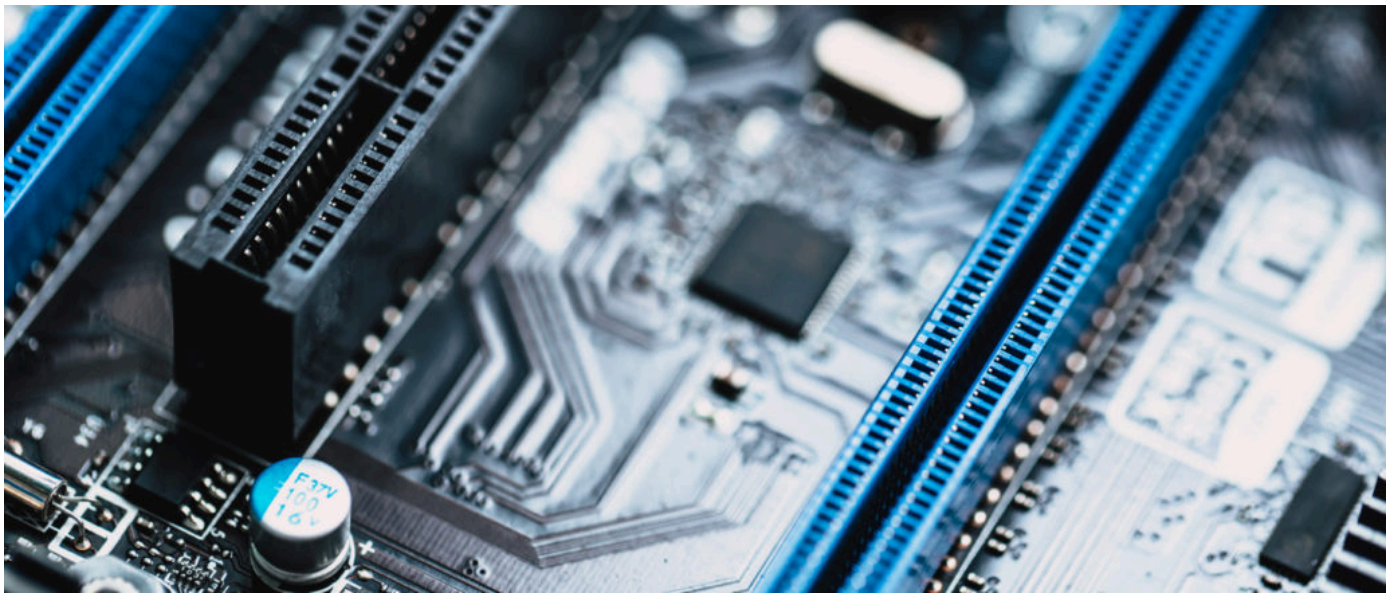
PCI-EXPRESS MUST MATCH THE CADENCE OF COMPUTE ENGINES AND NETWORKS

July 7, 2023 Timothy Prickett Morgan



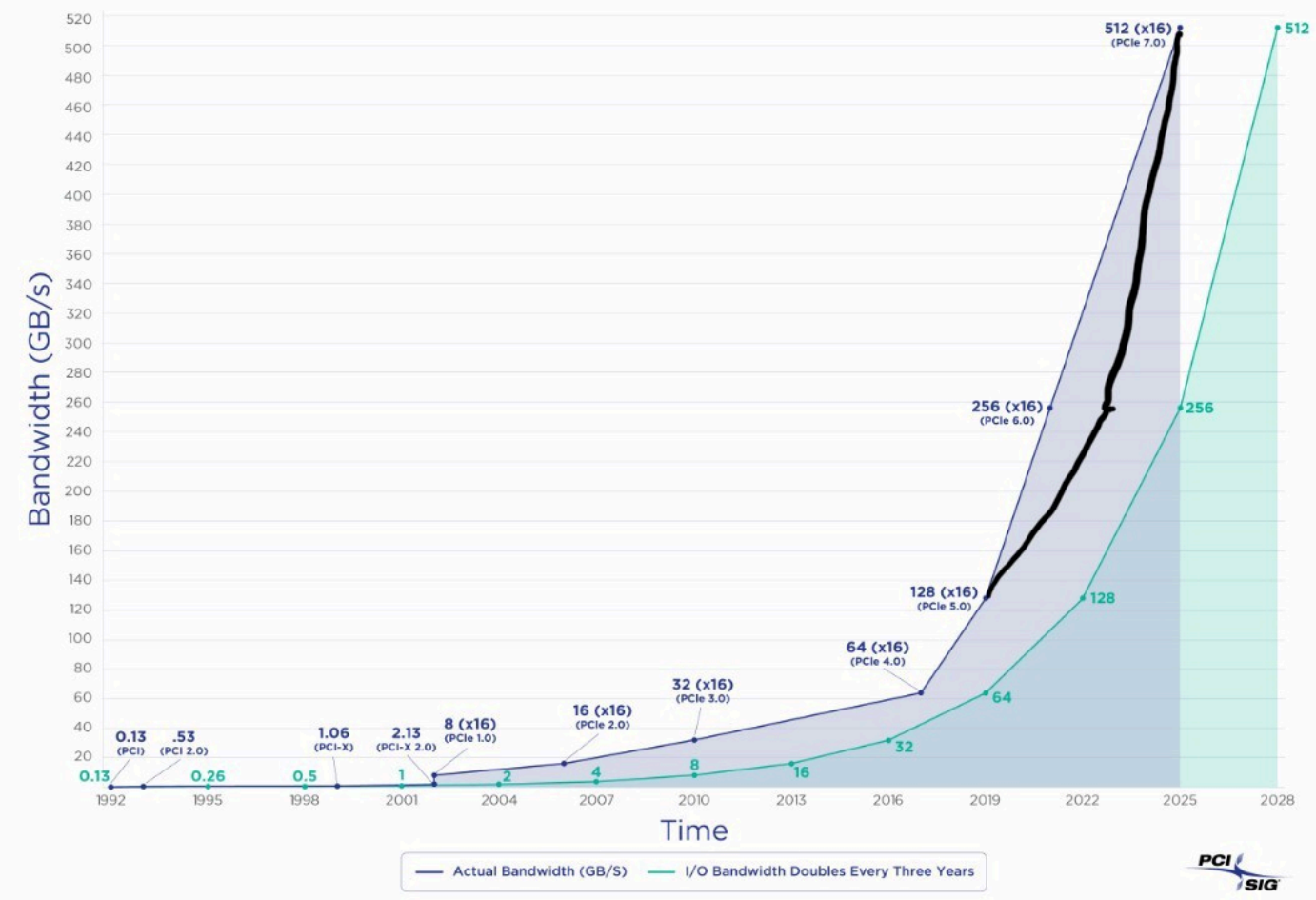
When system architects sit down to design their next platforms, they start by looking at a bunch of roadmaps from suppliers of CPUs, accelerators, memory, flash, network interface cards – and PCI-Express controllers and switches. And the switches are increasingly important in syste

designs that have a mix of compute and memory types and for clusters that will be sharing components like accelerators and memory.

The trouble is this: The roadmaps are not really aligned well. Most CPU and GPU makers are trying to do major compute engine upgrades every two years, with architectural and process tweaks in the year in between the major launches so they have something new to sell every year. Makers of chips for networking switches and interface cards in the Ethernet and InfiniBand markets tend to be on a two-year cadence as well, and they used to tie their launches very tightly to the Intel Xeon CPU launch cadence back when that was the dominant CPU in the datacenter, but that rhythm has been broken by the constantly redrawn roadmaps from Intel, the re-emergence of AMD as a CPU supplier, and a bunch of other Arm CPU makers, including at least three hyperscalers and cloud builders.

