Hewlett Packard Enterprise

HPE AI Day

Thursday, October 10, 2024, 1:45 PM ET

CORPORATE PARTICIPANTS

Antonio Neri - *President and CEO*

Fidelma Russo - *Executive Vice President & General Manager, Hybrid Cloud & Chief Technology Officer*

Neil MacDonald - *Executive Vice President & General Manager Compute, HPC & AI*

Shannon Cross - *Chief Strategy Officer*

PRESENTATION

Paul
All right. Good afternoon and thank you for joining Hewlett Packard Enterprise's first ever AI Day. We are presenting live from our Wisconsin manufacturing facility, and we are thrilled to host those who have joined us in person, as well as those who are listening to the webcast. Due to security requirements, we are unable to offer a live video feed for those joining online. However, the slide presentation accompanying today's discussion is available on the HPE IR website. In this session, you will hear from Antonio Neri, President and CEO, Fidelma Russo, EVP and GM Hybrid Cloud and HPE's CTO and Neil MacDonald EVP and GM Server. Following our executive presentation, we will take questions from the live audience, moderated by Shannon Cross, Chief Strategy Officer.

Before I pass it to Antonio, let me start with the disclosures. This event may include forward-looking statements involving risks, uncertainties, estimates and assumptions. If the risks and uncertainties ever materialize and the estimates or assumptions prove incorrect, our results may differ perhaps materially from those expressed or implied by such forward-looking statements. HPE assumes no obligation to update such statements. Please find more information regarding our forward-looking statements on our website at investors.hpe.com.

With that, please let me pass it to Antonio.

Antonio Neri
Thank you. All right. Good afternoon and thank you all for making the trip here to Wisconsin, and welcome to those attending online. Today, I will start by sharing with you how HPE is capturing the opportunity ahead of us. It is an opportunity driven by the powerful, disruptive force of AI, which is causing massive recalibration of organization IT strategies. I'm joined by Fidelma Russo and Neil MacDonald, who will do a deep dive into how our strategy positions us to serve the AI needs of growing enterprise segment, as well as how we power AI at scale.

By the end of the day, I'm confident that you will walk away convinced of three things. First, that our portfolio comprises the critical building blocks to deliver on the promise of both AI and hybrid cloud. Second, HPE’s innovation and expertise in designing, manufacturing, installing and servicing air cool and direct liquid cool AI systems of scale and for enterprises sets us apart. And third, HPE's unique, compelling AI value proposition across networking, storage, servers, services and financing position us to create sustainable value for shareholders. Those in the room will notice that we have artists who will capture our conversation in visuals, which we will share after the meeting. The team actually is an extension of our advisory services, and they have worked with us for many, many years and with customers to help their journeys in IT strategies. I hope their work will also help you visualize our unique differentiation.

Before you arrived, I met with Governor Tony Evers, who a few months ago took a tour of this facility. The States of Wisconsin has been a great partner to HPE to help this operation thrive. We look forward to continuing that partnership as we invest in exciting new capabilities, which those who were with us today got a glimpse of that. HPE has a global AI systems-at-scale manufacturing footprint. Our Wisconsin facility is one of the manufacturing hubs and one of the largest direct liquid cool factories in the world. Here, we designed and built the most complex air and direct liquid cool AI systems. This state-of-the-art facility, and the expertise that it houses, is a significant differentiator for HPE.

Since 2018 we have manufactured and deployed more than 200,000 direct liquid cool server notes. And since 2020, nearly 22,000, direct liquid cool networking switches, which in translation, is more than a million ports with our HPE slingshot networking fabric globally, and that enables supercomputing and generative AI. Neil will elaborate on these capabilities in his comments, as well as highlight an industry first direct liquid cool architecture that sets us apart from the competition.

Over the last six years, I have led a transformation to increase HPE's relevance with customers by allowing our portfolio to the powerful industry trends by pivoting our business to higher growth, higher margin areas of the IT market. AI has introduced the biggest technology disruption in a generation, driving what we call a new industrial revolution. AI will require enterprises to rethink their IT strategy and retool their IT estates. HPE's sizing the AI market opportunity to meet customer needs and accelerate shareholder value. We are doing this by innovating in networking, hybrid cloud and AI, which are all essential building blocks to deliver a unified technology experience and accelerate time to value.

We are executing our strategy in a very thoughtful and disciplined way, through a series of both organic and inorganic investments, as well as through curated strategic partnerships across the IT ecosystem. HPE has rapidly, rapidly expanded its AI portfolio with a focus on ease of adoption and sustainability, with an end to end cloud native experience. Our AI leadership is driven from decades of large-scale infrastructure and supercomputing expertise. We have been at this by almost 50 years when Cray built its first supercomputer system. HPE's expertise extends across the design, manufacturing, installation and management above air and direct liquid systems, including margin rich software and services to run and maintain these amazing systems.

Our services capabilities are a real differentiator. I don't believe anyone in the industry has the services expertise we have built from decades of leadership and learnings in this space. As we look ahead, the next generation of accelerated compute silicon will require more power density, which will generate more heat and risks creating a larger carbon footprint. Liquid cooling can cool down chips faster and more efficiently, as water contains three times more heat capacity than air. This allows to absorb more heat emitted by GPUs and the other components, such as CPU, memory and networking switches.

Direct liquid cooling innovation will become essential for the next generation of accelerated compute silicon and networking fabrics. And HPE is a leader in direct liquid cooling across both servers and networking, which relies on extensive intellectual property within the technology itself, as well as in the manufacturing and services processes. We have more than 300 patents in this space. Manufacturing these complex systems is both an art and a science, and those who walk with us today, you saw what it really takes. It requires really a cutting edge R&D and thoughtful choreography.

And again, this will be required as we move to the next generation of this innovation. And the reality is that the skill and the complexity will continue to grow. The AI infrastructure and services market is more than $170 billion opportunity for HPE, comprised of four key segments, hyperscalers and model builders, service providers, sovereigns and enterprises. The hyperscaler model builder and service provider segments have a lot in common. For starters, they are growing very, very quickly. The addressable market for hyperscalers and model builders is expected to grow at a 29% CAGR to about $66 billion through very large capex

deployments. For service providers, the market is growing at a 30% CAGR to roughly $44 billion through 2027 from a 2023 baseline.

These segments collectively comprise a relatively small group of customers who together use more than one million GPUs. They are either retrofitting large data centers or building new ones with tens of megawatts at each side, and soon to be hundreds of megawatts. Therefore, accessibility to clean energy will be critical. But as much as they have in common, there are some very important differences for each. Hyperscalers and model builders are training large language AI models on their own infrastructure, with the most complex bespoke systems. On the other hand, service providers, they are provided the infrastructure for AI model training or fine tuning to customers so they can place a premium on ease and time to deployment. Both customer segments benefit from our experience with large-scale infrastructure and deep data center expertise, and we expect that these customers will be the faster adopters of direct liquid cooling, which is a significant opportunity to HPE.

Customer in the third segment you see here in the slide, what we call the sovereigns, are building AI clouds to support government and private AI initiatives within sovereign borders. Countries are embracing AI technology for national competitiveness and security. We participate this market in two ways, both with our AI systems at scale and also with our supercomputing systems. These customers have compliance and regulatory challenges with the data which our hybrid cloud approach is uniquely positioned to solve. These customers need tens of thousands of GPUs per deployment. They consider the impact of these AI systems to their countries' carbon footprints, so they use renewable clean energy, and that's a top priority for them.

