ASSESSING INTEL'S DISRUPTIVE AI STRATEGY

SUMMARY

AI is the most disruptive technology in tech today, and in the long run it may create the biggest disruption in the history of computing. Part of this has already been felt in the silicon segment, where a number of major chipmakers have promised to deliver the most powerful or most usable AI functionality. From faster training times on servers to better user experiences on PCs, the new chips running on different architectures that are rapidly entering the market are beginning to deliver on those promises.

This disruption has caused Intel to realign its strategic direction. In 2024, the chip giant continues its reinvigoration under CEO Pat Gelsinger as it sustains ambitious plans to rapidly upgrade production nodes and the company's global supply chain to support a slew of new products. It will need those products to arrive on time, at scale, and with the announced performance to take advantage of the AI boom. Otherwise, it will not gain ground against its primary competitors NVIDIA, AMD, and Qualcomm.

The challenges in AI are significant, but with each passing season Intel seems to be winning over more believers that it has regained the knack for relentless execution that marked its ascendance for most of its history. Now we are nearing the midpoint of a year in which the AI PC presents a high-stakes challenge for many chipmakers and OEMs; at the same time, high-performance chips for AI training and inferencing are prized (and priced) like rare jewels. In this context, how well is Intel living up to its ambitions in the datacenter, in AI PCs, and with its plans for "AI Everywhere"? And what are the risks it faces?

XEON AND GAUDI RISE TO THE TASK OF AI IN THE DATACENTER

DEFINING THE DATACENTER MODERNIZATION CHALLENGE

The modernization of business through AI directly impacts the infrastructure being deployed in the datacenter. As data-driven AI workloads and applications gain traction in the enterprise, the underlying infrastructure must evolve to meet these needs. However, as IT organizations plan for capacity, they must account for the hard limits they will come up against, including available power and physical space.

Adding to these challenges is budget. While IT organizations are tasked with supporting AI and other initiatives, corresponding budget increases don't seem to accompany those mandates where GPUs cost in the tens of thousands of dollars each. Because of this, total cost of ownership (TCO) has never been a more critical and relevant metric for IT.

In response to these challenges, IT organizations look to deploy infrastructure that can handle the broad range of workloads that populate the datacenter, from virtualized instances to cloud-native to AI and beyond, with TCO firmly in mind. This begins with deploying infrastructure optimized with silicon—CPUs, GPUs, and accelerators that can support this broad range of compute requirements while staying within rigorous cost, power, and space requirements.

Xeon 6 — Efficiency and Performance

Understanding the wide-ranging performance requirements of the modern datacenter, Intel designed Xeon 6 processors to fully support these modernization efforts. At the core of Xeon 6 is a system on a chip (SoC) designed for the future. Features such as robust memory capacity and bandwidth, an increased last-level cache (LLC), a large amount of PCIe 5 I/O, and support for compute express link (CXL) can be used for both scale-out and scale-up workloads.

Populating the Xeon 6 SoC is one of two core families. The Efficient-core (E-core) was designed for compute density, meaning the highest number of well-performing cores in an efficient power envelope. Intel is positioning the E-core for a variety of workloads, including cloud-native application development, unstructured data, storage, networking, content distribution networks (CDNs), and the like. This is a platform that can also support highly dense virtualized environments.

While many industry watchers instinctively consider Efficient-cores as exclusively targeting the hyperscale community, Moor Insights & Strategy believes enterprise IT organizations can realize significant consolidation ratios by modernizing older infrastructure to E-core-based servers.

True to its name, the Performance-core (P-core) was designed for raw performance. HPC, AI, database and analytics, performant virtualization, HCI, and the like are all workloads that require high-performing cores to meet the needs of the modern business. This is where Intel targets its P-core. That said, not all high-performing workloads share the same characteristics. Some require fewer cores running at the highest frequencies, while others distribute work among more cores. Because of this, Intel has created several SKUs to support these diverse cases.

Acceleration is another requirement for workloads such as HPC, AI, and analytics. Because of this, P-cores take full advantage of the acceleration engines that Intel introduced in its 4th Gen Xeon Scalable processors.

Using acceleration engines to drive differentiated performance creates benefits for enterprise IT organizations. By designing discrete silicon to offload functions related to CPU-consuming jobs, Intel has enabled Xeon to deliver greater real-world performance while freeing up CPU cores to execute other tasks. For example, AI workloads running on Xeon benefit from AMX's ability to perform matrix multiplication and tensor processing in a highly parallelized fashion. By using dedicated silicon tuned for these functions, Xeon can carry out training and inferencing faster and more efficiently.