And then there is the PCI-Express bus, which has been all over the place in the past two decades. And while PCI-Express specifications have been released in a more predictable fashion in recent years, PCI-Express controllers have been faithful to the PCI-Express roadmaps but PCI-Express switches are well behind when it comes to product launches from MicroChip and Broadcom.

Sitting here on a quiet July morning, thinking about stuff, we think all of these roadmaps need to be better aligned. And specifically, we think that the PCI-SIG organization that controls the PCI-Express specification and does so through a broad and deep collaboration with the IT industry, needs to pick up the pace and get on a two-year cadence instead of the average of three it has shown in the past two decades. And while we are thinking about it, we think the industry would be better served with a short-cadence jump to PCI-Express 7.0, which needs to be launched as soon as possible to get I/O bandwidth and lane counts in better alignment with high throughput compute engines and what we expect will be an increasing use of the PCI-Express bus **to handle CXL-based tiered and shared main memory**.



We have tweaked this bandwidth chart from PCI-SIG, which does not show the PCI-Express 6.0 spec being released in 2022, as it was, but in 2021, which is incorrect.

Don't get us wrong. We are grateful that the PCI-SIG organization, a collaboration between all kinds of companies in the datacenter and now at the edge, has been able to get the PCI-Express bus on a predictable roadmap since the very late PCI-Express 4.0 spec was delivered in 2017. There were some tough signaling and materials challenges that kept the datacenter stuck at PCI-Express 3.0 for seven years, and we think Intel, which dominated CPUs at the time and dragged its feet a little bit on boosting I/O because it got burned with SATA ports in the chipsets used with the "Sandy Bridge" Xeon E5s that came out later than expected in March 2012. Rumors abounded about the difficulties of integrating PCI-Express 4.0 and PCI-Express 5.0 controllers into processors since then.

Generally, a PCI-Express spec is released and then within about a year or so we see controllers embedded in compute engines and network interface chips. So when PCI-Express 4.0 came out in 2017, we saw the first systems using it coming out in 2018 – specifically, IBM's Power9-based Power Systems machines, followed by its use in AMD "Rome" Epyc 7002s launched in August 2019. Intel didn't get PCI-Express 4.0 controllers into its Xeon SP processors until the "Ice Lake" generation in April 2021.

And even with the short two-year jump to the PCI-Express 5.0 spec in 2019, it wasn't until IBM launched the Power10 processor in its high-end Power E1080 machines in 2021 that it became available in a product. AMD didn't get PCI-Express 5.0 into a server chip until the "Genoa" Epyc 9004s launched in November 2022 and Intel didn't get PCI-Express 5.0 into a server chip until the "Sapphire Rapids" Xeon SPs launched in January 2023.

So it was really a three-year cadence between PCI-Express 4.0 and 5.0 *products*, as expressed in the controllers on the CPUs, even if the *spec* did a two-year short step.

We think that the specs and the products need to get on a shorter two-year cadence so the compute engines and the interconnects can all be lined up together. And that includes PCI-Express switch ASICs as well, which have traditionally lagged pretty far behind the PCI-Express specs for the 3.0, 4.0, and 5.0 generations that they were widely available.

The lag between PCI-Express ports and PCI-Express switches at any given generation are a problem. That delay forces system architects to choose between composability (which ideally uses PCI-Express switches at the pod level) or bandwidth (which is provided through a direct server slot). Systems and clusters need to be designed with both composability and bandwidth – and we would add high radix to the mix as well.

At the moment, there are only two makers of PCI-Express switches, Broadcom (through its PLX Technologies acquisition a number of years ago) and MicroChip. We profiled the MicroChip Switchtec ASICs at the PCI-Express 5.0 level way back in February 2021, which scale from 28 to 100 lanes and from 16 to 52 ports, but as far as we know, they are not shipping in volume. Broadcom unveiled its PCI-Express 5.0 chip portfolio back in February 2022, including the ExpressFabric PEX 89100 switch, which has from 24 to 144 lanes and from 24 to 72 ports. We are confirming if these are shipping as we go to press and have not heard back yet from Broadcom.

