Case Study

Infrastructure Solutions
Intel® Accelerator Engines

Deploy and Manage High-Performance Physical Servers with Cloud-Like Ease

IaaS provider offers powerful processors with built-in accelerator engines to boost the performance of business-critical workloads.

phoenixNAP

phoenixNAP is a global infrastructure-as-a-service (IaaS) provider for customers of all sizes, specializing in performance, security, and availability. Its offerings include:

- Data center colocation at six global locations
- Hardware-as-a-service options
- Traditional dedicated servers
- Private and hybrid cloud solutions
- Backup and disaster-recovery options

All solutions are available on an operating expense (OpEx)–based model to deliver advanced levels of flexibility and scalability, and offerings are personalized to each organization's needs.

Business needs today are highly dynamic, requiring organizations to adopt new workloads and technologies to support their business-critical operations. It would be impossible to talk about new workloads without mentioning artificial intelligence (AI) and machine learning (ML). The explosion of AI and ML is driving businesses to rethink their infrastructure strategies and investments. To ensure that an IT platform can support the demands of these emerging workloads, factors such as performance, security, and availability matter more than ever. But teams might not have the resources to optimize their infrastructures.

phoenixNAP helps organizations optimize their IT platforms to support data-intensive and computationally demanding workloads. Using Intel® technologies, phoenixNAP builds infrastructure solutions for a wide range of use cases such as advertising technology (adtech), making use of AI and ML to give customers a competitive edge. Intel's reliable, innovative solutions and engineering expertise are the key to this partnership. phoenixNAP's Bare Metal Cloud offering, for example, offers automated deployment of the latest 4th Gen Intel® Xeon® Scalable processors with Intel Accelerator Engines, which are designed to help increase application performance.

Initial tests showed performance and cost benefits for selected AI/ML workloads, with similar results expected for other types of workloads.

- The Bare Metal Cloud instances based on 4th Gen Intel Xeon Scalable processors beat the previous generation in speed of execution—in some cases they were almost 3x faster.[1,2]
- When comparing instances with built-in AI accelerators enabled, those with Intel® Advanced Matrix Extensions (Intel® AMX) outperformed those with legacy Vector Neural Network Instruction (VNNI) in each use case.[1,3,4]
- Results showed reduced throughput capacity costs ($/hour): about 43 percent less when using 4th Gen Intel Xeon Scalable processors, and 47 percent less when using 4th Gen Intel Xeon Scalable processors with Intel AMX enabled.[1,5]

AI and ML provide a competitive edge

Adtech software and tools are designed to effectively target, deliver, and analyze data to achieve success in digital marketing. In this highly competitive industry, adtech companies are leaning into AI and ML to get ahead.

Unique to digital marketing is the ability to customize digital advertisements in real time based on variables like user behavior, context, and data insights. Dynamic creative optimization (DCO) campaigns use algorithms and ML to analyze data and adjust messaging, images, and calls to action. This allows advertisers to deliver personalized and relevant ad content, which can lead to higher engagement and conversion rates.

Historically, digital marketing companies have relied on information captured by third-party cookies to improve audience targeting. As more web browsers prohibit the use of this tracking data, adtech can take advantage of AI to create look-alike models from small sets of user data, and then extrapolate to identify target segments.

Supply path optimization refers to a method of suppressing specific ad exchanges to ensure that ad inventory is placed at the most cost-effective price. AI is used to optimize real-time bidding by more accurately predicting the likelihood of user engagement. This can help adtech companies bid more effectively and improve ad performance.

Specialized hardware plays a crucial role in the inferencing phase of AI operations, enabling efficient and accurate processing of large amounts of data and complex computations. Adtech companies, however, are typically focused on software, and they might not have the expertise (or budget) to optimize their infrastructures.

