phoenixNAP
GLOBAL IT SERVICES

Strategic Solutions for AI Workload Acceleration and Cost Management

White paper by phoenixNAP
August, 2024

INTRODUCTION

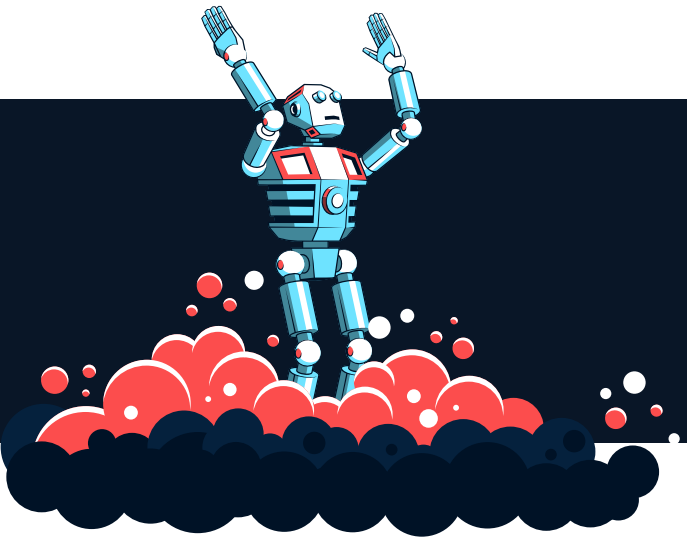
The effectiveness of Artificial Intelligence (AI) models across use cases has incentivized greater commitment to their development. According to Gartner, over 80% of enterprises will have used generative AI models by 2026[1]. This is a dramatic increase compared to less than 5% in 2023.

Despite efficiency-driving breakthroughs in data processing and analysis, the costs of training and inferencing AI models keep increasing.

This white paper offers an overview of the current artificial intelligence and machine learning (ML) landscape, analyzing the data processing and storage requirements of AI-driven applications. It provides recommendations for optimizing AI models' cost-performance ratio and proposes a cost-effective, turnkey infrastructure solution designed to boost the performance of AI workloads at scale.

I – THE CURRENT STATE OF ARTIFICIAL INTELLIGENCE

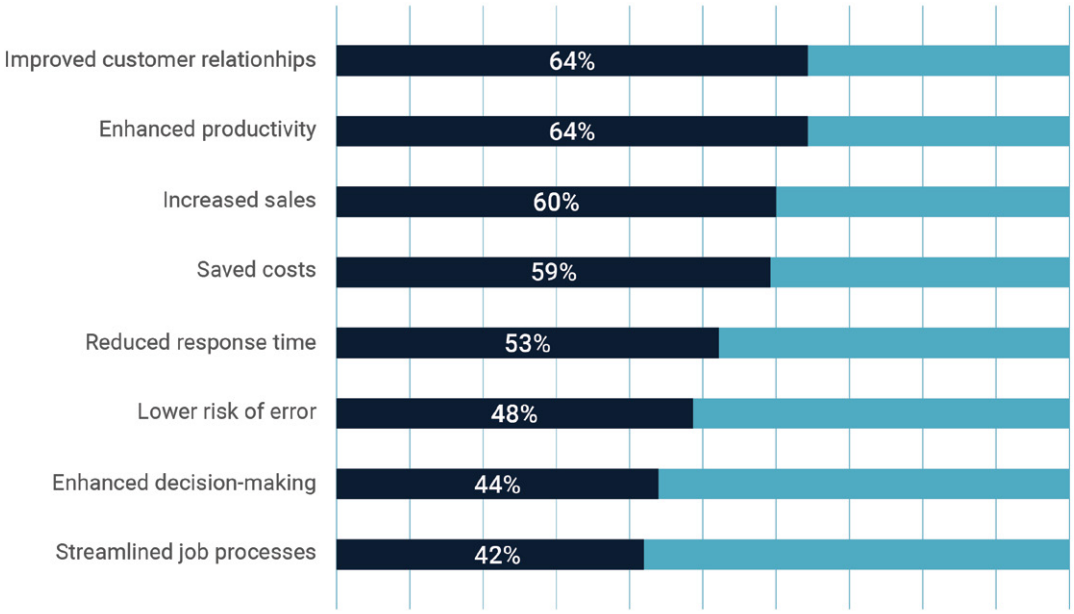
The growth of AI has been strong: the U.S. AI market size was approximately $48.51 billion in 2024[2], and it will likely expand at a CAGR of 28.83% from 2024 to 2033. The global AI market was worth a "mere" $15.7 billion in 2017[3], underscoring the rapid rise of this technology. **Improvements in infrastructure speed and scalability, emerging open-source technologies** like TensorFlow, PyTorch, and Hadoop, **and public awareness of language processing models** like ChatGPT have all contributed to this growth.



"The U.S. AI market size was approximately $48.51 billion in 2024, and it will likely grow at a CAGR of 28.83% from 2024 to 2033. The global AI market was worth a "mere" $15.7 billion in 2017, further illustrating the skyrocketing of this technology."

According to the recent Forbes Advisor survey[4], organizations that have adopted AI cite the following improvements to their day-to-day operations:

The Positive Impact of AI for Business Owners



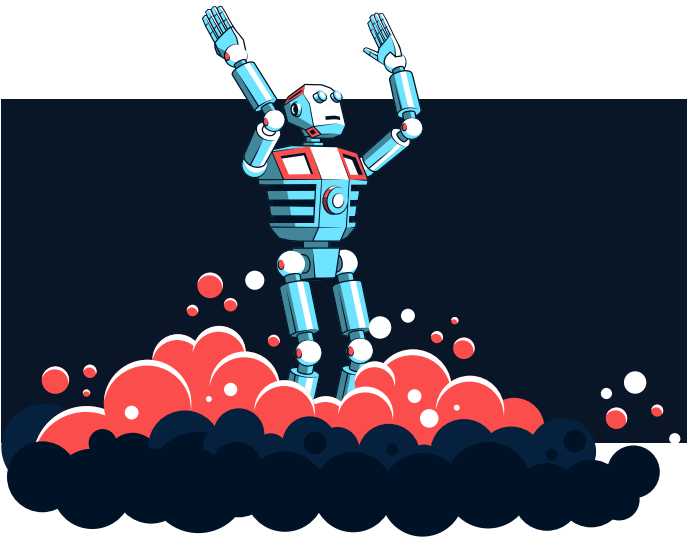
Data shows AI is present across industries, the driving factor behind its proliferation being the **versatility of its application and potential benefits**. Given its highly disruptive nature, we can only expect its impact to deepen.

