Microsoft AI webcast hosted by Piper Sandler

Company Participants

- David Carmona, General Manager, Artificial Intelligence
- Doug Burger, Technical Fellow, Distinguished Engineer. Cloud & AI

Other Participants

- Brent Bracelin, Analyst, Piper Sandler
- Harsh Kumar, Senior Research Analyst, Piper Sandler

Presentation

Brent Bracelin {BIO 2447337 <GO>}

(Starts Abruptly) Our live stream with Microsoft. My name is Brent Bracelin. I'm the Senior Research Analyst with Piper Sandler, covering the Cloud Software and Analytics space, tuning in from Bend, Oregon. I'm joined by Harsh Kumar, our Senior Research Analyst covering semiconductors, tuning in from Memphis, Tennessee.

We're very pleased to have three distinguished speakers from Microsoft joining this morning. Doug Burger, the technical fellow and one of the leading active researchers in computer architecture. Welcome, Doug; David Carmona, General Manager, Artificial Intelligence, at Microsoft; and then, Jonathan Nielsen is the Finance Director in Microsoft, Investor Relations. He is online, but not live streaming at this point in time.

Welcome, team AI at Microsoft.

Doug Burger

Thank you very much.

David Carmona {BIO 21505148 <GO>}

Thank you. Thank you, Brent.

Brent Bracelin {BIO 2447337 <GO>}

Before we dive into Q&A. Maybe, let's start with a brief description of your background, kind of role at Microsoft and maybe what location you're joining us from this morning.

Doug, I want you kick things off here.

Doug Burger

Sure. Hi, good morning everyone, and Brent, thank you for giving us the opportunity to participate.

I'm Doug Burger. I am a long-time, formerly a researcher in computer architecture. I was a professor at the University of Texas for 10 years, running a research group there, and we built CPU processors, new types of hardware architectures in silicon. I moved to Microsoft in 2008 and they started up research group in computer architecture in Microsoft Research. And my team and I built a number of new things, one of which was the Catapult FPGA platform, which forms the basis, measures smart networking and accelerates applications in Bing. And it's -- that was actually a large scale distributed system that we've built, not just the use of the chips.

And then about a year and three quarters ago, the company asked me to move most of my team into Azure to really start a new direction as our cloud grows and hardware innovations and full stack innovations become more and more important, and then more recently, I take it on a more active role in AI architectures and systems, and so that's really my focus now is AI infrastructure, the hardware and low software level.

And so, that's what we've been doing.

Brent Bracelin {BIO 2447337 <GO>}

Great. Very good. David?

David Carmona {BIO 21505148 <GO>}

Yeah, connected from Redmond. Actually, I'm connecting from probably a short walking distance from the office, but is useless these days. So, I could be anywhere. But yes, I'm in Redmond. I joined Microsoft almost 20 years ago and my background is software development. So, I was attracted at that time, Microsoft was very famous for Windows, right? But I was actually attracted to Microsoft because of the developer tools, which is like the DNA for Microsoft. So, I held a variety of roles all connecting, some of them in engineering, sign of the business, always connected around the developer business. So, I used to lead the business for Visual Studio, Visual Studio Online, Azure tools and so on. But then like five, six years ago, I noticed that something was changing in the industry and it's that there is a new software and that is called AI. So, I made a transition to AI, and now, I lead the AI Innovation team in the business side in Microsoft.

Brent Bracelin {BIO 2447337 <GO>}

Very good.

Doug Burger

I forgot -- Brent, I'm sorry, I forgot to mention, I'm in Bellevue, probably just a couple miles away from David. (multiple speakers)

Brent Bracelin {BIO 2447337 <GO>}

Well. That's the power of teams, right? We can all do this meeting virtually. So very, very cool. So what's tied into the kind of current seat of AI to kind of frame the discussion here. If you reflect back over the last year, what would you say where some of the kind of landmark AI innovations that occurred, either across silicon or systems or these new class of AI models? What were with the landmark things that investors should kind of pay attention to in the last kind of year?

Doug Burger

David, go ahead.

David Carmona {BIO 21505148 <GO>}

So, let's just start with the technology. So just recapping on build, so let's go to last week, right. So, last week, we were -- I'm unsure that we will go in more detail today, but one of the key things and it flew a little bit under the radar. So you have probably heard the term of the AI supercomputer that were announced last week, they change behind us.

So the implications of that are far beyond hardware. So the key thing that we saw with this AI supercomputers that you can train massive models, and by massive, I mean in the billions of parameters of where we were in the order of the millions of parameters. So that which can some very cool, that yeah, you can get better more or less more accuracy, what it really meant and that is the key change that is certainly something to be on top of is, how it's changing the way that you develop AI.

So, AI as a platform, what something that was not clear before. So, people, companies will have to develop AI models on a case-by-case basis. What we realized with this massive models is that they can be multi-task, they can be more generic than the traditional, very customized, vertical models for specific scenario.

So, that will bring a very interesting motion in the market, which is companies embracing this massive models and then customizing those models for their particular scenarios and trust, therefore, bringing like state-of-the-art AI to more companies, so that's certainly one thing that I would highlight To be on top, and that, I ensure you can provide more in there, but before that, let me just mention another thing also form last week that I think is very important, and I say concept of autonomous systems.

So we had this chip, everybody that is also shooting autonomous assistance with autonomous driving a lot of cases, but we're forgetting that autonomous systems are much broader than that. And last week, what we announced our bill was our first preview, so every developer now have access to that, to our autonomous systems platform. And it goes beyond motion control. It not only target robotics, but things like process optimization, things like machine calibration and a lot of scenarios that are much more, we are in the current situation for companies to address, so that's another one that we had like.

