the Exascale Report™

A Global Perspective

The voice of the emerging exascale community

http://TheExascaleReport.com
May 2013

THE WORLD NEEDS EXAFLOPS

Getting a Handle on the Next Big Thing

By: Mike Bernhardt
April 2013

A recent Exascale Report article on fabric integration *(Revolutionary Thinking; Evolutionary Technology, March 2013),* received many positive compliments from our readers. It also raised some interesting questions about the potential impact of Intel's plans for fabric integration on some of its competitors and other key players supplying products or technology for the discreet fabric infrastructure.

A revolutionary step in fabric integration, in theory, will enable new architectural approaches, and at the same time could turn out to be a disrupting force for several key players in the HPC community.

According to Intel's Raj Hazra, Vice-President Intel Architecture Group & GM Technical Computing, that's just a sign of how critical fabric is to the extreme scale discussion.

Over the past two years, Hazra has become solidly entrenched as one of the driving forces in the emerging exascale community, and has done a respectable job of rebuilding Intel's credibility among HPC stakeholders.

The Exascale Report was the first publication to challenge the Intel exascale announcement at ISC11 and report on the flood of negative reaction to it from the HPC community. [Intel Equipped To Lead Industry To Era of Exascale Computing]

Since that controversial announcement, Intel has maintained a steady march of progress in HPC, including technology and human resource acquisitions, several internal reorganizations, and a noticed effort to reconnect with the community. Today, under the leadership of Hazra, backed by the strategy, commitment and overwhelming resources of Intel Senior Vice President Diane Bryant, another powerful executive who is passionate about HPC, a new and quite credible exascale story is evolving.

The Exascale Report is pleased to bring you this exclusive interview with Intel's Raj Hazra.

Note: The editorial staff has taken certain liberty with translation of the audio interview file and our use of grammar and punctuation to better capture the intent and impact of this discussion.

The Exascale Report: *Fabric integration is an exciting topic that holds great promise, but we still have to face the same power, scaling, resiliency and financial challenges we've been discussing for the past several years. When we start to integrate fabric, where do we see the benefits?*

HAZRA: When we integrate the fabric structure for future HPC systems, there are a number of benefits we can discuss from an end-user perspective. Fundamentally it's power, cost and capability. First, by bringing fabric integration to work-class silicon, we give it the advantage of Moore's Law – an advantage it might not have elsewhere. The densest transistors in the world, the best performing logic transistors, are now also the transistors driving more complicated logic that drives the NIC and the interconnect. This has a direct impact on both the power requirements and the total cost.

The additional benefit is that by architecturally bringing the fabric closer on to the processor, we can now be much smarter and actually much more aggressive in terms of how to do certain things that the network has to do. In the past, the network had to be viewed as a device sitting out there, somewhere – it had to be commanded. By bringing it on to the processor and making it part of the processor's architecture, you can share system resources and do things in a much more architecturally savvy way. And it's not just doing the same things you did before and doing them faster. You can actually do new things such as making the fabric a part of your memory hierarchy decisions.

Intel's Raj Hazra

As the systems get larger at scale, the lines between the memory hierarchy as we know it today, such as caches leading up to DRAM, and the interconnect, which people have thought of as a different thing, those lines blur, and the reason is applications will want to see memory on the node or memory remotely as kind of one space.

So increasingly, off-node memory will have to be treated like memory and a level of intelligence has to be put into it such as: how do you pre-fetch it, or where do you keep it so that you have the least cost in accessing it. And, in doing all of those, interconnect becomes an integral part of that intelligence. That intelligence and that decision, initiates from the processor, or the processing, and therefore, the integration – bringing them closer together both physically and architecturally and allowing us to do things that we haven't done before.

TER: *So Raj, we had an article recently that presented the view that the U.S. is unsure about Exascale. Do you see any increasing support from U.S. funding groups or other sources at a level needed to move us down the path to exascale?*

HAZRA: It's a great question. I've spent a lot of my time thinking or acting on this area over the past year or so. Let's talk about what has changed and what hasn't since we started this debate three or four years ago. We all tend to focus on what's changed – the severe budget climate and the political administration have clearly changed. And to some degree the roles of some of the labs within the U.S. Government have changed. Certainly the geo-political climate in terms of competition has changed.