II – THE PRICE OF AI: RISING PERFORMANCE AND COST DEMANDS

The estimated compute costs for the final training run of large-scale ML systems in 2023 were over $1 million, compared to approximately $100,000 in 2018[6]. Such head-spinning price tags mostly apply to massive AI models deployed by large organizations. Consequently, SMBs struggle to find budget-friendly alternatives. As cited by Microsoft[7], over 50% of small businesses consider expenses the main obstacle toward AI adoption (44% when we include medium-sized businesses[8]).

Organizations usually run AI on-prem or in the cloud, typically preferring the latter option. However, it can be argued that **the cost of maintenance outweighs the respective advantages of these IT environments**. Running a cloud-native AI application can incur higher-than-expected cloud spend (a persistent issue for 82% of organizations[9]). Meanwhile, an on-prem infrastructure entails dizzying up-front investment, potentially exceeding $50 million for large-scale deployments.

Regardless of the preferred infrastructure, costs remain a major deal breaker.



Over 50% of small businesses point to expenses being the main obstacle toward AI adoption (44% when we include medium-sized businesses).

Why AI Infrastructure Requirements Are So High

To better understand why AI takes such a toll on IT resources, we need at least a cursory understanding of its creation process. Below is a broad overview of a machine learning (ML) model's pre-deployment procedure.

1

Preparation

Collect vast amounts of data.

•

Normalize the data to fit the used framework (e.g., TensorFlow or PyTorch).

•

Use feature engineering to select for appropriate variables (parameters).

2

Training

Analyze some of the ingested data to evaluate parameters.

•

Use the remaining data to test the algorithm.

•

Repeat the process until sufficient accuracy is achieved.

3

Validation

Confirm that the model meets performance thresholds.

•

Test the algorithm with unevaluated data.

•

Fine-tune the hyperparameters (learning speed, number of rounds, etc.) of the model.

4

Testing

Introduce unseen, real-world data to the model.

•

Test the model's inference capabilities.

•

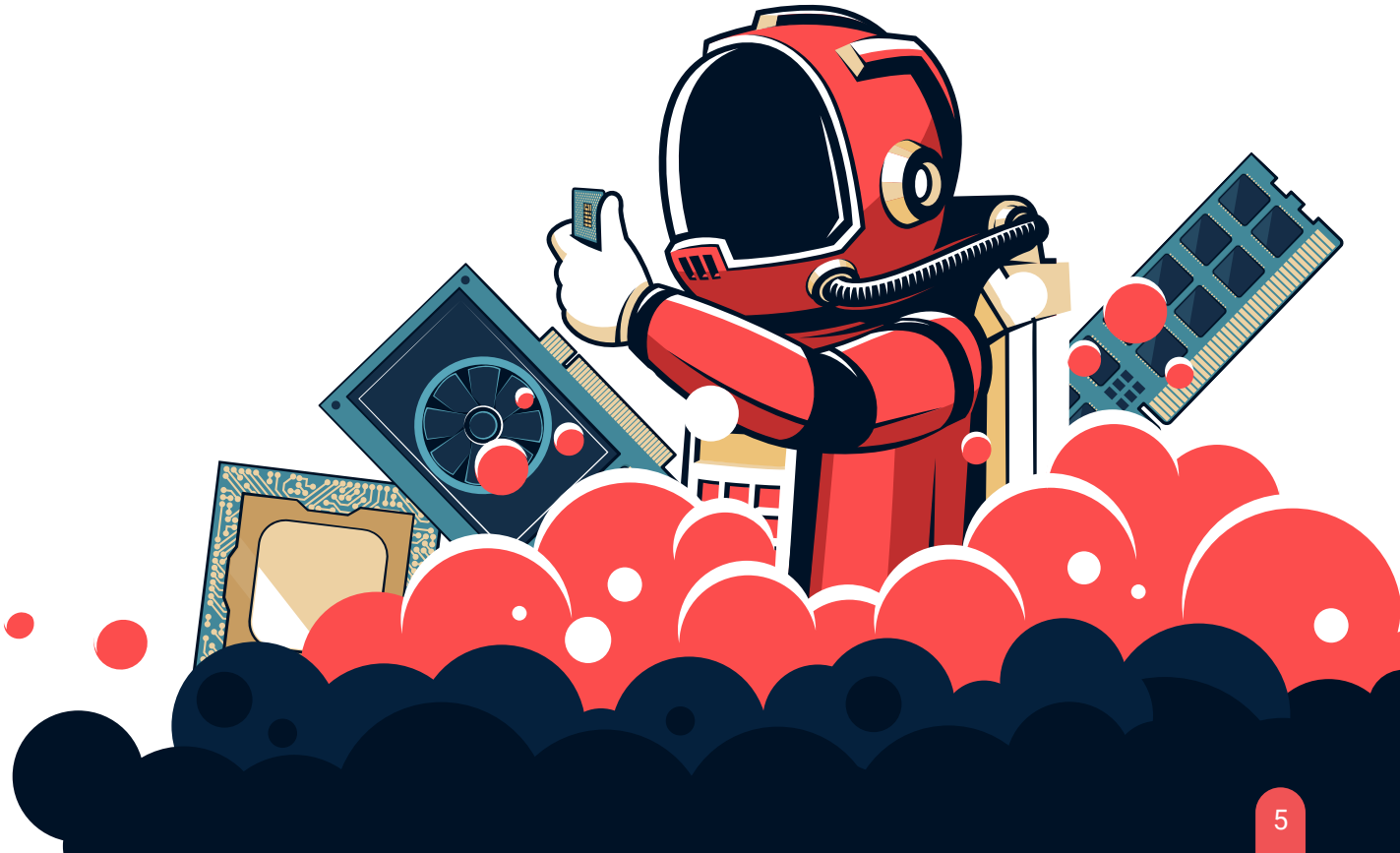
Replicate/compare to previous benchmark results.