Doug, I'm sure that you can add on top of both a lot.

Doug Burger

I don't know, that was pretty -- that was great, very complete, let's see. So a couple of things I might underscore with the -- in particular, with the AI supercomputer announcement is really what David said that the size of the models has been growing much, much faster than, I think, any of us predicted. I mean it was growing, but now and just a round number is about 10X per year, which is a growth rate that far surpasses Moore's Law, while it was active and we also have disrupted that was over 50 years.

And so those giant models, what we've seen so far is that the bigger we train them, the more they can do and that break through towards being able to train semi supervised, especially the transformers in natural language. About a year and a half ago, it really turned the field in that domain from a data-bound problem into a compute-bound problem, which is why we're now racing to build the supercomputers, so that we can train these models in a reasonable amount of time, days, weeks, months.

And it is very, very disruptive. When we look at the capabilities, these things are able to drive not just in terms of language capabilities, but like David said, multi-modal and what we're seeing is I think a very rapid infusion of AI across our product families. Now, obviously the supercomputers are expensive. The infrastructure to do them is, to train those large models is very different than what people built in the past.

And then, once you train them, you have to serve them into the models that are giant. Serving them in real time is also challenging and can be expensive. So, I think that that is a transition that I really don't think the industry has grasped yet or appreciated how fast this is going to drive the capabilities of AI and how different the infrastructure needs to be with these 10x requirements.

Before we go into the architecture, because it does sound like there are some unique things you're doing to create that AI supercomputer, I want to go back to the comment you made around just the size of the models kind of changing. Is this at the bleeding edge where it's kind of 1% of used cases? Or when you think about the size of these models and moving from kind of data bound to compute bound scenarios, is this more prevalent than maybe we appreciate it as outside looking in?

David Carmona {BIO 21505148 <GO>}

Yeah, Doug, I'll take it. So there are two used cases here that I think we have to separate very clearly, right. One, which is that 1% is companies that will be great in those massive models, and we don't believe that that will be every company in the planet. But interesting thing here is not actually building the model, it's that these models as we were mentioning before, they are more generic, they are multi-model, they are multi-task.

So you can centrally train those models, but then you can have companies with no such deep expertise or access to this compute capabilities, customizing those models for their own needs. Now, to give you an example, that's the way that we're actually doing in Microsoft. So, we transitioned from having a broader by broader AI approach, so you have Office creating their own models, you have Dynamics creating their own models, you have Bing creating their own models.

To having a culture where we centralized the creation of these massive models and then these teams can customize them, so we have features now in Office like auto reply or like document summarization that are coming from customizing this model without requiring that massive infrastructure and expertise. So, that's the two angles that I would separate that.

Brent Bracelin {BIO 2447337 <GO>}

And David, as you think about going from silo to centralized larger models. Is that accelerating the feature, like what's the big benefit. Is it the time to market with new AI functionality or is it more advanced AI function?

David Carmona {BIO 21505148 <GO>}

I would say it's several, right. So the -- first of all, let's go one by one. At least, I would mention three. So the first one super obvious is skills, right. So these are state-of-the-art models that now everybody have access to. So you don't have -- so for example, you will have the same level of almost, I would say, researcher's PhD creating those models, because you can't reuse it. So that's one super important that will say a lot to the democratization of AI in the industry.

The second one is data, so because these models are trained initially with huge amount of data that is again central. In our case, we're trending with all the index web being right, so huge amount of data that we're adding in there. Then, for customizing those models, you don't need that a big amount of data. So for example, in the case of Dynamics, they custom train that model within additional data that are coming in their case, they do things like customer churn identification, so they only custom train the model with that delta, with their domain, right. So very, very interesting.

And then, the third thing that I would say is compute power, so because the customization is using our technical transfer learning, you don't need to train the model from scratch, so you don't need that huge amount of infrastructure required to train the model initially.

Doug, am I forgetting any other thing that you will take?

Doug Burger

I think that captured it very well. I would add one point. You asked about the model size and you can think about the largest model that you can train in the billions and we talked a little bit about that in our announcement. You can view that as the front of the bad wave, that's all -- I wouldn't call it research, but it's really pushing the envelope in the systems and the algorithmic techniques and the data management.

And then, you find out what those models, what capabilities they have. And if you take one quick down in size, that was radically large six months ago and those models are now being used, just kind of throughout our infrastructure. So it's not -- I don't really think of it as a bimodal distribution where there is like a 1% thing and then there is everything else, the whole space is just moving very, very fast and there's the new thing, the biggest thing at the Edge.

And then, things that were huge one year ago, kind of now empowering all the products. And if you think about how fast this is happening and I think in the notes, I mean it takes you years to build custom hardware, silicon and then build the systems and deploy them, and so, if you need to make a major change in your hardware platform, that's years before you can do it. And yet, this huge shift is less than two years old and we're still pushing the envelope on the rate of growth.

David Carmona {BIO 21505148 <GO>}

Yeah.

Doug Burger

And so in it, and so we're kind of doing whatever we can with whatever we can get to drive the models up, and fortunately, we have made a lot of investments in the past that were benefiting from now. A second comment I would make, if you think about Microsoft from a strategy perspective in this, this goes onto something that David mentioned. We have a very successful public cloud with infrastructure that benefits our customers.

