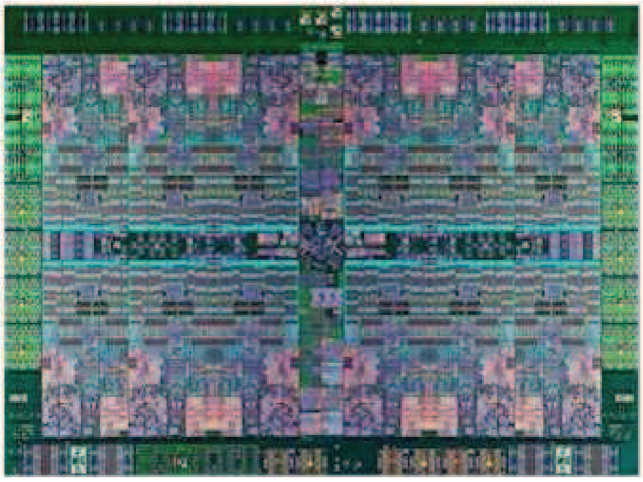


Like any other seller of processors and systems in the datacenter, IBM is in an uphill battle with Intel, whose Xeon processors have become the de facto standard in the glass house for everything except legacy workloads that are too costly and difficult to move. By selling off its System x server business to Lenovo, Big Blue has put itself in direct competition, wholly, with the X86 systems business for the first time in its history, and the company is hoping to leverage its Power8 processor and a slew of accelerators to make its Power Systems machines competitive with X86 systems on very specific workloads.

IBM is certainly not new to hybrid systems. The “Roadrunner” massively parallel supercomputer at Los Alamos National Laboratory, the world’s first petaflops-class machine, paired IBM’s PowerXCell multicore accelerator chips with AMD’s Opteron processors, with the Cell chips providing a lot of the floating point performance. IBM has learned a lot since Roadrunner was fired up in 2008, and one of the lessons that Big Blue learned is that the programming model for hybrid computing has to be simpler.

The Coherent Accelerator Processor Interface (CAPI), a new feature of the Power8 chips, aims to make the programming model easier by tightly coupling the processors and the accelerators – or even flash memory or other kinds of storage – to the main memory used by the processors themselves. This CAPI interconnect uses the on-chip PCI-Express 3.0 controllers of the Power8 chips and the bus they implement with peripheral devices in the system to link outboard accelerators to the processor complex.

At the moment, if you have an accelerator of some kind that is working in conjunction with a CPU, whether it is a an earlier Power chip or an X86 chip, data has to be moved to from the CPU’s memory to the accelerator, where it is manipulated in some fashion and the results passed back to the CPU for further processing or for use by the application as is. This back and forth of movement of data takes time and makes for some complex programming, too. What IBM does with CAPI is to provide a memory overlay that allows the accelerators to share the same memory space as the processors, and in this way data is not moved back and forth between the devices at all.

“Essentially, we are treating CAPI as a hollow core,” Fadi Gebara, senior manager at IBM’s Austin Research Lab, explains to *EnterpriseTech*. “There is a CAPI unit on the actual processor, and it has all of the plumbing and guts to speak to the rest of the cores as a participating member. It has no ability to actually compute anything, but it does have the capability of keeping the communication flowing nicely, taking care of queues and data handling. It funnels all of the commands to do compute out to the PCI electricals and there you can implement an accelerator of your choosing. The cool thing about it is that because it is a fully participating member on the Power bus, you can actually implement anything you want. It can be a computational accelerator, a random number generator, or a memory controller if you wanted a specialized DIMM, or it can act as a kind of a flash controller. The important thing to note is that there is no equivalent port or function in the industry.”

Well, not yet, anyway. But if it is a good idea, you can rest assured it will be emulated in some fashion or even licensed. If IBM can create a hollow Power8 core, other chip makers can create their own hollow cores and use the PCI-Express bus as a transport to more tightly couple accelerators to CPUs.

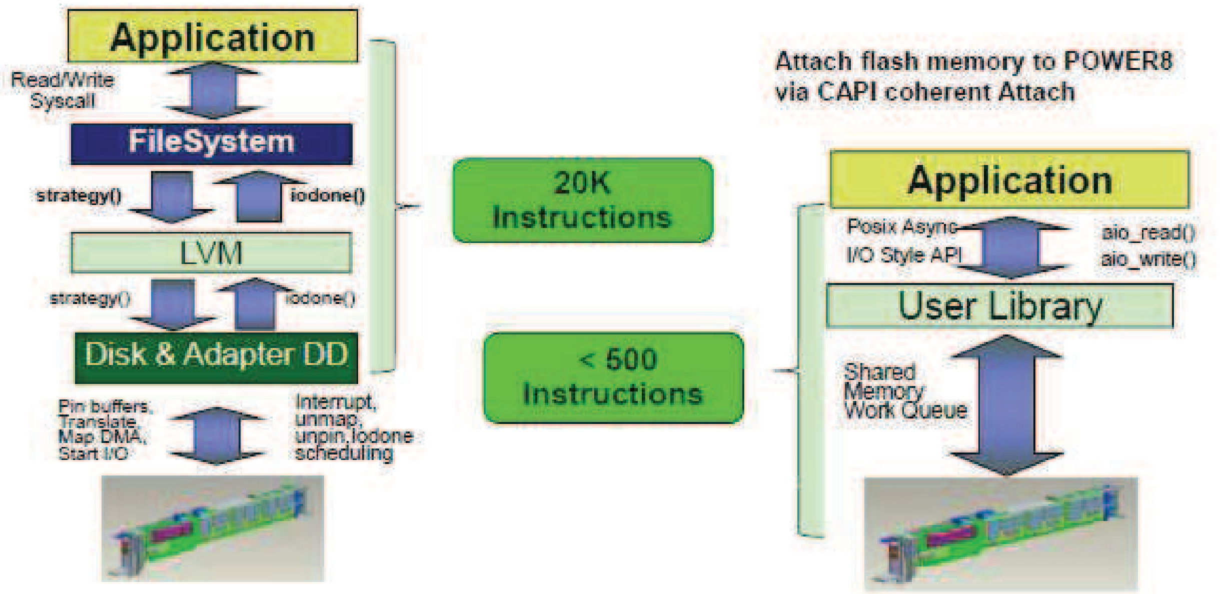
As *EnterpriseTech has previously reported*, IBM has begun shipping Power8 systems that make use of the CAPI interface in specific ways, but Gebara points out a number of other potential use cases.