Effective machine learning requires the following infrastructure resources:

- **Terabytes of SSD storage** for the training data.
- **Considerable IOPS capacity** to speed up hot data transportation time.
- **A high-speed network** to facilitate the necessary data transportation.
- **Sufficient RAM** to handle the immense volumes of data.
- **Exceptionally high compute** to perform thousands of algorithmic calculations.
- **GPU-powered parallel computing** to offload some of the processing strain on the CPU.
- **Appropriate power supply units and cooling solutions**.

To train an AI model similar to GPT-3, we would need around 175 billion parameters, >1TB of data in memory, and a computational capacity of 350 trillion floating point operations!

To sum up, the exceedingly expensive hardware necessary for AI modeling complicates efforts to cut costs during at-scale deployment.



III – STRATEGIES FOR RUNNING AI WORKLOADS COST-EFFECTIVELY

Optimizing the performance-cost ratio is crucial for making AI workloads sustainable. The following strategies will help you get the most out of training and deploying effective models.

1. Consider Compute Costs from the Start

Plan out the application's production pipeline and define acceptable expenses before development even begins. To take your strategy in the right direction, consider these critical elements:

- Planned time to market.
- Hardware/cloud costs.
- Software expenses for data collection, analysis, etc.
- Labor costs of developers, engineers, and data scientists.
- The complexity of the training data.
- Data curation – collecting, labeling, normalization.
- Training and maintaining the AI model.
- Number of applications or devices on which the AI will work.
- Legal fees and regulatory compliance.

2. Closely Monitor Performance

By continuously monitoring and analyzing the AI system's performance, you can more easily spot development bottlenecks. Focus on:

- Defining measurable performance goals with which to compare results.
- Establishing a framework to spot missing values, errors, biases, and outliers.
- Setting up routine reporting and alerts for errors or anomalies.
- Systematically evaluating performance metrics via validation or cross-validation.
- Defining a restore strategy should a critical error occur in the model.

3. Choose the Right Infrastructure

Depending on the specifics of your application, some infrastructure options make more sense than others. Below is an overview of viable infrastructure solutions, along with their strengths and weaknesses.

Hyperscale Cloud Services

+

Easily scalable

Minimal infrastructure management

Built-in backup and disaster recovery

-

Vendor lock-in

Noisy neighbors

Little control over environment

Lack of transparency leading to overprovisioning and cost overruns

On-Prem Data Center

+

Greater control over infrastructure

Highly customizable

More control over performance and security

-

Considerable upfront investment

Requires expertise and overhead to maintain

Difficult to scale up due to cost and space limitations

Resource waste when scaling down

Bare Metal Cloud

+

Cloud-like flexibility with greater control over IT resources

Dedicated resources managed via API, CLI, or Infrastructure as Code tools

Runs workloads directly on the hardware without hypervisors

-

Limited instance customizability compared to cloud deployments

User manages OS and applications

4. Focus on Data Quality, Too

The amount of data necessary to train an AI model depends on several factors, such as the algorithm used, the complexity of the problem it solves, and the number of dataset features. More data correlates with more accuracy because the model has more information from which to draw insights.

That said, placing a greater emphasis on attributes related to data quality (e.g., accuracy, consistency, and reliability) reduces your dataset without compromising on effectiveness. Data free of irrelevant attributes better steers the AI toward making correct conclusions.

3 out of 5 ML, AI, and data experts consider higher quality of training data more important than higher quantity of training data[10].

Adopt AIOps and MLOps

AIOps (Artificial Intelligence Operations) and MLOps (Machine Learning Operations) are workflow methodologies that streamline the production and monitoring of machine learning models. They assume characteristics typical of DevOps, such as CI/CD (Continuous Integration/Continuous Delivery) principles, automation, Infrastructure-as-Code practices, and containerization.

Adopting AIOs and MLOps yields:

- Greater **productivity** by automating data collection, model development, and testing.
- Enhanced **scalability** via easily reproducible pipelines and orchestrator-based scaling.
- Better **monitorability** with automated, real-time alerts for events like model drift.
- More **reliability** due to CI/CD and automation processes that remove human error.

Take Advantage of Future-Ready Hardware Technologies

Recognizing the potential of AI, hardware vendors continue to roll out groundbreaking products tailored toward the needs of AI-driven organizations. Riding on the coattails of this innovation can help you drastically improve the performance of your workloads.

For example, **Intel® 4th and 5th Gen Xeon™ Scalable processors** leverage more built-in accelerators than other CPUs on the market to easily accommodate demanding workloads. The most notable of these is Intel's **Advanced Matrix Extensions (AMX)**. It is an instruction set designed to cost-effectively accelerate matrix-oriented operations vital for natural language processing, recommendation systems, and image recognition models.

Intel's new CPUs also deliver other gen-on-gen improvements, including:

- **More Cores + Increased Multi-Socket Bandwidth**
  1.53x gen-on-gen performance gain for 4th Gen, 1.84x for 5th Gen.
- **Intel Data Streaming Accelerator (DSA)**
  Boosts the performance of storage, networking, and data-intensive workloads.
- **Intel Dynamic Load Balancer (DLB)**
  Dynamically distributes network data across CPU cores.
- **Intel QuickAssist Technology (QAT)**
  Accelerates encryption, decryption, and data compression.
- **Intel In-Memory Analytics Accelerator (IAA)**
  Improves analytics performance and accelerates database query throughput.
- **Intel Software Guard Extensions (SGX)**
  Isolates sensitive data in a secure enclave with hardware-based memory protection.