We also have this massive PaaS and SaaS business that really empowers our enterprise customers and consumers frankly within our office and teams and changes there, like David was saying, can really reach so many people and transform their business, and then those infrastructure changes that we made to get those models out to such a wide group of users that see value, we can then take to our infrastructure customers to use. And so, those two halves, I mean I think that's why Satya talks about the Intelligent Cloud, it's not enterprise software and cloud, it's one thing where benefits can slosh back and forth and empower everyone.

And I think we're really the top company that has the both of those halves in a very strong position.

Brent Bracelin {BIO 2447337 <GO>}

It is fascinating just seeing the pace of change in the space, that's hard outside looking in to really fully appreciate it, all we get is from consumers of the technology, we see that the advancements, right. We see small changes, the software is getting smarter, right. But it's hard for us to quantify how dramatic things are changing on the back end.

Let's drill down, and Harsh, you kind of weigh in here to the architecture of AI supercomputer. I'm just trying to understand you've Azure and you can kind of build your own kind of infrastructure to run your models on Azure, but now, you have this AI supercomputer that's different. So maybe walk through architecturally what's different about an AI supercomputer versus Azure, and why you need to kind of think that this needs to be kind of optimized versus kind of building it myself through my own APIs on Azure itself. So, walk us through just the architecture.

Doug Burger

David, do you mind if I take the initial crack at that one?

David Carmona {BIO 21505148 <GO>}

All yours, you're the architect.

Doug Burger

So first of all, I think it's a mistake to view this as Azure and not Azure.

Brent Bracelin {BIO 2447337 <GO>}

Yeah.

Doug Burger

Okay. There is a lot of common infrastructure across both sides. But if you think about like what you want for an IaaS customer, let's say a big enterprise IaaS customer, what you need is incredibly high reliability availability, you want great stability, you want predictable performance, you want access in many regions, it's really a planetary computer and people build their infrastructure on our computing utility and it has to stay up.

And then, now, if you go to the AI side and you're trading one of these large models, you're running a single program or job over a massive number of -- massive collection machines and it's okay for some of those notes to go down, right, because you can build resilience into that large deal distributing system and we have a lot of experience doing that, for example, with platforms like Bing, which has to stay up and respond, but it doesn't matter if one note drops out, and so you're building a distributed system that needs to complete the job as cheaply and as fast as possible, but the reliability can characteristics of the components are very -- can be different. You need, for example, a lot more networking to train these giant models because they're data intensive and the communication is global.

So your provision much more aggressive cross data center networking than you need to for an IaaS customer. And obviously, there is a different mix of hardware with silicon components because you're doing your massive amounts of linear algebra. But I do think that if you look at the high speed networking, the ability to communicate quickly, the power requirements that you need to drive this, so just raw performance. Those are all things that accrue to the core Azure business as well. There's just an extreme point and then that business will follow along. We will benefit from the technological advancements that we're making.

Harsh Kumar {BIO 3235392 <GO>}

Sorry, I had two questions. Earlier, I want to go back about five minutes ago, I think you mentioned or maybe David mentioned our transfer learning. Is that effectively, you would train this large data set and then you would give it to an enterprise data center to run. Are we talking about something even smaller than that like running it on a small chip at the Edge or are we still talking about kind of a data center type process here?

Doug Burger

David, why don't you take that one?

David Carmona {BIO 21505148 <GO>}

Yeah, both. So at the end, so these massive models, of course, they are, they are thought to be run in the cloud, but at the same time, we see many scenarios where through optimizations, the influencing of those models can be done at the Edge. So they go side by side. So, yeah, absolutely.

Doug, anything that you add on top of that. technically?

Doug Burger

So yeah, I think the way I view the transfer learning is there is building your base, your model which is super expensive and then there's flavoring it with specific data. So just like an analogy, you might be -- everyone on the call here are speaking English, but we all have slightly different vocabularies and nuances and choice of words, hence provide training in English more I could flavor it with some history from an individual to customize that model to the individual speech patterns.

And it's a lot, as David said earlier, it's a lot cheaper to do that flavoring than build the whole English model from scratch. And that's the training thing and then there's a question of, do you deploy in the Cloud or the Edge, and I think that's a separate question.

Harsh Kumar {BIO 3235392 <GO>}

And that's just absolutely incredible. My second question was on the preferred sort of computational mode. Is Microsoft a fan of FPGAs versus GPUs or you simply don't care? The common knowledge is that ASIC are cheaper, once you get the set sort of -- once you get the parameters down a little bit on what you want to accomplish, you always want to reduce the cost on computing down with ASIC, but I was just curious what flavor you guys prefer, Doug or David?

David Carmona {BIO 21505148 <GO>}

This one is for Doug.

Doug Burger

Imagine, I get asked this a lot and so I'm not going to say Microsoft believes in technology. Let me give you some framing that I think will help people understand the space. Okay. So if you think about what you're doing in AI for these training supercomputers, you're doing a massive amount of effectively linear algebra.

And so you need an outstanding system architecture for low agency communication, you need a robust software stack, you need resilience that this should in system level, like we said. And then once you get out of the silicon, you've got a number of trade-offs, you've got sort of the efficiency of the silicon, like how fast can do X just raw flops, you've got the future-proofing, if the model shifts, how well does the silicon map on to that shift.

And then you've got other innovations that you might have in the hardware that helped the model training that might help it to converge, like to save you cost. And so, you've got this X set of trade-offs and we actually invested multiple technologies. We disclosed what the actually bigger is built on because that part has been really, really great for training these large models.

