Transcript of Six Five Summit Navin Shenoy

Hi everyone, I'm Navin Shenoy and I'm responsible for the data platforms group at Intel. Really happy to be at the Six Five summit to talk about our fast evolving industry and the amazing opportunities ahead of us. I'd be willing to bet that we have all seen the profound effects, and acceleration of digitization in the last 15 or 16 months and in that regard I think it really is quite remarkable to think about this fact.

More than 1/2 of all of the data in the world has been created in just the past two years. Think about that for a second. There's tons of examples of this. The hundreds of thousands of photos posted every minute, the thousands, maybe even millions of video calls that we've all been doing at any given time. The billions of human beings that are using the Internet to shop and study and watch movies and hopefully avoid planes and hotels every day.

But here's the thing, out of all of the massive oceans of data that are being created, only a very small amount, probably less than 10% of the world's data has been processed to create any meaningful value or insight.

So to me that means that we are in the beginning of a new age where we are in the Golden Age of data.

Tremendous opportunity for us to do more with all that data.

At Intel, we've been very focused on helping our customers move and store and process the world's data to create more value. All of that unprocessed data that I talked about is, of course, untapped potential. Now, one of the keys to unlocking that data can be informed from history. In the late 19th century an economist by the name of William Stanley Jevons made the observation that technological improvements that increase the efficiency of coal use ultimately led to increased consumption of coal use in a wide range of industries, and it was considered a paradox because people at the time often associated efficiency with a reduction in usage, when in fact the exact opposite happened. More efficiency led to lower costs and created more usage. And of course this also holds true in technology industries.

And so, let me give you an example of this in the data center. This just plots the server processor growth overtime and you can see that since the introduction of the standard high volume server in the 90s, the server market grew and it grew until the early 2000s, and in that time frame we saw the invention of virtualization. VMware created this concept for X86 servers and they enabled IT professionals to better utilize physical resources by virtualizing more machines onto a single machine and that of course help reduce downtime and improve control and security. But some in the industry began to wonder, was the server market at its peak, and in fact...

You know, as you look at this chart here, you could see that the market did begin to stall out.

And that was partially because of the .com bust, but it was also because as servers were virtualized, IT departments didn't need to buy as many of them. So efficiency and utilization improved and temporarily we did in fact see less consumption, but Jevons Paradox ultimately held true.

And eventually the efficiency improvements led to more consumption. Now a lot more consumption. It's been quite extraordinary to see what's happened since that time. We've seen the market for servers and datacenters just absolutely explode, and it's now powering all of our digital lives.

And even more forces emerged to drive efficiency over the years. We saw, of course AWS and Azure emerge. Cloud players that perpetuated more efficient usage of computing. But as much progress as we've made, we've also, seen incredibly inefficient utilization of computing still to this day, right? The massive growth that we've seen in the industry is led to increasingly diverse workloads. Increasingly, large data centers, and that's put pressure on the underlying infrastructure. We're seeing challenges operating at scale. For example, there may be too much compute in one place, but not enough compute in another place for a given workload, and the exact opposite may be true for a different workload.

There may be challenges in moving data to the right place at the right time and keeping it secure along that path. Or there may be a need to accelerate certain workloads that work better on accelerators as opposed to general purpose microprocessors Now, at the edge you're seeing the need for lower latency, and so of course the centralized data center may be insufficient, and we may need to push out computing to where the data is actually created.

So of course, all of this has major implications on how the software landscape that runs on this infrastructure evolves. In order to make the data center usage more efficient the software industry has evolved. Instead of classic monolithic software applications, we have this new trend where applications are broken down into smaller service oriented components that run in their own containers called microservices. Each microservice itself is contained with its own load balancer and highly distributed and disaggregated architecture. The benefits of this approach include recovering from software crashes easier when a function in a container fails its workload can be redirected to a different container or microservice and continue running with little impact to the overall service. Or if there's high demand for resources by a single microservice, the system can make an appropriate request for more dynamically and automatically boosting computing resources memory to support that particular microservice. So All these highly complex, scalable and high performing microservices or demanding and creating a need for efficient orchestration, right? Those services need to be moved around in an efficient way. And when we look at real-world data from our customers, this is an example from Facebook. It shows that most of the compute across a wide variety of the workloads they care about is actually going towards overhead of workloads like moving memory or hashing and compression, and that means that our previous way of viewing compute in the data center where major workloads run on general purpose processors are just a portion of the overall compute in the cloud.

Right, the dark part of this graph is a smaller percentage of the overall task at hand, and the lighter part of this graph is a larger percentage, so we have to obviously solve this tax problem and so to solve this, there is a need for silicon solutions that act as a control point across the cloud infrastructure, to accelerate that overhead portion, the infrastructure functions you know, including network virtualization, storage virtualization, security or compression. Essentially a solution that will free up the microprocessor, the CPU, the general purpose solution, and the cores inside of that microprocessor to process data by handling these tax functions, these infrastructure functions.