The E-cores and P-cores will ship in two distinct platforms—the Xeon 6700 series and Xeon 6900 series. These different platforms are intended to further support the range of workloads powering the modern business. For example, as Intel launches the Xeon 6700E, the CPU core count will scale to 144 cores supported by eight memory channels and up to 88 lanes of PCIe 5/CXL 2.0. By comparison, the Xeon 6900E will support up to 288 cores for the densest compute platform on the market. While Intel has not yet provided memory and I/O support details, MI&S assumes these specifications will increase to enable balanced performance. The Xeon 6900P will follow in cadence, and it will address the most performance-hungry workloads. The 6700P will follow that, offering more diversity in packaging.

While Intel lost its footing in the datacenter CPU market in years past, its resurgence has been singular and significant. Starting with its 4th Gen Xeon CPU (code-named Sapphire Rapids), instead of simply adding more cores, the company focused on architectural differentiation by designing the acceleration engines that deliver application-specific performance optimizations. This design decision was smart because it enabled the company to deliver even greater performance in a way that can't be matched simply by adding more cores—so as the "cores war" reached parity, Xeon would have a competitive position that would be difficult to respond to. Which is exactly where Xeon 6 enters the discussion.

Gaudi — AI Leadership Without Compromise

Generative AI has been a datacenter trend with an incredible amount of hype for good reasons. It has the potential to fundamentally impact the way companies do business and the way people learn and consume—essentially, how they live. For all the enthusiasm in the silicon market for it, it is worth remembering that we are still in the

earliest phases of the GenAI game. While companies such as NVIDIA enjoy the limelight, there is still a long way to go and a big market to support.

Gaudi, Intel's AI accelerator, is built differently—designed and optimized specifically for enterprise AI. While the current offering, Gaudi 2, offers a compelling performance-per-dollar value proposition, the upcoming Gaudi 3 looks to lead the market in terms of raw performance as well as performance-per-dollar.

Gaudi 3 is built on a 5nm process, with a richness and balance of compute, memory, and networking to support the most demanding training and inference tasks. This balance is critical for fetching, processing, and moving data in the AI pipeline. Gaudi 3's specifications have translated into performance benchmarking results that position it as a leader. While benchmarks—especially those provided by vendors—should always be taken with a grain of salt, Intel's message is loud and clear: Gaudi 3 is a player.

Architecturally, Gaudi 3 boasts 64 fifth-generation tensor processing cores along with eight matrix math engines to deliver efficient performance. These cores are fed by 128GB of high-bandwidth memory with a throughput of 3.7TB/s and 96MB of SRAM with 12TB/s throughput. These capabilities, combined with 24 200Gb Ethernet ports, translates into performance: Based on its internal testing, Intel claims up to 1.7x faster time-to-train and an average of 1.5x faster inferencing than the NVIDIA H100 running common large language models.

However, what makes Gaudi 3 most compelling is the performance-cost consideration. As with the most recently announced competing products from NVIDIA and AMD, Intel has not yet published pricing for the Gaudi portfolio. However, Intel has publicly committed to being price-competitive with its rivals, and MI&S can comfortably say that the H100 has been priced in the $30,000 range. Given these two factors, MI&S sees Gaudi 3 as likely to have a strong value proposition. The performance-per-dollar we expect from Gaudi 3 would make it a compute platform that every enterprise IT organization should seriously consider as the foundation of its AI environment.

Gaudi's Market Fit

While NVIDIA, AMD, Cerebras, and others have targeted the cloud providers, enterprise IT craves an effective training platform that is performant, available, and affordable. This is where MI&S sees Gaudi 2 and Gaudi 3 gaining share. However, Gaudi is a bit of an anomaly in that it breaks the cost-performance perception of technology buyers by delivering a demonstrably performance-leading training and inference platform at a considerably lower cost.

A complementary way that Intel has delivered value is through the Intel Developer Cloud (IDC). It is here that enterprise organizations can equip themselves for their AI journey by gaining access to the tools, frameworks, foundational models, and other Intel technology to build out their environment—which can then be brought on premises.

Building and learning within the IDC on the latest Intel silicon allows a much smoother start for the AI journey. This service, while not unique to Intel, is invaluable for enterprise organizations.