And then, the second thing I would say is that I really don't like the term ASIC because it's -- in this space, it's kind of gotten overused, meaning it just means customer chip, right. And in the past, it meant something very different, right. You're building out a chip or a circuit for one purpose and what these things really are is their programmable engines that do linear algebra flexibly. And GPUs or FPGAs, we've talked about our engagement with graft core, there is a ton of start-ups and across Cloud and Edge, Intel, these are all programmable engines for AI-based linear algebra. We have different sets of trade-offs in the efficiency of the math units and the front end that feeds them and so there is just -- it's a new, what just like we had CPUs for many, many years, this is kind of a new class of programmable architecture that's emerging in a very rapidly evolving space. And again -- and so at the end of the day, GPUs or FPGAs or what you called ASICs, they're all chips that have a lot of math units on them, if our architectures could even.

Harsh Kumar {BIO 3235392 <GO>}

Thank you.

Doug Burger

What that ultimately will look like, I don't think we know or at least we don't talk about that side of the company.

Brent Bracelin {BIO 2447337 <GO>}

We'll go back to that. We'll kind of try to get a different angle on that because that's a big question. But my question, just as we think about the industry pace of change 10x, we're thinking technology changes around Moore's Law every 18 months, doubling transistors, but as you think about 10x every year, is there just a limited number of companies that can keep up with that pace of change? How nervous should investors be around things moving so fast that it's really going to be limited number of companies that can manage that pace of change? Any thoughts, David, as you just think about card rails that we kind of need to have in the industry and do we have efficient card rails there?

David Carmona {BIO 21505148 <GO>}

Yeah. I feel this is counter intuitive, but actually technology like this instead of limiting the access to AI to fewer companies, it does the opposite. So it's opposite to more companies. So let me elaborate on that. So we don't expect, I think we were mentioned before companies to building this massive models from scratch all the time, what we expect is the concept of AI as a platform. Just like what we had in the past with the Internet, with then, of course, mobile, and now, with the Cloud. What is the equivalent of that for AI. So if you look at how companies are developing AI today, they are, I mean, their console of a platform is, I'm sorry, Doug, this is going to hurt, but just leave it, in the sense that it is very close to having something very specific for my solution, so I have to be on top of the platform to have something customized for my solution, for my scenario.

And what we want to go is a place where AI can be good accessible like a platform, so you can use these AI models and you customize those models for your particular scenario. That's the big change that I've seen -- we will see coming very soon, right down. This applies also, I know, we're talking a lot about NLP, natural language processing, but this will apply to every other scenario.

So again coming back to the previous thing with autonomous systems, very, very similar pattern that we see with autonomous system. So would you say that creating an autonomous system is very limited to a few companies. Well, if you look at it from the angle of building all the components from scratch like deeper reinforcement learning, super complex technique that is not available for every enterprise to address. Our approach is very different, is to provide a platform abstracting on top of those techniques like reinforcement learning. So companies can address, can embrace autonomous systems by using their expertise on what they have expertise, which is their business, right. So the same comparison it applies to many other different facets on AI that we are pursuing across the company.

Brent Bracelin {BIO 2447337 <GO>}

Got it. It sounds like you're still with how you're moving from silo to more centralized models available to all of the different internal departments of Microsoft, you're going to start to expose some of those same models externally on a kind of Azure service basis. And so, you almost create a new -- a past layer for AI essentially, is that the really...

David Carmona {BIO 21505148 <GO>}

Exactly, exactly.

Brent Bracelin {BIO 2447337 <GO>}

And when was the timing around that? Is that something that's going to be like 2021, 2022 kind of opportunity where you're going to (multiple speakers)

David Carmona {BIO 21505148 <GO>}

I will go back to [ph]tax(\ph) position on it. So this is not like getting to the finish line. So this is going to be -- so we'll see value being delivered in this journey right down. The beginning was last week. So there is already a ton of value that companies got used today. Because of that dual approach that we have in the company, which is I would say, then three things. So when you look at our technologies, I would position them in three different states. So the first thing that we do is being very open about our innovation. So, for example, we open source these fragments, right. So that is, for example, last week, we open source the key framework -- the key distributor framework engine is called a deep speed to enable this training, right. So, that's something very cool, right.

But then we're going to a phase where we infuse these models into our growth and we already did that too, so you have a lot of features that are part of our products that customers can get access today that are used in that manner. And then, the third stage that it would highlight is how we also incorporate this into Azure for developers to also build on top of that right. And we also have for that things like Azure Connected Services which are there, these AI models that are served as a service, right. So they are very easy to use by developers. So with those three things, we cover the entire spectrum. And it's not that, hey, today, we are releasing everything, right. So, it's that as we move, we make progress in this journey. Things are getting infused into three main channels.

Brent Bracelin {BIO 2447337 <GO>}

Fascinating. And my last question before we shift gears to kind of (multiple speakers)

Doug Burger

Brent, I'm sorry. Could I add point onto that?

Brent Bracelin {BIO 2447337 <GO>}

Yeah.

Doug Burger

I think it's important for your investors. If you imagine today that I have a -- I started a new company and I want to become a public cloud vendor, I want to build another cloud business on my own infrastructure and compete with the big mega clouds. The capital I need to raise, to deploy, to have -- to build these giant data centers and have to be available around the world in many, many regions and the number of software features that I need to provision and the number of SaaS and PaaS services that I need to on board, it's just not achievable for a new entrant into the space, and especially when you look at some of the advantages that some of the big players have and I referred to our enterprise software business as well.

In the AI space, it goes directly to your question. You need to be able to build the giant, the biggest supercomputers that the world has built to train these models. You need to have really deep distributed systems expertise, you need a really, really strong AI team that understands this, and so, the capital to do that as well and the level of expertise and depth is also, I think, not quite at the point that you would have to struggle to try and start it in the public cloud, but it's very large and growing.