Our point is that PCI-Express switches have to be available at the same time that the compute servers, memory servers, and storage servers are all going to be created using chips that support any given level of PCI-Express. On Day One, in fact. You have to be able to embed switches in the servers and not lose bandwidth or ports or sacrifice radix to get bandwidth. We therefore need lots of suppliers in case one of them slips. This is one of the reasons why we were trying to encourage Rambus to get into the PCI-Express switch ASIC racket recently.

All of this is top of mind just as the PCI-SIG has put out the 0.3 release of the PCI-Express 7.0 spec.

Let's take a look at the projections we did for the PCI-Express roadmap a year ago when the PCI-Express 6.0 spec was wrapped up and PCI-Express 7.0 appeared on the horizon:

PCI	Spec	Data			Maximum Server Slot	
Specification	Released	Rate	Encoding	Frequency	Bandwidth	Type
PCI	1992	1.06 Gb/sec	*32b/34b*	33 MHz	133 MB/sec	32-bit Simplex
PCI 2.0	1993	4.26 Gb/sec	*64b/66b*	66 MHz	533 MB/sec	64-bit Simplex
PCI-X	1999	8.5 Gb/sec	*64b/66b*	133 MHz	1.06 GB/sec	64-bit Simplex
PCI-X 2.0	2002	17 Gb/sec	64b/66b	266 MHz	2.13 GB/sec	64-bit Simplex
PCI-Express 1.X	2003	2.5 Gb/sec	8b/10b	2.5 GT/sec	8 GB/sec	x16 Duplex
PCI-Express 2.X	2007	5 Gb/sec	8b/10b	5 GT/sec	16 GB/sec	x16 Duplex
PCI-Express 3.X	2010	8 Gb/sec	128b/130b	8 GT/sec	32 GB/sec	x16 Duplex
PCI-Express 4.0	2017	16 Gb/sec	128b/130b	16 GT/sec	64 GB/sec	x16 Duplex
PCI-Express 5.0	2019	32 Gb/sec	128b/130b	32 GT/sec	128 GB/sec	x16 Duplex
PCI-Express 6.0	2021	64 Gb/sec	PAM-4, FLIT	64 GT/sec	256 GB/sec	x16 Duplex
PCI-Express 7.0	*2023*	*128 Gb/sec*	*PAM-16, FLIT*	*128 GT/sec*	*512 GB/sec*	*x16 Duplex*
PCI-Express 8.0	*2025*	*???*	*???*	*256 GT/sec*	*1 TB/sec*	*x16 Duplex*
PCI-Express 9.0	*2027*	*???*	*???*	*512 GT/sec*	*2 TB/sec*	*x16 Duplex*
PCI-Express 10.0	*2029*	*???*	*???*	*1 TT/sec*	*4 TB/sec*	*x16 Duplex*

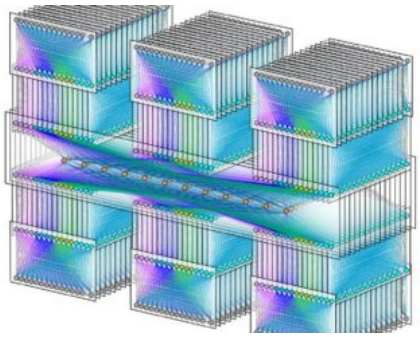
The PCI-Express 7.0 spec is not expected to be ratified until 2025, and that means we won't see it appearing in systems until late 2026 or early 2027. We think this wait is far too long. We need PCI-Express 7.0 to provide the kind of bandwidth accelerators need to chew on an enormous amount of data that is required to run a simulation or train an AI model. We need it matched up with a fully complex CXL 4.0 specification for shared and pooled memory.

We understand that it would be hard to accelerate PCI-Express 7.0 controllers and switches to market, and that all manner of products would also have to be accelerated. Compute engine and peripheral makers alike would be hesitant to not try to squeeze as much investment as possible out of their PCI-Express 6.0 product cycles.