Multiplying the impact of limited DevOps resources

phoenixNAP's Bare Metal Cloud offering allows organizations to make use of the power of physical hardware without many of the complexities—and costs—associated with it. Dedicated, preconfigured servers can be deployed quickly with a single API call. With instances available across the globe and at edge locations, each instance can be matched to unique compute, memory, storage, and networking needs. Because Bare Metal Cloud integrates with popular infrastructure-as-code (IaC) tools, DevOps teams can automate infrastructure management, leaving them free to focus on other priorities. Likewise, accounting teams won't be bogged down with management: Bare Metal Cloud offers simple, transparent pricing options.

phoenixNAP serves those who have specialized workloads and strict performance demands to provide optimized CPU performance while keeping costs down. This makes phoenixNAP an ideal IaaS provider for adtech companies—and any businesses that need to boost the performance of their workloads.

A performance-enhancing partnership

phoenixNAP Bare Metal Cloud offers automated deployment of cloud instances that use 4th Gen Intel Xeon Scalable processors with built-in Intel AMX accelerators. Inferencing performance of two AI models was tested to compare the performance of 3rd Gen versus 4th Gen Intel Xeon Scalable processors and to evaluate the efficacy of Intel AMX.

Two TensorFlow models were used in the testing: Bidirectional Encoder Representations from Transformers (BERT)-Large, an ML model for natural language processing (NLP), and Deep Interest Evolution Network (DIEN), a neural network architecture that can model the evolving interests of users over time. Testing evaluated real-time inferencing speed and throughput.

The first test compared 3rd Gen versus 4th Gen Intel Xeon Scalable processors—specifically, the Intel Xeon Platinum 8352Y processor and the Intel Xeon Platinum 8452Y processor, respectively—with an FP32 inferencing precision. As shown in Figure 1, the 4th Gen Intel Xeon Scalable processors performed better in each case.

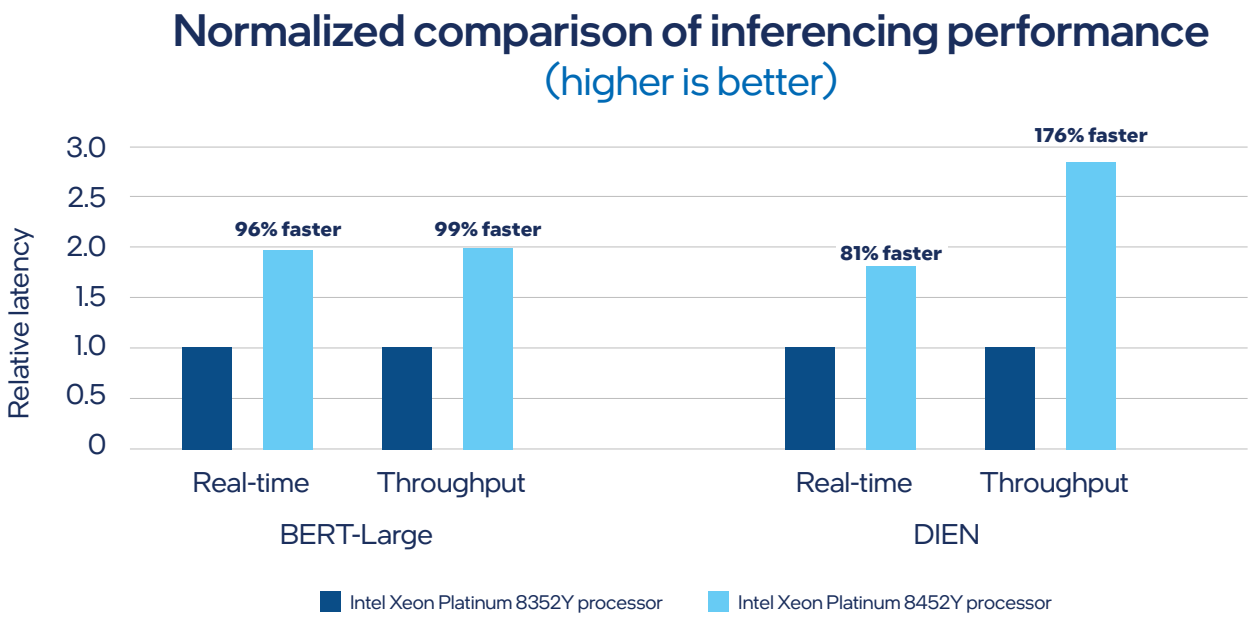

Figure 1. Normalized comparison of inferencing performance of 3rd Gen vs. 4th Gen Intel Xeon Scalable processors with FP32 inferencing precision[1,2]

As shown in Figures 2 and 3, testing conducted on 4th Gen Intel Xeon Scalable processors illustrated the efficacy of Intel AMX over legacy VNNI at varying INT8 inferencing precisions—in some cases 435 percent faster (see Figure 3).

