Intel Corporation

Intelligence, Performance, Visibility with Intel® Intelligent Fabric

CORPORATE PARTICIPANTS

Xiaojun (Shawn) Li
Sales Director, Next Wave OEM & eODM

Todd Koelling
Director of Marketing and Technology Planning

PRESENTATION

Shawn Li

Welcome, everyone to the Intel Network Builders Insights Series. I'm Shawn Li, Sales Director, Next Wave OEM & eODM at Network and Communication Sales Organization at Intel. And I am your host for today's webinar. Thank you for taking the time to join us today for our webinar titled “Intelligence, Performance, Visibility with Intel Intelligent Fabric”.

Before we get started, I want to point out some of the features of the BrightTALK tool that may improve your experience. There is a Questions tab below your viewer. I encourage our live audience to please ask questions at any time. Our presenters will hold answering them until the end of the presentation. Below your viewer screen, you will also find the Attachments tab with additional documents and reference materials which pertain to this presentation.

Finally, at the end of the presentation, please take the time to provide feedback using the Rating tab. We value your thoughts and we will use the information to improve our future webinars. Intel Network Builders Insights Series take place live every month, so please check the channel to see what is coming and access our growing library of recorded content.

In addition to the resources you see here, we also offer a comprehensive NFV and 5G training program through Intel Network Builders University. You can find the link to this program in the Attachments tab, as well as the link to the Intel Network Builders Newsletter.

Today, we are pleased to welcome Todd Koelling from Intel. Todd is the Director of Marketing and Technology Planning in Barefoot Switch Division at Intel. He is responsible for marketing and technology planning of the Intel Tofino Intelligent Fabric processors. Prior to this role, he was a segment market manager for communications service provider segments in the Intel Connectivity Group, where he also drove the marketing and ISV enablement for the Application Device Queues, ADQ, technology. Todd has more than 20 years’ Intel experience with a platform and solution focus, encompassing processors, chipsets, graphics, FPGAs, Silicon Photonics, ether network adapters, storage and application software.

Welcome, Todd, and thank you for taking the time to join us today. I will hand it over to you to start off. Thank you.

Todd Koelling

Thank you, Shawn, and thank you, everybody, for joining us today. As Shawn mentioned, I'm going to talk about Intel's vision for Intel Intelligent Fabric and how it brings intelligence, performance, and visibility to your network.

I've just got some of the standard disclaimers here. The only thing I would note is we are going to show some performance data. Obviously, that's going to vary depending on your configuration.

In terms of an agenda, first I'll talk about our vision for Intel Intelligent Fabric as a part of the Data Center of the Future. Then we're going to dive into some key technologies. This is all based on the P4 programming language. We'll talk about that and how we use it. I also want to talk about a key new technology under development for Congestion Control called Remote Priority Flow Control or Remote PFC. Then a key aspect of this entire solution is the In-band Network Telemetry and tracing the packets as they go through the network and decisions that can be based on those. Then I'll round it out with specific products and applications, as well as at the end, ways that you can contribute.

Back in June of this year, Intel announced our vision for the Data Center of the Future, and this involves various compute server nodes, as you can see around the edges here, in terms of general purpose compute, AI acceleration, computational storage. As part of that we introduced also the infrastructure processing unit, or IPU, the blue boxes there, and those are really designed to complement these Xeon processors by performing basic networking and communication functions, so that the processes can focus on more value-added tasks. Then in the center, connecting together these data center servers, you see Intel Intelligent Fabric that's going to pull all those together and help those to cross-communicate, and so this is what we're going to focus in on today. We're going to peel back that layer in the center there on Intel Intelligent Fabric and give you our vision for that.

But before we get to the solution, let's talk about the problem or the challenges. On the left here, you can see the problem I think we're all aware of is just the increase in traffic across the network, with 5G, IoT. Video Streaming, of course, is a major bandwidth hog, but also you have AR, VR, other new types of services and solutions coming into play, and so this just continues to drive the amount of traffic that we need to transfer, to analyze, understand, and transmit, and going up to 175 zettabytes as projected here in 2025.