How Does Intel Break the NVIDIA Stranglehold?

While not late to the game, Gaudi is certainly behind competitors, especially NVIDIA and AMD, in capturing market attention. However, it's worth repeating that we are in the very early going of the AI game, and there is a lot of market still to capture. This is especially true for on-prem training and inference.

Thanks to Intel's footprint in the enterprise and the strength of the IDC, MI&S believes that Gaudi 3 and Xeon 6 are positioned well. To ensure success, the company must continue to build on its strong OEM, ISV, and channel relationships, because it is through those relationships that AI strategies will be formed and executed. As with other specialized domains, AI will eventually become more familiar and be deployed like other enterprise workloads—aided by bundled solutions configured and optimized by the channel. By establishing itself early in the game for enterprise AI, Intel has a real opportunity to build a leadership position for itself.

Can the AI PC Do for Intel What the Original PC Did?

The First Wave of the AI PC Intersects with a New Refresh Cycle

With the arrival of headline-grabbing products, including Microsoft's Copilot+ and on-device AI PCs, there has been a flurry of activity from chip vendors in the PC space. The battle for AI supremacy is underway, with many chip vendors claiming various advantages ranging from AI TOPS to boasting about how many models their platform supports. Everyone wants to be the vendor with the "fastest" and "best" AI PC chips and claim the leadership crown, driven in part by the OEMs that want to claim the best AI capabilities for their devices. The jockeying within the industry will eventually bring us to the day when all PCs become AI PCs to varying degrees—at which point they all become simply "PCs" once again.

Until then, Microsoft and other ISVs must continue to look to different AI PCs from partner OEMs to bring AI to the mainstream for both consumers and commercial users. On top of the race for the AI PC, many businesses are looking to refresh their pre-Covid and early Covid systems. AI has quickly become a significant consideration for IT decision makers when buying a new PC, even if the purchase isn't immediate. An additional factor driving the refresh cycle is the impending sunsetting of Windows 10 support, which will happen before the end of 2025; it could be easier for companies to replace those older systems now rather than wait.

Intel's Ecosystem Play for the AI PC Is Well Underway

Intel has shown its partners and the industry a clear roadmap for building PCs around Intel Core Ultra chips. The first Intel Core Ultra processors, code-named Meteor Lake, have already shipped more than 7 million units in AI PCs. Intel's roadmap is strong, with Meteor Lake's mobile successor, Lunar Lake, coming ahead of schedule in Q3 of 2024. Just after that, Arrow Lake is slated to continue the company's rapid roadmap execution with a broad release of mobile-to-desktop AI PC products in Q4. Intel claims it will ship inside 40 million AI PCs by the end of 2024, spanning more than 230 designs. Intel also claims that its AI PC platforms already support more than 500 AI models, setting the stage for rapid AI deployments. Security is quickly becoming a killer app, and AI-enhanced security is already emerging from industry leaders, including CrowdStrike, McAfee, Microsoft, and TrendMicro, to name just a few.

Lunar Lake Is Intel's Next-Gen AI PC Architecture

As Microsoft has communicated, the next generation of AI PCs, now dubbed Copilot+ PCs, will need to reach at least 40 TOPS of NPU performance. Intel's Lunar Lake builds on the foundation of the software ecosystem that Meteor Lake has established and adds more performance and battery life. Intel is already claiming significant estimated CPU improvements against the competition, saying that its new CPU cores are faster than AMD's Ryzen 7 8840U and Qualcomm's Snapdragon X Elite. These claims are tied to Intel's new CPU architecture, which includes new Lion Cove Performance-cores and Skymont Efficient-cores.

Into this new generation, Lunar Lake brings an NPU capable of up to 45 TOPS for AI inference and a new GPU with more than 60 TOPS of AI performance, totaling more than 100 TOPS of platform AI performance. Intel's new X^e GPU architecture, code-named Battlemage, delivers significantly faster AI with its enhanced XMX units and 50% faster graphical performance. These performance improvements are joined by big reductions in platform power usage thanks to the advanced low-power island based on

the low-power tile in Meteor Lake. Intel claims up to 30% lower power than AMD's last-generation Ryzen 7840U, almost identical to the Ryzen 8840U. Intel also claims that its Lunar Lake platform uses 20% less power than Qualcomm's 8CX Gen 3, the only current part it can compare against since Snapdragon X Elite systems won't ship until June 18. However, the 8CX is a two-year-old chip running a much older process node and might not be a great comparison.