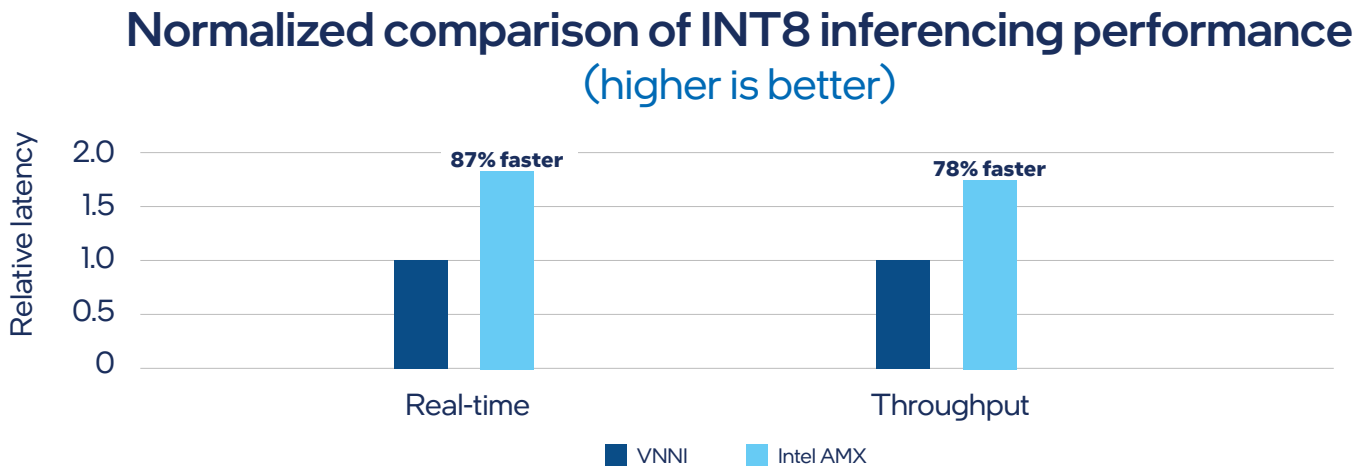

Figure 2. Normalized comparison of INT8 inferencing performance of VNNI and Intel AMX on 4th Gen Intel Xeon Scalable processors[1,3]

