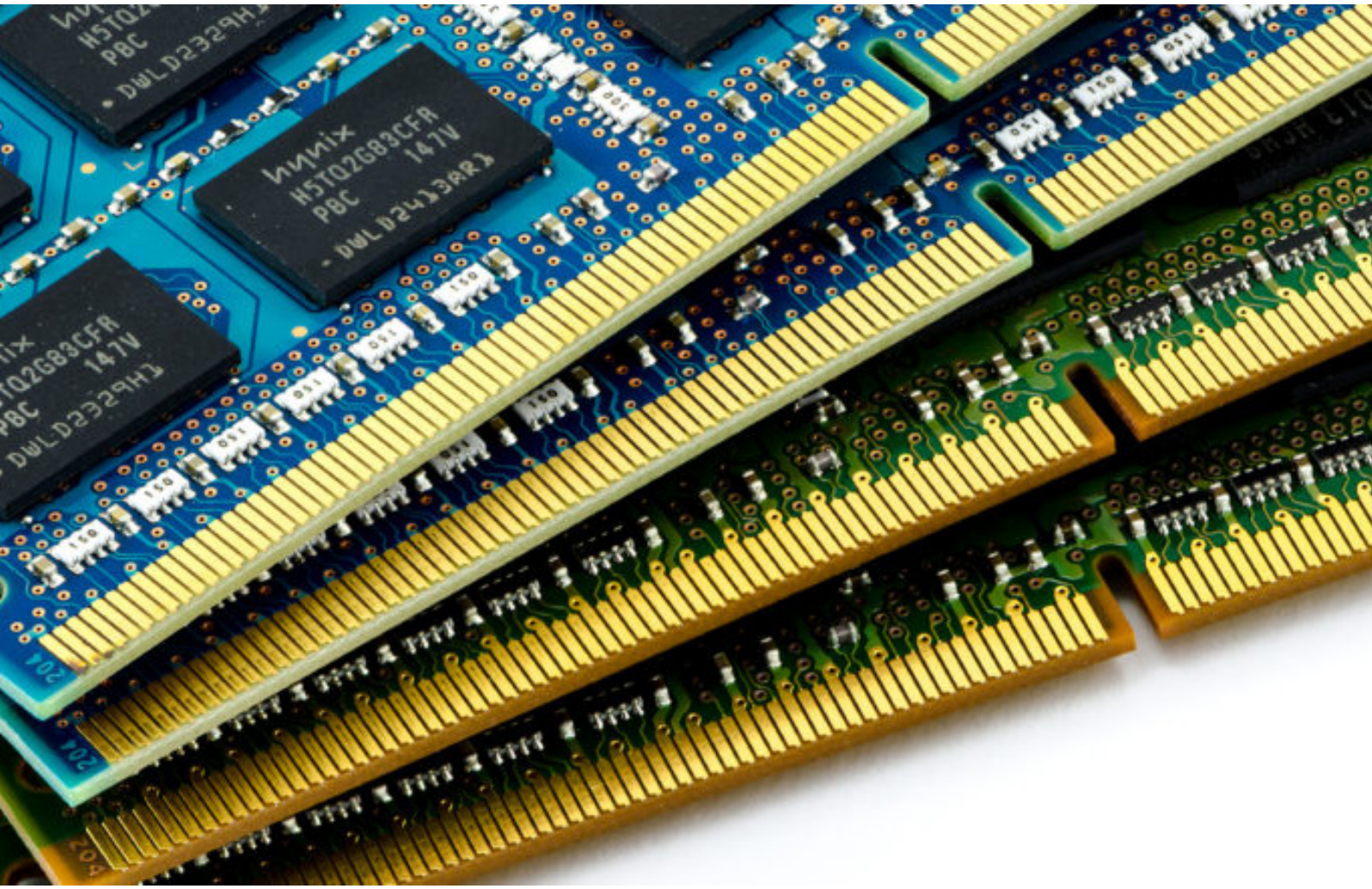
THENEXTPLATFORM



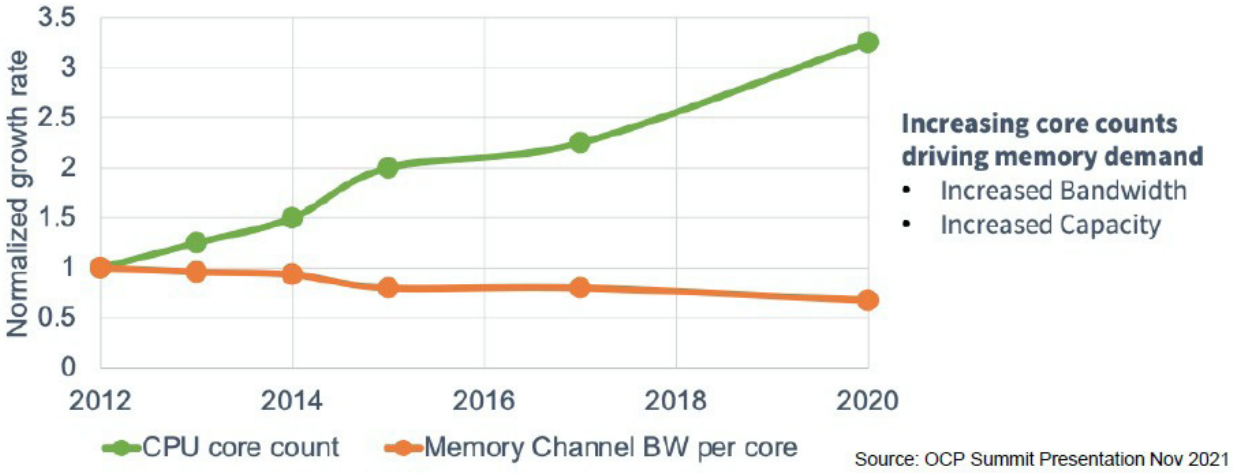
MASHING UP CXL AND GEN-Z FOR SHARED DISAGGREGATED MEMORY

If you are impatient for not just memory pooling powered by the CXL protocol, but the much more difficult task of memory sharing by servers attached to giant blocks of external memory, you are not alone. Memory fabric creator IntelliProp is right there with you.

And that is why IntelliProp, which has been a custom silicon design shop for the past two decades, has spent the past few years mashing up PCI-Express, CXL, and Gen-Z technologies to create its Omega Memory Fabric, also giving us a preview of what commoditized CXL memory sharing will look like in the future as various technologies, including Hewlett Packard Enterprise's Gen-Z protocol and IBM's OpenCAPI protocol, are merged into the CXL specification.

Hiren Patel, who is chief technology officer at IntelliProp, was one of its co-founders when it was established in 1999 as a custom silicon designer, and became its president in 2002. The dot-com bust was not a fun time for chip design in some ways, so Patel did a project lead stint at Infineon and then some contract engineering work at Broadcom and Fujitsu while assuming the CTO role at IntelliProp as it worked on various projects. Patel was given the added responsibility of being chief executive officer at the company in 2019 when he was also in the heat of things with the consortiums behind the Gen-Z, OpenCAPI, and CXL protocols, which were all trying to solve different aspects of the disaggregated memory puzzle. For the past three years, before Gen-Z was absorbed into CXL in November 2021, Patel was president of its consortium, in fact.