How Could Lunar Lake Eclipse the Competition?

Now that Microsoft has set the terms of engagement, there is a race to deliver an AI PC that has a performant CPU, runs efficiently, and has an NPU with more than 40 TOPS of AI performance. This entire AI PC race aims to meet Microsoft's Copilot+ PC requirements while also delivering a platform that is competitive with Apple's M-series chips. For example, the 45 TOPS in Lunar Lake is more than the 38 TOPS in Apple's recently announced M4 and should come close to matching Qualcomm's Snapdragon X Elite—at least on paper, since there won't be any direct comparisons for the foreseeable future.

By improving on Intel's successful Meteor Lake across the board, Lunar Lake delivers all of the characteristics of what a competitive next-generation AI PC should have while also delivering the compatibility of x86. That means Lunar Lake won't require any emulation for legacy apps. The timing of Lunar Lake may even present a challenge to Qualcomm's current dominance of the Copilot+ PC, although there is no clear timeframe for when Lunar Lake will get Copilot+ PC software. Lunar Lake will also benefit from Intel's rapid scale-up of the software ecosystem it has built for x86 and AI.

Who Will Want a Lunar Lake AI PC?

Because of the flexibility and diversity of the PC ecosystem, there is no one-size-fits-all solution. This means that Lunar Lake–based AI PCs will deliver different benefits to different types of users. For example, Lunar Lake could attract enterprises using on-device AI capabilities to take advantage of GenAI without sacrificing security or privacy. Creatives can also use Lunar Lake to quickly and cheaply iterate GenAI without heavily relying on the cloud. Additionally, some people may want a laptop that's comparable to a MacBook, but don't want an Apple product. These users may be attracted to Lunar Lake's raw CPU, battery life, and AI performance potential. Lunar Lake systems could also attract users who only want "thin and light" systems.

Lunar Lake's Effect on Intel's Business

This launch should help bring Intel closer to competitive parity with Apple and Qualcomm. With Lunar Lake, Intel continues to build on its 3-D packaging expertise, which it is expanding to satisfy Intel Foundry customer demand. Lunar Lake also demonstrates that Intel's Client Computing Group executes reliably and with competitive products. Indeed, its arrival reaffirms that Intel has turned the corner on product execution. Intel's Lunar Lake delivery should also buoy investor sentiment until Xeon 6 and Gaudi 3 can start shipping in volume.

How Lunar Lake Fits into the AI PC Landscape

Lunar Lake is an AI PC for users who want something genuinely competitive with Apple's MacBook. It will target premium thin-and-light notebooks in over 80 different designs. Intel's Arrow Lake should deliver a broader performance across mobile and desktop. That said, we still need to see how Lunar Lake will compete with Qualcomm's Snapdragon X and Apple's M4 series. AI benchmarks will be difficult to come by early on, but Intel and Qualcomm will hopefully participate in MLCommons' MLPerf benchmarking. Still unknown is how Lunar Lake will fit in among other rumored Arm processor players such as MediaTek, NVIDIA, and Samsung, which will possibly come to market in 2025.

Intel's "AI Everywhere" Approach — What It Really Means

AI Is Just Getting Started

As with any hot trend in technology over the years, the term "AI everywhere" is being used . . . everywhere. With that in mind, it pays to filter out as much of the hype around AI as possible to get to the truth. In this case, the connection between what happens with AI in the cloud, in the enterprise, and on our devices is direct and real. Day in and day out, serious companies are finding genuine performance improvements, whether in their products or in their operations, by the thoughtful application of AI.

We at MI&S see this every day in the companies we cover, in one vertical after another. And the executives we talk to confirm what our own assessment tells us: The world has not even scratched the surface of what AI can do, let alone what AI everywhere will do. So although AI may seem to be at the apex of its hype, the real-world application of it is just getting started. One Intel executive characterized the current market situation as "the bottom of the first inning," and that seems about right to us.

THE IMPORTANCE OF INFRASTRUCTURE AND ECOSYSTEM

One way Intel sometimes characterizes itself is as a “platform company”—which accurately conveys much of its philosophy and history. Intel is a platform-building, system-building organization, whether we’re talking about its global supply chain, its partner ecosystem, the Intel Developer Cloud, the gargantuan fabs it is building in far-flung regions, or the grand project of bringing the x86 microarchitecture into the age of chiplets. Or its approach to AI.