But that's just the traffic side of things. The other thing that many customers I talked to are really heads down and focused on are the software aspects, the move to network function virtualization, NFV, and SDN, and cloud native architecture using container-based microservices. This just takes a big focus right now to implement that. On top of that, we have AI entering the picture and other changing workloads that really are spawning the need for growing network optimization, increased distributed workloads, as well as storage. Distributed or pooled storage are changing the architecture. And really, operational challenges, figuring out exactly what's going on in the network is increasingly difficult. And of course, the security is always a concern from end-to-end. We need to pay attention to that and make sure our network is very secure, and do all the above, while not increasing our network CapEx and OpEx too much, or if anything, reducing those for service providers.

So, quite a bit of challenges here, but the point is, it's not just increasing bandwidth, but the network also needs to get smarter to deal with this new environment.

So, to address this, Intel's put together our vision for what we call Intel Intelligent Fabric, and first let me define what I mean by fabric in this context, and as shown in the picture in the center here, what we mean by fabric is leaf-spine topology, no particular one in mind here, but in general, a leaf-spine topology with optimized connections. And what's unique about this that we're bringing to bear is a full suite of Intel products and technologies. So, at the heart of it is the Intel Tofino Intelligent Fabric processors. We'll spend a lot of time in depth on those today, but also the infrastructure processing units, or IPUs I just mentioned, ethernet adapters, Xeon processors, various accelerators, including graphic accelerators and FPGAs, and these are all interconnected with the high-speed optical Silicon Photonics. That's more on the chip or the silicon side, but also on the software and industry standard size. There's many things that we are targeting to help bring this together and make it easy to implement, such as the P4 programming language. The SONiC is an open source networking OS, and we recently worked with the industry to roll out the IPDK, which is probably new to most people here, but the Infrastructure Programmer Development Kit. This provides a convenient environment for programming using the P4 language, these various targets across the network.

And the goal of all of this is really shown on the lower left here to just bring about ease of use in terms of implementing or moving workloads, addressing the massive bandwidth problem we just talked about, adding increased AI, self-monitoring, self-analyzing, self-healing networks, enhancing the security, keeping that robust while we're doing all this, and then likewise, addressing those CapEx and OpEx issues by improving the density, power, and cost of this setup.

Now, if we put all these elements together, the benefits fall into what I call three benefit vectors. So, these are things that we offer today, and we will continue to improve and bring new features on, in these various buckets or areas. So, the first is intelligence, the second is performance, and then the third is visibility and control, which is really what makes this solution unique through the programmability and in-band network telemetry.

So, let's start with intelligence, go a bit deeper there. The first key thing is that this is a fully customizable P4 programmable pipeline, that you can program all these network elements now, making the network as a programmable platform across all the switches, the ethernet adapters, the IPUs, and so on. Also, we offer intelligent packet processing for accelerating AI and ML workloads. So, to be clear, we're not using AI within the switch to accelerate the switch, but we're using the switch to accelerate specifically ML training, machine learning training, as a good example. You can use the FPGAs for very heavy duty workloads to expand the table and buffer sizes, and we offer security features with the Xeon processors such as Software Guard Extensions, Intel SGX, or Trust Domain Extensions, Intel TDX.

In terms of performance, we're continuing to beef that up. We're up to 25.6 terabits per second with the new Intel Tofino 3 Intelligent Fabric processor, and that uses very high-speed 112 gigabits per second SerDes, but it also offers 56 gigabits per second flavors to make it an easier transition from today's 56 Gig speeds and existing implementations. High-speed Silicon Photonics offers optical connections, and then we've also optimized for specific hyperscaler use cases with the Intel Fabric processors.

Then, as mentioned earlier, visibility and control is key here, just watching these packets as they work their way through the networks, finding out how they got there, did they get delayed, any issues we might see, and then once we understand that, then we can get in and address those, and that uses In-band Network Telemetry, or INT, and on top of that Intel has a tool called Deep Insight Network Analytics Software that provides reports that allow you to go in and remedy that, and increasingly, we see the use of AI to help automate that and provide insights into what's happening.

You can do traffic monitoring and steering for enhanced security, filtering out packets you don't want in the switch, coming into the switch, or likewise, steering around places maybe you don't want to go for security reasons, and then a lot of this, it's not just processing the data, but hearing it from the Intel IPUs, as well as the ethernet network adapters.