But Patel and the engineers working at IntelliProp saw the opportunity for disaggregated memory way back in 2017, when the Gen-Z group started talking in detail about switched memory fabrics. Gen-Z was inspired by the research and development work done by Hewlett Packard Enterprise for The Machine, a concept system that is still the inspiration for memory-centric system design and that led to the founding of the Gen Z Consortium back in October 2016. And soon thereafter, IntelliProp got to work designing chips for creating pooled and shared external memory that derived from the Gen-Z standards.



But then a funny thing happened. Everyone with a protocol to do memory coherence joined the CXL Consortium in 2019 – in fact, they did it at our Next I/O Platform event in September that year– and by early 2020 Intel declared a truce in what might have turned into Bus Wars II, and then Gen-Z IP was merged into CXL a year ago and OpenCAPI memory technology from IBM was merged into CXL in August of this year. And now CXL 3.0 and beyond gets to be what Gen-Z and OpenCAPI (and CCIX from AMD and Xilinx) were trying to do with coherent memory plus the fairly limited asymmetric CXL approach Intel was talking about initially.

And now, yet another vendor is getting ready to take on the server memory bandwidth wall and help reduce memory costs by extending server memory over the PCI-Express bus and by pooling and eventually sharing it over the wire from a memory server that is functionally equivalent to a NAS or SAN. But things will help lower overall memory costs while boosting memory capacity and bandwidth. This is true because a memory DIMM based on skinny memory chips is a lot less expensive per unit of capacity than one based on fat memory chips, so if you can double the memory channels, you can double the bandwidth while keeping the capacity the same and also lower the price for that extended capacity.