And so just by definition, that's going to limit the number of players. We were able to do this and Microsoft positions that we want to empower all our customers, like David said, can give access to this technology, but the number of people, the number of organizations that can build that infrastructure from the ground up, I think, is going to be small.

Brent Bracelin {BIO 2447337 <GO>}

That (inaudible) and kind of leads into my next question around OpenAI. Maybe could you just walk through that, opening our relationship, you put, obviously, a bunch of money there, a lot of really great talented engineers there, kind of everything in OpenAI, so maybe just walk through that Microsoft OpenAI relationship, what is it, and how does it relate to the AI supercomputer announcement?

Doug Burger

I only say one -- I'll say two things, and I think David should weigh in.

I want to be judicious about commenting about a close partner. So number one, OpenAI is among the most ambitious organizations in terms of what they're trying to do with AI and their mission. And so, they, like our internal teams, push the infrastructure very hard. Okay. I think, and given the scope of their ambition, the fact that they chose to enter into this Microsoft partnership, I think it's a very, very strong vote of confidence for our infrastructure in our roadmap. That I think you can look at and say, wow, one of the most ambitious AI organizations on the planet was willing to sign this deal and commit to our infrastructure, that should tell you something about our roadmap, which we don't publicize.

Brent Bracelin {BIO 2447337 <GO>}

Very helpful. Anything else, David, do you want to weigh in there?

David Carmona {BIO 21505148 <GO>}

Yeah, I think coming back to Doug's points before, right, in the sense that in order to serve these massive models, you need to deal just to throw a number in there. So, the AI supercomputer that one-off last week ago is actually, it became immediately the number five supercomputer in the world. Right. So that's a sign of the scale that you need there.

So to deal a system like that and the other companies that I do all the time is that you should look at this has the formula one, right. But then, all the innovation that you're doing that formula one, you make it available to the typical card that we use every day. So the same thing we're doing with this investment.

So it's not only about creating this AI supercomputer to build this massive model that we believe will be the foundation for that AI platform that we were mentioning, which is only will be available for a few vendors, who can really have the capital on that, but then, the other thing is how we are thinking all of that innovation infusing it into Azure for everyday Azure, right. So, that's the other big aspect to consider in this.

Doug Burger

That's right. And just to underscore that point, Brent, I'm sorry, I'll sign up in a minute, think about what David just said, just largest supercomputer in the world with the models growing 10x per year, just think about the implications of that for a minute and where this is likely to be.

Brent Bracelin {BIO 2447337 <GO>}

Version 1 fifth largest super computer in the world. We're trying to keep up with the modeling kind of appetite that's growing 10x a year, very, very, fascinating. Harsh, you had a question?

Harsh Kumar {BIO 3235392 <GO>}

No, I was outweighed actually. I had a question on quantum computing, but I'll wait for you to finish.

Brent Bracelin {BIO 2447337 <GO>}

Let's shift gears to kind of more of a enterprise reality. The enterprise reality is we're going through a pandemic. We have work from home, it feels like this is an environment where you could see things and AI product start to pause and slow, so what are you seeing kind of in this post pandemic environment relative to kind of AI, the appetite for AI, is it slowing, is it accelerating, just walk us through what you're seeing there?

David Carmona {BIO 21505148 <GO>}

Do you want to take this one, Doug?

Doug Burger

Go ahead.

David Carmona {BIO 21505148 <GO>}

(multiple speakers) Yeah, please be on top of that. Yeah. So there is something, I have to say even before the pandemic, we were seeing a change in the conversation with enterprises. So I remember, maybe four, five years ago, every conversation that you have with our customer was about, hey, what is this AI fee, should I be on top of that, right. Then, the next thing then finally companies realize, okay, this is big, I need to embrace AI if I want to transfer my business, but the next thing was how do I get started. And we saw a lot of adoption in the enterprise, but in the first phases of that adoption, right, so many pilots, many proof of concept, et cetera. From that time to now, we've been seeing that shifting in the market to really how do I have funding back in the business, how do I move beyond the proof of concept, how do I connect AI with business outcomes.

So that conversation was really happening, and this pandemic, this global crisis, what it has done on this conversation is accelerating it even more. So, we've seen our customers just cutting to the chase, telling us, hey, what I need now is putting AI into action and I need it now. We usually have that conversation in two fronts, right. One is for the current crisis of (technical difficulty) AI tool let me deal with the disruption, buying experiencing today but then, even in the longer term, we know that we're never going back to the previous normal. It's going to be a new situation with a lot of economic uncertainty with a lot of care on how do I optimize the revenue, how do I optimize my operation, et cetera in my company. So how AI can help me there.

So in the first one, let me maybe, I think your question was more on the first one, right. Is it more today what our companies doing with AI so how is AI helping in the disruption that we are experiencing and we see three primary scenarios. So, these are the three used cases that we actually saw our race on the demand from our customers, and we see that across the industry, so not for a particular vertical.

The first one, which is very obvious is of course customer service, so that is the most visible one, right. So, companies that were, in many cases, piloting using AI to streamline their customer service, they are accelerating those projects like, we see examples of companies that have been working for 18 months on a project and then with the current crisis, they put it in production in two days, right. So that the acceleration that we see in some cases, and that goes in many cases, definitely customers support, so you have like a perfect storm with more request from your customers. But at the same time, your operations are being impacted by the crisis, but in other cases, we're seeing broader scenarios like really being able to identify your best customers, being proactive, personalizing your service to those customers. So we see that across a number of customers and I'm happy later if we have time to go into some examples for companies, and that's one scenario. The other one that we see hugely happening in this crisis of course business process optimization. And I'd like to separate that into, what we see that, the first thing that customers are noticing is that processes that were very established in the past, now they have been disrupted dramatically. So, think of supply chain, right, think of forecasting demand. So, all fraud detection, customer churn, those processes are dramatically different right now. So they need to put in place processes that are more agile, that they are more flexible to those changes, and AI is a solution that is helping in that area. So that's one thing.