A key part of this is also the software and to address this we're providing an end-to-end software architecture that provides a consistent programming model across a variety of data plane targets throughout the network. So, for example, you can use this to program host-based servers using DPDK and Intel Xeon processors. The IPUs come in two flavors. The Mount Evans, this is an IPU ASIC we actually co-developed with Google, and that's one flavor of IPU. The other is FPGA-based IPUs that are totally configurable. An example of that would be the Oak Springs Canyon IPU, and of course, we have the Intel Tofino IFP-based switches or intelligent switches. All of these can be used using this common P4 programming front-end.

In terms of the control plane, another big part of this is the SONiC open source networking OS. This is a project through OCP, and it is an open source networking OS for the control plane, and just this past November at the OCP Summit, we announced that, in conjunction with our partner Microsoft, we have begun contributions to the Switch Abstraction Interface, or SAI, Packet Test Framework, and that will grow to about a thousand tests over time as we move forward.

We also recently announced at the Intel Innovation event in October 21 that a new product at the heart of this Intel Intelligent Fabric, and that is the Intel Tofino 3 Intelligent Fabric processor to address this explosion in 5G and IoT data that we talked about earlier, as well as the distributed workloads, and accelerating AI, and providing this cloud-to-edge visibility. The Tofino 3 is ideally suited for that, and it brings the intelligence, performance up to 25 terabits per second, and visibility and control necessary to do that. And we talked about data center and the Intelligent Fabric, and certainly it's really targeted and optimized for that, but people may think of, well, is that just cloud and hyperscale data centers? Well, no, it can be cloud or edge data centers. It's also well suited for high performance computing, or HPC, or comms service providers moving to cloud-based technologies. So, there's a variety where this-- a variety of uses for this focused technology. And the overall theme here is moving beyond just this programmability, but a switch to intelligence with Intel Tofino 3 IFP.

So, next, I'd like to dive deeper into some of these technologies and flesh out for you what we mean by the intelligence.

So, first, a great analogy I like to use is for the P4 programming languages, is looking how different compute architectures over time have unique languages that work well for them. So, I think everyone's obviously familiar with C++, and being used for CPUs. You can

also use OpenCL for GPUs, MATLAB for DSPs, or TensorFlow now for the machine learning and AI type of processing. The good analogy for that then would be P4 for the networking programming. So, this is an open source language. It is run by the ONF community, and it can be used for, basically, programming all of these different elements of the network.

It has a very broad range of support. Intel is a key player and leader in this, but this community has grown to over 4,000 developers and 100-plus member organizations, and again, it's across multiple industries and levels of implementation here, OEM system integrators, researchers, all digging in and cooperating in the community to develop these capabilities. And if you're interested further, I recommend you go to P4.org, and you can learn more about the community and see more examples of how this language is being used.

Next, I want to dive deep here and roll up our sleeves in the technical portion of this presentation and talk about a new technology we called Remote Priority Flow Control, or Remote PFC.

Now, to start with, let's look at the challenge here, or at least one of the challenges in the data center in particular related to Congestion Control, and that's what we call Incast Congestion. This is a case where you have a many-to-one traffic pattern. So, as illustrated on the diagram on the right, we have various servers sending traffic over to a dedicated pool of servers on the right. A good example of this would be RDMA-style workloads with distributed or pooled storage. So, we have multiple hosts accessing this storage on the right.

Another example would be AI workload updates, where we have multiple people implementing the workload and then we need to do some updates to that model, so we send it to a central server to calculate the new model, and then we have to send it back. Sending all of this traffic up into that limited number of servers can really create what we call the Incast Congestion through many-to-one style traffic pattern. Now, there are some ways to address this. RoCE v2 has what's called DCQN to help address that, but it's an end-to-end approach and it can be slow and take time. IEEE has an existing standard called PFC, or priority flow control, and that's when the target servers here, if their queues get full, then they can send a message back to the previous layer in the network to stop sending, but this can create head-of-line blocking or some other side effects that may not be anticipated, and there is also possibility of packet loss as a result. So, typically, this is just implemented in the ToR. If we get implemented all the way across, then that would help. But, like I said, it gets a little complicated to do that, so people typically just implement it in the Top-of-Rack switch right now.

