

Microsoft Build 2025
Scott Guthrie
Tuesday, May 20, 2025

SCOTT GUTHRIE: Well, good morning, everyone, and welcome to Build. Yesterday, Satya talked about how we're offering the lowest-cost, highest-scale infrastructure for both cloud and next generation AI workloads. Today, I'm going to dive deeper into the infrastructure that's powering everything you've seen over the last two days and give you a behind the scenes tour of some of the innovations and optimizations we're bringing to Microsoft Azure.

Today, Microsoft has over 70 Azure regions live worldwide, more than any other cloud provider. We're continuing to dramatically expand our data center capacity and provide new data residency options so that you can put your services and your applications closer to your customers everywhere around the world. This year alone, we'll add more incremental data center capacity than we had across all of our data centers in Azure just three years ago, and much of this data center capacity is optimized specifically for AI. We've taken AI-optimized innovation to an entirely new level. Let's zoom in on one of our new AI-optimized regions that will go live in the coming months.

(Video segment.)

That video was one of our new Azure AI data centers, and we're building out multiple other data centers of this same size and the same design in parallel around the world. These data centers run the most advanced AI supercomputers, and that in turn enables all of you to leverage even more intelligent AI models at even lower cost. We recognize also that with this power and scale comes responsibility.

Microsoft is now one of the largest buyers of renewable energy in the world, already sourcing 34GW of renewable energy as well as other zero carbon sources, and we're on track to meet our goal to have our data centers powered by 100% renewable energy by the end of this calendar year. This includes the data center you just saw in this video, as well as the others that we're building out around the world.

We're not just using renewable energy; we're also using innovative approaches to further optimize the sustainability of this AI infrastructure. One example is the use of wood in the construction of our data centers. Now, wood may sound kind of quaint and strangely old-fashioned, but we're actually starting to use a hybrid data center construction approach using cross-laminated timber. This timber is a fire-resistant wood material that we expect that will reduce the embodied carbon by 65% compared to typical building structures. We're piloting alternatives to concrete foundations, with the goal of lowering the embodied carbon by more than 50% compared to other sources of traditional concrete.

Now, let's double click into the data center building, which will run a single large cluster of hundreds of thousands of interconnected NVIDIA GB200 GPUs. GB200s are the latest NVIDIA GPUs and are the most advanced in the world. Microsoft was the first cloud provider to bring online the first server, the first rack and the first data center running NVIDIA GB200s. This provides the most advanced AI training and inferencing platform in the world. We have multiple customers, including OpenAI, already running production workloads on this infrastructure today, and there are some cloud providers like AWS that are still haven't launched a GB200 offering.

An interesting characteristic about the NVIDIA GB200 is the density at which these GPUs run. They're literally packed together in a single rack. There are 72 GB200 GPUs in a single rack, all interconnected in a single NVLink domain, and this allows you to train and run a much larger AI model than previous generations of hardware. But the challenge when you put lots of heat generating components together in a small space is that you need a lot of cooling, and these AI systems are now so power dense that you can't use traditional air cooling, which is what almost every data center has historically leveraged.

Instead, you need to use liquid cooling to efficiently dispense the heat. On the left-hand side here of this picture, you can see what a server rack looks like with liquid cooling. You notice all the hoses. These hoses are continuously circulating in cold liquid into the servers, which cools them down and then extracts the hot liquid out. The picture on the right is of the trunk pipes in the ceiling above the racks in the data center. These circulate the cold water for the entire row of racks and then extracts the hot liquid from the data center. This hot liquid then exits the data center and goes through a cooling loop, where we cool the liquid down before we then feed it back into the data center on a continuous closed-loop system.

You can see in this picture here one of the liquid cooling systems attached to the data center. Each of these 80 units contains a 20-foot fan that blows cold air on the liquid as it circulates. If you want to get a sense of the size and scale of each of these fans, look at the giant trucks in the background of this picture and look how small they are compared to each of those fans.

Now, the great thing about a closed-loop system over traditional evaporative cooling techniques is that the water isn't wasted. It'll continually circulate between the servers and chillers to dissipate the heat without requiring any additional water supply. In fact, all new Microsoft data center designs going forward will use the zero-waste water cooling method.

Now, as AI models get more intelligent and AI workloads get bigger, they also need more GPUs interconnected and operating as one. One of the really differentiated things about this data center is that it operates as one contiguous, interconnected cluster of hundreds of thousands of GB200s. Our network had to be hyper-optimized to enable this type of interconnection. As context, there's enough network fiber cables inside this one data center to wrap around the world four and a half times. This is the type of scale and AI infrastructure optimization that supports the most demanding AI workloads in the world.

Now, what makes this even more impressive is the fact that we can combine multiple Azure AI data centers around the world using our AI wide area network, or WAN. And this AI WAN

supports up to 400 terabytes of network bandwidth, making it the fastest and most scalable AI WAN in the world. And this enables large, large scale distributed training across multiple Azure regions, allowing you to use one giant AI supercomputer for your jobs.

Now, data is the fuel that powers AI. Along with all these CPUs, you also need a lot of storage, a lot of databases and a lot of compute systems that can be used with it. Every one of our Azure AI regions has exabytes of storage and literally, millions of CPU core compute cores that are co-located together.

And this highlighted part of the data center here runs some of these systems. And if this building looks long, it's because it is. It's actually five football fields in length. And this picture here shows you now what's inside part of this building. And Azure, again, as mentioned, provides high performance, exabyte level storage.

This allows customers like OpenAI to use the storage with training data and run those GPUs at super high velocity. In fact, we can now drive over 2 million read or write operations and transactions per second on a single blob storage account within Azure. You can also create millions of blob storage accounts, and collectively, this provides unparalleled storage performance to any AI training or inferencing workload.

