INTEL'S AI STRATEGY, OFFERINGS, AND DIFFERENTIATION

INTRODUCTION

As attention to AI grows, technology vendors must tackle the challenge of transcending AI hype to create effective AI-driven solutions and convey the value of them for their customers and ecosystems. As the world's largest chipmaker, Intel provides not only chips, but also related platforms, systems, software, development tools, and foundry services to create broader solutions for the entire computing landscape, from datacenters and the cloud to the network, client, and edge devices. The company is enhancing its efforts with a savvy approach to promoting open standards and the desire to increase the trust of both independent developers and its wide range of customers.

Based on our knowledge of Intel's market and our conversations with senior Intel leaders, this analysis explains the viewpoint of Moor Insights & Strategy (MI&S) on Intel's AI vision, strategy, target customers, products, and market differentiation.

INTEL'S AI VISION

Intel's sweeping AI vision, which it conveys with the tagline "Bringing AI Everywhere," is appropriate for a company of its stature. The company believes that AI can improve not only business productivity and entertainment, but every aspect of human lives, from health care to environmental protection. As an integral part of its approach, the company has implemented measures to ensure the responsible use of AI to reduce bias and mitigate other potentially harmful effects. Intel knows that today's trendsetting large-scale generative AI represents only a small fraction of AI's potential and believes that AI in all forms should be accessible and usable anywhere — with full transparency, privacy, security, and trust.

As Intel CEO Pat Gelsinger told Patrick Moorhead, CEO, founder and chief analyst of MI&S, "I think Intel will have a great opportunity as we compete at the high end [of the AI market], but [also] deliver it in volume — AI everywhere for every application on every device for everyone." This ambition is in line with the major corporate renovation that Gelsinger has led since he returned to the company as CEO in 2021. To better position itself against competitors and to take advantage of AI and other market

opportunities, Intel is massively expanding its production facilities and supply chains in Asia, Europe, and the United States.

INTEL'S AI STRATEGY

Intel is one of the biggest players in computing. AI will become a pervasive part of compute in the next few years, so it fits that Intel should be involved in every aspect of it, from the datacenter and the cloud to the network, the edge, and the PC. The history of modern business also shows that most big transformations in technology have benefited from high degrees of openness, for example, the open standards underlying the Internet. Unlike many closed alternatives in the market today, open standards make it easier for everyone to develop complementary new technologies — in other words, to innovate — which makes Intel's commitment to an open platform a wise approach. I should note that "open" hasn't always won (Apple is a prominent exception), but more times than not, it has.

Intel's openness allows its customers to leverage its breadth of silicon IP, platforms, systems, software tools, and foundry capabilities, which are enhanced by the major architectural shift that has brought AI capabilities to Intel's chips. This allows device manufacturers to create new functionality and to gain the efficiencies that come from moving AI compute as close as possible to the data and to end users.

The company is also wise to cultivate the community of external developers in its ecosystem. For many years, Intel has engaged extensively with independent software vendors, system integrators, startups, the open-source community, and academia. In fact, Intel is so strongly identified as a hardware company that it does not often get the credit it deserves for its long history in software — both writing its own and supporting code written by others. Yet Gelsinger describes today's computing environment as "software-defined, silicon-enhanced — because it is the software layer that is defining the user experience [and] the algorithms." Intel's leaders are vocal about developers' critical role in innovation, and the company has invested to support developers at a level that matches that mindset.

INTEL'S TARGET MARKETS FOR AI

Intel serves a broad global customer base with everything from high-end AI training accelerators such as Gaudi and its latest 5th Gen Xeon server processors (specifically designed for AI) to the smallest high-efficiency CPUs for laptops and low-power edge devices. While NVIDIA has drawn more attention in 2023 because of the prevalence of

its GPUs for training large-scale AI models, Intel remains a dominant force for CPUs in datacenters, the datacenter edge, PCs, and everything in between. Intel is extending its reach by using its own AI accelerators (specialized high-performance ASICs), GPUs, and NPUs in tandem with CPUs to provide comprehensive platform solutions at scale.

Gelsinger has talked about meeting the needs of the "AI continuum," which includes not only training the large language models (LLMs) of OpenAI, Meta, IBM, Hugging Face, and the rest but also training smaller models, performing AI inferencing at every scale, and providing compute for other forms of AI such as machine learning and deep learning. As we saw with machine learning, as generative AI grows, inference will be done much more frequently than training. As more inferencing and other AI workloads shift to on-premises settings and even individual personal computers, more of them can be executed using multi-core CPUs with integrated accelerators from Intel, playing to the strengths of Intel's huge installed base and long-term focus on lower TCO. The company expects these trends to drive demand for its products in the datacenter, cloud computing, networking, edge computing, and PC markets.

Challenges Faced by Intel's Customers

Intel's OEM and ODM customers face explosive demand for AI functionality. AI promises to revolutionize nearly every aspect of the computer industry, and manufacturers are pressed to remain competitive while meeting the rising expectations of users. This pressure naturally extends to developers at these companies, who want to increase their productivity by being able to choose whichever tools work best to solve a particular challenge.

End customers for Intel's chips range from hyperscalers offering cloud services, to IT departments running datacenters and networks for their companies or the datacenter edge for retailers and manufacturers, to individual PC users who want better personal computer functionality for productivity or play. In the datacenter market, there is increasing demand for data-intensive clusters to handle large AI training and inferencing workloads. The market for AI in edge computing is possibly even riper, with demand for edge and client inferencing even higher than it is for datacenters. Given the complexity of edge environments, Intel believes — and we agree — that this burgeoning opportunity must be addressed by heterogenous multi-platform solutions. (This is another argument in favor of Intel's open-platform approach.) And delivering on the promise of the AI PC — which could be even more transformative than the adoption of wireless broadband connectivity — will require serious, sustained innovation in both hardware and software.