So, to help address this, we're working with IEEE, and coming up with a new technology we call Remote Priority Flow Control, or it's also-- the name is still being finalized. It's also I think most recently been termed source PFC, but regardless, this technology approach is getting some really good traction.

And so to try to explain it quickly here, we have our senders on the left, multiple senders, multiple traffic flows, going through this first Top-of-Rack switch on the left, then to a second Top-of-Rack switch, and then ultimately to the receiver. So, we're just showing these two layers right now to illustrate the issues. And so what happens today with PFC, if we look in the center there, the number one, we get all that incast traffic, and with PFC, at some point, we realize, uh oh, our queues are full, we got to stop. So, it'll send a message back to the host in the previous node in that Top-of-Rack switch number one, and say, hey, please stop sending. And so, OK, that stops the flow, but basically, it stops everything, and so it ends there. But with Remote PFC, or source PFC, we take what came in as an L3 packet, we convert that to an L2 frame, and we can send that back to specific senders and not only tell them, hey, stop sending, but for how long to stop sending. So, if there's someone who's being a real traffic hog, we can go ahead and tell them to stop so that others can get through. In the meantime, it actually does work in congestion with the RoCE v2 DCQN... I'm sorry, DCQCN, and you can see that on the right, so we can still implement that in conjunction there with number four and number five. So, using this system, we have a much faster and more targeted end-to-end Congestion Control.

Now, where does the intelligence come in? So, the key here that I want everyone to understand today is in this intel Tofino 2 layer, you can see the microscope there looking at what's happening, and the key there is in that ToR, using the switch silicon, we can set a trigger threshold. So, rather than waiting until all the buffers are full, we can read that queue depth, and then calculate a target queue depth. So, we can program that in there, and then we can tell the specific sender to stop and for how long they need to stop. Now, it's just a rough estimate. It's not super precise, but it's working very well in the tests that we've run so far. So, this is an example here, in this

number one where the microscope is that-- I'm sorry, the magnifying glass, where we can add that intelligence and improve the performance of the network.

OK, but don't just take my word for it. Let's look at some real data here, and so here we have a case, we'll look at the test setup on the left first, where we have multiple senders here. So, we have a sender in red, and then multiple blue senders generating hundreds of terabits per second of traffic. OK, so they send that into the Top-of-Rack one, switch number one, and that's illustrated in the gray there, and then that goes across to a Top-of-Rack two switch and then it gets to the appropriate receiver in red or the receiver in blue. And so, if we look at the queue depth and how long it takes to transmit that data, we can see a comparison now of the existing IEEE PFC, priority flow control, on the top set of graphs versus the Remote PFC, this new approach with the intelligence that I talked about. And so, with that-- because the problem here is that all this traffic from the blue senders, the 20:1 incast, can overwhelm the traffic in the red on the 4:1 incast.

OK, so the key, if we look at these charts, if we dig into that, this shows the queue depth, so lower is better here. You want a lower queue depth, so we're not clogging up the buffers, and then you also want it to complete sooner. So, if we look at that, the ToR1 to ToR2 port, the one in gray there, you can see that our queue depth increases. Now, that's still well within the realm of what the Torfino 1 or 2 can handle, but it does get backed up and then it takes longer to transmit using PFC. With Remote PFC, by slowing down those blue senders, asking them to pause at times, then you see our buffers barely get full, and then we complete much sooner. Rather than the 50 seconds, it takes only about 37 seconds. Likewise, when we get to the Top-of-Rack 2 switch, going out to receive 2, you can see that initially with PFC, it bumps up, it comes down a level, but it still takes a while. The queues are much lower filled when we get to receiver 2. OK, and that's the bulk of our traffic, but the big difference then is in the red there for receive 1, which before that was getting clogged up with all the traffic from the blue senders. So, you can see there that without PFC, it takes about 50 seconds, and the queue depth fluctuates, but with Remote PFC, the queues stay relatively less full, and it completes in only about less than 25 seconds. So, it's almost half the completion time. So, this is just a good example of how using this Remote PFC, you don't get bogged down with the high traffic on other secondary traffic flows.

