PROWESS

Technical Research Study

How Your Business Can Be Prepared to Embrace AI Today

Prowess Consulting examined the Intel® AI hardware and software portfolio to uncover AI solutions that can accelerate innovation and help transform your business for an increasingly AI-powered future.

Executive Summary

Today's artificial intelligence (AI) landscape is more multifaceted than ever, offering solutions from a range of technologies tailored to diverse business needs. Generative AI tools like ChatGPT® and DALL-E have captured the most attention, but traditional predictive machine learning (ML) and deep learning (DL) algorithms are the tried-and-true AI workhorses that many businesses count on to deliver business value. While artificial general intelligence (AGI) remains an enticing future prospect, there are traditional and generative AI solutions you can implement today to accelerate innovation, increase customer engagement, improve user productivity, and boost operational efficiency.

In this report, commissioned by Intel, Prowess Consulting examines how you can successfully deploy a best-practice AI implementation for your organization. In the face of such a dynamic market, our key recommendation is to select solutions from trusted AI partners who are aligned with open and industry standards. This is because an open AI platform facilitates flexible hardware configurations and seamless code deployment. Whenever possible, use prebuilt AI frameworks and libraries that are verified to run optimally on your hardware.

Here are a few of the benefits we believe an ideal AI solution should deliver:

- Accelerate innovation
- Modernize the data center
- Span cloud, core, and edge environments
- Secure data, users, and systems

The usages presented in this report illustrate AI's impact across multiple domains, from fraud detection and customer recommendations to supply chain optimization and confidential database searches. The server configurations are built on AI-accelerated hardware and open-/industry-standard software from the Intel® AI portfolio.

Digging into the AI Boom

No doubt you have read about the groundbreaking developments in content creation being achieved by generative AI tools such as ChatGPT and DALL-E. The exploding popularity of these emerging tools is motivating organizations across all industries to investigate the potential of AI technologies to drive innovation, engage customers, improve operational efficiencies, and pursue business growth.

The sophistication and usability of generative AI are grabbing headlines these days, leading to industry debates about the possibility of achieving AGI in the not-so-distant future. However, fantastic predictions aside, the bulk of business value being delivered by today's AI technologies comes from traditional, predictive ML and DL algorithms.[1] One thing this ongoing evolution of traditional and novel technologies calls attention to is that AI is not a monolithic, one-size-fits-all solution; rather, it's a collection of distinctly different technologies suited for different tasks (see Figure 1).

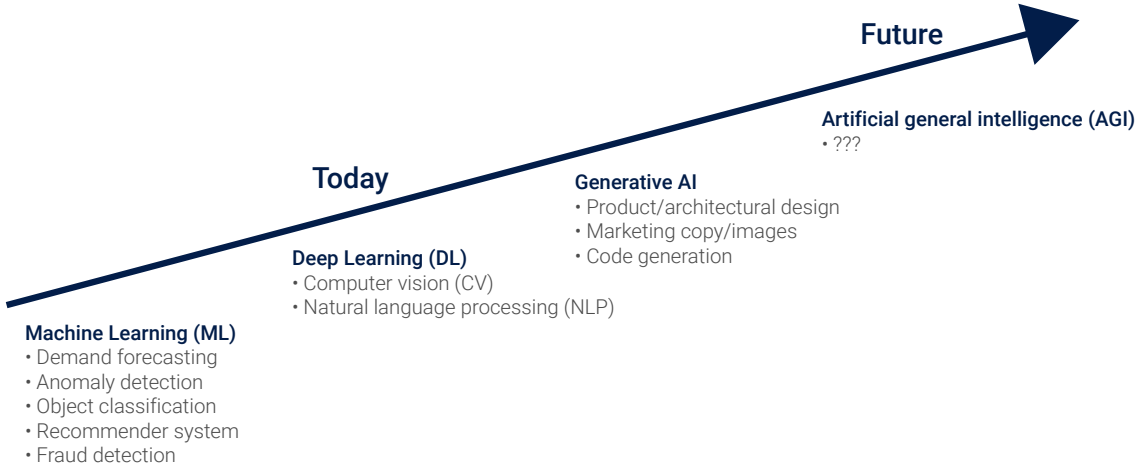

Figure 1 | The ongoing evolution of AI technologies

Why You Need an AI Strategy

Looking past the hype, Prowess Consulting believes there is legitimate business value to be gained from implementing AI in your organization. And we are not alone in this opinion. According to a 2023 Gartner survey, one-third of respondents reported that their organizations are already deploying non-generative AI technologies.[2] A Forrester Consulting study found that three-quarters of the people surveyed were using generative AI to solve specific business problems.[3]

Here are a few business benefits you could get from today's AI solutions:

- **Accelerate innovation and transform your business.** After a few years of struggling to cash in on the promise of the latest technological advances, companies are choosing AI to give their businesses an edge against the competition.
- **Modernize your data center** with the goal of increasing revenue, improving communications, and accelerating workflows. ML and DL are mature AI technologies with a long history of delivering business value. And they keep getting better.
- **Plan and implement a best-practice strategy across your organization.** AI is deployable anywhere—in the cloud, at the data center, and on the edge.
- **Grab an opportunity to upgrade security.** Include a security audit as part of your AI planning and choose solutions that provide multilayered, end-to-end security features.

Removing Roadblocks to Value

We would be remiss if we did not point out a potential impediment to achieving the benefits listed above. The reality is that AI can be challenging to deploy successfully. There are a few reasons for this difficulty. As AI technologies evolve, they require improved infrastructure capabilities to support their expansive possibilities. Your upgraded software and hardware might require more IT expertise to administer. In addition, when AI is implemented and accessed across more of your organization, you might see licensing and administration costs go up. Maintaining global security is another concern as you extend AI into the cloud and out to the edge.

> "[Generative AI] must be done effectively, meaning that while doing it soon enough is important, doing it at the enterprise level and with the right vendor is also critical."
>
> — Forrester Consulting, 2023[3]

To help smooth your road to AI success, we strongly recommend choosing a trusted partner who can give you the best value now and in the future. For example, it would be extremely unusual, not to mention time-consuming, to build AI models from scratch. A much more sensible approach would be to choose optimized frameworks and pretrained models that have been error-checked and verified to run on your servers.

We cannot overemphasize the value of an AI partner who lets you build using open and industry standards. Open-/industry-standard hardware and software support heterogeneous computing, which can help maximize the return on your IT investments. With such an approach, you can mix and match hardware into a combination that best addresses your compute needs without having to completely rewrite code every time you switch hardware brands. And open-standard infrastructure also lets you efficiently write code once and deploy everywhere using open-source tools such as oneAPI. This is a critical capability for executing your AI across the entire enterprise and helping to eliminate application silos.

Planning a new AI strategy offers an excellent opportunity to assess your current security technologies and policies. We recommend including security as an essential criterion when deciding on an AI solution. Make sure your partner offers technologies that protect your assets and expertise that addresses your needs.

The next section presents some examples of how you can use AI to benefit your organization. Each example includes a bare-metal server configured using Intel® hardware and software components. This selection of scenarios is by no means a complete catalog of all the ways you can harness AI for your business. Rather, we suggest using these scenarios as comparison points or as a launchpad to help you in your AI journey.

As you explore these AI examples, remember that a best-practices AI strategy follows a continuum rather than trying to reach an endpoint. Think of AI as an end-to-end deployment that spans your enterprise: data center, cloud, and edge/client. The AI pipeline stages include data prep, model creation, training/re-training, fine-tuning, inference, and deployment. Your tasks for designing, testing, deploying, and managing your AI pipeline are part of an ongoing DevOps cycle.

Exploring the Business Value of AI

Each example in this section includes all the hardware and software components you would need to build a bare-metal server configured to work in a live production environment. This real-world applicability allows us to use performance results to quantify the business value provided by a particular AI configuration.

Detecting Credit Card Fraud

This is an example of an ML server that banks could use to detect fraudulent credit card transactions. Speed to insight is critical in this scenario to minimize or avoid customer losses caused by fraudulent purchases. The server runs compute-intensive inference workloads that process massive volumes of credit card transaction data as quickly as possible. Table 1 shows an example of this server configuration.

Table 1 | An AI pipeline for detecting credit card fraud[4]

Industry and task	Dataset	Learning types (in order)	Model	Output/ result	Intel® software	Intel hardware
Banking and finance—detect credit card fraud	Credit card transactions	1. Unsupervised clustering 2. Supervised ML	DBSCAN clustering LightGBM ML	Fraudulent transactions detected.	Intel® Extension for Scikit-learn Python® daal4py API Intel® Distribution for Python Intel® AI Analytics Toolkit (AI Kit)	3rd Gen Intel® Xeon® Scalable processors

ML and DL are mature AI technologies that organizations have been using for years to detect anomalous and malicious behavior. For this scenario, the ML/DL pipeline is quickly configured using the prebuilt Intel® Distribution for Python® ML framework, the Python daal4py API, and the Intel® Extension for Scikit-learn DL library. Using these prebuilt frameworks helps eliminate writing command codes and helps ensure error-free deployment. After deployment, the Intel® AI Analytics Toolkit (AI Kit) is used to boost clustering performance by up to 22%.[4]

As mentioned previously, you can get enormous labor-saving advantages by building your AI pipeline with open-standards-based software. An open ecosystem facilitates interoperability across hardware and software. It reduces the amount of recoding needed when it's time to scale out or upgrade. Intel builds all its toolkits, including Intel AI Kit and the powerful Intel® oneAPI Base Toolkit (Base Kit), on the open-source, industry-standard oneAPI programming model.

Predicting Customer Purchases

In this example, compute-intensive ML inference workloads processing large datasets are used to accurately predict customer purchases based on customers' shopping behaviors (see Table 2). This type of recommender system can help improve customer satisfaction. For example, if a customer chooses an air filter for their specific model and year of automobile, the recommender system can suggest any additional parts, such as a replacement gasket or cover, that are required for proper installation.

Table 2 | An AI pipeline for predicting customer purchases[5]

Industry and task	Dataset	Learning types (in order)	Model	Output/result	Intel® software	Intel hardware
Retail and e-commerce—predict/recommend customer purchases	E-commerce purchase history transactions	1. Semisupervised learning 2. Supervised learning	K-nearest neighbors algorithm Random forest classifiers	Customer purchases predicted.	Intel® AI Analytics Toolkit (AI Kit) Intel® Extension for Scikit-learn	3rd Gen Intel® Xeon® Scalable processors

This is an ML server built from a prebuilt DL framework, the Intel Extension for Scikit-learn. The Intel AI Kit is used to optimize performance, increasing batch training performance by up to 40% and batch inference performance by up to 95%.[5]

Reducing Product Manufacturing Costs and Time-to-Market

Table 3 shows the configuration for a traditional DL server used to run computational fluid dynamics (CFD) simulations. Computer simulations can lower manufacturing costs and accelerate time to market. A CFD simulation can be used to replace physical product design and testing processes, which can help reduce material waste, improve precision, and lower testing times.

Table 3 | An AI pipeline for reducing manufacturing costs and time to market[6]

Industry and task	Dataset	Learning type (in order)	Model	Output/result	Intel® software	Intel hardware
Engineering and manufacturing—reduce manufacturing costs and time to market	TFRecord file of 3,001 images with random geometric shapes and the profile of a fluid flowing around it (boundary)	1. DL	U-Net architecture	2D velocity vector and boundary information converted to a fluid profile image.	Intel® Optimization for TensorFlow™ Intel® oneAPI Deep Neural Network Library (Intel® oneDNN) Intel® Neural Compressor	3rd Gen Intel® Xeon® Scalable processors

Along with scikit-learn and PyTorch®, TensorFlow™ is one of the most popular prebuilt open-source frameworks for ML.[7] Intel works with Google to co-develop the Intel® Optimization for TensorFlow, an ML framework that includes upstream optimizations for Intel® Xeon® Scalable processors. These optimizations help the fluid profile calculations run as efficiently as possible on the Intel Xeon Scalable processors powering this server.

Intel® oneAPI Deep Neural Network Library (Intel® oneDNN) and Intel® Neural Compressor toolkits help boost DL performance, delivering up to 131% higher training performance after optimization for this scenario. The server also achieves up to 167% faster inferencing after the workloads are optimized and quantized to INT8 arithmetic.[6]

Employing On-Time Delivery Forecasting and Delivery Status Tracking

Numerous studies conclude that customer satisfaction is highly influenced by shipping transparency en-route and reliable on-time delivery.[8] Compute-intensive and memory-intensive ML and DL algorithms can be used to accurately forecast and track supply-chain information. The configuration depicted in Table 4 is for a traditional ML/DL server built on preconfigured XGBoost and scikit-learn frameworks.

Table 4 | An AI pipeline for forecasting on-time delivery and tracking delivery status[8]

Industry and task	Dataset	Learning types (in order)	Model	Output/result	Intel® software	Intel hardware
Supply-chain management—predict/track shipment delivery times	Anonymized dataset from e-commerce company	1. Supervised ML	Voting model regression Voting model classifier	Wait time and likelihood of delay predicted.	Intel® Optimization for XGBoost Intel® Extension for Scikit-learn Intel® AI Analytics Toolkit (AI Kit)	3rd Gen Intel® Xeon® Scalable processors

After configuration, the Intel AI Kit is used to improve real-time classification time by up to 87% and batch classification time by up to 98%.[8]

Improving Natural Language Interactions and Safeguarding Sensitive Information

This is an example of a generative AI server used to deliver keyword search results from massive datasets containing highly confidential and proprietary information. While this is not an out-of-the-box AI solution, it demonstrates Intel's expertise and support in co-developing a state-of-the-art generative AI solution.

Table 5 | Server configuration of an AI pipeline for delivering keyword results from secured datasets[9]

Industry and task	Dataset	Learning types (in order)	Model	Output/result	Software	Intel hardware
Business consulting—improve accuracy of semantic keyword searches and secure confidential and proprietary information	Confidential and proprietary data	1. Generative AI 2. Secure model training 3. Secure inference	Domain-specific foundation model/large language model (LLM)	Retrieve and summarize information via semantic keyword search.	Proprietary production-ready hybrid cloud-scale software	4th Gen Intel® Xeon® Scalable processors and Intel® Gaudi® DL training and inference processors

This solution, jointly developed by Intel and Boston Consulting Group (BCG), exemplifies the four business benefits that a best-practice AI implementation can deliver: accelerating innovation, modernizing the data center, implementing an enterprise-wide strategy, and upgrading security. BCG users reported noticeable improvements in search accuracy, worker productivity, and job satisfaction following the implementation of generative AI. BCG attributes its success to choosing Intel as its AI partner.[9]

> "Generative AI is an emerging and dynamic space, which means organizations must pick the right technology to power their GenAI journey. The technology must be enterprise-grade from day one and allow for privacy, security, ease of use, and scalability."
>
> — Suchi Srinivasan, BCG Managing Director and Partner[9]

Intel AI Hardware

The Intel AI portfolio includes a variety of hardware, including CPUs with built-in AI accelerators, data center GPUs, and training and inferencing processors (see Figure 2). The hardware is designed to provide AI and general-purpose computing across your enterprise and to work in core, edge, and cloud environments.

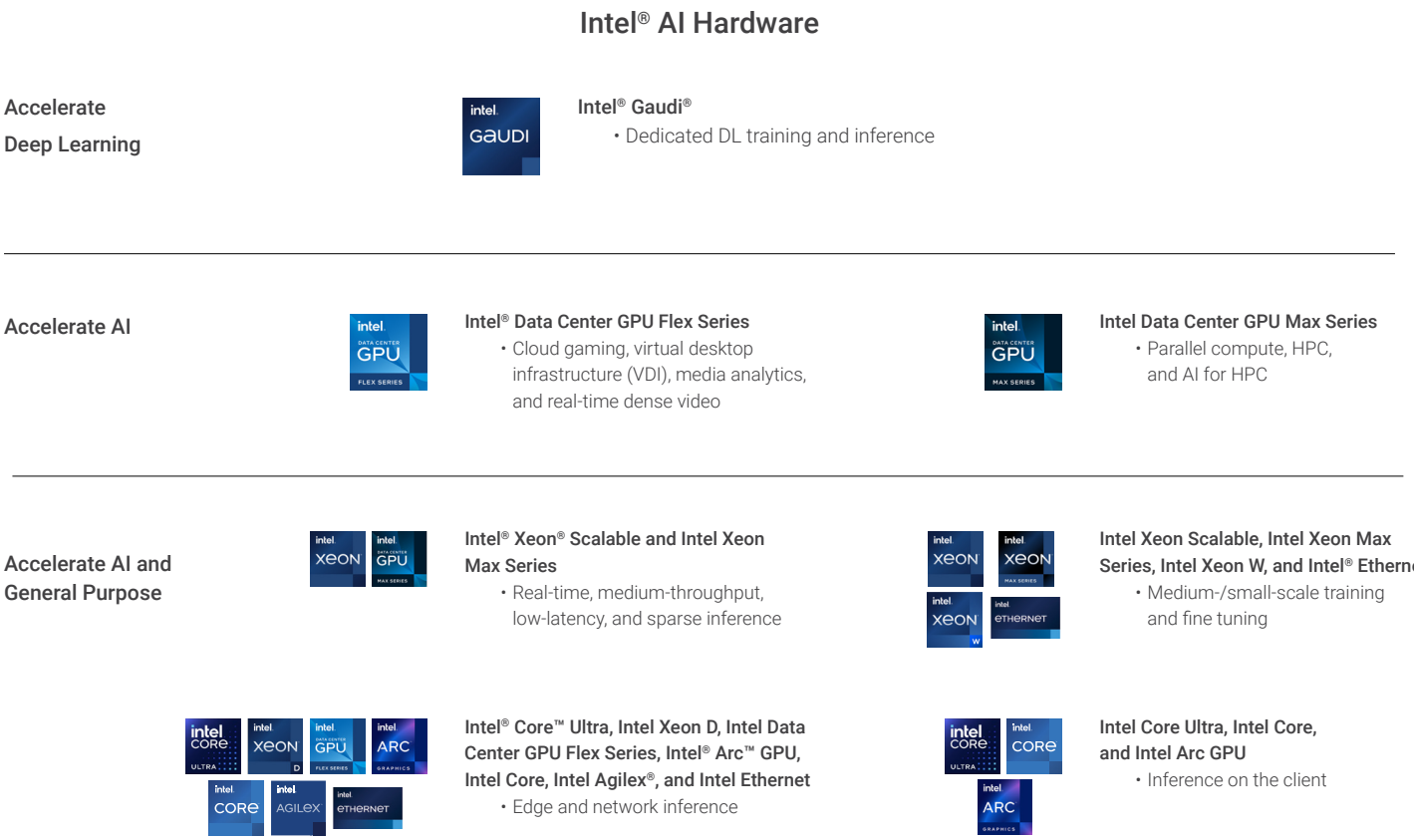

Figure 2 | Intel® AI hardware includes CPUs, GPUs, and dedicated DL processors

CPUs with Built-in AI Accelerators

All the AI server configurations in this report are powered by 3rd or 4th Gen Intel Xeon Scalable processors with built-in Intel® AI Engines that are designed to improve the performance of AI training and inference. Intel's newest AI engine, Intel® Advanced Matrix Extensions (Intel® AMX), is built into 4th Gen Intel Xeon Scalable processors. Intel's latest-generation processors use Intel AMX to deliver high-performance ML inference and training without requiring additional hardware. Intel AMX does this by accelerating matrix-multiplication operations for BF16 and INT8 data types, processing these parallel computations similarly to a GPU.

Intel AMX delivers up to a tenfold performance boost compared to the previous-generation accelerator built into 3rd Gen Intel Xeon Scalable processors.[10] This enhanced performance means that for medium-scale training and inference models, a CPU with built-in acceleration might be all the AI hardware you need to meet customer service-level agreements (SLAs).

If you decide to go with a CPU-based AI solution, Intel testing indicates that the Intel Xeon Platinum 8462Y+ processor with Intel AMX outperforms the 4th Generation AMD EPYC™ 9354 processor for inference workloads by more than 7x, using ResNet®-34 models and meeting an SLA of less than 100 ms.[11]

The 4th Gen Intel® Xeon© Scalable processor with Intel® AMX delivers up to 7x higher inference performance than the 4th Gen AMD EPYC™ server processor.[11]

Intel Xeon Scalable processors also come with built-in Intel® Security Engines, which allow you to implement a confidential computing strategy that protects sensitive and confidential information. Intel® Software Guard Extensions (Intel® SGX) allows you to share AI models from within hardened enclaves without exposing data, systems, and resources to unauthorized access.[12]

Dedicated AI Hardware

Our analysis focused on CPU-based server configurations; however, we should note that Intel offers CPUs and dedicated processors designed to enhance AI performance in network devices, edge devices, clients, and the cloud.

For example, the Intel® Gaudi® DL training and inference processor offers a cost-saving alternative to the NVIDIA® A100 Tensor Core GPU, which soared up to as much as $10,000 in February 2023—if you could find one for sale.[13,14] It can take a lot of GPUs to train generative AI models like GPT-3, which has 175 billion parameters. NVIDIA reports that it took 1,024 A100 GPUs just over one month to train a GPT-3 model.[15] By comparison, MLPerf® Training 3.0 results show that an ML server powered by Intel Xeon Platinum 8380 processors and 384 Intel Gaudi2 processors was able to train a GPT-3 model in 311.94 minutes.[16]

3rd Gen Intel® Xeon® Scalable processors accelerated with Intel® Gaudi®2 processors trained the GPT-3 model in 311.94 minutes.[16]

Intel AI Software: Tools, Frameworks, Guides, and Code

Figure 3 shows that Intel's software offerings run the full spectrum of AI implementation, from engineering codes and recipes to pretrained models to completely prebuilt frameworks to deployment tuning and optimization tools. Intel's AI software strategy is designed to provide expertise, support, and software that helps ensure that you benefit from the full potential of the company's hardware, no matter what stage of AI implementation you're at.

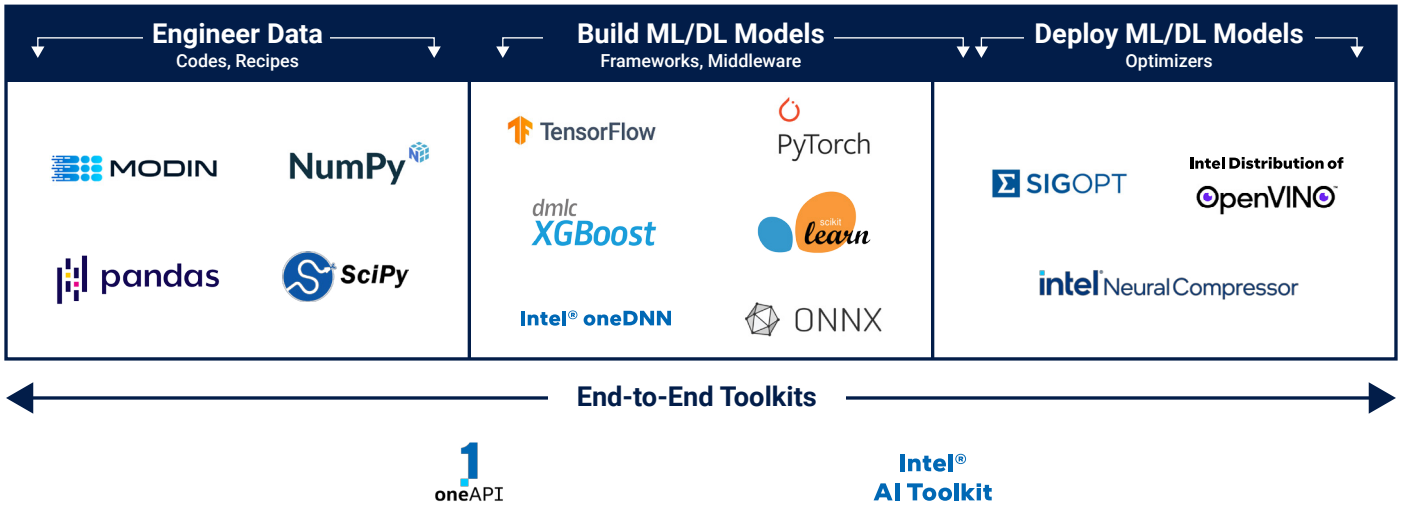

Figure 3 | Intel's end-to-end AI software offerings

Intel collaborates and partners with major AI software developers with the goal of simplifying how their apps are used. For example, Intel offers developer tools and resources that companies can use to scale their AI solutions wherever they are needed and however they are being consumed. These resources include upstream optimizations, pluggable extensions, optimized distributions, and open-source-based toolkits.