Still, as PCI-Express 6.0 is put into products and goes through its rigorous testing – which will be needed because of the new PAM-4 signaling and FLIT low-latency encoding that it makes use of – we think the industry should start accelerating and match up to the CPU and GPU roadmaps as best as possible and to get onto a two-year cadence alongside of them.

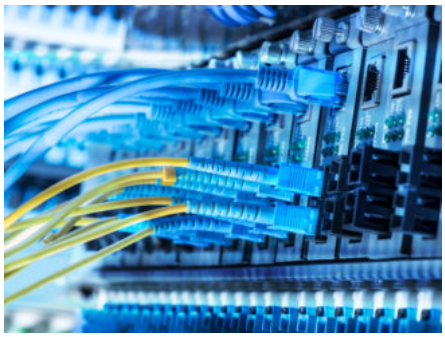
Get the components in balance and then move ahead all at once, together.

RELATED ARTICLES



GETTING META: ABSTRACTING AND MULTISOURCING THE NETWORK LIKE AN FBOSS

DATACENTER IS THE NEW UNIT OF COMPUTE, OPEN NETWORKING IS HOW TO AUTOMATE IT



SUPPLY CHAIN EASING CREATES ETHERNET SWITCHING BOOM

10 COMMENTS

q^8 says:

JULY 8, 2023 AT 12:51 AM

In my gastrointestinal opinion, those higher PCIe+CXL levels have the potential to catapult the server scene into a terroir of organic growth, a healthier computational microbiome, and long-lasting freshness. Notwithstanding current AI's appetite for quickly chewing-through and lossily digesting copious servings of mystery data, without properly savoring it, leading to possibly bloated software guts, bubbly (super)market bursts, generalized discomfort, and zombie hallucinations. The prompt ingestion of a composable fiber substrate (PCIe6.0+/CXL3.0+) seems to be just what the doctor ordered, for sustained cycles of regularity and enthusiasm, in the server industry's occasionally fluctuating bowels! q^8

REPLY

hoohoo says:

JULY 8, 2023 AT 10:55 AM

Thanks for the laughs! Bravo!

REPLY

emerth says:

JULY 8, 2023 AT 11:08 AM

As an aside... won't someone think of the desktop market? It's not chump change.

We need a bifurcation in the PCIe version shipped with server and with desktop/notebook CPUs. PCIe5 has driven prices up and expansion options down on the desktop. We used to get 4 or even 5 x8/x16 electrical PCIe3 slots on a mid price mobo, then a pair of x8 PCIe4, and now anything other than the single x16 PCIe5 option is horrifically expensive. Imagine what a mobo will cost once PCIe7 is the standard connect to an AMD or Intel CPU!

Certainly ever higher bandwidth and ever lower latency PCIe is a boon for building data centers and clouds, but the other end of the market is not served well by this at all.

REPLY

Laurence 'GreenReaper' Parry says:

JULY 8, 2023 AT 12:21 PM

If I recall correctly, NVIDIA recently pushed back it's forthcoming architecture. Since they're also having problems keeping up in the CPU and switch arena, wouldn't it make sense to move everything to a three-year cadence instead? That way we might not need to re-evaluate systems so often, too.

REPLY

peacekeeper44 says:

JULY 9, 2023 AT 11:17 PM

I am having a hard time seeing how the comparison of PCI-E standards keeping up with the CPU/networking cadence makes any sense. Just because all the peripherals having a generation leap every 2 years does not mean they have doubled their capabilities in that generation. However every time PCI-E moves to the next standard they are doubling. CPU's graphics cards and storage solutions take a long time to saturate the new standards. PCI-E 5 storage solutions are nowhere near maxing out the spec. PCI-SIG could delay their new specs and still be ahead since they double their bandwidth every time while pushing more energy efficiency. Meanwhile all the CPUs and graphics keep pushing power usage to make marginal gains. PCI-E specs are not the bottleneck to progress in the processing and moving of data.