So, what they seek, they seek our expertise in data center and system design, installation and maintenance services. And what we see right now is about 15 countries which are taking the lead, obviously, including the United States. So, the sovereign AI market opportunity is projected to grow at 18% CAGR to approximately $19 billion, again, through 2027. And then finally, there is the enterprise segment, which is moving from what we call experimentation to adoption and is ramping quickly. The enterprise addressable market is expected to grow at 90% CAGR, which would represent a $42 billion opportunity over the next three years.

Our HPE strategy innovation is very well suited to the needs of these segments. AI at the core is a data intensive hybrid workload. Enterprises must maintain data governance, compliance, security, making private cloud an essential component of the hybrid IT mix. HPE has redefined the hybrid cloud space with our HPE Greenlake cloud, which delivers a unified cloud native experience that is hybrid by design. It offers complete industry leading software defined and cloud native infrastructure, software and services that nearly 37,000 unique customers are benefiting from today.

The enterprise customer AI needs are very different from the other three segments I described earlier. Their focus is to drive business productivity and time to value. As a result, they put a premium on simplicity of the experience and ease of adoption. Very few enterprises will be their own large language AI models. A small number might build small language AI models. They typically pick a large language model off the shelf that fits the needs and fine tune these AI models using their unique, very specific data, using also vertical name blue agents, as we call it. We see most of these were loads on premise and co-locations where customers control their data, given their concerns about data sovereignty and regulation, data leakage and the security of AI public cloud APIs at this stage.

In addition to model fine tuning, enterprises are also aggressively deploying inferencing to accelerate time to value. We are seeing a growing amount of inferencing happening at the edge, where the data is generated and processed real time. This is a particular true and vertical such as retail, manufacturing, hospitality and healthcare. We're addressing these enterprise segment needs through a unique partnership with Nvidia. This summer, we jointly announced Nvidia AI Computing by HPE, a portfolio co-developed AI solutions and joint go to market integrations that enable enterprises to accelerate adoption of generative AI.

Our leading enterprise AI offer is HPE private cloud AI, which we made available in early September. It is a full turnkey private cloud stack that makes it easy for enterprises of all sizes to develop and deploy generative AI applications. With three clicks and less than 30 seconds to deploy, HPE private cloud AI integrates Nvidia accelerated computing network and AI software with HPE's AI server, storage and cloud services, all delivered through a unified experience under the HPE GreenLake cloud. It also includes Nvidia name edge of blueprints for multiple generative AI use cases, including industry-specific use cases.

But storing and protecting data is a critical requirement to enterprises to maintain that data sovereignty and compliance, but it requires a modern software defined and cloud native approach, both on premises and in the public cloud. Our HPE GreenLake cloud enables enterprises customers to access what we call a high performing HPE Alletra Storage MP for all data protocols with a flexible consumption model on premises and in the public cloud. For each of these segments, HPE has developed a very compelling value proposition, which both Fidelma and Neil will elaborate in their remarks.

Now, we as a company continue to curate a robust portfolio of IP that intersects these powerful IT inflection points to increase our relevance with our customers and enhance our value for our shareholders. One critical IT inflection point, driven by AI and hybrid cloud, is networking scale. While the key differentiation, obviously, and the major technology driving today is the need to scale more accelerated computing around AI, one of the biggest opportunities beyond servers is networking, with a combined market size of $135 billion by 2027 all inclusive. Customers are confronting the reality of a significant network complexity created in the last two decades.

An increase in the number of connected users, devices, applications, and the explosion of the data we have seen both structure and unstructured, has created significant complexity, performance challenges and cyber security risks. As the network complexity grows, the network must anticipate risks and threats on an expanding and more distributed attack surface. And with increased importance of sustainability to the organizations, the network must also operate more efficiently. And finally, AI's data distributed nature requires a secure, modern networking foundation to connect pools of data for training and inferencing.

What this means, means deploying AI at scale would require a massive step change in the network architecture. And since acquiring Aruba Networks, which I did in 2015, we have invested in building a robust networking portfolio, first by targeting the intelligent edge with a leading HPE Aruba networking campus and branch offering that's mobile first, cloud native and AI driven. And through the acquisition of Cray, which others here in this site, you have seen, we innovated the industry first directly cool interconnect fabric of switches and network interface cards that allowed us to achieve what we call the exascale supercomputing breakthrough. And now, with a pending acquisition of Juniper Networks, we will be able to complete the full, what I call, modern stack networking portfolio of unique intellectual property, featuring what we call the industry-leading secure, AI-native network, purpose-built with AI and for AI.

As model builders and hyperscale as-a-service providers scale their AI clouds, the need for high performance fabric will increase, and AI cloud data centers will also need to be connected to each other, making the interconnect optics a critical requirement to reduce latency and cost. So, the combination of what we already have in HPE with our HPE slingshot, HPE silicon photonics and with Juniper industry leading high performance one data center switches and security offerings, will position HPE to deliver a true complete suite of modern, secure AI networking solutions at scale. We will leverage this innovation as the core foundation to deliver fully integrated hybrid cloud and AI solutions. At the edge, the combination of HPE Aruba and Juniper Mist AI will provide enterprises with the most innovative AI native campus and branch and IoT solutions, leveraging best in class AIOps to deliver the best user experience and the network operating experiences.

And the results will be less time, less cost, less resources and less energy, with an automated AI driven solution stack. And by combining the power of AI ops with our integrated approach to security between the network and the security layers, will help organizations protect devices, data, end users, wherever they are. Our HPE SASE, the Secure Access Service Edge solution, built through the combination of our Silver Peak and Axis Security acquisitions, delivers already a modern, cloud, native secure service edge protection for the distributed enterprise. With our complete purpose built portfolio of high performing networking silicon infrastructure operating system, security software and services, we are positioned to meet the needs of any organization today and in the future.

This new leading networking portfolio will be unmatched in the industry. It will enable HPE to participate in the full $135 billion TAM opportunity in the networking space, which will generate significant shareholder value for the years to come. Once the pending transaction closes, we will be able to share more details about our HPE networking strategy and our product roadmap with Juniper CEO Rami Rahim, who will lead our networking business going forward.

Now, I would like to invite Fidelma on stage, who will talk about the role hybrid cloud plays in AI and a new category of enterprise AI. Neil will follow Fidelma to speak about how HPE delivers AI at scale, so Fidelma.

Fidelma Russo
Good afternoon, everyone. I'm Fidelma Russo. I'm the Chief Technology Officer here at HPE, and I also run the Hybrid Cloud Business Unit. So, I'm delighted to have the opportunity to discuss generative AI in the enterprise with you today. And as Antonio said, hybrid cloud is critical for enterprise AI success. Now, we all know that gen AI will be transformative, and there's already been billions of dollars invested. But one question really lingers, is, when will it pay off? And for the world gen AI investment to pay off, adoption must not only be successful across the model builders, the sovereigns, the fortune 50, but broadly across the enterprise and the public sector market.

And so if you think about it, we need to increase adoption by about 1,000x. It's a big goal. So, what does that mean? So, it means that the number of enterprises adopting gen AI successfully must grow 100 times, and the number of use cases and applications deployed within those environments needs to grow tenfold. Today, however, only about one in 10 AI proofs of concept ever make it to production, and this is due to the complexity of the AI lifecycle, from training to tuning to inference and the hybrid nature of the workload. So, at HPE, we are meeting customers where they are, and our mission is to make gen AI dramatically simpler for enterprises of all sizes, so they can quickly realize the promise of a really game changing

technology. And when enterprise AI adoption happens at scale, it will require a significant overhaul of traditional network, server and storage infrastructure.

So, this new build-out will also drive more software, more services, which creates a big opportunity for HPE to expand our gross margins. So, so far, based on different discussions with customers, the most common use cases that are emerging in the enterprise, including financial services using AI for personalized offers and claims processing, pharma companies focused on drug discovery, manufacturing companies looking to accelerate their product development and increase their factory automation. Now, latency considerations for inference will also drive some of the next wave of decentralized deployments towards the edge.