The other thing of course is specifically because of the economic situation, we see customers looking for cost savings right down, that in business processes, that means a lot and we see many customers that are using AI to shorten their business process, to streamline business processes and to make them more productive. So that's the second scenario.

And the third one that we see a lot because of another perfect storm, right now, is employee productivity. So I think it's, of course, (technical difficulty) but it goes beyond that. So we see another perfect storm with very, very, very complex situation that employees have to deal now from the business point of view, but there an impacted productivity because of the current situation. So we see companies looking at AI to how they can in the short term increase that employee productivity and we see many, many examples of those. So those are the three.

Doug, I don't know if you want to add something on top of that, but we can go deeper into examples, if you want to, later.

Doug Burger

Yeah, I thought that was super interesting. It was even interesting for me and I mean the company, so thank you, David. I would just underscore what David said. What we're seeing is that this crisis is accelerating company's desire to do digital transformation partially because they need to and partially because if you're going to optimize the process that you deferred for a long time, you might as well just do it, do it right. And we -- Satya has invested enormously and having Microsoft BD trying to be company that will help people solve problems through digital transformation.

And so, I think it's opportunity with our infrastructure and our services to really help our customers.

Brent Bracelin {BIO 2447337 <GO>}

And as you think about the types of kind of AI used cases that seem to be kind of resonating most right now, you give three examples. Is it kind of NLP? Is it text-speech, computer vision, all of the above? Is there one type of AI that's really like, oh my god, it's taken off much faster than I would have thought.

Doug Burger

David, go ahead.

David Carmona {BIO 21505148 <GO>}

I think all of them, but if I have to pick a favorite right now, that's NLP. So, NLP is right now in a very hot moment because of this acceleration being especially relevant to an area that was very tricky such as NLP. So, NLP is very complex, it requires generalization. It is very, very complex to solve, and with this new state of their models, we see huge potential in that. But in the examples that I said before, NLP is there, but we see all the techniques of AI across the board.

Harsh Kumar {BIO 3235392 <GO>}

David, I can vouch that it's been about nine months since I have spoken to a live person without speaking to a machine first.

Brent Bracelin {BIO 2447337 <GO>}

I think we all experience that. Absolutely. Can't -- we have about 15 minutes here left, what I think it'd be helpful is to maybe get an update on Project Brainwave. Obviously, Doug, I think the last time we met, we kind of was about three years ago and get a kind of a deep dive, love to understand, and maybe Harsh is probably better here to go down this, but I'll turn it over to Harsh to get maybe an update on Project Brainwave.

Harsh Kumar {BIO 3235392 <GO>}

Sure, sure. Why don't -- David or Doug, why don't you give us a 30-second or one-minute overview on what you're trying to accomplish with Project Brainwave. And then, I can jump into some questions.

Doug Burger

David, do you want me to take this?

David Carmona {BIO 21505148 <GO>}

Doug, that's for you.

Doug Burger

Okay. So we, the project is continuing, it's continuing to go super well, we're using it, continue to use it at scale worldwide to serve the models that we trend. Like I said, we definitely have a mix of technologies in our infrastructure, and with that program in particular, we're able to do -- incorporate innovations fairly rapidly to really push what the models on the serving or inference side are capable of doing.

I think we have really focused on our internal businesses because that's where the need is pressing if you think about the need to serve these giant models, and the latency in the cost, the expense, that's where a lot of innovation is required. I don't want to talk too much more about it, but we are going to be seeing more publicly this year.

So that would be a chance to get updated then.

Harsh Kumar {BIO 3235392 <GO>}

And one term that we hear a lot and we obviously see as in Hollywood movies and stuff like that as well is quantum computing. I was curious since you guys are some of the top leading experts in the country on AI, I was wondering, is it quantum computing even applied AI in some manner, or is it useful for completely different things? And if so, how do you guys feel about the potential for that maybe commercializing in the next five, 10, how many every years if it's out that long?

Doug Burger

I would -- David, do you mind if I start with this?

David Carmona {BIO 21505148 <GO>}

Yeah. You take it.

Doug Burger

So, there is certainly an overlap between the two in the research community looking at doing machine learning with quantum algorithms and quantum-capable algorithms. What I would say right now, my personal view, I'm not speaking for the company here, is that initially the two technologies, AI supercomputers and quantum systems, will be targeting different applications, but the things you can solve with a lot of the quantum algorithms are just very different problems than what we're doing with these large-scale AI models. Eventually, there may be some convergence, but where quantum, I think, will do best initially is in problems that have huge compute requirements, but not a lot of state, simply because of the difficulty of keeping larger massive state in the superimposed domain. And if you look at the amount of state we're driving to train these models, it's massive. So, I think of them is disruptive technologies that are initially attacking different classes of problems. And then theoretically, there is possible for overlap in the longer term.

Harsh Kumar {BIO 3235392 <GO>}

And Doug, when you say state, what exactly do you mean? I'm not familiar with that.

Doug Burger

Just information, bits.

Harsh Kumar {BIO 3235392 <GO>}

Okay. I got it.