Another way to look at this is the completion time, looking at the P50, the median latency, or the P99.9 tail latency, which is typically of most concern. You see you've got about an 8% reduction here, again, but the thing is underneath this, the red completes much faster than that blue one. The blue is about the same, so that keeps our average latency higher, but the red is completing much faster. And then there's a huge reduction, a 73% reduction in our median latency. So, this shows very tangibly the benefits of using that approach.

All right, so that was Remote PFC, and then the last thing I want to talk about is just highlighting the In-band Network Telemetry, and how we can follow each packet's journey from beginning to end, from our source hosts into the destination host, and we can use the INT. Basically, we're going to add header data to the packets, and then we can use that to trace them as they go through the system. So, for every packet now, we can determine how did it get there? Why is it here? How long was it delayed, and if it was delayed, why was it delayed? So, this is a great visibility feature that can improve your network performance.

OK, oops, it jumped too quickly there. OK, so those are some of the technologies. Let me roll that now into how that's manifested in different products and applications. So, first of all, in terms of a roadmap, these are the Tofino Intelligent Fabric processors. The original one was the Intel Tofino and that had up to 6.4 terabits per second, 25 Gig SerDes, that's in production now. The Tofino 2 just entered production mid this year, and that has up to 12.8 terabits per second and 56 Gig SerDes, and then the Tofino 3 Intelligent Fabric processor that we just announced doubles that bandwidth from 12.8 up to 25.6. It does use a modular chip design, meaning for the higher-end bandwidth it has two chips connected with an EMIB chip-to-chip bus interface, and as I mentioned earlier, it boosts the SerDes up to 112 gigabits per second, but it also supports 56 gigs, so that makes an easy transition from Tofino 2. And so all of these, again, bring this intelligence with P4 programmability, the AI/ML acceleration, and the high security boosts the performance, including not just the throughput, but also the power optimization with these hyperscaler-targeted use cases, and then the visibility and control through the In-band Network Telemetry.

There's a wide range of applications these can be used for. So, again, they're all based on P4 and you, or the community, or we have various ISV, and system integrator partners can help develop that really customized pipeline. So, you get the maximum utilization out of the chip, and this can be used for all sorts of different applications, not just enhancing your routing and switching, but other specific applications like a network packet broker, broadband network gateway, load balancing, and so on. So, there's a wide variety of ways to innovate and utilize these very customizable, high performance products.

Then just looking a little bit more into the Deep Insight. This is a tool that Intel provides that sits on top of the INT or utilizes the INT, and so you can use this for the path and latency tracking, congestion analysis, packet drop analysis, topology discovery, UI reporting, and real-time notifications, and then also data retention and historical analysis. So, this is a great tool to enhance that performance and give you additional insight into that Deep Insight capability.

Let me also use this juncture to note that we also, from an open source standpoint, introduced what we call host INT for Linux servers, and this takes the P4, In-band network technology, In-band Network Telemetry capabilities and then opens that up to the Linux community. So, there's multiple ways that you can attack this problem and gain insight to improve your network performance.

So, to bring this to a summary, exciting new technology and approach here, starting with our vision of the Intel Intelligent Fabric and then implemented using P4 on the Intel Tofino Intelligent Fabric processors. Ways that you can get involved or contribute are joining the P4.org community. If you haven't already, I recommend you go check that out. In addition to the website, we host periodic workshops where people can exchange ideas and develop different sorts of applications. I mentioned earlier, the OCP SONiC networking OS is an open source approach, and of course, there's also commercial solutions available as well from many of our partners.

We have what we call the Intel Connectivity Academy, where you can go and you can take classes and learn about P4 and how to implement these concepts, and we also have a very strong relationship with the academic and university community around research in what we call the Intel Connectivity Research Program. We recently just did a roundup of all the contributions of people in this program and found over 60 in-depth technical papers on how to use P4 in a networking environment. And then for more information on the Intelligent Fabric processors, you can visit intel.com/ifp and it provides details on the products and the software that I've mentioned here today.

All right, so that brings us to our conclusion. Shawn, let me direct it back to you then for any questions that may have come in.