So, a good example of this is a conversation I had with a very large airline last week, and they are looking at using inference at the edge in digital kiosks. Because I'm sure as you've all noticed, when you go and you check in, it takes some time, and that's because the computing is not at the edge. It's going back to the center and coming back out. And so, these real time applications don't run well enough for us as humans to really engage with them in the way that people want them to do, and Gen AI and inferencing at the edge can help that. But with this decentralization also comes the need for high performance networking and allowing you to connect the locations and transfer the data as quickly as possible.

So, we also see many common horizontal use cases spanning customer support, legal compliance and contract reviews, invoice and expense management, business productivity and content creation. And these examples illustrate the potential for broad adoption across many verticals. So many of these use cases prefer on prem and hybrid deployment models because they rely on sensitive or propriety data or on data that's scattered around enterprises, and it makes it incredibly difficult to move that to a public cloud. Data sovereignty laws and regulations, plus the need for control make a private and hybrid cloud approach preferable.

So HPE, we are also an enterprise, and internally, we are adopting gen AI across multiple parts of our business. So, our IT organization has developed a number of bots to improve productivity, especially around things on password reset, root cause analysis, renewing, security protocols. We use copilot across our development teams for software code development and debugging and also for identifying and resolving security issues. And our services team uses AI for case augmentation, guided troubleshooting, anomaly detection and auto case resolution. And a really interesting factoid that I want to share is that we started our internal chat HPE on the public cloud, but with our usage growth projected to be 3x to 5x next year, we are moving more use cases to our internal HPE private cloud AI, because we have significantly less cost, we get better performance, and we can better support use cases with sensitive data or compliance requirements. We are also at the forefront within our product development of using AI in our products to improve their security, provide predictive analytics and proactive remediation through centralized visibility across the infrastructure and business apps.

Now, based on many customer engagements and our own internal projects, it's clear to us that the three biggest adoption barriers to AI in the enterprise are time to value, how long does it take to realize that, data management, how do I manage all of that data and risk and compliance. And as we scan the market, no solution addressed all three of these barriers, and all of them came with great complexity and long deployment times. This is why, in June, we announced Nvidia AI computing by HPE and our first offering, which is a fully integrated, turnkey system for Gen AI called HPE private cloud AI, represents a new category of AI systems for the enterprise.

So unlike competitors, who focus on reselling GPUs and servers and custom solution integrations that require consultants and many months of work, we took a very different path to create the first AI engineered system. And while others can kind of look at it, deliver the parts for the car, maybe some instructions, we give customers a fully assembled and tested car that's really ready to drive. And the biggest advantage we have comes from the co-engineering of HPE's software, Nvidia's AI software, including Nvidia's NIMS, inferencing microservices, and that's all delivered through our HPE Greenlake private cloud control plane. Our solution creates a differentiated experience for customers and creates opportunities for HPE to grow higher margins with our software, our storage and our platform capabilities. We have already filed 12 patents associated with HPE private cloud AI, and we expect many more to follow.

I was on the road last week, and I spent a lot of time with customers, and the number one reason they want HPE private cloud AI isn't cost. It's because they don't have the expertise to build and run these kind of AI systems in the enterprise to support their business quickly, and their data science teams. And the simplicity of our offering is our strongest unique selling proposition. Since the launch in June, we've seen significant demand from enterprises, and we're really pleased with it. So, the customers also realize that while cost isn't the number one thing, we offer superior TCO versus the public cloud, and our studies show that HPE private cloud AI is up to 75% less expensive than the public cloud.

So, let's dig deeper into how we address the top enterprise Gen AI adoption barriers and why we have so much customer interest. First, it's all about time to value, and our approach involves unique innovation with Nvidia. It involves expanding the partner ecosystem to terminate on private cloud AI, and it involves a cloud like try and buy model and HPE services to deliver the last mile. We deliver a turnkey experience with three clicks to deploy, and in contrast today, it takes enterprises anywhere from two to six months to get started with their AI projects on prem. It's available in four sizes, small to extra-large, involves different combinations of networking, servers, storage and GPUs, and we go from solving simple inferencing problems to tuning and training small models with enterprise data. Enterprises can invest and they can grow capability and they can scale over time.

In summary, the takeaway is we take the guesswork out of the infrastructure with this approach. Out of the box, we offer the full Nvidia AI enterprise software suite, along with a carefully curated set of HPE software and application frameworks called AI solution accelerators, and they let you streamline building, managing and scaling AI workloads. So, let's take an example. Customers don't need to build out a RAG workflow from scratch, it's out of the box. They can preconfigure and deploy digital assistance in under 30 seconds. This allows your data scientists and application developers to start their projects within a matter of minutes, versus a matter of months. Now, we recently introduced our unleash AI partner program in September to build a rich ecosystem of partners, including IFCs, system integrators and service providers. And all of this provides enterprises with a wide range of options for curated, pre-validated partner solutions so they can quickly deploy their applications with the confidence that they will work seamlessly within the enterprise.

So, in addition, we're also in the process of deploying private cloud AI at seven Equinix locations worldwide, and this allows you to get started with low risk. Customers can sign up anywhere from two weeks to three months, in a simple consumption model, perform their experiments. And then when they're ready, they can deploy at Equinix, other data centers, or come on prem. And finally, our advisory and professional services and system integrator partners like Deloitte are helping customers define and implement their business process logic that they need to change to make this productive in the last mile.

So that's the first one time to value. We all know that data complexity was the biggest barrier in digital transformation, and this is even more true when enterprises want to adopt gen AI. Data silos exist across the enterprise and a modern approach to accessing, storing and protecting the data is critical. Data has gravity, and the best approach is for AI to move to the data, not for the data to move to the AI. And so, in order to make that happen, we include our data fabric software in private cloud AI. This helps you address data silos by creating a modern data lakehouse capability, allowing customers to connect to the box with all of their distributed data sources, whether it's on prem or in the public cloud.

And with that built in data fabric, enterprise data sources can be accessed from one place, globally, and we now help enterprises discover, access and prep data for their applications. While this sounds familiar and sounds like data bricks, we differentiate with a hybrid by design architecture. This is the first of its kind, without needing to copy data from on prem or the public cloud, and we support a variety of formats across a multitude of vendors. The technology is already deployed at scale in many large financial and manufacturing organizations, and it drives high growth margin, high software margins and renewals.

So, edge to cloud networking with built in security and technologies like SD WAN will also play a critical role to provide the high performance connectivity that you need to access your on prem and public data sources. So, we've talked about accessing the data, but AI also requires you to modernize your storage infrastructure, and data sets are scaling from terabytes to petabytes and up to exabytes for model builders. So, to keep up with the significant volume and performance demands of Gen AI, the unique disaggregated architecture of Alletra MP Storage allows customers to independently scale for capacity and performance. And customers who deploy the Alletra MP architecture are AI ready, and they have a storage system that provides the performance necessary to feed the processing power of the GPUs and the scale to consolidate, prep and process the largest AI data sets. This Alletra MP architecture can be used as a data storage solution for multiple workloads, and it is the storage foundation for HPE private cloud AI.

So finally, we provide continuous data protection with near instantaneous data recovery times with our virtual product, and this gives you the ultimate protection for training data for your models and for your applications. With our approach customers and enterprises, you don't need to figure out separately your storage architecture, for AI model pipeline and your data management process. We bring it all together. So, this means that a broader range of enterprise customers can focus on time to value from Gen AI.