All of those things are true. But the few things that haven't changed are the following:

The world will need exaflops.

Whether it's to address problems emerging in science or national competitiveness, whether it's national security, or it's just this explosion of big data class applications that have a strong economic benefit, beyond just grand discoveries in science, the world will need exaflops.

The second thing that hasn't changed is that we will need exaflops at a cost-manageable and useable envelope. That is, it's not a matter of, "ok, the world will need exaflops so let's just keep building bigger and bigger scale out machines and we'll give them exaflops." This inflection in our technology and innovation to get to exaflops has not changed. We still can't afford a 500MW power plant sitting next to a data center in order to reach exaflops.

The third thing that hasn't changed is the momentum of multiple nations in seeing HPC as an integral fiber of the mantra: “to compete you have to compute.” This is now a national agenda in many nations. Each day that number keeps growing. So, facing that, the question around the world no longer is do we need exaflops, the question is what’s the best way to get there?

Looking at the nations around the world that have announced their intent to go after high-end computing -- whether they say it’s exascale or whether they call it multi-hundred petascale, whether it’s India or China, Russia, Brazil or Japan – they are all talking about the need to invest and their discussions are all rooted around an economic necessity.

They’re not saying let’s have an exascale just to have the top rated machine in the world. It’s now national competitiveness and driven by industries. Europe is one of the more recent entries into this foray after what some would say was a decade or two of having watched the game, and not participating as strongly as some of the other geographies.

But they’re all coming in because it is absolutely essential to get it done.

I do think in the U.S. in particular, even though we don’t have an exascale program yet, and in spite of multiple efforts to get bills on the floor of Congress, the heartening thing is that no one is really vociferously asking why we need exascale now. They are asking **how** we would build one. How would it serve some of the broader uses beyond science and national security and make it economically feasible to fund?

Some of this is the budget climate we live in today. It would have been easy to say, “Let’s just forget this.” But no one is forgetting it in Washington, D.C.

Everyone is struggling with how to do it. And this is where I think there is a tremendous opportunity for public / private partnership.

We (Intel) have said this from day one. This is not about the government agencies funding everything on their own. And this is not about industry taking all of the risk on their own either. Both from a financial perspective, but just as important, from a technical perspective, to form that partnership for building the ‘right’ exascale is going to be even more critical.

We are moving very slowly due to the sequester climate. To go down to Washington, D.C. and talk about these things when fundamental agencies are cutting back on basic services, well that’s a hard conversation to have.

But I think there is an across-the-board realization that this is not something we can ignore – or if we do, we ignore it only at our own peril. And that realization is a heartening thing.

TER: *Have any economists taken a look at how exascale or exaflops-level computation could change things?*

HAZRA: Interestingly, we have multiple views of this. There are some studies, especially those run by the DOE, around what the impact might be if we could solve some of the grand challenges in science.

For example, if you could predict hurricanes down to 30 minutes, how much less budget would FEMA need? How much life and property could you save?

Using Katrina as an example, if you look at the amount of damage that was created, how might it have been different if we had a much earlier warning? How much was that worth then in terms of dollars?

People have done specific studies of natural disasters or even financial meltdowns, and if you could predict these and avoid some of them, how much would you save? And these are staggering numbers.

As another example, think about carbon sequestration. As the world is moving toward more of an awareness and maybe an economic system around managing carbon footprints, and if carbon becomes an equity to be traded, you are going to need a very good verification system or you are going to be at a disadvantage.

So these are all things people have studied and said there is economic value - and in all of them HPC or exascale computing plays an important part.

But let's expand on the discussion of economic impact. When you talk about an exaflops at 20 MW, people immediately start to see big data centers – cyber security missions. One thing that excites us here at Intel is that you could do this – you could build a petaflop in a rack, or in half a rack. Twenty kilowatts gives you a petaflop, so maybe you build a half petaflop in a 10 kilowatt rack, which would fit under a desk.

We look at the tremendous TAM (Total Available Market) expansion possibilities. When we look today at barriers, at the number of small and medium industries in say manufacturing or at supply chains around oil and natural gas or climate modeling, they could actually do more by using modeling and simulation.

Then we look at the barriers to why they can't, and it's simply because supercomputing and HPC still isn't personal. It's not personal in terms of cost. It's not personal like 30 years ago the way the PC was. People would resist and say they've never learned to use a computer. But if I come home every day and it's on my desk, I'll learn how to use it.

