FAQs When Choosing an AI and Analytics Infrastructure with 3rd Gen Intel® Xeon® Scalable Processors

1. What infrastructure resources have an impact on AI and analytics performance?

Many people believe the only infrastructure resources that matter for AI and analytics are compute resources. And those resources are critical for accelerating time to insight.

But too often, other resources that have a major influence on AI and analytics performance are ignored—such as memory capacity, storage performance and capacity, and networking performance.

These infrastructure resources are critical because compute resources rely on rapid access to data, and any time cores are idle, performance is compromised.

That's why Intel has taken a comprehensive, end-to-end approach to AI and analytics infrastructure innovation. These innovations:

- Greatly increase memory capacity so more data can be closer to the processor
- Deliver a range of SSD options to enable an optimal balance of breakthrough storage performance and cost-efficient capacity
- Accelerate I/O to and from servers with networking solutions that enhance performance, reduce latency and offload network tasks from CPUs

2. What is AI training and inferencing?

Artificial Intelligence, or AI, includes a broad range of techniques to derive actionable intelligence and insight.

Deep learning is a subset of AI that uses massive data sets to create a multilayered neural network. Then, the neural network is used to autonomously recognize patterns and drive optimal actions.

AI training is the creation stage of the neural network. AI inferencing is the autonomous recognition stage that uses the trained neural network.

AI training generally occurs in a data center or cloud, as it involves enormous volumes of data, requiring massively parallel, high-performance processing power.

AI inferencing is most often deployed in edge applications, where near-real-time recognition and action are required to optimize outcomes.

Common AI inferencing use cases include: computer vision, image recognition, speech recognition, anomalous behavior recognition and more.

These inferencing use cases have many practical applications, such as:

- Smart digital signage – Using computer vision to recognize if customers are paying attention to signage content, and combining it with anonymous demographic information to measure and customize content
- Chatbots – Using speech recognition to enable personal assistance, self-help and interactive digital experiences
- Financial fraud detection – Recognizing anomalous financial behavior to proactively prevent illegal or unauthorized financial transactions

- Cyber threat prevention – Recognizing anomalous network or application behavior to prevent malicious attacks from accessing sensitive systems, applications or data

3. What are the advantages of choosing a CPU-based infrastructure for AI?

A key consideration when choosing your AI infrastructure is whether to go GPU-based or CPU-based.

In the early days of AI proliferation, GPUs were considered the technology of choice for many AI tasks, including AI training and inferencing, because they had a performance advantage over general-purpose CPUs.

However, deploying a GPU-based infrastructure for AI workloads generally meant deploying and managing a siloed infrastructure for AI needs, separate from the general-purpose infrastructure designed to run your many other workloads.

Recent Intel CPU and platform innovations have greatly accelerated AI workloads, such as training and inferencing, enabling IT to deploy, operate and support a common infrastructure for their AI and data-centric workloads, as well as the vast array of other applications that drive their business.

And that means improved IT resource utilization and efficiency, and decreased CapEx, OpEx, data center complexity and sprawl.

4. Why are 3rd Gen Intel® Xeon® Scalable processors a great choice for AI inferencing and training?

Intel® Deep Learning Boost (Intel® DL Boost) was introduced in 2nd Gen Intel® Xeon® Scalable processors to significantly accelerate AI inferencing performance, enabling organizations to run demanding inferencing workloads on their general-purpose infrastructure without compromising performance.

3rd Gen Intel Xeon Scalable processors are extending Intel's leadership in CPU-based AI acceleration by integrating bfloat16 support within Intel DL Boost.

These are the first general-purpose CPUs to offer bfloat16 support, which further speeds AI training performance. But what's new and even more exciting is how bfloat16 significantly accelerates AI training performance.

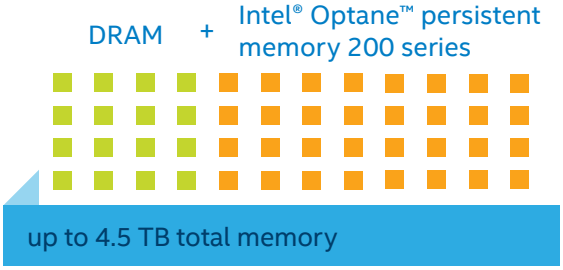
Popular AI frameworks and libraries have been performance-optimized by Intel software engineers to take advantage of the latest Intel Deep Learning Boost features to maximize AI performance on 3rd Gen Intel Xeon Scalable processors.

Now you can run more AI workloads on your general-purpose, Intel technology-based IT infrastructure to reduce IT costs and complexity, while improving utilization and efficiency.

5. How does Intel® Optane™ persistent memory 200 series enhance AI and analytics?

When you think about AI and analytics workloads, what immediately comes to mind when you consider their infrastructure requirements? If you're like most people, you think about how compute-intensive those workloads are.

But what really separates AI and analytics from other performance-sensitive workloads is how data-intensive they are. They require large volumes of data, and the speed at which that data can be accessed by the processor has an enormous impact on overall performance.



Until recently, DRAM was the only choice for system memory. And while DRAM delivers outstanding performance, it has not kept up with Moore's Law when it comes to density, capacity and costs. Thus, it has not kept up with processing performance.

And that's where Intel Optane persistent memory 200 series comes in. Supported by 3rd Gen Intel Xeon Scalable processors, this breakthrough memory innovation greatly expands system memory capacity, enabling up to 4.5TB of memory per processor socket[1].

The expanded capacity means you can utilize much larger datasets for analytics. And, unlike DRAM, it can persistently store data—enabling up to 225 times faster[2] read access compared to mainstream NAND SSDs.

Applications, such as SAP, use the increased capacity and persistence to accelerate analytics and enable new, in-memory database possibilities. The persistence also enables significantly faster restarts after scheduled maintenance.

1) 6 x 512GB Intel Optane persistent memory (3,072 GB) + 6 x 256GB DDR4 DRAM (1,536 GB) = 4,608 GB total memory per socket.

2) Intel Optane persistent memory idle read latency of 340 nanoseconds. Intel® SSD DC P4610 Series TLC NAND solid state drive idle read latency of 77 microseconds.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure.