intel

Sandra Rivera
Executive Vice President
GM, Datacenter and AI Group

Leading in AI: A holistic approach that is uniquely Intel

Few companies are better equipped than Intel to conquer the production era of AI.

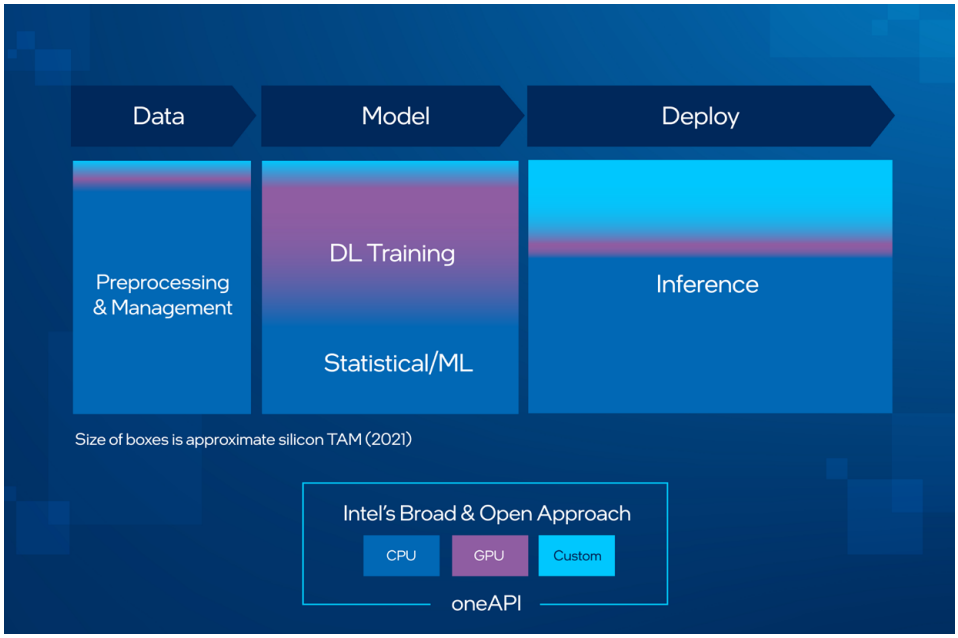
Executive Summary:

- Artificial intelligence is the fastest-growing compute workload, and it's growing in complexity, increasingly requiring more compute, power and bandwidth.
- We're also at an inflection point: AI is moving beyond the data center, and as we move into the production era of AI, it's clear that its future is in the wild.
- AI's proliferation from the cloud to client to edge requires a holistic approach that Intel is uniquely equipped to provide.
- Intel's AI strategy is to accelerate adoption by lowering the barrier to entry for customers. By leveraging the success of Intel® Xeon® and Intel's robust portfolio of architectures and embracing an open software ecosystem, we will be able to not only lead in AI, but to influence broader industry trends to make AI more accessible for everyone.

At its core, artificial intelligence (AI) is the ability for machines to recognize patterns and make accurate predictions based on them. AI models continue to become more sophisticated and complex, and they increasingly require more compute, memory, bandwidth and power.

AI is the fastest-growing compute workload and one of the four superpowers that Intel believes will have a transformative impact on the world. Though it was born in the data center, I believe AI's future is in the wild. The production era of AI on the client and at the edge is finally here, and for AI to proliferate from cloud to edge, the community needs a more open and holistic approach to accelerate and simplify the entire data, modeling and deployment pipeline. Our strategy is to repeat what we've done for other major tech transitions throughout our history: open it up to more customers, speeding the democratization of AI and accelerating wider adoption.

Few companies are better equipped to lead the world into the next era of AI: Harnessing our broad ecosystem, leveraging open software and, crucially, delivering an array of architectures (from CPUs and GPUs to ASICs and beyond) to cater to the vast number of AI use cases will enable us to shape the market and pave the way for ubiquitous, open AI.



An array of architectures infused with AI

When it comes to AI, many jump to deep learning training and GPU performance. GPUs get a lot of the attention because training tends to be massively parallel. But that is just a part of the AI landscape.

A major portion of practical AI solutions involve a mix of classic machine learning algorithms and small-to-medium-complexity deep learning models that are well within the capabilities of modern CPU designs such as Xeon.

Today, the AI data flow pipeline runs mostly on Xeon, and we're making Xeon run even faster with built-in acceleration and optimized software. With Sapphire Rapids, we are targeting to deliver up to a 30 times total AI performance gain over the previous generation, and we're making Xeon even more competitive by bringing more AI workloads onto Xeon to reduce the need for discrete accelerators. For some Intel products, like Xeon, AI capabilities and optimizations are not a new concept, and we plan to expand on this approach, building AI into every product we ship, whether it's in data center, client, edge or graphics.