So when things are personal, you learn to use it. You learn to get value out of it. And HPC has never been there because, well, who's going to pay a million bucks a megawatt?

Or, we can try to go to the cloud and move tremendous amounts of data.

So the exciting part of exaFLOPS is that while it allows you to build those great big machines, it also lets you democratize HPC by essentially building half a petaflops, which today would be in the top 100 supercomputers, under everybody's desk. And it would be done at both a capital and a TCO cost that changes the game.

The economic discussion is fundamental. That's 300,000 potential users just in manufacturing in the U.S. that we've looked at. If they were to use any level of HPC or extreme scale computing, how much value would they generate and what would it do to the HPC market?

TER: *What role do you see Cloud or Big Data playing in terms of helping us realize exaflops?*

HAZRA: I actually am less skeptical of the connection of Big Data, HPC and Cloud. And to understand why, you have to look outside the traditional HPC area for a change.

If you look at traditional HPC, you get this view of a massive data center. Think of standard usage patterns at facilities like Oak Ridge or Livermore. You come in, you have your offices nearby, you run your code – or you remote log in to a grid – it's still a very tight association, with the capability being the data center.

Now stop looking at that for a second and take a look at what's happening in public cloud.
For public digital services, not so much private company data, or functioning of IT systems, well, that's on fire. That is the way. Despite the data movement security challenges, the TCO benefits of a cloud-based delivery mechanism for those kinds of services is unparalleled. Public cloud is

growing faster than anything else in the data center in terms of volume.

And as you see more things getting connected into it – for example, smart cities that use cloud as a back end, data is all going there as well.

So in many ways, the things you need to put in the recipe for analytics is already coming together, including a delivery mechanism for services and a business model for those delivery mechanisms. The accumulation of multiple sources of data as multiple services makes use of that. What's missing is potentially an HPC capability that does the analytics and provides the basis for new services.

And to me, that's what's going to drive the forced major injection of HPC into the cloud. It is not so much HPC as a service or an infrastructure, and I know Amazon and others are there and people do use it, but the next quantum jump is not going to be, "Hey let's do HPC in the Cloud!" It's that everything we do in the cloud will have a high performance analytic back-end as more and more people, with all of the data coming in for the services that exist today, want to get a higher level of value.

No one thinks of Facebook as an HPC user, because today, we store our pictures, we say hello to our families, we can track what's going on in our kids' lives, and so forth. But you know the next set of things that are going to happen on those kinds of services is correlations. New business models will run targeted ads. Or as an example of a benefit to end users, the ability to determine the location of all your friends. It's the social graphing thing and that requires high performance capability because they are not solving what used to look like graph problems.

Or let's take a traditional HPC problem like star search. To me, Big Data is simply about the fact that more of the world is digitized, and people want to gain value rather than leaving that data in an un-monetized heap. And that's simply going to increase as more of the world gets digitized.

The second is – the cloud is here to stay – because it is one of the best ways to deliver 'at scale' services.

The third is, HPC will have to become an integral part of that infrastructure because what you want to do with those services that you deliver to more and more people, with increasing value from the data that is there – all of this requires high performance computing.

TER: *How do we deal with the overall data transfer and the storage component of that?*

HAZRA: That's an interesting question. One of the big challenges of Big Data is for the first time, whether it's social data or machine-to-machine data, it is quite clear that the old model of, '*there is compute – move data to it'* is not always going to work. You need to move compute closer to where the data is because sometimes it's cheaper to move compute than to move data.

And there are multiple models of it – and that's already happening in single systems of clusters today – where rather than moving data expensively from the file system all the way across the network to my node, I will just move the MPI ring so I do computation closer. Storage systems are getting built and file systems are getting built with more meta data so you can actually do more effective processing at the edge instead of having to bring all that data in. Our Lustre work is headed down that way.

So, even in single systems, those concepts, and they're simply just being applied and distributed, infrastructure as well – move the computation to where the data is. Sometimes that is driven by data policy. For example, if you look at an oil and natural gas company that is prospecting in a nation that does not allow the data to be taken out, it doesn't matter that their data center is sitting in Houston. So they have a grid mechanism by which they can do an infrastructure as a service, sometimes down to the local geography.