In the Data Engine for Analytics setup, IBM is using a card from Nallatech that has an Altera FPGA that has the CAPI port actually programmed into the FPGA logic so it can speak to the “hollow cores” on the Power8 chip. In this case, the Power8 processor is offloading the processing of the Gzip algorithm to compress and decompress data stored in a Hadoop cluster running on the nodes. The Gzip algorithms runs about ten times faster on the FPGA card than it does on a Power8 core.

When used in conjunction with IBM’s General Parallel File System (GPFS), which has been rebranded Elastic Storage in recent months, this Gzip compression allows customers to build Hadoop clusters with about a third less raw data capacity without sacrificing data protection. IBM is still doing benchmark tests to show how its complete stack – including the Platform Symphony Java messaging software and its BigInsights implementation of Hadoop – all come together on this setup to significantly boost Hadoop performance.

The Data Engine for NoSQL bundle is a special CAPI-enhanced system that runs the Redis NoSQL in-memory database, and it specifically allows for data to be dumped from main memory to flash memory. The Nallatech FPGA card in this case has the CAPI port programmed on it and also an interface to IBM’s FlashSystem 840 all-flash arrays, allowing for data to be passed back and forth between the CPU main memory and the flash in a very quick manner.

You don’t have to run Redis to get the flash acceleration feature, this is just IBM’s first implementation that makes use of it. The following chart shows why this more direct approach to the attachment of flash memory is important:



- Issues Read/Write Commands from applications to eliminate 97% of instruction path length CAPI Flash controller Operates in User Space
- Saves 10 Cores per 1M IOPs

In a normal system with drivers running to allow the processor to communicate through a PCI-Express controller and then out to the flash, it would take around 20,000 instructions. With the CAPI flash controller operating in the user space of the operating system, you can cut out about 97 percent of the overhead and do it in less than 500 instructions. Because of this massive reduction in latency, you can, as Gebara puts it, “treat flash like slow memory instead of fast disk storage” from the point of view of the applications. This is essentially what IBM is doing with the Redis NoSQL bundle above.

Another possible use of CAPI is to communicate with offload engines that create random numbers, which are a key component of all kinds of simulations but are especially troublesome for Monte Carlo simulations at the heart of financial modeling systems.

“These simulations need to generate a ton of random numbers to fill out the form in which they want to run simulations against. If you do this in a standard system or in software, you will quickly deplete all of the random numbers at a given time. The system can only produce a random number every so often, and if you are running these massive simulations that need gobs and gobs of random numbers, you run out.”

By offloading the random number generation to a CAPI-connected FPGA, you can generate a lot more random numbers and therefore accelerate the Monte Carlo simulations.

The same ideas hold true for encryption and decryption and CRC error checking. Another possible use for CAPI acceleration that IBM is examining is for the acceleration of regular expressions, which are the common ways that log files and other semi-structure data are analyzed and picked apart. Any data sorting, parsing, searching, or filtering routines that might get bogged down in the processor can be programmed into the FPGA. And importantly, once DSPs, GPUs, and other kinds of accelerators are equipped with CAPI ports, they can be used in a similar manner. The tough trick now is for IBM to work with accelerator suppliers to get CAPI ports on their cards. The FPGA was the easy one, since it can be programmed.

Share this:

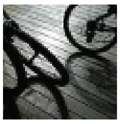
33 16 6

Categories: Slider: Front Page, Slider: Systems, Systems
Tags: accelerator,CAPI,DSP,FPGA,GPU,IBM,Power8

About the author: Timothy Prickett Morgan

Editor in Chief, EnterpriseTech Prickett Morgan brings 25 years of experience as a publisher, IT industry analyst, editor, and journalist for some of the world’s most widely-read high-tech and business publications including The Register, BusinessWeek, Midrange Computing, IT Jungle, Unigram, The Four Hundred, ComputerWire, Computer Business Review, Computer System News and IBM Systems User.

5 Responses to *Inside IBM’s “Hollow Core”*

1. 

   Tom Boulet October 28, 2014

   System was sold to Lenovo.