Upstream optimizations, for example, are incorporated into the latest releases of some of the world's most widely used AI, ML, and DL configurations. When you install prebuilt frameworks—such as PyTorch, TensorFlow, and many others—you automatically get optimized performance on Intel CPUs, GPUs, and Intel Gaudi processors. These optimizations can automatically deliver huge performance gains, in the 10–100x range.[17]

> "We have heavily optimized these [TensorFlow, PyTorch, scikit-learn, XGBoost, Ray, Apache Spark] frameworks and libraries to help increase their performance by orders of magnitude on Intel platforms … 10–100x software AI acceleration."
>
> — Intel[17]

To bridge the gap between software releases, Intel develops AI toolkits and libraries designed to provide additional performance and enablement capabilities. All Intel AI software is built on the open- and industry-standard oneAPI model to help ensure stability, interoperability, and performance for the latest advances in AI technology.

Conclusion

Driven in part by the explosive growth of generative AI, an increasing number of organizations are using or plan to use AI to benefit their businesses. Properly implemented, AI can help improve user experiences, streamline supply chains, reduce product time-to-market, and detect credit card fraud. One way to effectively harness AI's amazing capabilities is by building with open-standard-based hardware and software. We recommend working with a trusted partner who can help ensure that your AI solution delivers on the promise of transforming your business. Our exploration of the Intel AI portfolio indicates that it offers solutions aligned with these recommendations.

We also believe the time to act is now—to stay competitive, drive growth, and ensure your organization's success in an increasingly AI-powered future.

Learn More

- Begin exploring the Intel AI portfolio by visiting "Accelerate Time to Insight with Artificial Intelligence and Deep Learning."
- Learn more about Intel's AI software solutions at "AI & Machine Learning"
- See more research reports by Prowess Consulting.

[1] McKinsey & Company. "The economic potential of generative AI: The next productivity frontier." June 2023.

[2] Gartner. "What CIOs Need to Know About Deploying AI." June 2023.

[3] Forrester Consulting. "Maximizing Business Potential With Generative AI: The Path To Transformation." Commissioned by Grammarly. July 2023.

[4] Intel. "More Efficient Credit Card Fraud Detection." Accessed August 2023.

[5] Intel. "Efficiently Automate Retail Purchase Prediction." Accessed August 2023.

[6] Intel. "Computational Fluid Dynamics: Calculate the Velocity Profile around an Object." Accessed August 2023.

[7] hackr.io. "Best Machine Learning Frameworks (ML) for Experts in 2023." January 2023.

[8] Intel. "Increase Retail Order-to-Delivery (OTD) Time Forecasting." Accessed August 2023.

[9] Boston Consulting Group. "Intel and BCG Announce Collaboration to Deliver Enterprise-Grade, Secure Generative AI." May 2023.

[10] Intel. See claim [A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[11] Intel. "4th Gen Xeon Outperforms Competition on Real-World Workloads." July 2023. See slides 4 and 14 for details.

[12] Intel. "Confidential AI Data Intel Security Solution." Accessed August 2023.

[13] CNBC. "Meet the $10,000 Nvidia chip powering the race for A.I." February 2023.

[14] Business Insider. "Nvidia GPUs are so hard to get that rich venture capitalists are buying them for the startups they invest in." June 2023.

[15] NVIDIA. "Scaling Language Model Training to a Trillion Parameters Using Megatron." April 2021.

[16] Based on MLPerf® v3.0 training results published on June 27, 2023. **Intel:** Intel® Xeon® Platinum 8380 processor, Intel Gaudi2 processor, PyTorch® release 1.13.1a0, GPT-3 = 311.945 minutes. Source: MLCommons. "June 27, 2023—Training: v3.0 Results." June 2023.

[17] Intel. "Software AI accelerators: AI performance boost for free." Accessed August 2023.

PROWESS

The analysis in this document was done by Prowess Consulting and commissioned by Intel.

0923/230190