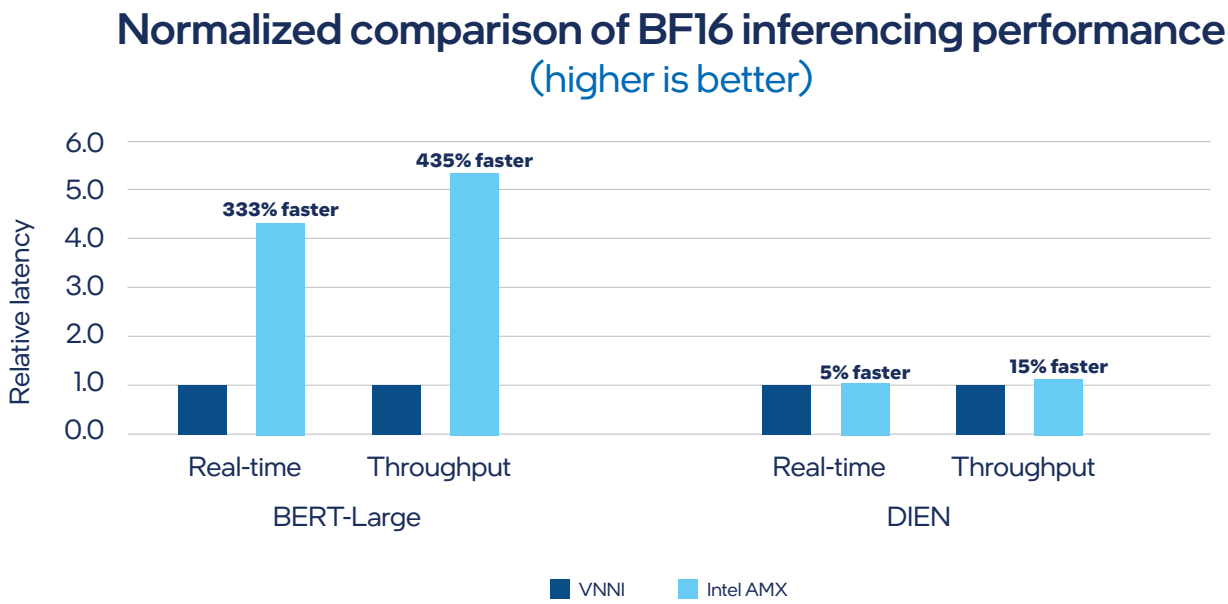

Figure 3. Normalized comparison of BF16 inferencing performance of VNNI and Intel AMX on 4th Gen Intel Xeon Scalable processors[1,4]

Price performance of Bare Metal Cloud

To evaluate the price performance of Bare Metal Cloud, the cost per hour for each instance was considered.

- In the inference real-time use case involving BERT-Large (with FP32 precision), throughput performance capacity costs were 43 percent less when using 4th Gen Intel Xeon Scalable processors, compared to 3rd Gen Intel Xeon Scalable processors.[1,5]
- In the inference real-time use case involving BERT-Large (with INT8 precision), throughput performance capacity costs were 47 percent less when using 4th Gen Intel Xeon Scalable processors with Intel AMX enabled.[1,5]

Workload-optimized solutions from phoenixNAP

phoenixNAP solutions are designed to automate IT processes while providing agility and cost control. Backed by Intel Xeon Scalable processors and Intel Accelerator Engines like Intel AMX , the phoenixNAP Bare Metal Cloud solution offers both performance and cost benefits for common AI/ML workloads, such as those seen in adtech. It is expected that similar outcomes will be observed in other domains like content sharing or e-commerce. Working with Intel helps phoenixNAP stay competitive by offering an optimized infrastructure to accelerate the performance of critical workloads while also helping reduce costs.

Learn more about phoenixNAP.

Learn more about 4th Gen Intel Xeon Scalable processors.

[1] Testing by phoenixNAP as of 03/10/2023. Configurations: **4th Gen Intel Xeon Scalable processor–based system configuration**: Dual-socket Intel Xeon Platinum 8452Y processor; 256 GB DDR5 4,800 megatransfers per second (MT/s); 2 x 2 TB Solidigm DC-P4510 NVM Express (NVMe); operating system (OS): CentOS STREAM 8; kernel: 6.2.2-1.el8.elrepo.x86_64; BERT-Large: precision: FP32, bfloat16, INT8, pre-trained models version 2.10.0; DIEN: precision: FP32, bfloat16, pre-trained models version 2.10.0. **3rd Gen Intel Xeon Scalable processor–based system configuration**: Dual-socket Intel Xeon Platinum 8352Y processor; 256 GB DDR4 2,933 MT/s; 2 x 2 TB Solidigm DC-P4510 NVMe; OS: CentOS STREAM 8; kernel: 6.2.2-1.el8.elrepo.x86_64; BERT-Large: precision: FP32, pre-trained models version 2.7.0; DIEN: precision: FP32, pre-trained models version 2.7.0.

[2] Normalized comparison of inferencing performance of 3rd Gen versus 4th Gen Intel Xeon Scalable processors with FP32 inferencing precision.

[3] Normalized comparison of INT8 inferencing performance of VNNI and Intel AMX on 4th Gen Intel Xeon Scalable processors.

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

[5] Calculations based on phoenixNAP Bare Metal Cloud pricing: Intel Xeon Platinum 8452Y processor–based instance (d2.m6.xlarge) costs $1.92/hour and Intel Xeon Platinum 8352Y processor–based instance (d2.m2.xlarge) costs $1.73/hour as of date of testing.

Printed in USA 0623/KF/PRW/PDF Please Recycle 355470-001US