So, this now brings us to the last barrier to fast adoption. Anytime you move an application from POC to pilot to production, customers need to address risk and compliance. AI ups the ante because enterprises must worry about a whole new set of risks, such as model drift, bias, ethics and hallucinations, as well as a rapidly changing government regulatory environment. And private cloud AI includes sophisticated guardrail capabilities delivered by software IP from Nvidia and HPE, which provide robust governance, access control tracking for your data models and your AI powered workflows.

OpsRamp, our observability software is integrated into the solution. This gives you complete visibility and control over your data, your models, your software and your hardware infrastructure. In addition to many of the ITSM integrations that OpsRamp supports, we now integrate with CrowdStrike to automate and ensure security across the AI stack. We recently closed the acquisition of Morpheus Data, and we're rapidly integrating it into the GreenLake

platform to give customers a common management experience, whether you're on a private cloud or on a public cloud, a true hybrid experience.

And this accelerates our unified hybrid cloud platform. It allows customers to deploy and orchestrate AI applications in a consistent way, and it ensures visibility across the lifecycle using OpsRamp and its AI copilot. In summary, we have a robust strategy to eliminate barriers to Gen AI adoption in the enterprise. We have formed key partnerships, and we have made significant investments in GreenLake cloud. Our solutions and services empower enterprises to unlock the full value of Gen AI, and they differentiate from the competition.

Now, as I highlighted earlier, AI is fundamentally a hybrid workload, and HPE GreenLake is the industry's leading hybrid cloud platform. This platform approach allows us to innovate quicker. It drives improved customer experiences, and it creates cross sell opportunities of other HPE software, infrastructure and services which will grow revenues and drive margins expansion over time. We have created a new category of AI systems with Nvidia to tackle the top three adoption barriers. We offer superior TCO versus the public cloud by up to 75% and we deliver a differentiated solution against our competitors.

Ultimately, the exponential increase in AI adoption will not only accelerate our leadership in hybrid cloud, it will allow us to monetize our GreenLake cloud platform, attach more HPE infrastructure, software and services, and grow gross margins. Now, the strength of innovation in our server business and unique capabilities like this facility and our leadership in directly cooling while directly addressing today, the challenges of the model builders, service providers and sovereigns, will become essential to the enterprise in the future.

And to tell us more, I'd now like to welcome Neil MacDonald to the stage.

Neil MacDonald
Thank you, Fidelma. Good afternoon, everyone. I really appreciate you taking the time to join us today at AI Day. My name is Neil MacDonald, and I'm the general manager of the Server business unit at Hewlett Packard Enterprise. For those of you who could join us here in person, I hope that you have enjoyed the tour of this HPE manufacturing facility, where you saw firsthand the expertise that we put into each system that we built. You also had an opportunity today to visit the museum where you witnessed the history of Cray and the timeline of innovation. But now I want to talk more about the future and how that history of innovation is enabling us to capture the opportunity with generative AI, and how that is driving innovation in compute at HPE.

I'm going to focus on a few key areas and then get to the very heart of our differentiation at HPE. First, I'll cover some technology trends that are occurring and impacting in the market. Then, I'll discuss HPE's leadership in direct liquid cooling. Finally, I'll discuss our leadership in frontier class excess scale systems, the kinds of customers that require that level of performance, and why they're coming to us. Now earlier today, Antonio set the stage for our deep dive into the transformative landscape of networking, hybrid cloud and AI.

As you all know, the AI market is experiencing explosive growth. That will mean massive demand for compute, for storage, for networking and for fully integrated systems. It's still very early in enterprise, generative AI adoption, but certain industries like financial services, media, healthcare, retail, manufacturing, are all making significant investments to pursue generative AI strategies. All of that interest is driving the need for incredible computing performance, massive power and cooling and innovation in connectivity and in storage, to tie it all together, while

delivering reliability at scale. Those are challenges that we at HPE are well positioned to solve for our customers.

But to appreciate the compute power that's required for generative AI, let's begin with a few basics, CPUs, GPUs, accelerators all use lots of transistors to process data. These transistors each generate tiny amounts of heat. When they're operating, that heat can add up, causing the chip itself to heat up. Excessive heat can impair or limit the performance of that chip, and cooling is therefore crucial to reliably operate the system at its highest performance. But it's important to consider how performance and power have changed over time.

You go back less than 20 years to 2007, a typical server in a data center consumed around 330 watts of power, the whole server, and it used processors which had on the order of magnitude of 200 million transistors. Data centers cooled these servers using heat sinks and fans that moved cooler air over hotter metal to extract the heat energy away from the chips and into the data center. And once in the data center, other equipment circulated the air to chillers to get the heat outside of the facility. That's quite an inefficient process, but it was enough to remove the heat generated by servers of that era.

Since then, there have been many, many improvements in chip fabrication, and those have increased tremendously, the density of transistors on a single chip that enhances processing speed, but it also increases the power consumption and the heat production from technology. A single accelerator today is already at 100 billion transistors. Compare that to only 200 million back in 2007. And now, a single server loaded up with accelerators will typically consume over 10 kilowatts, compared to that 330 watts just 17 years ago, and that is expected to double in the next year.

So, consider the implications of all of this. A data center running 10,000 servers, they're all air cooled, they'd be consuming 10 kilowatts each, so about 100 megawatts in total, but that's just the server. Over and above the power consumed to run the devices, the air cooling is also going to consume power, and that could be as much as another 80 megawatts of power, which could cost $56 million a year and produce 268,000 tons of Co2 equivalent as a result of that energy use. Clearly, this volume of heat and power is outstripping the capabilities of air cooling, so what are the alternatives?

Well, clearly, using this much power for cooling isn't sustainable, and it behooves us all to minimize that energy use. This is where HPE's legacy of innovation uniquely positions us to deliver critical solutions for today and for the future. As you saw earlier on our timeline, HPE has over 50 years of expertise and experience in solving some of the industry's greatest computing challenges, and heat mitigation is no different. Let's consider some of the cooling technologies that are available to meet some of these data center needs and constraints.

One option beyond air cooling is liquid to air cooling. We can deploy rear door heat exchangers and self-contained liquid to air cooling racks. In those approaches, liquid coolants are used to absorb heat from the chips, the CPU, the GPU, and then that heated liquid is transferred to a radiator, where it is cooled by air before being recirculated back into the system. For example, the adaptive rack cooling system is a variant of liquid to air cooling but runs liquid cooling through the entire set of systems in the rack, not just the processors. We use these liquid to air methods to cool some of our ProLiant and Cray XD systems, and they extend the amount of energy we can deal with in a rack.

Another option, though, is hybrid cooling, in which liquid cooling cold plates are used to cool the GPUs and the CPUs, while fans are used to cool the rest of the system and infrastructure. Those cooling methods can be used on new or existing systems, but they're not going to be sufficient in the future for the most demanding AI workloads. The most effective way to cool is 100% fanless direct liquid cooling, which we have pioneered for cooling our leadership class, Cray EX systems in the fastest computers in the world. It's the cooling technology that has helped us achieve seven of the top 10 on the green 500 list. This method pumps cooling fluid through cold plates, through the entire system, totally eliminating the need for fans, and as a result, it can also reduce the cooling portion of a data center's utility cost by up to 90%.

Why does that work so well? Water is just much more efficient at transferring heat than air is. Think about this. Think back to the last time you burned your finger. Did you blow on it? Did you stick your hand in a refrigerator? No, you ran your hand under a tap of cold water. Why? Because cool water transfers the heat much more effectively, and it works just the same in a data center. The efficiency will make 100% fanless direct liquid cooling essential for AI data centers, because these systems are being built with the very highest density of transistors operating at the highest level of performance and creating extreme amounts of heat.