REPLY

Timothy Prickett Morgan ☆ says:

JULY 10, 2023 AT 9:28 AM

I'll say it this way. Because PCI-Express can't keep up and is not a reasonably good switch that keeps pace with the bus, NVLink Switch had to be invented. And because of that, all accelerators did not have a native, cheap switch architecture from the get-go. I am seeing what would have been possible if this had been done right from the beginning. Also, I think if people could have more bandwidth and more PCI-Express lanes on CPUs today, they would gladly use it. but since this is not happening, PCI switches would be very useful to allow the pooling and sharing of these components. Static configurations are not ideal for either memory or accelerators.

REPLY

l8gravely says:

JULY 10, 2023 AT 9:12 AM

How many people have ever even *seen* a PCI-E switch and have it hooked upto more than two servers? It's a mythical beast that I'm sure exists, but not in the everyday world. The market just isn't there for a switch vendor. It's bad enough to build an ASIC able to digest 16 lanes of PCIe 5.0 right now, so building one to do 64 of them with what, full crossbar switching? And at what latency?

Unless the desktop market needs this, the server/AI market is just going to have to wait, since they don't drive the numbers. Also, how many servers actually need this as well outside of hyperscalers or research labs? I just don't see this happening any time soon.

REPLY

Timothy Prickett Morgan ☆ says:

JULY 10, 2023 AT 9:24 AM

I am thinking forward, when we have CXL memory and accelerators as a standard part of a server. The reason no one uses the switches is because there is an impedance mismatch. The latency of PCI is a hell of a lot less than Ethernet, and I think is a better way to make pods of stuff that is composable.

I could care less about what desktops need. That can't be the gating factor, and has not been in servers for quite a number of years. If it was, there would not be the kind of GPU compute we currently have in the datacenter, just to give one example. The desktop-first thinking is what has gotten us in this mess as far as I am concerned.

REPLY

Paul Berry ☆ says:

JULY 10, 2023 AT 11:13 AM

I think it is super cool to think about composable servers, using shared memory bloc' rearrangeable accelerators, and possibly numa cpus. However, I'm not super optimis

who will construct farms of servers in this fashion and use them effectively. Even if you can rearrange your memory and compute capacity, can software really make use of it on the fly? If you have to restart an application to make use of it, then isn't is simpler to move the application to the hardware, rather than the hardware to the app? Isn't that the kubernetes, cloud, buzzword future we've all been promised? Sure, maybe a few of the supercomputer applications will be able to reserve the hardware in time to use it (batch processing rides again), but I don't see who else.

REPLY

Paul Berry says:

JULY 10, 2023 AT 11:02 AM

I have definitely seen Pcie switches, though they are far from widespread.

The PCI sig needs constant improvements, because they limit the top end of what pcie can do. That said, it's not clear everyone need adopt newer versions so rapidly.
The consumer space, even high end gamers, probably don't need a new pcie spec for desktop peripherals nearly so often. Given the cost and trace length limits, it's not that useful. Usb has largely displaced pcie for most low-perf uses, but there's no reason a desktop shouldn't be configured with a single high speed interface, and a few low speed slots.
In the enterprise space, I'm not sure quite what cadence is required for NICs and slot-resident accelerators. Higher than consumer to be sure.
The one I'm not sure about is between chiplets on a package, or at least between sockets on a tightly integrated module. Is Pcie the industry standard QPI now? If yes, then it needs to support those levels of speed.

REPLY

LEAVE A REPLY

Your email address will not be published.

Comment

Name *

Email *

Website

POST COMMENT

This site uses Akismet to reduce spam. Learn how your comment data is processed.

ABOUT

The Next Platform is published by Stackhouse Publishing Inc in partnership with the UK's top technology publication, *The Register*.

It offers in-depth coverage of high-end computing at large enterprises, supercomputing centers, hyperscale data centers, and public clouds. Read more...

NEWSLETTER

Featuring highlights, analysis, and stories from the week directly from us to your inbox with nothing in between.

SUBSCRIBE NOW