Shawn Li

Great, thank you. And we have some questions here and let me start it. Question one is, Tofino 3 IFP targeted only at data center? Can it provide a benefit in any other deployments?

Todd Koelling

OK, great question because I spoke of it today in the context of the Intel Intelligent Fabric, which is clearly targeted towards the data center and the hyperscale data centers, cloud data centers at that, but that technology certainly translates to other places as well. So, first of all, we have not only cloud data centers, but edge data centers as well today, so going to be using both of those. You have high performance computing, which uses a very similar structure, and more and more, we see the telecommunications or communications service providers moving to these cloud native container-based networks, and so it can likewise benefit that. So, it really is a wide range of technologies or end use cases that I think it's suited for.

Shawn Li

Great. So, another question. Why do you call the Intel Tofino 3 an Intel Intelligent Fabric processor? Is it really just a switch?

Todd Koelling

OK, good question and something we're quite excited about. It is a switch, it performs your basic switching functions, but it's a switch and more, and that's what we want to emphasize. It's really bringing intelligence to the switching functions, and some examples of that are the P4 programmability. You can fully customize your pipeline. Also, we didn't get into deep today, but the AI/machine learning training acceleration. Similar to our incast example, you can have multiple-- as you're developing models, you can have multiple senders and you can disperse that across multiple nodes, and so we can very-- rather than sending each packet one at a time, we can bundle those up and then send it to the worker nodes individually. Also, the hyperscale use case power optimizations, again, another way of adding intelligence, as well as the Remote PFC that we talked about in-depth in our example today. So, it really is taking the switching to a new level, and so that's why we've dubbed it the Intelligent Fabric processor.

Shawn Li

Great, thank you. And question three. What's new with Tofino 3 apart from the higher throughput you mentioned?

Todd Koelling

Yes, so first of all, so the architecture, the base architecture in Tofino 3 is the same as Tofino 1 and Tofino 2. It's a match-action pipeline architecture, which is good because then there's commonality. I think what's new is that the higher throughput, as I mentioned here, the 112 Gig SerDes, and then, for example, like the AI technology, you can prototype on that with the Tofino and Tofino 2, but we really see that maturing and coming into play with Tofino 3.

Shawn Li

Great. The next question, is there any architectural difference between the Tofino 2 and the Tofino 3?

Todd Koelling

OK, yes, I just touched on that in my previous one, but to reiterate, yes, they use the same architecture, this match-action pipeline architecture, which has multiple match-action units with it, and that's really what's unique about it, so we retain that same architecture, which makes it easy to program across all the three products. So, that is actually the same. It’s some of the speed and software enhancements on top of that that make Tofino 3 different.

Shawn Li

Great, and the next question. When will the Tofino-- well, just a second. Does the-- just a second. When will the Tofino 3 IFP be available?

Todd Koelling

Yes, good question. We just said production in the future on that one slide, but let me give everybody a specific data point, is that we will have customer samples for that in Q2 of 2022, so Q2 of next year. So, the customer samples are just around the corner, and like I said, for many functions, you can start prototyping now with Tofino 2.

Shawn Li

OK, available in Q2 2022.

Todd Koelling

Q2 2022.

Shawn Li

Good, thank you. And the last question so far, Todd, does the high speed of Tofino from 6.4 to 25.6 impact on the performance, flexibility, or programmability?

Todd Koelling

Yes, the main benefit there is going to be the-- just that raw throughput, right, having the 25.6 terabytes. One way you can look at it is as you go from the edge further into the network core, to the cloud and data center, the further in you get, the more traffic there's going to be, and so the more-- the more bandwidth that is needed, the more throughput is needed. And so, for example, we really see Tofino 3 primarily being deployed in the hyperscale data centers initially. You really don't need that 25.6 out at the edge. That's not necessarily needed. So, Tofino 2 and maybe even Tofino 1 is adequate there today. So, can really see the Tofino 3 starting in the cloud data centers, in the wireless core, and then eventually moving its way out as the traffic just keeps going up and up.

Shawn Li

Wow, good. Good, and that's all the questions I received from the audience. And thank you for joining us today. Please do not forget to give our team a rating for the live recording so that we may continuously improve the quality of our webinars. Thank you again for joining us today. This concludes our webcast. Thank you.