Doug Burger

How many bits of state do you hold in one computer. Is it a 1,000, is it a 10,000? It's not the petabytes of data we're operating on in the Azure for computer space.

Harsh Kumar {BIO 3235392 <GO>}

But it sounds like we're far away, it's the best way to think about quantum computing is at least a couple of years away, if not more.

Doug Burger

I don't want to take a position on the quantum roadmap. I would say that for large-scale AI, I think we're pretty far away in the quantum space.

Harsh Kumar {BIO 3235392 <GO>}

Brent, do you have anything?

Brent Bracelin {BIO 2447337 <GO>}

Yeah, so let's shift gears a little bit maybe and talk about kind of some real world used cases. We've covered a lot of ground here, you gave us some scenarios, how people are using it, but you just recently announced NDA, you announced all sorts of kind of new used cases, I think, at build last week and two weeks coming up to build around some new large enterprises, where AI was Microsoft AI, Azure AI. I think there's like 20,000 enterprises that have deployed Azure AI at some point in the last year. But walk through how people are actually using it, other interesting scenarios that you come across that you'd like to highlight for us to take the technology in the real world scenarios?

Doug Burger

This is a question for David.

David Carmona {BIO 21505148 <GO>}

Yeah. I take that one, Doug. I would use the frame that I was using before, but I would add one component. So the frame that I was using before was those three-key scenario. So like I can go through is like canonical examples in there and even bring some of the real customers that are in those three scenarios, but then I would add an additional one, which is the post COVID-19 moment, right.

So that's the moment where we will see companies reimagining again in this completely new normal that will experience there and I can bring some examples of things in that area as well. So let's start with today. Let's start with customers that are using AI today for those three scenarios. So we started with customer service. Customer service is a very easy one because it's weird to find a customer that is not now using AI for their customer service, and in general, their customer engagement, right. So, we have many examples. I think the most exciting one that we announced last week was actually in the healthcare area. So, in Microsoft, as you know, not only we provide the platform for customers to create, for example, bot to the user with customer service, but we also provide SaaS solutions that they can use directly for those solutions for those scenarios. So, one in particular is called the Azure Healthcare Bot Service, and it is a vertical bot that health organizations can use today for their customer service, in this case, their patient service. So, what we have seen in the last month is an exponential usage of that service. So just to give you a sense, we have seen 1500 health organizations between public health organizations or health providers, et cetera, that are implementing new projects. So think of that, 1500 new projects that were put in production, specifically for COVID-19 management. Those bots that are being deployed in just these few weeks, they have a reach of 30 million people.

So think of the scale of that, right, so -- and they provided a huge tool to really unload, to reduce some of the load that we were seeing in health professionals like doctors and nurses to have a first level interaction with patients. It was a self-assessment tool that patients can take in order to be redirected then to the right resource, right. So that would be one example that I will bring.

In business process optimization, we have seen -- I don't know where to start there, because again, every company is doing a lot in that area. So one, I think you mentioned that one, but one of the ones that we announced last week was FedEx. Like FedEx is the perfect example of business process optimization.

So they had this massive amount of data, so they have a very granular data on their shipments in FedEx. So, with this agreement, they are going to have AI on top of that, so they can have better intelligence on what is going on. So they can not only identify trends that are happening or products that are happening, but also to optimize that supply chain right in their organization. So really, really a perfect example of business process optimization. If I could see everywhere, I can think of IndiaLends is another customer. IndiaLends is underwriting a credit platform in India for like 50 banks in India and what they did in their case is using AI for their credit approval system and they were able to decrease the time -- the processing time -- the internal processing time for credit approval 50%, which in this times imagine the impact of that is having the ability to process twice the number of credits, right, in these -- in a moment like this, which is absolutely critical.

Manufacturing, we have seen also many examples. I think I mentioned this one before, where we see companies are starting to use autonomous systems, not for the moonshot of autonomous driving, but things like motion control or process optimization in manufacturing that are getting big results from day zero, right. And we see many customers in that front. What is going to bring in there. Well, you tell me a business process optimization. Actually, we saw it also in healthcare where we have such a limited resourcing, as we have right now. It is critical to things like medical supplies, even hospital beds are optimized with things like AI. So very, very typical.

The other one that I mentioned was employee productivity. So let me bring just a couple of examples in there. So one generic, we see that a lot is AI working with humans to outman their productivity. So, just to give you an example of that, Reuters. So in the case of Reuters, you have -- again, you have their journalists that are doing what they do, which is writing articles. And in this case, they were using AI supporting them, so they can attach relevant videos to those articles. And not only, they have to go through that very manual process of finding relevant videos, but they also -- by using AI, they also increase the average completion rate 80% with this technique, right. So, not only better employee productivity, but then a better result.

The other one that I would mention that is very connected to that is not only don't think of employee productivity only removing like tedious tasks or repetitive tasks. It's also about making better decisions. So the other example that I will bring in here is Team Rubicon. They are managing their more than 100,000 volunteers in the US for COVID-19. So think about the massive scale of the solution required for that. I mean in their case, they are using AI to really identify and optimize the deployment of volunteers across the US. So, very important work that is really about making better decisions. So, those were just a few examples on the first part, how companies are using AI in the response rate.

But let me just use a couple of examples on their imagine phase, right. Because I think that is the big conversation that we should how, what is going on after after this health crisis over and it turns into an economic crisis, right. And, we see three key things happening in there. So let me use an example in there.