Remember that data center that we mentioned earlier, with the 10,000 air cooled servers, each consuming 10 kilowatts and all the extra energy that was needed for cooling? That example is exactly why 100% fanless direct liquid cooling matters. A 100% fanless direct liquid cooled infrastructure based on Cray EX technology could reduce the cost of cooling power needed in that example from $56 million a year to $2.1 and drops the annual Co2 equivalent produced from 268,000 tons to around 10,000 tons, with respect to the cooling. Our patented 100% fan-less direct liquid cooling innovation and our expertise in system design enable us to create combinations of cooling technologies for the most compute intensive applications.

We pioneered that technology in the frontier system three years ago, the world's first system to breach exascale performance. And since then, we have refined it. And today, we are announcing the HPE 100% fanless DLC system architecture. Our high efficiency, high performance system design is the foundation for our cooling system innovation. We're delivering the industry's first 100% fanless, direct liquid cooling system architecture with proven exascale capability, and it's built on four pillars.

The first pillar is an eight element cooling design, encompassing not just GPU and CPU, but the whole server blade, the local storage, the network fabric, the rack and cabinet, the pod in the cluster and the CDU. The second pillar is the high density, high performance system design, complete with rigorous testing, monitoring software and the on-site services to support successful deployment of these sophisticated computing systems. The third pillar is an integrated network design based on a dragonfly topology and connected directly with copper, because copper is not only less expensive, it is also more power efficient.

The fourth pillar is our open system design, to enable us to offer our customers flexibility in the choice of accelerator technology to meet the unique needs of each customer's AI workloads. And of course, these innovations have to create value for our customers. The 100% fanless DLC system architecture delivers a number of unique benefits. Specifically, it yields a reduction of up to 90% in the power used for cooling relative to an air cooled environment. But it also reduces the energy used for cooling by 37% from a hybrid direct liquid cool environment, which uses a combination of liquid and air.

These benefits result in lower utility costs and lower resultant carbon footprint from the energy consumed. It also reduces another kind of pollution at the same time, the noise generated in the data center environment. But because we can support two times the power and server cabinet density of our competitors, we consume half the space. That means that the size of new sovereign systems can be reduced in terms of real estate, and that is not a trivial difference for our customers. Our architecture also features this performance optimized integrated network fabric. And inside the system, the fabric connects components directly over copper. That reduces the cost of expensive optical connections in the fabric, but it also reduces the power for these connections by at least 50%.

Now, this level of innovation requires different factory processes, different capabilities in your facilities than the standard air cooled systems. And the manufacturing footprint in this plant is one great example of our capabilities around the world, and it's also a sign of our anticipation at HPE of the importance of liquid cooling, and the investments that we made ahead of our competitors to scale this capability. That investment has resulted in the shipment of over 200,000 100% fanless direct liquid cooled nodes.

So, we explained the architecture, and we've shared some of its benefits, now let's think about who needs this scale of performance and the associated cooling capabilities, both today and in the future, and why they need it. Model builders require the most powerful systems to train frontier class models, with enormous data sets and the world's largest model parameters. These systems must run reliably, or model training will fail. These systems involve tens or hundreds of thousands of accelerators.

As GPUs continue to scale performance, large systems for this segment will have to be 100% fanless direct liquid cooled in order to reliably run and efficiently run in terms of power consumption on these very large node counts. We have been building systems at scale here at HPE for years, as the people here in this room with us today have witnessed in our facilities here. The model creation, though, isn't just about the computer. It's not just about the AI cluster. You also need data storage infrastructure that has a different architecture than in the past, networking connectivity between accelerators, networking connectivity between nodes, networking connectivity across data sources and across sites. HPE gives model builders the required flexibility for all of these networking needs, including our unique 100% fanless direct liquid cooled Slingshot fabric. The network is liquid cooled.

We're excited about the future prospects for offering our customers additional flexibility with solutions from the forthcoming Juniper acquisition and from our participation in the ultra-ethernet consortium and the ultra-accelerator link technology forums. This will enable our most sophisticated users to have the choice and flexibility to create the generative AI infrastructure that meets their business needs. But to run inference on these trained models for their customers, our model builder customers will also need large scale inference infrastructure. And while the architectural needs are somewhat different than model training infrastructure, density and economics are going to drive a need for liquid cooling and reliability at scale here as well.

Now, some model builders are taking the approach of leveraging third party service providers for infrastructure, either completely or partially. This has created a market of tier two, tier three service providers who deliver generative AI infrastructure services for third parties. These service providers sell on the one hand, to model builders, but also to startups and innovators lacking the capital to build their own infrastructure and to others experimenting with generative AI. These service providers are quickly growing their infrastructures in order to provide the

means for others to develop their AI models, to train them, to tune them and to use them for inference. And these service providers are working to rapidly expand their customer bases.

To do that, they need to be able to provide the latest technology. They need to be able to deploy it quickly. They need it to run reliably, and like every form of service provider, they have to be cost competitive in their cost structure to deliver it. In June, we expanded our partnership with Nvidia, and we promised time to market on new accelerators that will help our AI service provider segment stay competitive in their market. Service providers will also benefit from our innovation and our industry collaborations I mentioned in networking and also in storage required to underpin these services, such as, once again, the 100% fanless direct liquid cooled Slingshot fabric.

Like the model builders, our portfolio of high performance network fabrics will support flexibility for our service provider customers and enable their rapid deployment of generative AI infrastructure. But for all service provider businesses, efficiency is critical. We expect 100% fan-less direct liquid cooling to become pervasive in the service provider market in future, just as we expect it will with model builders who keep their infrastructure in house. Our complete portfolio, our track record, and deploying generative AI infrastructure at scale reduces risk and improves time to market for service providers. And we can also bring them a rich set of services to manage infrastructure on their behalf, including facilities management.

Finally, to accelerate their deployments and to avoid the delays associated with physical data center build outs, we can also deliver our infrastructure in modular data centers, which shorten that time to deployment and simplify their time to market. As you've heard, this tier two, tier three service provider market represents incremental opportunities for HPE, not just in Compute, but in networking connectivity and in storage solutions. Interestingly, there's a class of enterprise customers, sophisticated enterprise customers, who are deploying generative AI technology at scale, and they are resembling tier two and tier three service providers, because they're running a large scale internal infrastructure to provide AI services to their internal users.

Just as we do for service providers and for model builders, HPE can bring the power of our experience to accelerate these implementations and de-risk these enterprise investments. This enables these service provider-like enterprises to focus on achieving competitive advantage from using generative AI by getting productivity gains, business process improvements and enhanced customer experiences. But the majority of enterprises need a simpler approach to generative AI. And as Fidelma described earlier in detail, we are delivering that with HPE private cloud AI. Because we have this rich experience in system design and liquid cooling and support globally, we have an ideal position from which to help customers across these segments.

You heard the list of firsts in the tour of the Cray museum earlier. We have the top two supercomputers in the world, and industry leadership in 100% fanless direct cooling systems which will become increasingly necessary in the generative AI infrastructure build out. Enterprises don't need this today, but as generative AI models grow, and as enterprises test and deploy, their needs for cooling, power and flawless systems are only going to grow. As history has shown us, and as I laid out earlier, with the progression of the technologies involved.

In addition to model builders, service providers and enterprise customers, sovereign governments and national organizations are also recognizing that generative AI is key to their economic and scientific futures. Sovereign organizations want to control their own destiny by owning the infrastructure and controlling sensitive data without relying on another country or on a commercial third party. In the United States, the FAST initiative asserts the importance of

being able to create large language models in order to protect the US economy and national security.