Reinforce your competitive advantage by leveraging these innovative technologies, **preferably as a service to reduce TCO.**

For more information on Intel AMX and how to use it to optimize your AI development pipelines, refer to **phoenixNAP's Knowledge Base article**.

Leverage the Latest GPU Technologies to Supercharge AI Compute

GPUs can handle extremely complex tasks by performing thousands of operations concurrently. Although initially developed for rendering graphics, they tackle complex models like neural networks with similar competency. Since GPUs provide parallel environments for simultaneous data processing, they deliver **record efficiency for processing training and inference workloads**.

IV- BARE METAL CLOUD: POWERFUL, AFFORDABLE FOUNDATION FOR AI AND ML ACCELERATION

Bare Metal Cloud is a service that allows users to rent dedicated physical servers while maintaining the performance and scalability of the cloud. As a best-of-all-worlds solution, it **reduces processing time and latency while offering greater customization and control over IT resources.**

Advantages	Cloud	Bare Metal	On-Prem	BARE METAL CLOUD
Simple management	✓	✗	✗	✓
Control over infrastructure	✗	✓	✓	✓
No upfront investment	✓	✗	✗	✓
Easy access to high-tier resources (hardware, security, power, cooling)	✓	✓	✓	✓
Quick deployment at scale	✓	✗	✗	✓
No "noisy neighbors"	✗	✓	✓	✓
Transparent pricing	✗	✓	✓	✓
Customizability and flexibility	✗	✓	✓	✓
Fast troubleshooting	✗	✓	✗	✓

phoenixNAP's Bare Metal Cloud (BMC) is such a service. Providing access to dedicated resources with cloud flexibility, this Infrastructure-as-a-Service solution is a single-tenant, highly scalable, and customizable environment consumable on an OpEx model.

The platform includes features like:

- Global and edge deployment in minutes via API, CLI, or WebUI.
- Preconfigured server instances powered by Intel's latest Xeon Scalable CPUs.
- Cutting-edge CPUs, DDR5 RAM, and high-performance NVMe drives.
- Access to systems with high-density, discrete dual Intel GPUs connected via Intel's high-speed Xᵉ Link Bridges.
- Management via popular IaC tools (Terraform, Ansible, Pulumi).
- On-demand access to terabytes of Network File Storage or S3-compatible Object Storage.
- Data encryption in transit via Intel Software Guard Extensions (SGX).
- A wide range of OSs, including CentOS, Ubuntu, Windows, ESXi, and Proxmox.
- One-click deployment of production-ready K8s clusters via SUSE Rancher integration.
- Flexible hourly billing and bandwidth models with discounts for reservations.

Bare Metal Cloud Success Story: How SpyFu Eliminated AI Performance Bottlenecks and Cut Cloud Costs by 50%

The creators of a popular AI-powered marketing revenue engine run mission-critical workloads on phoenixNAP's BMC and enjoy enhanced performance while cutting cloud bills in half. **Read our case study** to find out how!

Consuming 4th and 5th Gen Intel Xeon Scalable CPUs via Bare Metal Cloud

phoenixNAP's BMC offers preconfigured server instances powered by 4th and 5th Gen Intel Xeon Scalable CPUs with built-in AMX accelerators. The combination of Intel AMX and BMC's high-performance, API-driven IT infrastructure allows you to:

- Deploy enterprise-scale environments optimized for extracting value out of vast datasets in minutes.
- Leverage tools like Terraform and Ansible to automate deployments and scale AI infrastructure.
- Increase PyTorch performance by up to 14x for real-time inference and training.
- Accelerate matrix operations to supercharge AI application speed and accuracy.
- Enjoy lightning-fast throughput and response times thanks to all-flash storage and NVMe drives.
- Reduce time to insight with one-click access to CPUs and workload acceleration engines.

Bare Metal Cloud's **d3 server instances** offer API-driven, preconfigured servers for your demanding AI and ML workloads deployable in minutes across the U.S., Europe, and Asia.

Bare Metal Cloud Gen-On-Gen Benchmarks: Inferencing Performance on 4th Gen Intel Xeon CPUs

In 2023, Intel tested the real-time inferencing speed and throughput of two AI models on phoenixNAP's Bare Metal Cloud to compare the compute of their underlying processors. The CPUs used were:

1. 3rd Gen Intel Xeon Platinum 8352Y
2. 4th Gen Intel Xeon Platinum 8452Y with AMX accelerator

The test was performed using two TensorFlow models:

1. Bidirectional Encoder Representations from Transformers (BERT)-Large, an ML model for natural language processing.
2. Deep Interest Evolution Network (DIEN), a neural network architecture for dynamic modeling of user interests over time.

Performed with FP32, INT8, and BF16 precision, the three tests showed a **substantial performance increase at better costs** when using the 4th Gen Intel Xeon Platinum 8452Y processor with an AMX accelerator. Moreover, the Intel AMX accelerator proved significantly more performant than legacy VNNI (Vector Neural Network Instructions).