The first one that we talk a lot today was AI that scale. So I know that we talk a lot about AI scale in the concept of the model. But there is a bigger motion that will happen in the enterprise, that is really applying AI at scale in their business. So, it's moving from that pilot phase to re-infuse it into everything that they do. And for that, the key thing in there, I will bring an example right now, is how you need to scale the usage of AI across your business units and moving beyond your technical units. So we see that business being more connected to the AI transformation than it was before. That is something that will happen certainly as we move into this phase. And an example, I would bring in here of course the key thing. I think the key lesson here is that we have gone through that before with software development. So remember this conversation 10 years ago we were having it for software. And we knew how to do that. It was called DevOps, and it was all about bringing developers to technical units and the business together in a combined life cycle. We don't have that for AI. So you have technical units that are working in silos in this pilot, but they are not being infused into the business. So the equivalent for DevOps in AI is called MLOps and we see that as a huge trend moving forward.

And an example, I will bring in here that we just published is the Department of Transit in Vancouver. In their case -- so think of this case, we talk a lot about the size of the models, but what about the number of models that you need to transform an industry, to transfer your company. In the case of this company, they have 18,000 models. So think of the number of models, think of managing those models. There is no way that you can do that with just a silo technical unit doing that, right. So they brought together the business and the technical units with a common MLOps process that is across all of that. So that will be one. I know that I'm talking a lot here, but let me just go through the other two, because I think it's important.

The second one is regarding...

Brent Bracelin {BIO 2447337 <GO>}

We've got one minute drill and we have to we've to conclude.

David Carmona {BIO 21505148 <GO>}

One minute, I'll do it in a half-a-minute. But the second one that I want to mention is how the next step in this transformation is to empower the business. So we have talked a lot about technical units developing this AI, but the next step and we made very big steps in Microsoft is to empower the business. So they can also apply AI, right, with things like the Power Platform, with Dynamics 365. We see that as a reality moving forward, and then, one example that I will bring here is Novartis. So Novartis, a pharmaceutical as you know, so we just -- like a year ago, we signed an agreement with them where they are basically empowering their 50,000 employees.

So, almost half of their employees with AI, but don't think of it as using AI, but more as applying AI with flexibility with freedom to their processes. So you can have researchers that are researching new drugs or new treatments or new vaccines and how they can be augmented with AI by using in their processes to drug manufacturers any cellular matter expert. So that's a huge motion that we will see moving forward.

And the third one, and this would be my last one, is let's not forget our Responsible AI. So we see Responsible AI, we have this conversation for many years are really and we are seeing also a huge shift in the conversation front, hey, what are the challenges of AI, what are your principles of Microsoft to now, AI under a huge risk as I accelerate my adoption of AI, help me to implement that AI Responsible, so I can mitigate those risks.

So we have many examples in here, but just to read one, TD Bank, the bank in the US from Canada. You can see how they're using Responsible AI to mitigate things like bias or things like adding more transparency to the models that they are deploying.

Brent Bracelin {BIO 2447337 <GO>}

Thank you, David. I just -- you guys have so much to talk about AI. Jonathan's extended this conversation for another hour, I'm just kidding. Last question for me, Doug, as you think about the pace of change here, tax in the last year on the modeling side, what are you most excited about, thinking about the AI industry and Microsoft opportunity in the next three years? One thing, what's the one thing you're most excited about?

David Carmona {BIO 21505148 <GO>}

Are you thinking, Doug?

Doug Burger

David, I need to give that some thought.

David Carmona {BIO 21505148 <GO>}

I keep going back to the concept of really putting AI into action. So the concept of democratizing AI, I know that it's not so technical, but there's a lot going on to empower that motion that we're doing, because it goes beyond just the infrastructure of the software, it's really about our comprehensive solution from research to the three clouds, to our tools, right. So, really, really powerful if we get there in the next year.

Brent Bracelin {BIO 2447337 <GO>}

Driving those business outcomes for sure for you. Yes.

Doug Burger

Yeah. I'm going to give a little bit of a squishy answer, but it's a true answer. You very rarely get the opportunity, and this is a personal view, to work on things that will really change the world in a positive way. And I think the capabilities, we're going to be able to generate, just from my perspective with the hardware and the systems work, you know can help really solve global problems.

I mean that's like that is once in a lifetime opportunity to build something that allows us to really make meaningful shifts on personalized medicine, climate change, efficiency, security, turning back this concept of responsibility. So for me, I just really feel driven by a mission and a purpose to really move the needle and make the world a better place, it's really -- it's a blessing.

David Carmona {BIO 21505148 <GO>}

Great, Doug. That was -- that was so nice from you, and you should have said at the beginning, so I adjust my answer.

Brent Bracelin {BIO 2447337 <GO>}

Well, listen, we're out of time. Really appreciate, David and Doug, sharing your views on the current state of AI here. As always, it's a pleasure, and thank you so much for sharing your thoughts.

David Carmona {BIO 21505148 <GO>}

Thank you.

Doug Burger

Thanks.

Brent Bracelin {BIO 2447337 <GO>}

It's a good opportunity to chat with you all. Great job. I should tag team with you more often.

David Carmona {BIO 21505148 <GO>}

It was a pleasure, as always. Bye.

Brent Bracelin {BIO 2447337 <GO>}

Take care, all. Bye-bye.

This transcript may not be 100 percent accurate and may contain misspellings and other inaccuracies. This transcript is provided "as is", without express or implied warranties of any kind. Bloomberg retains all rights to this transcript and provides it *solely for your personal, non-commercial use. Bloomberg, its suppliers and third-party agents shall have no liability for errors in this transcript or for lost profits, losses, or direct, indirect, incidental, consequential, special or punitive damages in connection with the furnishing, performance or use of such transcript. Neither the information nor any opinion expressed in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP.*