In the United Kingdom, the AI research resource program is providing extremely large machines such as Isambard AI at the University of Bristol, to ensure sovereign AI capability and ease of access to both research and industry. And in Japan, the Ministry of Economics trade and industry has announced that it would be investing $226 million to develop a generative AI focused supercomputer, part of a billion dollar initiative, and HPE is very proud to have been awarded that contract. While sovereign entities do have deep expertise in data science and in computing, they collaborate with us at HPE for our expertise in system design, manufacturing, installation and a full service experience to enable their performance, their scale, the reliability, the resiliency and their efficiency, but not everybody does this.

Some of our competitors are giving customers the pieces from which they could build a spaceship in all of its complexity, whereas we build, test, deploy and support an integrated spaceship experience on behalf of our customers and with our customers. We have a long track record in supporting sovereign infrastructure projects that encompass compute, storage, networking and services. We are inspired by the opportunity to apply our expertise to this broader sovereign technology deployment landscape with generative AI. We're honored to support the sovereign generative AI deployments that we have already delivered, and we're excited about the upcoming deployments and opportunities.

OK. So, let's pull this all together. Investment in generative AI is expanding very rapidly. Its performance requirements are massive and cooling innovations are critical to success. Enterprises need to ride the generative AI wave to remain competitive in their industries. Service providers are quickly entering the AI market and scaling. Sovereign organizations are expanding their AI needs and their specialization, and this all creates an incredible market opportunity for HPE that spans compute, networking, storage and services. HPE is committed to leading the way in AI. Silicon innovation is driving a requirement for liquid cooling, while the increasing size of models and the mass deployment of inference are only going to require more reliability at scale.

All of this can leverage decades of experience at HPE in designing, manufacturing, operating and servicing frontier class systems, leveraging 100% fan-less direct liquid cooling and delivering system reliability and performance at scale to make us a reliable, trusted partner for our customers. We've deployed tens of thousands of GPUs for model builders. We've delivered generative AI infrastructure for tier two and tier three service providers around the world. We've delivered the highest performance frontier class sovereign systems in the world, and we are delivering now sovereign AI infrastructure with systems deployed in the Americas, Europe and soon, Asia.

At HPE, we previously led enterprise customers through multiple technology transitions, like the adoption of client server in Unix, and the widespread deployment of X86 servers across the enterprise. Leveraging all of that experience that we've gained and our knowledge of enterprise needs, we're doing it again with the adoption of generative AI in the enterprise. We're excited about the future, and we look forward to driving continued success across our model builder, service provider, enterprises and sovereign AI segments, encompassing not just compute, but networking, storage and services.

Thank you all very much for your attention. Now, let me hand it back to Antonio.

Antonio Neri
Well, thank you, Neil, and thank you, Fidelma. I hope all of you by now have a vivid picture of the HPE opportunity in AI, as the technology continues to disrupt the current landscape and honestly, unlock countless possibilities for many organizations. And through a very intentional strategy, we have carefully curated a portfolio which we will deliver on the promise of AI and also hybrid cloud, because hybrid cloud is a core foundation to deliver on the AI promise. The pending Juniper acquisition will further strengthen our capabilities with a leading AI native, modern network foundation.

Our innovation and expertise across an entire portfolio of designing, manufacturing, installing and servicing AI systems set us apart. And together, HPE strategy portfolio and differentiate the capabilities position us to profitably capture the significant AI market growth opportunity and deliver value for shareholders. And now what I would like to do is invite Shannon, Neil and Fidelma back here on stage. So, we are going to conduct a Q&A, and I'm sure we're going to get quite a bit of questions and excited, again, to have you participate today here in person, and thank you for those who are online attending the audiocast.

(Audio Gap)

So, Shannon, you're going to lead this?

Shannon Cross
Yeah.

Antonio Neri
Perfect.

QUESTION AND ANSWER

Shannon Cross
OK. Hi, everyone. We'll now move into Q&A before we start, I'd like to refer you back to the risk statement that Paul mentioned at the start of the session. I'd also like to remind you that we don't provide inter-quarter guidance. Therefore, please focus your questions on our technology and our long-term strategy. Also, as we are webcasting, please state your name and company prior to asking the question. So, with that, we will take the first question.

Antonio Neri
I think Adam was first.

Shannon Cross
OK.

Antonio Neri
Sorry. Tim. Yeah, Tim. Tim was first.

Shannon Cross
I know. Got to be faster.

Tim Long
I got the best seat here, so I should go first. Thank you. Tim Long at Barclays. Appreciate all the information. Just wanted to get your sense when looking kind of at the different verticals, a lot of

different strategies on how they deploy the technology and consume it. What do you think direct liquid cooling or fanless or the different iterations means for each of the customer bases, and some of them like to vertically, use ODMs and do it themselves? Do you think what's happening with or what's going to be required with direct liquid cooling will change the dynamic for some of those customer sets where they might need to change the approach and go more towards an HPE integrated solution compared to the way they're currently procuring? Thank you.

Antonio Neri
Sorry, Neil.

Neil MacDonald
So, if you think about the problem statement for model builders, it's about the speed with which they can train the largest models relative to their competition. And in order to do that, they have deliberate computational power. With the constraints on energy supply, there's an incentive to do that as efficiently as possible and deploy the highest portion of the energy that they secure to useful computation. That drives a need for greater efficiency. These customers are already designing infrastructure at data center level. That's already the design focus. This gives them some other choices at data center level of how to go deliver it, by leveraging the work that we've done on 100% fanless direct liquid cooling and the resulting efficiency that they can gain.

The same is true for the service provider market, who also have the same needs and are in very, very competitive situations about the cost of the service they can deliver to their end customers. Sovereign organizations building dedicated, generative AI infrastructure are already deploying today on 100% fanless direct liquid cooled infrastructure and getting those efficiency gains. The enterprise landscape has several different segments, and as I mentioned, there are some enterprises who resemble service providers. And perhaps more modest scale, spanning the very largest enterprises in the world. There, the same motivations exist to be as efficient as possible in the infrastructure that's procured and to manage the costs of running it, including minimizing energy wasted and cooling. But more broadly, as Fidelma covered, the majority of enterprises have a broader set of more fundamental needs around enablement and simplicity in accelerating their journey with generative AI.

Antonio Neri
So, Tim, I think the other part of that is that as we go to this liquid cooled transition, which is a must, our expertise is going to be more seeked by those customers who need data center scale capacity in designing and servicing those systems, versus just thinking the traditional way or the hybrid way, which is not sustainable. And therefore, our engineering capabilities is a very attractive attribute for some of those customers going forward.

Shannon Cross
OK.

Antonio Neri
OK. I guess we're gonna go to here, and then we go--

Shannon Cross
Yeah, we'll go over here.

Antonio Neri
And then we'll go over there.

Aaron Rakers
Thank you. Aaron Rakers at Wells Fargo. I guess maybe first on that question, I just want to maybe be really succinct. Do you think your 100% direct liquid cooling architecture gives you a deepening opportunity to participate in some of these hyper scale build outs, one? Two, you're leaning in on Slingshot, right? And I've asked you many times about Slingshot. Can you talk about where we're at, as far as integrating that, where we're going, as far as going from 200 gig to 400 gig, and really, how that stacks up relative to this inertia round if it's UEC or moving from 800 to 1.6T in the network? Just walk me through where you're at on Slingshot.

Antonio Neri
Yeah. And I would like Neil to answer the Slingshot because it's all under his shop today. And obviously, with the pending acquisition, Juniper, we think we can go faster in many of those transitions because the amount of talent and additional resources they can bring in addition to other IP that we're going to use down the road. But on the hyperscalers, I will say the following. The division of hyperscaling AI is shifting. It's not just the three large clouds. Model builders on its own, they are becoming hyperscalers, right? And so, the reality is that we already today support one with a different configuration. And then down the road, we believe we're going to be more attractive to some of those model builders for what they're doing, because some of those will be trying to optimize the infrastructure to the way that run the workload on itself. And therefore, I believe we think we're going to get much more advanced at that.