Normalized Comparison of FP32 Inferencing Performance (Higher Is Better)

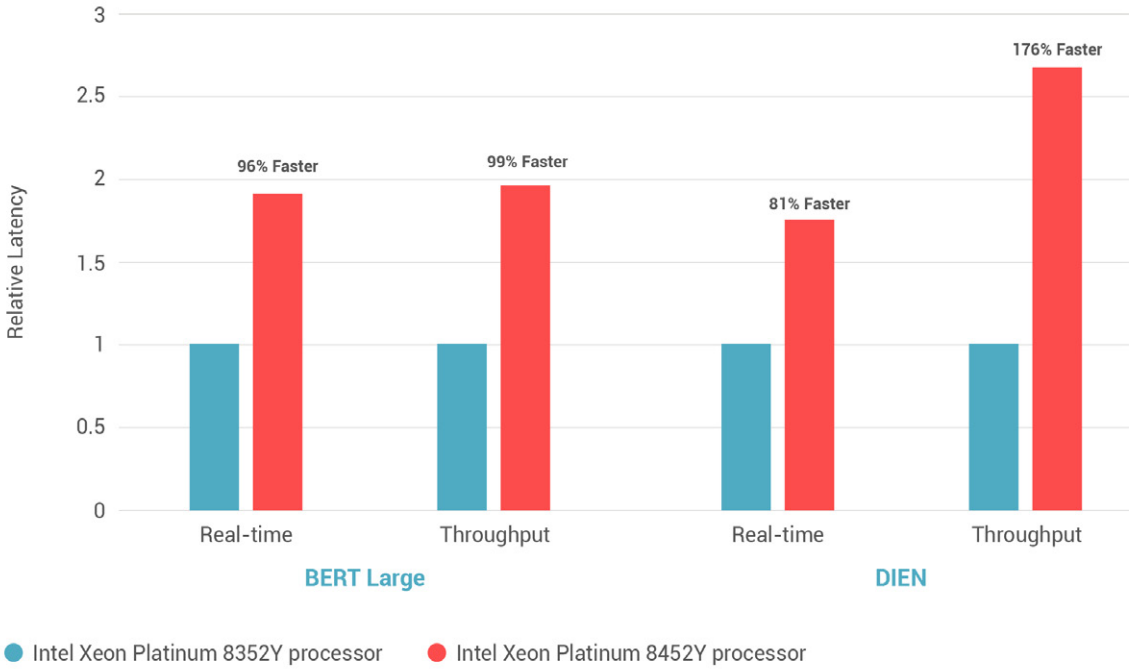


Figure 1. Normalized comparison of inferencing performance of 3rd Gen vs. 4th Gen Intel Xeon Scalable processors with FP32 inferencing precision

Normalized Comparison of INT8 Inferencing Performance (Higher Is Better)

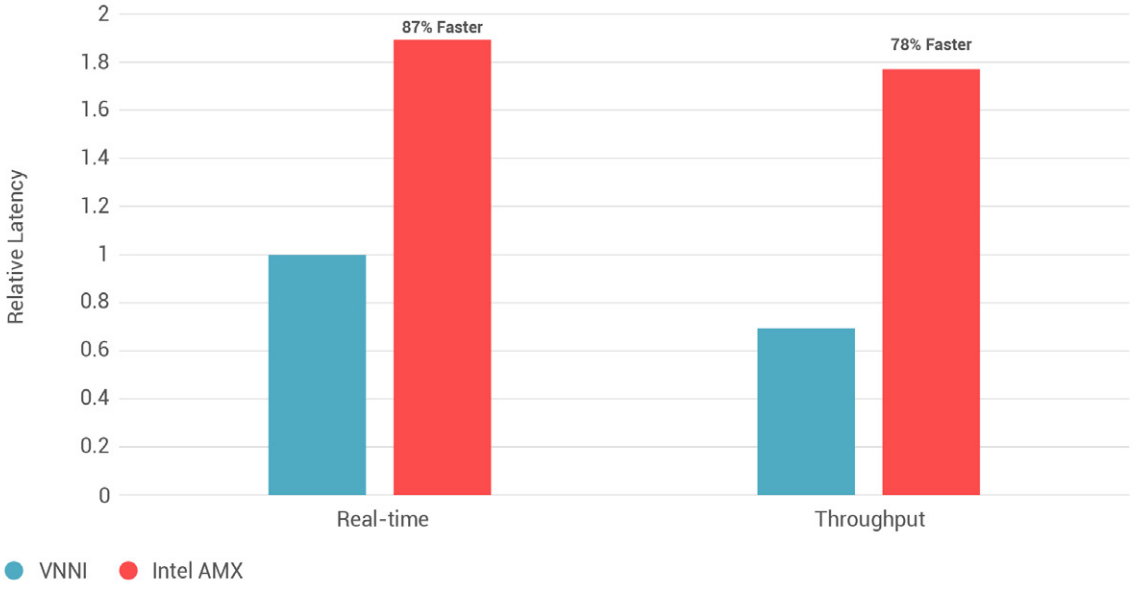


Figure 2. Normalized comparison of INT8 inferencing performance of VNNI and Intel AMX on 4th Gen Intel Xeon Scalable processors

Normalized Comparison of BF16 Inferencing Performance (Higher Is Better)