This is all familiar to John Spiers, who we talked to along with Patel and who has a long history in the storage business. Before founding IntelliProp, Patel was a senior engineer at disk drive maker Maxtor during the dot com boom. (Maxtor was merged into the Seagate Technology conglomerate a long time ago.) Prior to IntelliProp Spiers co-founded a number of technology companies including

venture backed Lefthand Networks (acquired by Hewlett Packard Enterprise) and NexGen Storage (acquired by Fusion-io). Since then he has held top leadership roles for IT infrastructure providers including Pivot3 and Liqid.

"I started out in the storage industry when NAS and SAN didn't exist and people just stuck disk drives in their servers and had under utilization and over utilization and all kinds of performance bottleneck problems." Spiers tells The Next Platform. "The same thing is happening with memory, and when you disaggregate it and share it in a pool, you need management, security, high availability, failover, multi-pathing, data encryption, and additional features and capabilities that are part of tier one storage, is what we are really bringing to the table because we understand this space."

When 50 percent of the cost of a server at hyperscalers and cloud builders is main memory, and they are 50 percent of the server market, it is no surprise to see companies trying to create memory servers or memory area networks or network attached memory – all analogs to what happened in enterprise storage – using a mix of the CXL 1.1 protocol on CPUs over PCI-Express 5.0 and either Gen-Z or OpenCAPI memory. Frankly, we can't wait for both PCI-Express 6.0 and CXL 3.0 to be ready, and that is why IntelliProp is making interface and switch chips that will bring its Omega Memory Fabric to market soon, and then intersect faster interconnects and better CXL protocol releases infused with Gen-Z and OpenCAPI technology further down the road.

"If you look at Gen-Z, that was really the fabric that allowed external expansion of memory and had all the features of security, high availability, multi-pathing that allowed you to scale memory outside the server," explains Spiers. "Whereas CXL was really focused on endpoint devices connected via CXL and is more focused on expansion in the inside of the server. We are ahead of the market because we have already implemented the Gen-Z stuff in conjunction with CXL, whereas the CXL spec itself has not."

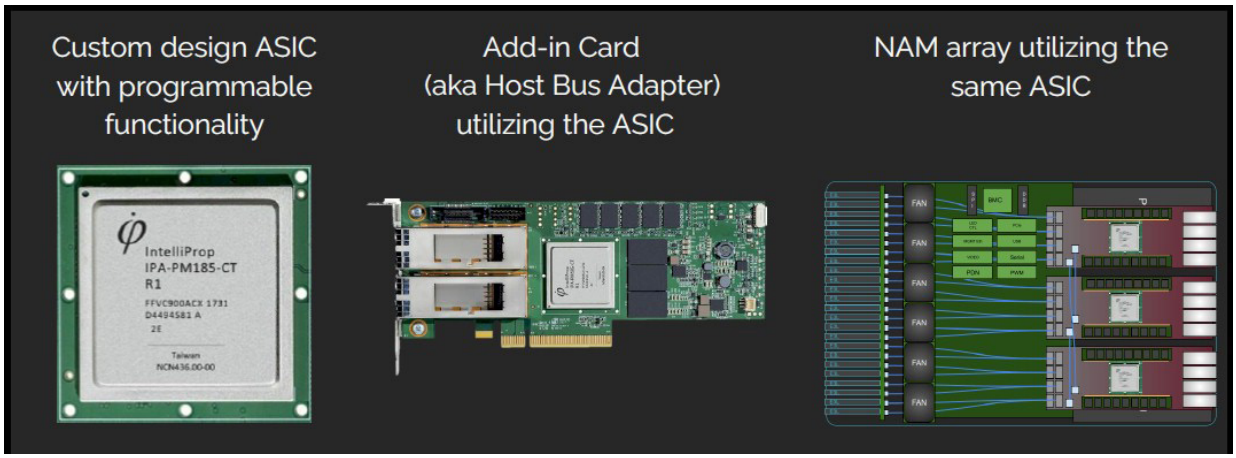
The switch to disaggregated, composable memory and not just extending directly attached memory is an important distinction. While CXL memory extension is useful for some, as Patel correctly points out, this just means system architects will be tempted to overprovision memory even more inside of a particular machine because they have effectively double the memory – half on DDR controllers and half on PCI-Express controllers.

"You can't keep adding more DDR memory ports to a big ASIC anymore," says Patel. "It just takes up too many pins. You can use CXL and lower the pin count and get memory expansion you get inside of the box. You pay some latency for it, of course. But Gen-Z was thinking outside of the box, and our ASIC is a memory expansion device but we also have a fabric port coming out of that chip and then we have media boxes and switches and switch hubs that let you start tiering main memory."