Now, when you have a storage system that can support this level of performance, though, and scale, the next question is, how do you take advantage of this performance from within your compute VMs? And that where our Azure Boost technology comes into play.

Azure Boost is our IO accelerator and management system integrated into every new Azure server. It offloads server virtualization processes typically done on the CPU of the server onto custom built silicon, enabling faster storage and networking performance. And it's what's giving Azure now industry leading IO performance for mission critical workloads. We can drive over 400,000 network connections per second to each VM, 6.6 million IOPs for local storage and 800,000 IOPs for remote storage. And Azure Boost now comes standard in every new Azure server, regardless of whether it's being used for AI or non-AI workloads.

We're also doing more silicon innovation with our Azure Cobalt silicon, which is now the industry leading ARM 64 price performance CPU in the cloud. Cobalt based VMs provide up to two times the performance of other VMs with .NET apps, and it's 20% less expensive than AMD or Intel based VMs. And developers from Databricks, Snowflake, Elastic, Adobe and Microsoft Teams are already taking advantage of Azure Cobalt in production today.

We walked through a handful of the innovations and optimizations that we're doing across our Azure AI infrastructure. All of these combined are really engineered to be a whole system that's unleashing incredible AI innovation. And to give you a sense of just how impressive this one Azure AI data center will be, once it goes live in a few months, it'll actually be 10x the performance of today's fastest supercomputer in the world. This step change is what's going to accelerate the next wave of AI model development and the next wave of AI innovation. And this is just one of many data centers we're bringing live around the world right now.

Now, AI infrastructure like this also is going to enable exponential reduction of AI cost. If you look at where we were just even two years ago, the price of, say, GPT-4 has plummeted by 93%. And this innovation, model improvement and cost reduction is what's going to power a new era of AI apps and AI agents. It's going to enable all of you in this room and watching online to integrate AI into every workflow and build transformative AI solutions that weren't possible before.

And when I think about the canonical app for this new era, there's one app that stands out in particular, and that's ChatGPT. ChatGPT has over 500 million weekly active users, and it's the fastest growing app in history. And ChatGPT is built entirely on Azure, using the exact same Azure services that you can use as well, services like Azure GPU VMs, Cosmos DB, Azure Kubernetes Service, Azure Postgres and Azure Storage.

I've talked about all the innovation happening within our Azure AI infrastructure. Now, let's walk through some of the work happening in some of these other services that are also needed to build great AI solutions. Let's start by talking about the data tier.

ChatGPT needed a database that would enable petabytes of data storage, trillions of database transactions, and can support tremendous growth. And to do that, ChatGPT uses Azure Cosmos DB.

Cosmos DB is a globally distributed, multi-model database service that delivers turnkey scale out with guaranteed millisecond latency and uptime. And with Cosmos DB, we've built a database service that can automatically replicate data to any Azure region around the world to give users lightning fast performance regardless of wherever they're accessing an application.

In ChatGPT, as users interacted with the app, conversations, prompts and metadata are all stored using Cosmos DB. And this enables ChatGPT to maintain context across sessions for 500 million users, delivering a natural user experience with low latency and high reliability at truly global scale. And with real-time replicas of the database distributed across Azure regions around the world, ChatGPT can put data closer to its users, resulting in much faster responses.

Cosmos DB allows you to elastically scale your storage and performance throughput with zero application downtime. You can start with gigabytes of data and then scale it up to manage petabytes of it, and you can start by processing, just say, 100 operations per second, and then scale to millions of operations per second around the world.

And best of all, with Cosmos DB, you only pay for the storage and performance throughput that you use. And it's what's offering ChatGPT the flexibility and scalability to handle its unprecedented user growth and trillions of database transactions. Cosmos DB delivers incredibly fast response times and five 9's of availability, meaning ChatGPT is demanding performance and uptime needs. It's been essential in terms of ChatGPT's growth and success.

I've covered how ChatGPT scales their data tier with Cosmos DB. Let's talk about the application layer and the compute that's used in it.

Now, ChatGPT needs to be able to scale their application tier across more than 10 million compute cores around the world. What's amazing is they only have a dozen engineers that actually manage that whole process. And that's where another one of our key services comes into play, which is our Azure Kubernetes Service.

ChatGPT is built on top of AKS, which provides a highly scalable Kubernetes service for cloud-native applications. AKS is a fully managed Kubernetes service available in every Azure region, and it streamlines operations at any scale. AKS offers automated deployments, auto healing, automated patching, and built-in security safeguards. And this is what enables applications like ChatGPT to scale without significant operational resources.

And AKS truly gives you cloud-native scale. You can now deploy your solutions either on a single AKS cluster, or you can also now use our new AKS Fleet Manager, which enables you to have any number of AKS clusters around the world and employ standard policies and standard operational models, so that your operations team doesn't have to scale linearly with the number of compute nodes that you're running.

It's another example of unique innovation that we're driving at Microsoft, and we're doing that innovation, both in our own services as well as by contributing to open source projects. And we're proud that last year, Azure became the largest cloud contributor to open source CNCF projects around the world.

And this slide here just shows some of the examples of projects that we've made significant contributions to. ChatGPT is taking advantage of many of these services, as can all of you.

And collectively, all this technology is powering a new generation of AI apps and agents. Everything that you've seen at Build so far this conference, yesterday and today, is all running on top of this Azure infrastructure. And we're seeing great companies like these building differentiated AI solutions, and we'd love to see your logo on the wall this time next year as well.

I hope you have a wonderful rest of Build and look forward to seeing what you build. Thank you.

(Applause.)

END