OK. And then let's remember also the run and the servicing is as important as the design and the manufacturing, and that's why the expertise is not just optimizing the system architecture with fanless direct liquid cool, but it's also the runtime, when you run these models. One of the key important factor is that you start the model, and you want to finish the runtime of the model. If you start and stop, start and stop, you're wasting time, energy and I will say, cycles for all their needs. And one of the things that HPE know had to do very well is running from beginning to the end. And some of those techniques are what we call internally checkpoint techniques that has a lot of software associated with that, including the contention management on the network, which is one of the key elements that we already use in our Slingshot.

But what is the next transition here, Neil?

Neil MacDonald
So, we continue to accelerate our work around slingshot and the development of that technology. You talked about line rate speeds, and of course, one of the key things to understand is that Slingshot's intrinsic efficiency delivers great performance from the line rates that can't be delivered at the same line rates by traditional ethernet. We enjoy an environment where there's a lot of innovation in networking fabrics. And as I alluded to in my remarks earlier, there are multiple different fabrics involved in building these infrastructures, fabrics between accelerators, fabrics between nodes of accelerators, fabrics from those clusters to the storage and fabrics between sites and locations.

We're going to continue to innovate across all of that, and Slingshot will play a part in some of those network fabrics, as will other networks that we leverage in partnership with our technology partners. When you think about the key, though, the key is that if you want the efficiency in the infrastructure, you can't just partially liquid cool a server and leave a whole set of the server infrastructure air cooled, and all of the switching air cooled, and all of the storage air cooled. So, a key principle of our 100% fanless direct liquid cooled system architecture is that the infrastructure in its entirety become 100% liquid cooled, so that we can gain all of these benefits

for our customers and deployment, both in capex and in terms of the run cost of the infrastructure. So, Slingshot continues to play a role in that, as do our other fabrics.

Antonio Neri
And I would say also, Aaron, that there's multiple trends, right, 200, 400, 800, those are all going. And what I'm excited is that, obviously, when I think about the entire data center architecture and the interconnects between data centers, right? And you saw my chart that Neil at the end had the chart, graphically trying to simplify things. You need the pipe inside the data center. That's a big opportunity for a routing perspective. And obviously they are leading when an 800 gig is going to be essential. Juniper already has the first 800 gig router in the market. Then when you go to the data center, fabric itself, before you go into the rack, that's another 800 gig opportunity, and obviously, Juniper has been working on that.

And then when you go to the rack, to the Neil's comments, right, you're going to have multiple fabrics. But the fact of the matter is that our HPE Slingshot is going to play a massive role, also to give customers some flexibility. Then one of the things I'm really excited that we haven't spoke here, but Silicon Photonics. Silicon Photonics going to play a role in the interconnect side and in Fidelma's HPE Labs, one of the key technologies that we have been working for a number of years is Silicon Photonics. And so, the ability to capture that part of the ecosystem through the Silicon Photonics foundation is going to be very important as well.

Shannon Cross
OK.

Antonio Neri
All right. You want to go here? Here and there. And we go to Daniel there.

Asiya Merchant
Asiya merchant, Citi. Just wanted to ask, I mean, it's very impressive facility here. I get it that you do probably DLC at your product facility as well. The model builder opportunity and the other markets, do you feel like the growth rates that are projected there can be efficiently supplied through these two facilities? Is there a need for more investments to kind of go after that and those markets in a big way, especially the tier two, tier three providers. And then for Fidelma, the storage opportunity, how should we think about the attach for Alletra in the model builders, as you guys are scaling into that vertical? Thank you.

Antonio Neri
Yeah, I think we're going to continue to scale the capabilities, whether it's here or other location. In fact, we started the European years back because we had to meet the needs of customers closer what they are. And the reality is that over the last six years, we have made significant capital investments, which you saw throughout the tour, right? And more is needed as we go forward, as the number of systems move from what is today, a hybrid approach to a complete 100% liquid cool. And so, we have the capacity to do that here in this site. That was one of the reasons why I met with the governor this morning.

He has been an incredible supporter of making sure that the regulation in the state allows us to get access to the right power. A lot of the power that we get today is clean power here. Obviously, we have the river next to it, which the great state of Wisconsin, Michigan, has a lot of water around it. So, these are things that we consider. But remember, as Neil said, who are the fast adopter of that technology is going to be mostly the sovereigns, because they already are there, and they're going to go to the generative AI and then model builders, and then we'll find a

way after that. And so, we will invest and scale that investment as needed as we go forward. That's not the issue. That's not the constraint that we may have. And the reality here we can build many, many systems concurrently. And you see that the number of equipment, both on the floor and testing beds and the like, with the capacity that we had, which is, as I said, between the two combined sites, already 86 megawatts. Next year, we're going to exceed easily the 100 megawatts. We will make those a capital investment. That's not the issue. But to Neil's point, we anticipated that many, many years ago. So, which means we have a head start, because building those pipes and the pumps takes a long time.

And then, Fidelma, you want to talk about storage?

Fidelma Russo
So, I mean, in terms of the need across all of the verticals is to feed the GPUs as fast as possible. OK, because you have to keep these things fed. And so that has forced kind of a rethinking around storage architectures and beat us and then manage the data. So that's the need for disaggregated storage. And so, we have things like the data fabric, which I mentioned, which is GPU direct certified. And so really, what that says is, I can pass the test that feed the GPUs as fast as possible and really operate, really, the next tier of memory. But as we look forward, in terms of model builders, like we're working with Neil is, how do we take our Alletra MP architecture and continue to scale it? How do we find interfaces like Fast Object? Because as things move from file to object, which is really the protocol that things want to get consumed in, going forward, we're going to see that adoption all the way from the enterprise up to the model builders.

Shannon Cross
Great. Meta?

Antonio Neri
No.

Shannon Cross
Oh, sorry, and then we'll get you. Sorry, Mehdi.

Antonio Neri
Yeah, either way.

Shannon Cross
Yeah.

Antonio Neri
All right.

Mehdi Hosseini
Thank you. Mehdi Hosseini from Susquehanna. Two questions. If I were to think about the component level, you have kind of a fragmentation DGX, HGX and MGX. And how do you map that out from a component level to your private cloud? How should I think about who is driving what? And then as a follow up, which I don't want to get into financials, but as you think about the scaling your private cloud, how does monetization come in? It seems like you're trying to offer various services along with products, but I'm not so sure if I understand how you go about capturing some of the TAM that you highlighted and how the monetization happens.

Antonio Neri
Yeah, I don't know. I think you both have to answer that question from an enterprise perspective.

Fidelma Russo
Yeah. So, from an enterprise perspective, the way we approach it, if you look today, and it's all about systems, OK? So today, when enterprises deploy workloads, they buy servers, they buy compute, they buy networking, they buy software. And the premise of actually going to the public cloud a number of years ago was, hey, we make all that simple for you. And so, and now, you have to bring AI to the data, and especially on prem, you have to think about it as more like an instance on prem, and it's made up of carefully curated servers, networking and storage, and Neil will talk more about his, like, the server provider.

And then, how do we monetize it? We monetize it by the value that's in all of the software, so all of the automation, all of the quick start, all of the NIMS that are deployed, the fact that you have a ready to run stack, and when we talk to customers, they really get that value proposition. So that's the monetization process. And then on top of that, it's advisory services and deployment services on prem.

