2022,[13] strengthening the recommendation to standardize on Xeon – and the next generation will provide even greater performance

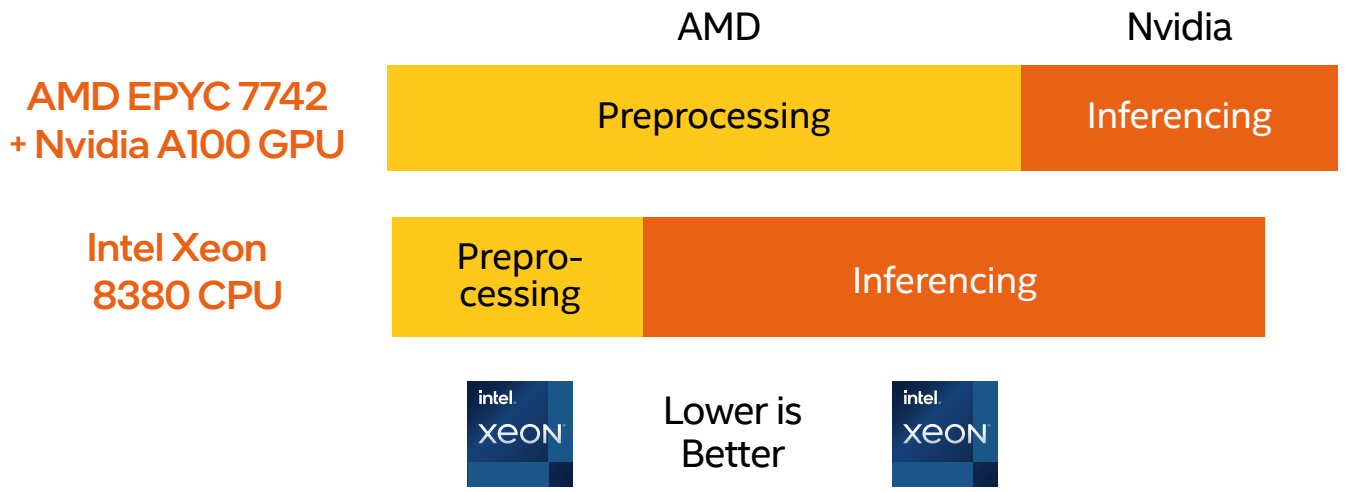
06 Intel delivers End-to-End AI Performance

Optimize your workload for the Intel Xeon Scalable processors you already have installed to get better end-to-end performance without introducing delays or burden. Leverage the Intel-based technologies you know

Complexity of non-Intel integration challenges will result in extended latencies

End-to End Document Level Sentiment Analysis (DLSA)[14]

Preprocessing can dominate time to solution, and the GPU is typically idle



Intel® Xeon® Scalable processor systems are lower cost **(up to 17%)** without the added GPU complexity[15]

07

Intel® Open-Source Software Avoids Lock-in

Write once, use anywhere with Open-Source software. DL/ML framework users can reap all performance and productivity benefits through drop-in acceleration without the need to learn new APIs or low-level foundational libraries as many AI frameworks are already running on Intel.

Maintain flexibility with oneAPI and OpenVINO.

Intel's end-to-end portfolio of AI tools and framework optimizations for customers is built on the foundation of the open, standards-based, unified oneAPI programming model and constituent libraries.

Utilizing OpenVINO allows developers to write once and deploy anywhere with tools designed to optimize and deploy DL inference models.

The Intel® oneDNN library is being adopted widely – oneDNN provides the building blocks for deep learning applications with very fast performance across x86_64 processors, and provides a wider breadth of performance optimizations for developers.

Along with developing Intel-optimized distributions for leading AI frameworks, Intel also up-streams optimizations into the main versions of these frameworks, delivering increased performance and productivity to your AI applications even when using default versions of these frameworks.

Faster Performance vs. Prior Generation:

TensorFlow Intel optimization:

Up to 11x higher batch AI inference performance on ResNet50 with 3rd Gen Intel® Xeon® Scalable processor.[16]

Visit the **performance index page** for additional 3rd Gen Intel® Xeon® Scalable optimizations

08

Intel's Extensive AI Portfolio

AI is a complex and varied ecosystem. Intel® provides a product portfolio of performance hardware and Open-Source software to achieve evolving AI needs with maximum performance and cost efficiency for any workload.

Intel offers the broadest AI portfolio for customers, including CPUs, FPGAs, VPUs, ASICs, forthcoming discrete GPUs and more, allowing us to position the right hardware for any customer use case.

No matter the AI deployment type, the Intel portfolio provides the hardware and software capabilities you need for success:

Dedicated Training:
Intel® Xeon® Scalable and Habana® Gaudi® today; Intel Ponte Vecchio GPU coming soon.

Data Science:
Intel platforms with Intel® Optane™ PMem offer large memory for data science workflows.

Best Effort Training:
Leverage reduced cost and complexity through Intel® Xeon® Scalable for intermittent, off-hours training cycles.

Open Source Software:
Don't limit business trajectory – take advantage of productivity benefits without the need to learn new APIs or libraries.

E2E Performance:
Intel® Xeon® Scalable can deliver competitive workload performance with better perf/$.

AI Inferencing:
Intel® Xeon® Scalable hardware delivers for high-performing inference with improved TCO.

1 https://www.embeddedcomputing.com/technology/ai-machine-learning/ai-dev-tools-frameworks/the-evolution-of-ai-inferencing

2 Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2021.

3 https://pages.awscloud.com/GLOBAL-ln-GC-400-OTH-INFR-IDC-Intel-Whitepaper-Realizing-Business-Outcomes-learn.html

4 https://techdecoded.intel.io/resources/one-line-code-changes-to-boost-pandas-scikit-learn-and-tensorflow-performance/#gs.bzkn2n for workloads and configurations. Results may vary.

5 https://www.hpcwire.com/off-the-wire/habana-labs-announces-turnkey-ai-training-solution-habana-gaudi-platform-and-ddn/

6 https://www.crn.com/news/components-peripherals/intel-takes-on-Nvidia-with-habana-based-aws-ec2-instances

7 MLPerf results for Training v1.0 published on June 30, 2021. See https://mlcommons.org/en/training-normal-10/

8 See [43] at www.intel.com/3gen-xeon-config

9 https://www.slideshare.net/IntelSoftware/software-ai-accelerators-the-next-frontier-software-for-ai-optimization-summit-2021-keynote-249477197

10 See claim 4 at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/intel-optane-persistent-memory-200-series/ for workloads and configurations. Results may vary.

11 See [44] at https://www.intel.com/3gen-xeon-config

12 See Key100 Sandra Rivera AITI001 Pradeep Dubey Slide 37 at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/innovation-event-claims/

13 See Key100 Sandra Rivera AITI001 Pradeep Dubey Slide 37 at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/innovation-event-claims/

14 See www.Intel.com/InnovationEventClaims, AI001, Meena Arunachalam, #25, for workloads and configurations (this is the chart)

15 See [100] at https://www.intel.com/3gen-xeon-config

16 https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/, Footnote #118

0722/DCS/MG/PDF