A major challenge faced at different scales is running AI modeling and inferencing locally without accessing the cloud. This allows a PC user to implement AI functionality much faster while maintaining security and increasing privacy, and without being dependent on a network connection. On the commercial side, running AI on-premises makes it more relevant for businesses by allowing them to use tailored models with their proprietary data for their own use cases. This reduces the time and expense of sending workloads to the cloud and back while maintaining data sovereignty, governance, and security. More than that, it rightsizes workloads. As Gelsinger observes, running inference for a trillion-parameter model on the biggest supercomputer in the world cannot achieve anything practical for most organizations, whereas running a well-chosen billion-parameter model, fine-tuned on proprietary data and running on on-prem servers or PCs, could be highly effective.

INTEL'S AI-DRIVEN PRODUCTS

Since the 1980s, Intel has been synonymous with CPUs. Yet while Intel's advantages as a CPU maker for everything from servers to laptops are clear, it has also invested heavily in other semiconductor IP, especially for AI workloads. These include accelerators — such as the Gaudi family that has grown out of Intel's acquisition of Habana Labs four years ago — as well as GPUs for computationally intensive and more flexible AI tasks, NPUs for the most efficient, longer-running tasks, and workhorse FPGAs and ASICs to handle specialized functions.

All of these designs fold into a disaggregated architecture that we believe is the future of all chipmaking, in which multiple IP blocks (sometimes called chiplets) are included in a single chip, allowing exactly the right block to be used for a given process. By design, these advances are closely tied to the new production nodes that Intel has introduced in the past three years, such as the Intel 4 node for the Meteor Lake family of chips, which reached the market under the Intel Core Ultra name in December 2023. When Gelsinger recently said, "We are ushering in a new age of the AI PC," he was referring specifically to computers enabled with these processors.

Intel's latest performance numbers show its Gaudi AI accelerator beating NVIDIA's bestselling A100 GPU in raw performance, throughput, time to train, cost per token, and power efficiency. While Intel's chips have not overtaken NVIDIA's high-end H100 GPU on raw performance, Intel says that it now beats the H100 on price-for-performance,

providing the only viable alternative to NVIDIA for generative AI.[1] The company hopes that Gaudi will be competitive in LLMs as well. Intel has also announced plans to use 4,000 Gaudi2 accelerators along with Xeon server processors in a new AI supercomputer; the anchor customer for the machine will be Stability AI, maker of the Stable Diffusion image-generation model.

In support of its focus on developers, the company also offers early access to the latest AI infrastructure and solutions via the Intel Developer Cloud. There, developers can easily access not only cloud-based accelerator, CPU, and GPU compute, but also a variety of toolkits (such as oneAPI and OpenVINO), libraries, and AI foundation models. This helps them develop and test software, as well as determine which configuration of silicon IP blocks will work best for running the software in production.

Differentiation for Intel's AI Offerings

Intel understands that AI is not a one-size-fits-all proposition, so it is developing chips and solutions based on the needs of different segments. Its commitment to open platforms and heterogenous architectures will appeal to many device makers and software developers who want to avoid the proprietary lock-in of other vendors.

Given the ubiquity of its microprocessors across the computing landscape, it comes as no surprise that Intel is ready to deploy CPUs for AI tasks large and small. But what separates it from any other chipmaker is that it also covers all the bases with AI accelerators, GPUs, NPUs, FPGAs, ASICs, and development tools.

Many people likely are not aware just how thoroughly Intel bakes AI into everything it does — not only its hardware and software, but also, for years now, the production processes, packaging, power delivery, and other technologies in its world-class fabs and soon-to-be foundries. This focus on using AI to improve quality and yields benefits Intel as well as the chip design companies that use its foundry services.

Moor Insights & Strategy Assessment of Intel's Approach to AI

Notwithstanding the industry hype, AI was not born in November 2022 when OpenAI introduced ChatGPT. More than that, the rise of generative AI does not mean that data science, machine learning, deep learning, and other older areas of innovation are

[1] See "New MLCommons Results Highlight Impressive Competitive AI Gains for Intel," Intel, 27 June 2023; and "Intel Shows Strong AI Inference Performance," Intel, 11 September 2023.

suddenly irrelevant. Technology builds on itself, and we believe that AI will have an outsized and enduring impact on other technologies — and people everywhere — much like previous breakthroughs for the Web, e-commerce, social media, mobile technologies, and the cloud. Intel shares this belief.

We believe that, besides drawing on its world-class — and improving — production capabilities, Intel is operating in a sweet spot where hardware and software intersect. It is enhancing that position through its parallel commitments to open platforms and secure, responsible, trustworthy AI. Even better, its recent production and core semiconductor overhaul supports a renewed focus on bringing its products to market faster.

All of this means that many more computer makers and end customers are going to turn to Intel for AI functionality to create more value from the 75–90% of enterprise data that is still on-premises a decade and a half after the birth of the cloud. In this context, there will be myriad opportunities for inferencing at every scale and using a range of AI compute options, and that's before we consider all the demand for more traditional CPU-driven functions (data processing, data management, and so on) created by new AI projects. Beyond that, the AI-related opportunities for harnessing Intel's flexible chiplet-based architecture are boundless. Manufacturers including Acer, ASUS, Dell, HP, and Lenovo are already taking advantage of these capabilities in ways Intel never anticipated.

The best technologies lead to positive changes not only for businesses and consumers, but for society as a whole. That is what Intel is enabling with its portfolio of chips and software and its end-to-end approach to AI, from the cloud to individual devices and everything in between.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
Intel commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.