Antonio Neri
So, think about it this way, there is a day zero, day one, day two. Day zero, obviously upfront advising, like they are doing here, not defining what that journey is from a strategy use case and IT architecture perspective, then pick the right recipe. In our case, most of the enterprise customers will adopt the simplicity of deployment that Fidelma just went through. Because the reality, when you look at these four config, Fidelma is up to how many hundreds GPUs? You can go higher. But the reality going to start with 64, 72 or 128, and then go from there. Now, if you have a large financial institution, that may want a pod of 500 GPUs at the time, because they're doing risk modeling or some sort of large language model. That, we have done with a DGX solution, right?

And by the way, we introduced that solution last December, at the HPE discover Barcelona, and we have sold, obviously, a number of that. But as we go down market in the traditional enterprise nomenclature, as we know and you saw, I thought Fidelma did a fabulous job showing you the use cases, the size and the scale needed is much smaller because you're either doing a fine tuning or you're doing a small language training model, or you're doing an inferencing on the same solution and or a RAG training environment. And that's with few 100 GPUs, you can go buy like we do.

In our environment, we have very small amount of GPUs to do all the use cases Fidelma described, because we took the LLL model, and we brought it on prem. But now I get better accuracy by giving very specific data that's relevant to us with potentially less even parameters, which is more accurate sometimes. So have to find that balance, but obviously we can offer that flexibility to enterprises.

Fidelma Russo
And I think one of the things as I've spoken to a lot of enterprises over the last couple of months, there's enterprises that will want just a private cloud, but many of the larger ones, they have catalogs that they provide to their users, and some of them are at the tippity top, which look more like the service provider and will employ those kind of discrete systems, which is, I want to get the fastest that I can possibly get. But then underneath it, you will find in the same

enterprise, they've also got those horizontal use cases that they will then deploy using those and so it's kind of a range within the enterprise.

Neil MacDonald
The way I think about the majority of the enterprise market is they've got four challenges in getting the benefits of generative AI. The first is the technology stack here is completely alien. It doesn't look anything like the technology stack that even well-experienced enterprises have in dealing with kind of classic enterprise cloud. So that's a whole new learning area. Then they've got to figure out, as they bring that together, how do you apply raw, generative AI technology to a particular use case, or class of use cases. And trying to do that from scratch on top of piece part infrastructure, and then piece part software that you've assembled and curated, has a big learning curve, a lot of time, a lot of energy that most enterprises do not have the bandwidth or the talent to deal with. And that talent, of course, is in very high market demand.

The third piece is that in order to use that productively, you have to connect it to data. Generative AI is fundamentally a data driven opportunity. And taking a generic, generative AI model and trying to apply it in your business context is not going to succeed. It has to be infused through some more training, some fine tuning or other development deployment models like RAG. It has to be infused with that enterprise data, but that enterprise data is scattered all over the place, and so the next problem you have as an enterprise is, how are you going to get to all of that data.

Given that your competitors are getting after generative AI, if they're moving faster than you in solving these three problems, you're at a big disadvantage, because you can't get to number four, which is getting the business benefit and productivity or customer experience. What Fidelma has done with HPE private cloud for AI is solved the first three of the four problems, so the enterprises can go focus on getting the value, because we provide the curation of all of the infrastructure as an integrated offer. We provide the integration with the models and the AI solution accelerators. We provide the integration of our data fabrics and other tools to solve the data access problem. So, three quarters the enterprise problem is solved. And if you compare that to an enterprise who's going to try to just buy piece parts and integrate, they're going to go so much faster with that solution. I hope that answers--

Antonio Neri
And that has a value oriented approach, which obviously, over time, includes the services on the back end, right, profitable growth. I think we have Meta, and then we have--

Shannon Cross
Well, yeah, we only have time for one more, unless we go quickly.

Antonio Neri
Yeah, that's fine. We need to go quickly.

Shannon Cross
I know, people have planes..

Meta Marshall
So just in terms of, kind of, maybe you laid out kind of the four points of the four hurdles, but is there a fifth hurdle of customers kind of waiting for applications that can help solve kind of maybe the next step of how to kind of productize this? And are you seeing kind of partnerships that you can form on that?

Fidelma Russo
So that's why, I mean, if you if you look at the Unleash AI program that we put together, it's basically, you can have all the infrastructure in the world, but if you don't have an application that can run and solve your use case, then you've got something, but we can't really drive it. So, we're working with a whole bunch of ISVs to qualify on the platform, on private cloud AI, so that it's ready to run. It's qualified out of the box. And what you have to do is integrate it into your environment, versus figure out, how do I put these things. So, we kind of try to sell in terms of use cases, the sizing is in terms of use cases, and so that's what we're building out.

Antonio Neri
So that ecosystem is so important.

Fidelma Russo
Is critical.

Antonio Neri
Is critical, including the SIs on the front end of that. And also, on the back end of that is of distribution and value sellers. That's why this offer is so perfectly suited, because for them what value they add, in many instances, we are deploying now the private cloud AI in their lab so they can bring customers to test different applications as they go through the mobile development and deployment. All right, Daniel.

Daniel Zhu
Hi, Daniel Zhu on behalf of Toni Sacconaghi at Bernstein. So today we've seen and heard a lot about HPE's differentiated capabilities and liquid cooling. We go back to slide seven in the presentation. In 2027 we're still expecting the AI server TAM to be 70% air cooled. So just wanted to ask, what gives us confidence that HPE can gain share in air cooled servers specifically? And it would also be helpful if you could help us dimension what percentage of the backlog is Cray EX and what percentage is liquid cooled?

Antonio Neri
So, on the latter, we are not going to comment on that, because that was not the purpose of today. And again, Shannon talked about the fact that we are not providing any inter-quarter or guidance. But on the other aspect, I mean, fundamentally, that's why we broke the TAM. So, you understand the progression, but when you look at the CAGR of that progression to direct liquid cooled, is the fastest growing, and that's why I'm so excited to see the three fastest growing underneath is the server liquid cooled, is the networking fabric, which goes from a couple of billion dollars to $15 billion just for AI as a part of the $135 billion, and then obviously the advisory piece that I have been asked.

But as Neil showed that slide, we can do all form and factors. And you saw on the floor, Daniel, and we put you in front of one of the system, that was super-hot and a lot of air blowing, and we took you to the amazing fanless direct liquid cool infrastructure, there was no air moving everywhere. And so, the point is that we can go with the customers based on what choice they want. But remember, next year we have a massive transition in the silicon itself because of the power density. And some of those silicon don't offer any more air cool solution. You have to direct liquid cool. And so, by design, you're not going to be on 100% air cool. You're going to be at least in the hybrid. And what we are saying is that in some cases, we can improve, depending on how they're trying to achieve that by another 37% on top of that. And so that's why HPE is uniquely positioned to address both as we go through that transition, which is a point of

differentiation, and be able to manufacture here on the same location and be able to test it and scale it.

So, all right. Well, I think that's it. All right. Well, I want to thank, first of all, all of you for making the trip to Wisconsin, for taking the time with us last night and today. I hope you got a sense of how amazing this transition is, how HPE is uniquely differentiated as we go through these transitions to meet all the customer needs. And those are online, because at some point I asked someone in the back and say how many people online, there were almost 800 logged on the call. So, thank you for attending today's call, and I hope to see you soon.

Shannon Cross
Great. And from a logistics standpoint, first of all, take a look at what we've done here. We will also put that up so people--we'll either send it to you, or we'll put up on the website, so you'll have that, but it's a nice walk through of our (inaudible) so thank you.

Antonio Neri
And thank you, Adrian and Sarah, for bringing it all to life.

Shannon Cross
And then we are going to have the bus leave as close to 2:30 as possible. I know some people have flights as they need to catch. So, use the restroom, meet us outside, and again, thank you very much for coming.

Antonio Neri
Thank you.