Spiers jokes that IntelliProp has created a CXL/Gen-Z host adapter that is equivalent to running CXL 6.0. . . . The IntelliProp ASIC converts CXL to Gen-Z and then uses normal 802.3 Ethernet (either copper or optical) as a transport for disaggregated memory, and then converts back to CXL to attach to CXL memory from Samsung and SK Hynix. The Gen-Z switch ASIC that IntelliProp has created has a latency of about 35 nanoseconds through the switch, and the trick is that Gen-Z never did encrypt the headers, only the data in the packet, so the packets just flow lickety split through the switch. With CXL switch fabrics running over PCI-Express switch fabrics, the whole packet, including its headers, is encrypted, according to Patel, and each switch in a hop has to unencrypt a complete packet to try to figure out where to send it.

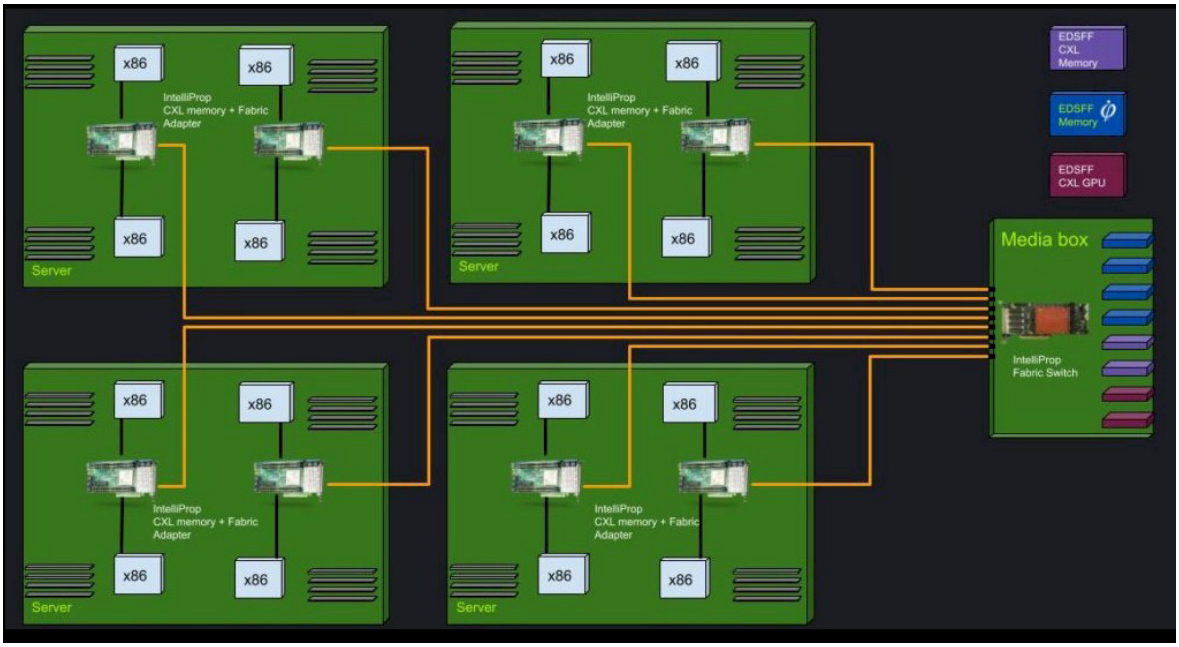
Eventually, CXL will take the best of OpenCAPI, Gen-Z, and anything else and fix such issues. Eventually.

In the meantime, you can start here:



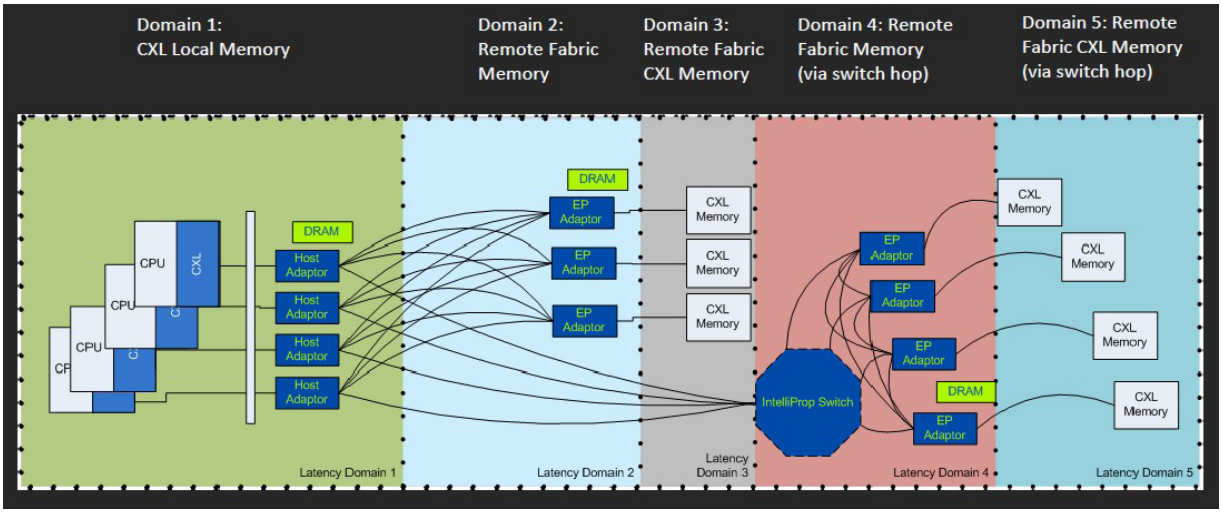
There is the CXL/Gen-Z converter and switch ASI implemented on a host adapter card that can be used as an endpoint in a server and a network attached memory (NAM) appliance, shown on the right.

Here is a drilldown into the NAM:



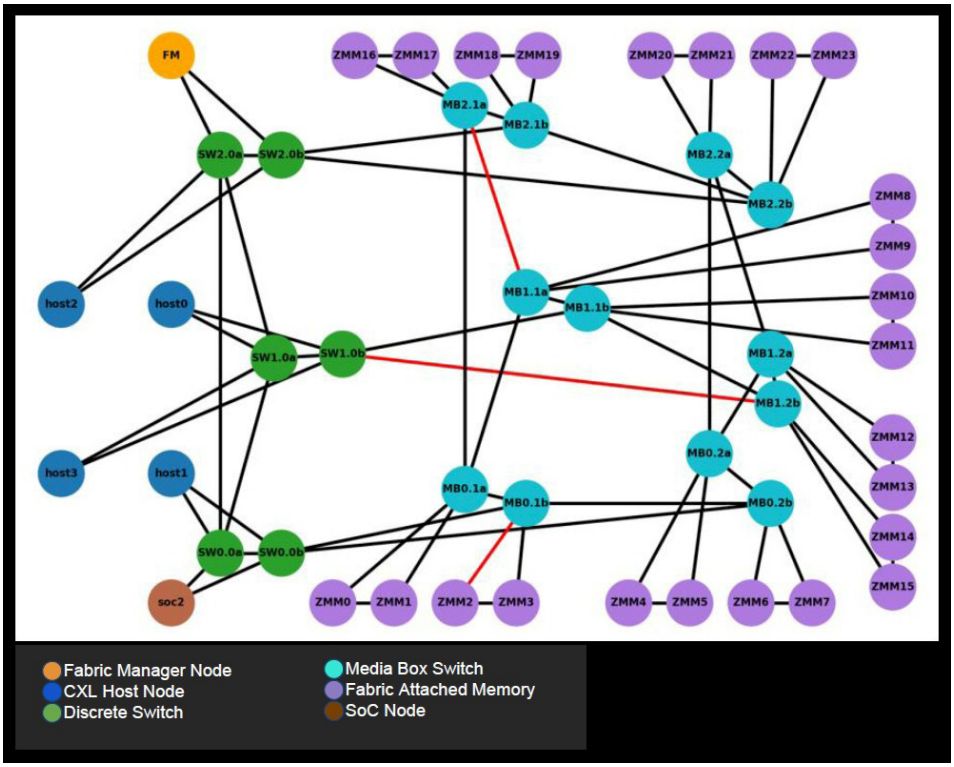
The neat thing about external memory servers is that the CXL-Gen-Z fabric that IntelliProp has created doesn't care if the physical memory on the other end of the 802.3 wire is DDR3, DDR4, DDR5, or HBM stacked DDR memory. The latencies and bandwidths will play out as they do, but customers could choose very cheap memory or the very expensive stuff, depending on their capacity and bandwidth needs, out there on the memory appliance.