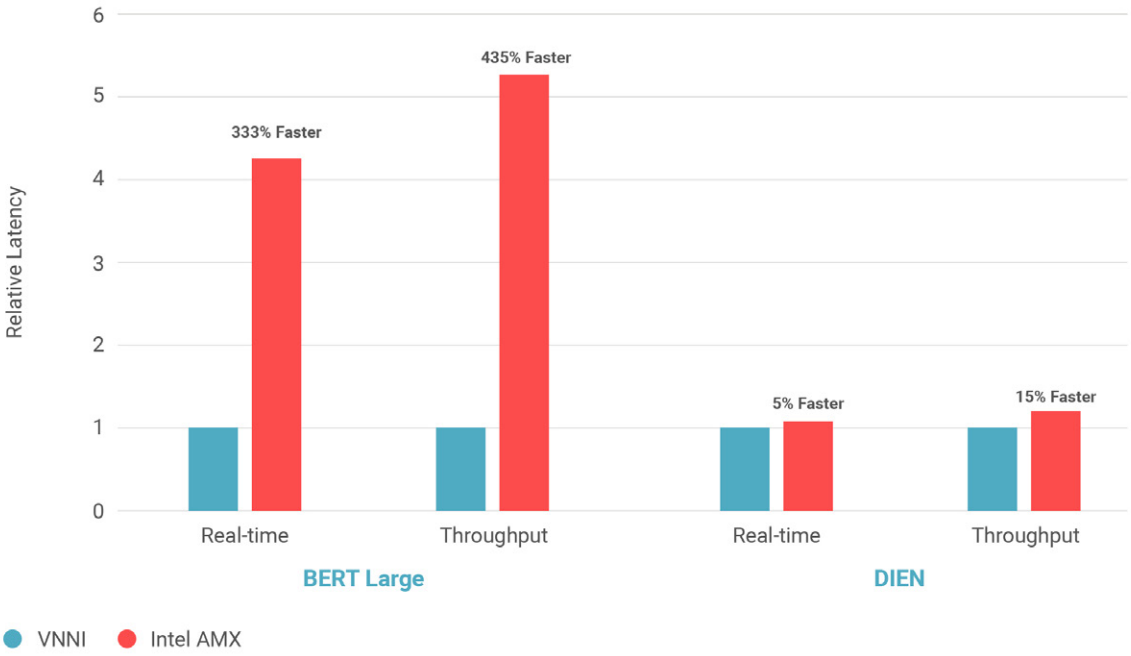


Figure 3. Normalized comparison of BF16 inferencing performance of VNNI and Intel AMX on 4th Gen Intel Xeon Scalable processors

Cost-Per-Hour Comparison

The testing demonstrates that 4th Gen Intel Xeon Scalable processors are more cost-effective compared to 3rd Gen Intel Xeon Scalable processors. Here's what we found:

1. In the inference test involving BERT-Large (FP32 precision), throughput performance capacity costs were **43% less when using 4th Gen Intel Xeon Scalable processors**.
2. In the inference test involving BERT-Large (INT8 precision), throughput performance capacity costs were **47% less when using 4th Gen Intel Xeon Scalable processors with Intel AMX enabled**.

For the best results, **combine these AI accelerator-packed processors with 56-core Intel MAX 1100 GPUs** by leveraging our Bare Metal Cloud's **d3.g2 instances**. Spin up API-driven dedicated server instances with Intel's highest-density discrete, dual GPUs in just a few clicks for faster, cheaper AI training and inference.

- **108 MB L2 cache**
  High-capacity secondary cache memory to skyrocket workloads.
- **56 GPU Cores**
  Intel's highest-density GPU featuring 56 Xe cores.
- **48 GB of HBM2E Memory**
  Ensures lightning-fast data access for large and complex ML/DL models.
- **Intel Xe Matrix Extensions (XMX)**
  Performs up to 256x Int8 operations per clock for quicker AI training and inference.

Ray Tracing Acceleration
Delivers faster scientific visualization and animation with 56 ray tracing units.

Shared Code Investments
Optimizes CPU and GPU performance without duplicating your development efforts.

Intel OneAPI
Simplifies cross-architecture programming using a data-parallel language and APIs.

Wondering how to set up an ML environment on a BMC GPU instance? Read our **Knowledge Base guide** on deploying ML-optimized instances, installing the Intel oneAPI Base Toolkit, and setting up PyTorch or TensorFlow!

Training and Inference Benchmarks: Comparing Intel MAX 1100 GPUs and 3rd & 4th Gen Xeon Scalable CPUs

phoenixNAP tested training and inference performance on Bare Metal Cloud instances with Intel MAX 1100 GPUs and Dual Xeon Gold 6442Y CPUs, comparing the resulting numbers with those of BMC deployments powered by 3rd and 4th Gen Xeon Scalable CPUs only.

The server configurations leveraged OneAPI (oneAPIBase and oneAPIHPC toolkits) to maximize the performance of the inferencing and training runs, which were conducted using ResNet50 and BERT-L models. We contrasted the output from the GPU instances with that from the following Xeon Scalable CPUs:

1. 4th Gen Intel Xeon Platinum 8452Y (with AMX both enabled and disabled)
2. 3rd Gen Intel Xeon Platinum 8352Y

The results show that **Intel MAX 1100 GPUs consistently outperformed the 3rd and 4th Gen CPUs**. The GPU-powered instances processed more data in less time at a greater throughput, creating opportunities for faster, more cost-effective insights from AI models.

ResNet50 – Real Time inferencing (1024 on INT8)