So the notion of where you have your data and where you compute it is going to become increasingly more important. And that's going to lead to architectural innovations at multiple scale.

Just draw the contrast. Microarchitects at Intel worry about when you have, say for example, 60-plus cores and you have multiple memory controllers, how do you make sure there is fairness and access to each of these memory controllers so data is moving efficiently? While they are doing that, there are solutions architects wondering if (for example) I have 25 data centers across the world at a company like AirBus that has multiple development centers , multiple data repositories and test centers. How do I have a workflow that makes this data look unified, meaning it's there when I need it. Across that scale – it's the same problem. The challenge is *Data Choreography* so compute and data live as close together as possible.

TER: *So what storage schemes will do best in this new world, for example, SAN, NAS, etc.?*

HAZRA: Before we get hung up on labels such as NAS or DAS or SAN, we need to look at basic assumptions about how storage is functioning.

Storage is very much a hierarchy just like memory is today. Maybe there is more sophistication in memory hierarchy because of all the work that's been done in caches and so forth, but storage is and continues to be a hierarchy.

One of the things that changes or evolves any hierarchy is when you have fundamental technological innovations that change either the nature of each level of the hierarchy or they change fundamentally the capacity per cost equation. And that's really what's happening in storage.

I'll give you a simple example. If you can have sufficient non-volatile storage on a node itself in an HPC machine, so that you don't have to go over to the file system, and then non-volatile storage can be accessed even if the node is down, then your entire resiliency architecture changes from how you did it before with network attached storage.

So I'm not saying network attached storage goes away. But this is a particular example of where you can do things differently than moving it all over the network and dumping it into a file system. You now have the technological capability to store a node's worth of 'state' in a checkpoint window and be able to get it out, even if the node went down.

We are at the doorstep of those kinds of breakthroughs, and when those happen, yes – our view of how the storage hierarchy works will change.

Some of those hierarchies are also changing because of usage patterns. This whole notion of 20 years ago – there was tape and there was disk and there was memory – that's all changing. Today, just the amount of data being generated and the frequency with which different kinds of data is being generated is creating more levels of hierarchy. Very cold storage. Facebook can't afford to lose any picture I upload, but it is quite clear that I don't look at all pictures equally – or my friends don't look at all my pictures with equal frequency. Now if I only had 10 pictures on Facebook, it probably wouldn't be a big deal for Facebook to keep all 10 pictures in the highest level of the hierarchy availability. But collectively when "us" – the users – have 300 million pictures of data that we are uploading, the creation of these very efficient storage hierarchies becomes very, very important in an economic sense.

TER: *How about "the next big thing?"*

HAZRA: I think it is a collection of things that will make the next big thing. The traditional challenges of power efficiency, resiliency, application scalability – the things we all talk about when we talk about extreme scale computing -- they don't go away.

But I think what we are starting to see is how those are becoming interconnected at the system level and solutions are emerging that are really at system scale.

Take for example our interconnect strategy. Is it a power strategy? Is it a resiliency strategy? Or is it a strategy to let applications scale more effectively and efficiently? The answer is all three -- the memory hierarchies and the storage hierarchies, we are talking about are all three.

So I think the big breakthroughs are going to come when individual areas of investigation are going to connect at the system level to create system architectures that have the best trade off of these sometimes conflicting goals.

The next big thing I watch out for is not a new circuit or a new kind of core or processor or memory technology. It's a combination of all of those and a system architecture that lets me express what I can do as an application developer on a system with a collection of those capabilities. Which is to say I actually think of it as a system-level co-design problem.

The next big thing?

System level co-design.

People have talked about co-design, which is a valid concept, but they usually do it as co-design of an application for an open MP model on a processor. I think this is co-design of system software and hardware. But more importantly co-design even amongst the hardware components.

Typically in co-design – you look at the level above it. This is where you need to look sideways. How do we get the memory person and the storage person and the compute person to work together in order to create a balanced system?

In fact, upwards co-design has already been established. Co-design sideways is what's going to win the day for us.

And just one more closing thought. I think we are understating the economic benefit of everything we've been talking about. The benefits of exploration as we go down this path will have far-reaching effects that will impact all of us. How do we possibly calculate the far-reaching economic benefits of saving lives, saving property, and changing the world in which we live?