For the deep learning training that truly performs best on GPUs, we want customers to have the freedom to choose the best compute for their AI jobs. Today's GPUs are closed and proprietary, but we have a domain-specific AI processor in Habana Gaudi, and a general-purpose GPU in Ponte Vecchio that will be based on open industry standards. We're pleased with the progress of Gaudi, with the general availability announcement by Amazon Web Services (AWS) in the fourth quarter of 2021 that Habana Gaudi-based DL1 instances perform up to 40% better than existing GPU-based instances on price performance, and with early Gaudi experience testimonials.

Harnessing an established ecosystem and capturing more customers

Specific models, algorithms and requirements change depending on use case and industry. For example, an autonomous vehicle company needs to solve problems spanning perception (using object detection, localization and classification), high-definition mapping and route planning with actions that need to adapt to a dynamic environment. On the other hand, a chatbot for a technical support application needs to understand the technical jargon of the particular company and industry to answer questions in a relevant manner. AI hardware and software needs also vary by customer, segment, workload and design points. Device, embedded and client AI require low-latency inference systems in constrained power and thermal envelopes, and many, but not all, require assistance with low-code or no-code tools. There's also a growing demand that AI developed in the cloud be edge-aware, so that solutions developed in the cloud can be deployed at the edge (or vice versa).

All of these factors are driving innovation across the board, from the data center to the network and out to the edge, and influencing hardware architecture at the system level, including high-bandwidth and high-capacity memory, fast interconnects, and intelligent software.

The largest growth market within the end-to-end AI workflow pipeline lies within the model deployment and AI inference phase. Today, more than 70% of AI inference is run on Xeon, and one of the fastest-growing AI inference use cases is the intelligent edge, where Xeon has established a strong foothold.

I have spent the past eight months engaging with key customers and learning more about their needs and workloads. These conversations have given us insight into what some of the most influential customers (like cloud service providers) need and have shown us how strategic partnerships can help inform key areas of our portfolio. There are tens of thousands of cloud instances running on Intel today, and it is growing faster than any other architecture. There are hundreds of billions of lines of code written on x86 and hundreds of millions of Xeons installed throughout the industry. Intel is uniquely positioned to propel the industry horizontally with industry standards and vertically in segments like automation and healthcare, where needs are more specialized.

An open software stack for AI developers

We realize hardware is only one part of the solution, and we're taking a software-first mindset with our AI strategy. Software first includes secure AI software components that enable users to leverage Xeon's unique software and security features, like confidential computing via Intel® Software Guard Extension (Intel® SGX), which protects critical data and software while it's in use. Intel SGX is the industry's first and most-deployed hardware-based trusted execution environment for the data center, and our Xeon roadmap includes more confidential computing technologies that will extend our leading position.

We've spent years optimizing the most popular open-source frameworks and libraries for our CPUs, and we have the broadest portfolio of domain-specific accelerators that are built on an open standard to make code easier to port and avoid lock-in — and yet we have more to do to grow our position and move forward. We want to enable open AI that spans from cloud and data center out to client, edge and beyond.

While enabling Intel optimizations in AI frameworks by default is critical to drive broad silicon adoption, we need to meet the needs of all kinds of AI developers. That includes framework developers working at the bottom of the software stack to low-code or no-code subject matter experts working higher up the stack, and all the engineering and operational elements of deploying, running, training and maintaining AI models (MLOps). Though their roles are very different, every stage of the AI workflow shares common desires: get from concept to real-world scale fast, with the least cost and risk. This means choice, and open solutions based on common frameworks that are easy to deploy and maintain.

We have built BigDL, which supports large-scale machine learning on existing big data infrastructure, and OpenVino™, which speeds and simplifies inference deployment to many different hardware targets with hundreds of pretrained models. I know that with consistent standards and APIs, composable optimized building blocks for developers working at the lower levels of the AI stack, and optimized and productized tools and kits for low-code developers, AI developers will thrive with Intel. Our continued investments in AI accelerators and security will enable us to make these critical elements of compute pervasive across all our customers, market segments and products.

Ubiquitous AI powered by Intel

Artificial intelligence is already transforming industries, and it has the potential to improve the lives of every person on Earth, but only if it can be deployed more easily and at scale. Lowering the barrier to entry for AI requires the right collection of technologies infused with AI. Our approach is a winning formula that will accelerate the next era of AI innovations:

By helping to define the developer environment through our open source efforts, we will be able to develop and influence customer solutions that impact entire industries. We forecast Intel's AI logic silicon TAM to be more than $40 billion by 2026. We are tackling this opportunity from a position of strength, and I'm excited for what the future holds.