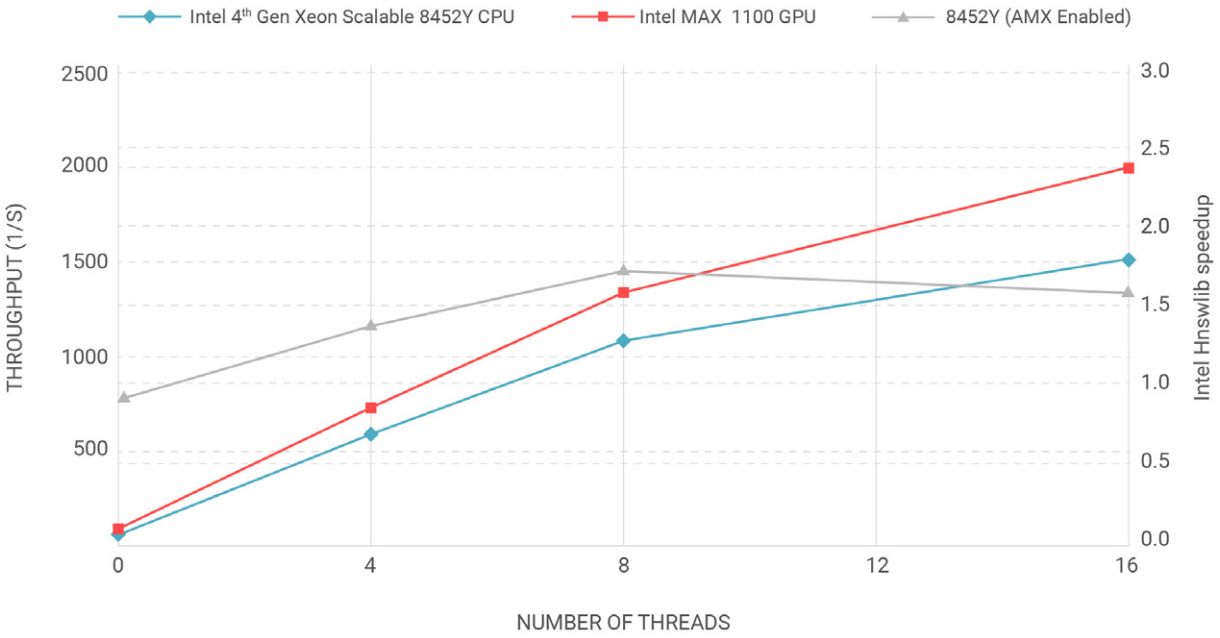


Figure 4. Comparison of real-time ResNet50 inferencing performance with INT8 inferencing precision: 4th Gen Intel Xeon Scalable 8452Y processors (with and without Advanced Matrix Extensions (AMX) enabled) vs Intel MAX 1100 GPUs

Training Time Comparison for ResNet50 Model (Lower Is Better)

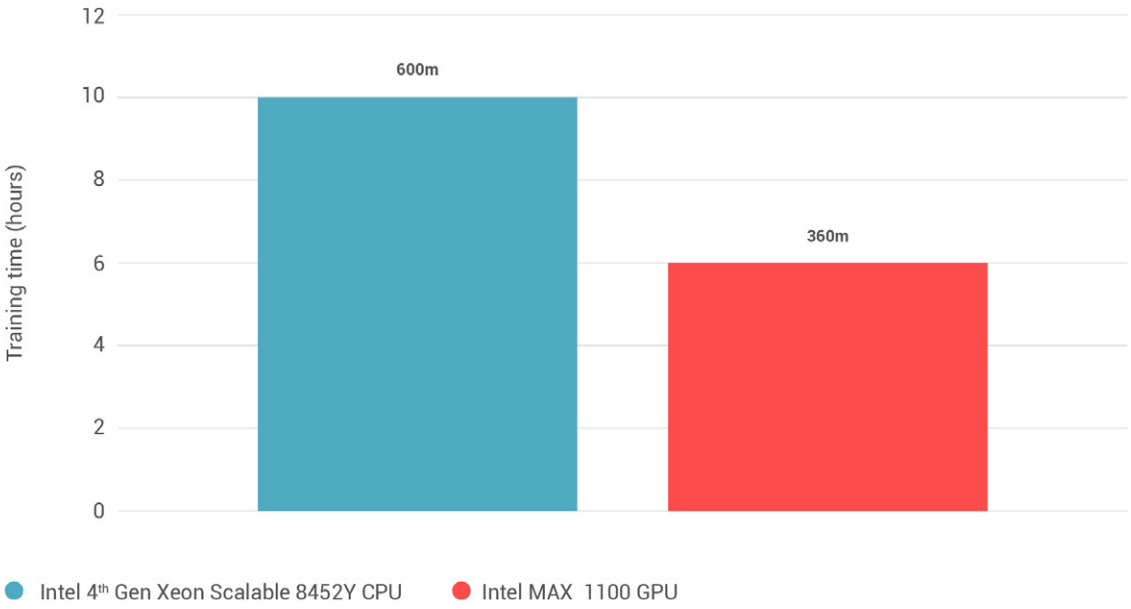


Figure 5. Comparison of ResNet50 training time between 4th Gen Intel Xeon Scalable 8452Y processors and Intel MAX 1100 GPUs

BERT-L Performance (Higher Is Better)

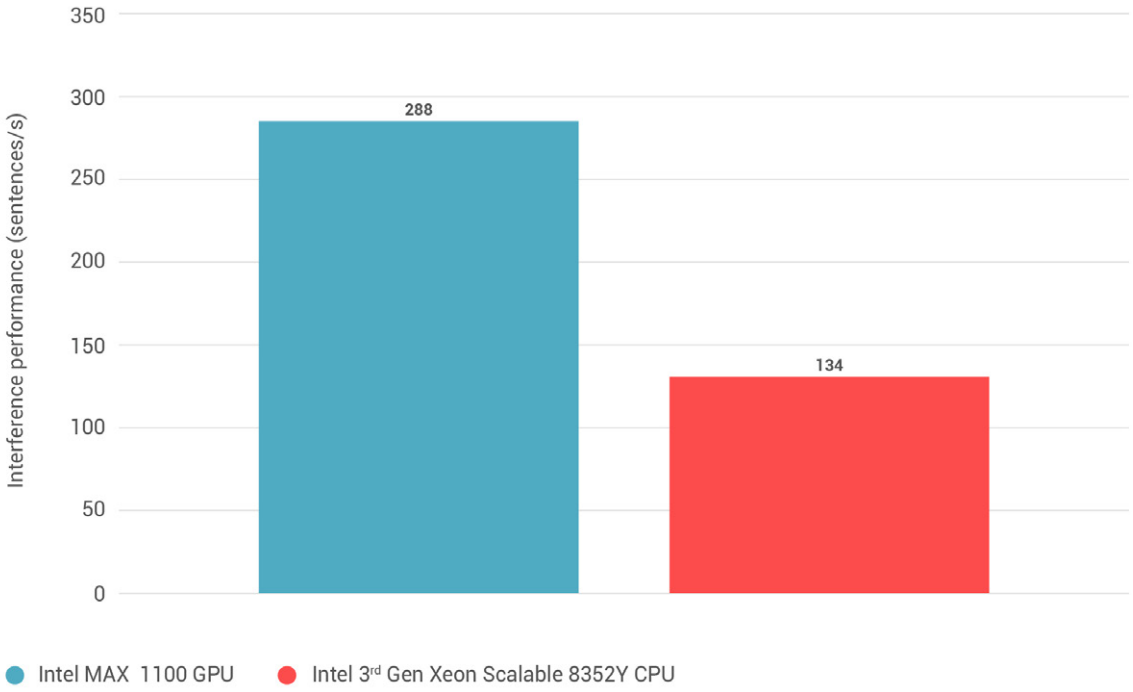


Figure 6. Comparison of BERT-L inference performance (sentences/s) between 3rd Gen Intel Xeon Scalable 8352Y processors and Intel MAX 1100 GPUs