Companies just finding their feet on the AI journey must first consider the underlying infrastructure that enables their AI-powered applications to deliver value, from the datacenter to the network edge to laptops and other devices. With its platform-oriented approach, Intel is well-positioned to deliver the building blocks of that infrastructure, starting with silicon that is designed for performance and efficiency and extending to its partnerships with the OEMs that make the hardware powered by Intel’s chips.

Great hardware means nothing if customers cannot use it to its full potential, which is where Intel’s embrace of third-party developers and its close ties to both ISVs and channel partners come to bear. As touched on earlier, Intel also touts its support of 500-plus LLMs for AI—far more, it says, than AMD or Qualcomm. In many cases, it is the quality and fit of the LLM that set the parameters for AI performance in the enterprise, so Intel’s accommodation of such a broad range of them is another testament to the robustness of its ecosystem. That is a big selling point for customers and a major asset in competition.

INTEL’S MARKET POSITION

Speaking of competition, perhaps Intel's biggest challenge in the datacenter is the leadership positions held by its largest rivals, NVIDIA and AMD. Intel’s relative disadvantage is perhaps most significant in AI accelerators, but the market is so nascent that it's important to note that this disadvantage may be more in perception, not necessarily in technology.

While Intel has plenty of ground to cover in building awareness and positioning for Gaudi, the essential components for competing hard are there: a well-designed part with a performance advantage and a significant price/performance advantage. Further, the company has that unmatched enterprise IT ecosystem—which is going to be critical for its success. Finally, MI&S believes that IDC is an underrated resource that will prove to be invaluable for enterprise organizations learning and deploying AI across the business.

On the datacenter CPU front, Intel's work to rearchitect Xeon for acceleration will be its critical differentiator. Delivering workload performance gains through its acceleration engines will pay dividends for enterprise customers as they use these processors in the real world. As Intel achieves core parity (and leadership) with Xeon 6, this differentiation can be the new measurement of CPU performance.

Intel has a long history of being the market leader in PC processors and still holds a dominant market share. That has enabled it to remain competitive as a company even when its products have not been as competitive individually. Lunar Lake changes the dynamic for Intel at a time when it has more competition than ever, giving it a fighting chance to stave off both Arm and x86 competition in the AI PC era.

CONCLUSION

The momentum we see from the company makes us think that the old Intel—the execution juggernaut—is back in place. Leading indicators have demonstrated its turnaround: aggressive architectural roadmaps that have been executed to a T, accompanied by a process node plan that is unprecedented, yet has likewise been executed against.

In the datacenter, silicon performance is no longer about cores, frequencies, and billboard specifications that are irrelevant in the real world. It's about how fast models can get trained, how quickly business users can get to the right decisions, and how quickly organizations can meet the needs of their customers. Xeon 6 and Gaudi 3 are aimed squarely at those targets.

With the advent of the Copilot+ PC as defined by Microsoft, the PC enters an entirely new era where AI is fundamental to Windows and the user experience. Lunar Lake looks especially well-positioned to enable these new capabilities while still delivering competitive battery life and responsiveness. Intel also has Arrow Lake coming shortly after, giving its PC ecosystem more choices and increased performance levels for power users who want to access all the benefits of the new generation of AI PCs.

The line from silicon to real-world value has never been more direct, and each major silicon player has responded. NVIDIA, Qualcomm, AMD—these are serious competitors that are fighting hard to gain any advantage. Even so, we see Intel in a strong position thanks to its architecture, ecosystem, manufacturing, and channel presence.

This paper is published on the eve of the Computex expo in Taiwan, which promises to be a showcase of rare caliber for the PC industry's wares. AI is sure to dominate the event, and we will be particularly attuned to news from Intel, its ecosystem partners, and its competitors.

The past year and a half of AI mania has brought changes that few could have foreseen, and it is a safe bet that the coming years will be much the same. We will continue to watch Intel closely to see whether it keeps up its drumbeat of sound decisions and steady execution. We believe that the maturity of its technology and the quality of its strategic approach should see it through any upheavals that the market may bring.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR(S)

Matt Kimball, Vice President and Principal Analyst, Servers
Anshel Sag, Vice President and Principal Analyst, Mobility & VR
Tim Walker, Research Director

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.