Because of the Gen-Z tech sitting in the middle, the remote memory is hot pluggable, which is very cool, and memory sharing is already enabled, which means multiple servers can access the same dataset, which is the Holy Grail coming officially with CXL 3.0.

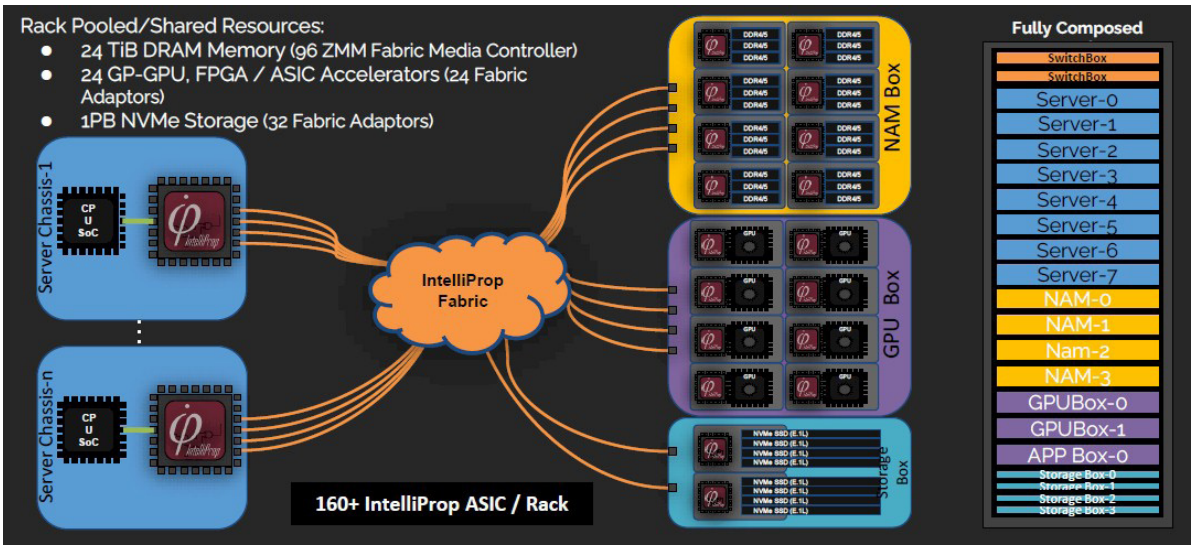


The Omega fabric has it "unofficially" ahead of CXL 3.0 with CXL 1.1 endpoints thanks to Gen-Z in the middle, and that means, for instance, you can have a memory appliance with a single dataset on it for an AI inference that fans out to many servers, and with lower latency (up to 10X lower, says Patel) than with InfiniBand or Ethernet with RoCE. Patel says that running AI deep learning relationship models, the Omega fabric can scale across dozens to hundreds of host nodes, each with multiple GPU memories that are linked by that fabric.

You can see above why, with this memory topology and several tiers, you need a fabric manager:



Here is, conceptually, what a full rack of composable memory across CPUs and GPUs plus NVM-Express flash storage might look like:



This rack of composable gear has 160 IntelliProp chips in it, and that is what Patel and Spiers are excited about.

The IntelliProp ASIC was tested since this time last year in an FPGA version, and after some fund raising, the real chip is expected to start sampling in 2023 and go into production in early 2024. That seems like an eternity in the IT business, but it isn't.

The IntelliProp ASIC is a multipurpose device with four ports, each running four lanes running at 25 Gb/sec, for a total of 400 Gb/sec coming out of each ASIC. If you cross connect them in a non-blocking manner, you can make a 12-port or 16-port Gen-Z switch, and while this might not be enough to be a full top-of-rack, in the first iteration of the Omega fabric, it was important to just design one device. In the future, it stands to reason that there will be a beefier Omega fabric switch as well as a multiport Omega adapter as distinct ASICs.

It also stands to reason that this switch ASIC will be comprised of building blocks from the adapter, all within a socket and eliminating some of the hops across the chips. The chips are implemented in 12 nanometer processes, which

is fine for 25 Gb/sec signaling, says Patel, but moving to 56 Gb/sec signaling will require stepping down the SerDes to 7 nanometer processes. The FPGA implementation ran at 400 MHz, and the goal is to run the real ASIC at around 1.6 GHz, and a future chiplet version of the ASIC might have 64 ports and run at 2.4 GHz or even 3.2 GHz, having much higher radix and much lower latency.

We look forward to seeing this in production and how it changes the economics of memory in the datacenter. And we expect a lot of different ways to skin this datacenter memory cat.