BERT-L Training Performance (Lower Is Better)

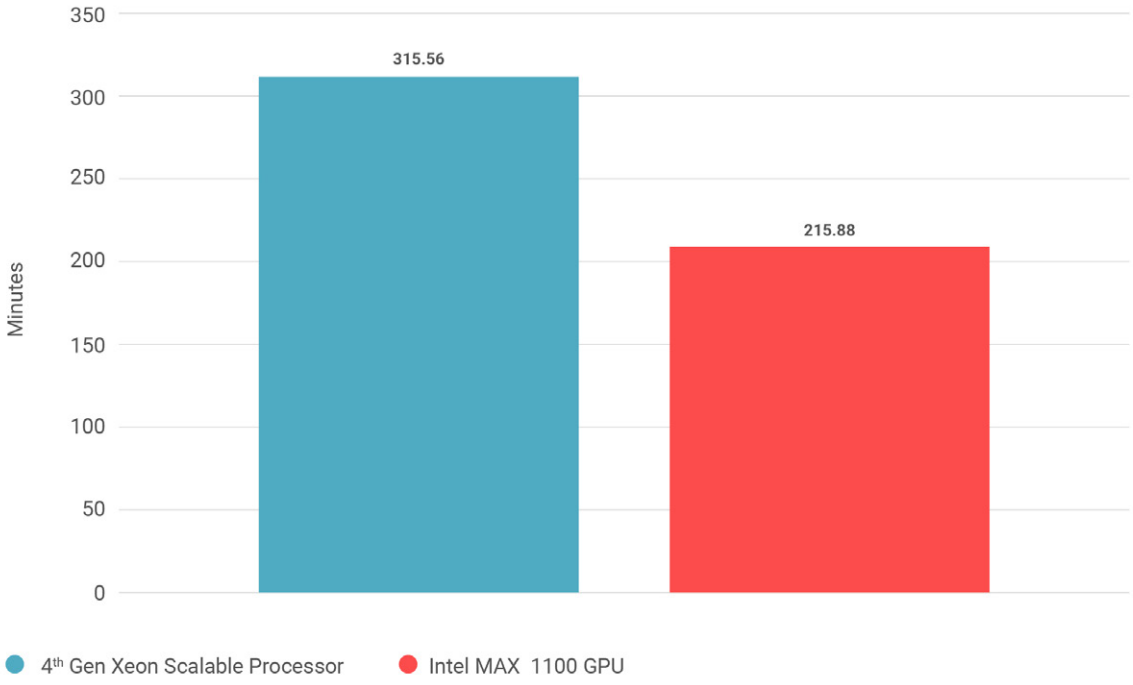


Figure 7. Comparison of BERT-L performance using 4th Gen Intel Xeon Scalable processors and Intel MAX 1100 GPUs

To summarize, **Bare Metal Cloud deployments powered by Intel MAX 1100 GPUs help you to greatly optimize AI workloads**, especially inference on pre-trained AI models and large data set retraining.

V - HOW TO DEPLOY AN AI-OPTIMIZED BARE METAL CLOUD SERVER IN UNDER 5 MINUTES

Deploying a preconfigured Bare Metal Cloud instance takes mere minutes. To spin up a Bare Metal Cloud server for your AI workloads, simply:

1. **Log in to the BMC portal** (or create an account if you don't have one).
2. Click the 'Deploy New Server' button on the Home or Servers page.
3. Select a data center location for the server.
4. Choose your preferred billing model.
5. Pick a desired configuration.
6. Select a preferred operating system.
7. Enter the name and description of your instance.
8. Add a saved public SSH key or assign a new one to the server you are deploying.
9. Configure public IP allocations if you need to access your server directly from the Internet.

If you prefer more in-depth instruction, **refer to this guide** from our Knowledge Base.

Creating an account on phoenixNAP's Bare Metal Cloud is free and only takes a minute. Doing so allows you to browse BMC instances and AI-optimized configurations. **This video** shows you how to get your account up and running!

CONCLUSION

Artificial intelligence is here to stay and is quickly finding its home across industries, helping to drive innovation and propel growth across a range of verticals. However, cost remains a barrier for businesses looking for the infrastructure to support increasingly complex AI workloads at scale.

With as-a-service access to 5th Gen Intel Xeon Scalable processors and 4th Gen Intel Xeon Scalable CPUs coupled with future-proof Intel GPUs, phoenixNAP's Bare Metal Cloud grants you turnkey access to a workload-optimized IT environment that dramatically accelerates AI training and inference while minimizing your TCO.

Resources:

1. 2023 Gartner Hype Cycle for Generative AI
2. Precedence Research, U.S. AI Market Size and Growth 2024-2033
3. ResearchAndMarkets AI Global Market Outlook (2017 – 2026)
4. Forbes Advisor Survey
5. AI Now Institute
6. "Trends in the Dollar Training Cost of Machine Learning Systems" by Epoch
7. Microsoft "Small Business State of Mind" Report
8. Small Business Now Report by Ascend2 and Constant Contact
9. Flexera 2023 State of the Cloud Report
10. iMerit 2023 State of MLOps Report

phoenixNAP | Global IT Services

phoenixNAP is a full service IaaS and colocation provider delivering programmable, OpEx-friendly infrastructure solutions from strategic edge locations worldwide. Focused on innovation, cyber security, and compliance-readiness, phoenixNAP delivers scalable and resilient cloud, dedicated servers, colocation, HaaS, and availability services.

Contact us at **sales@phoenixnap.com** or
call **1.877.588.5918** to find your ideal IT infrastructure solution

phoenixNAP® GLOBAL IT SERVICES