And we call this silicon solution a new unit of computing the infrastructure processing unit or the IPU. It's an evolution of our smart NIC product line that when coupled with a Xeon microprocessor, it will deliver highly intelligent infrastructure acceleration. And it enables new levels of system security control isolation to be delivered in a more predictable manner. FPGAs can be used to attach for workload customization and overtime these solutions become more and more tightly coupled. So blending this capability of the IPU with the ongoing trend in microservices is a unique opportunity for a function based infrastructure to achieve more optimal hardware and software to in effect solve that problem that I described earlier, right? Deal with overhead tax and more effectively orchestrate the software landscape on a complex data center infrastructure, so by managing infrastructure in this manner IPUs can scale to have direct access to memory and storage and create just more efficient computing right? Jevons paradox applies. So while the term IPU may be new to you are focus in this critical area of system infrastructure is not new. We've been working with our top customers over the past several years, building up this capability technology by technology, and we've gained a strong position in the market along the way. We've already deployed IPU's using FPGAs and very high volume at Microsoft Azure, and we've recently announced partnerships with Baidu, JD Cloud and VMWare as well. We're going to have much more to say about our vision and a road map for IPU's at our Intel on event in October, on both the hardware and importantly on the software side. Now another way to attack the dynamic shift that we're seeing in the hyperscale data center and to more efficiently process the workloads that we're seeing, is to accelerate those workloads. And we're doing that in two ways today. First of all, we're taking the general purpose microprocessor, and we're adding acceleration to it, right? The Xeon microprocessor is now embedding accelerated functions to run specific workloads faster and at the same time we recognize that XPU's have emerged. Discrete application accelerators such as the Habana, Gaudi and Goya solutions, and so let me talk a little bit about that dynamic. The launch of our latest 3rd Generation Xeon processor platform earlier this year was groundbreaking in many ways the 3rd Gen Xeons are the only mainstream X86 server processor with built in AI acceleration and so this capability is led to tremendous improvements generation on generation over 4X improvement for deep learning type of workloads. In addition, for more than a decade, Intel has led the industry in reducing the cost of cryptographic algorithmic innovation with innovations that we've added to Xeon such as Intel's AESNI acceleration instructions.

The 3rd Gen Xeon processor builds upon this leadership position with built-in Intel crypto acceleration with breakthrough performance across a host of important cryptographic algorithms, including public key cryptography, symmetric encryption, and hashing, and we're excited that our customers are seeing up to 4X cryptographic performance improvement and also the 4X improvement in deep learning from the prior generation. So this really is just an amazing feat through CPU built in acceleration. But of course some of our customers demand more and they need more and as a result we made some decisions to build purpose built, dedicated, discrete accelerator's, particularly for AI training and inference. In 2019 we acquired a company called Habana Labs, and the partnership with Habana labs has just been tremendous, and we've seen great momentum for their solutions in the market.

Starting with and most importantly with the AI training solution called Gaudi, we announced a very important partnership with Amazon Web Services at the end of last year, and we're very close now to coming to market later this year with our deep learning training workloads using the Gaudi solutions at AWS for a wide variety of applications such as personalization language processing, object detection and classification. The AWS EC2 instances they're going to leverage 8 Gaudi accelerators and maybe most excitingly for customers is these solutions are going to deliver up to 40% better price performance than the current GPU based EC2 instances for machine learning. So very exciting and anticipate seeing these instances come to market very soon.

The Habana solution will be the first non GPU accelerator offered for training models at AWS as well as for many other customers overtime and importantly developers now have an opportunity to get a head start at building solutions based on this solution at developer dot Habana dot AI is a wide variety of information on how to get started with that if you're a developer.

So another solution we're looking at has been driven by the rapidly expanding cloud environment pushing out to the network in the edge. Right, a critical issue is to address the network itself so that it becomes more efficient to deliver better outcomes for data.

Right, we know that it's no longer sufficient to have a centralized data center processing the world's information. Data has to be processed closer to the edge, and so telecommunications service providers have recognized this, and they've begun to think about the network itself as almost like a cloud in and of itself, right? How do you make the network resources available accessible to different applications, right? How do you make the network more cloud, like more virtualized, more programmable more agile. We have worked together with our telecom service provider customers over the last decade, starting with the transition to network function virtualization for the core network and now for the first time we're bringing this cloud like capability to the edge of the network to what's called the RAN or the radio access network.

The old way of designing networks was to use largely proprietary fixed function equipment in the radio towers and base stations that we all drive by all the time, and this was great for performance, but not ideal for costs or flexibility and the new way is to essentially use the server technology that we've seen in the cloud and in the enterprise such as a Xeon Processor and we've combined that with software for the network such as Flex RAN to run those network workloads for the RAN on cloud like technology, and this is ultimately going to help make the network management more programmable with virtualization at the heart. Just like we saw in the data center and we're going to continue to push this technology even though we're in the early days of this transition to what the industry calls VRAN and I believe that will see very exciting evolution and development here. Just like we saw in the cloud for the RAN. Eventually we can envision seeing the cell phone tower infrastructure, you know, kind of like a mini data center at the edge and that will in turn provide more efficient services, avoid the need to take lots of data back to the central data center and that will lead to lower costs, better performance and a better user experience. Super excited to see leading operators such as Verizon, Dish Wireless, Telefonica, SK Telecom, Rakuten and many others harnessing approving cloud architecture and building out their 5G networks with us. There's many examples of progress that we've made and of course many other partners have been working closely with us to get these deployments into place with the operators. This is a very exciting time for the telecommunications industry transformational I think we'll look back on this time, much like we look back on this time like the early days of the cloud so watch for us at the Mobile World Congress 2021 alongside many customers, where you're going to hear more on how Intel is helping to unleash the power of 5G from the silicon to the software and driving this capability all the way out to the edge.

So tying this back together, our vision for the data center of the future will require a new intelligent architecture for the cloud all the way out to the edge general purpose CPU's and IPU's and XPU's will work in concert to run microservices more efficiently with shared memory and storage and it will be enabled with open source software frameworks to unleash the next generation of microservices with scale and efficiency. And if we do a good job of innovating relentlessly here, we will see dramatic improvements in efficiency and performance. And Jevons Paradox will lead to another wave of explosive growth for all of us.

In fact, I believe we will see the biggest and fastest buildout of computing infrastructure in human history over the next decade. Thank you again for having me here and thank you to our partners and our customers for collaborating with us. The Golden Age of data is here. Let's get after it.