Primeur Weekly! Magazine

2014 Another year on the road to Exascale - An Interview with Satoshi Matsuoka and Thomas Sterling - Part I

23 Feb 2015 Leipzig - *For the fifth year we interviewed Satoshi Matsuoka and Thomas Sterling on how far we are on the Road to Exascale. Five years ago we thought we would be half way by now. Even last year there was a lot of optimism. Today, more and more it seems likely the date when we will reach exascale will slip by a number of years. Exascale computing is not only about processing, it needs big data, really big data. To get all the data in and out of the system requires new ways of memory are being developed. What will be the winner is not yet clear, but interesting trends are emerging. The dominant computer architecture paradigm with a central processing unit where data moves in and out, is not very suitable for exascale or knowledge processing. Hence new architectures with new associated run time systems are needed. There is also excitement about low energy HPC using for instance ARM processors. Concentrating on low energy alone is missing the important point of ARM: that it is an open extensible architecture. The ultimate use of exascale systems is for knowledge processing, what could lead to a real artificial intelligence. Would such an intelligence pass the Turing test? Of course not, it would laugh at something as human as a Turing test. These are the main topics discussed by Satoshi Matsuoka and Thomas Sterling in this interview, conducted at the ISC14 conference.*

This interview is published in four different episodes:

Part 1. Introduction - Exascale hardware developments

Part 2. Exascale software developments

Part 3. Knowledge processing and intelligent machines

Part 4. International Exascale Collaboration

Exascale hardware developments

*Primeur Magazine:*Welcome to another year on the road to exascale. Actually, when I listen very carefully to Thomas' presentation earlier today, we can be done in one minute, as simply said "nothing has changed".

Thomas Sterling: I think my point was: there is the*appearance*that nothing has changed. That is: only at the zero derivative. I think we actually are at a point of inflection. At higher levels there is a significant shift in change towards the future. That was the point I was trying to make.

*Primeur Magazine:*So at a higher level, there is more thought about what exascale systems should look like and how we can make them. What is the general view then today?

Satoshi Matsuoka: In the advances in the overall endeavour towards exascale, we now have some real machines. We also have technology trajectories that are more clear. Not just concerning the timeline, but also the expected performance of the devices and their availability, their performance/power characteristics, etc. We now are starting to get a very clear picture of what the machine might look like, and it might not be just one view but rather several different views. But if you look at, for example, some of the recent talks being given by Peter Kogge, based on the Exascale Report, what we find is that exascale is still hard; and in fact much harder than people expected.

In some ways it may even be more difficult than Peter predicted. This is becoming much more clear, and as such people are lowering the expectations of the extrapolations of the existing technologies to exascale. So what was predicted was to reach exascale by something like 2018, but that was pushed back to 2020 and now the whole thing moved to 2022 or later. Also at the same time, because of this push back - and we do not even know if the push back will even allow us to do exascale in 2022 - people are starting to add more innovations in various levels of technologies. One representative type of endeavour is the often proposed use of novel types of memory at different hierarchies of the system. People are realizing that DRAM power would be significant. People worry about compute scaling, but DRAM is even worse because of technology reasons.

So even if people design systems that can have an exaflop/s of computing power, with DRAM only memory per performance could be minuscule to the extent that it is not usable by many of the applications. The emerging technology option now is to harness the use of novel memory with much lower power characteristics, allowing us to have capacity and performance at the same time, and perhaps lower cost. But this is a very different technology from conventional machines, and we have to master its usage. As such, we are at an inflection point where we are starting to see these innovations, which is good, because this is HPC heading these essential aspects of the IT evolution, not just for exascale; so it may have propagational effects to the whole IT industry.

Thomas Sterling: I fully agree with that. I just might extend it. I just want to comment, because Satoshi is too modest. The defining work that has been done in the last couple of years, is the work that he and his colleagues have been doing on the TSUBAME, through strong empiricism done both on system and on driver applications. He has shown what the possible envelope in the energy versus performance capabilities can be. This is exposing in a way that is undeniable how difficult useful, and I underline useful, exaflop/s is. I am still enough pessimistic to believe that some nation will decide for a 1000 Petaflop/s Rmax machine, just for the sake of taking that credit. But I do not think the community will allow them to get the stature from such a futile effort. It is a stunt machine. It will not be the first one. Unfortunately, more than one nation, including my own, at one time or another has committed to that, in this regard. The community has a very tough time in admitting to it. This is why I am supporting and emphasizing Satoshi's comments, in that, even for the last 60 years, we maintained a priority of focus that was first use, then memory, then communications.

We have known for some time that both in the area of time and in the area of energy the priorities are exactly the opposite. First and foremost is data movement, secondly it is memory capacity and memory bandwidth, and only thirdly, a distant third, is processing. In fact the numbers that people often quote on processing are often misleading. They say: look, 48% of the budget is dedicated to processors. But most of that power in the processing chips is dedicated to the memory hierarchy, not to the actual computation itself. With three layers of caches, really a large part belongs to the memory hierarchy. Then you get much more realistic numbers. I think the point of inflection which is slowly being recognized privately by the industry vendors, publicly by the users and independent system designers, such as the Tokyo Institute of Technology and others, is allowing the community now to consider alternative structures. For these - and now I wave my own flag - it are innovative execution models that allow them to be used effectively and help define the programming interfaces that are likely to emerge in order to address the challenges that Satoshi has identified.

*Primeur Magazine:*So what will the memory structure look like? Can we already say something about that?

Satoshi Matsuoka: At the device level there are already competing technologies. There is, of course, DRAM and do not count out SRAM. Some people claim that, as semiconductor feature sizes get smaller, actually SRAM could be more advantageous, because it is a switching device and not a capacitive device, and at some point the former may use less area and energy than the latter.

Then there are new non-volatile memory devices. There is Flash memory of course, and there are a lot of innovations being done like 3-D structures, predominantly for consumers but also enterprise is starting to adopt it. Then there are more advanced devices, like STT-RAM and PC-RAM, as well as Resistive RAM. It is not clear which of these devices will become the overall winner, because it is not just about the characteristics, but also the manufacturability costs and availability. All these kinds of factors need to come into play to be a replacement for DRAM and flash memory.

But supposing that some device is far more durable and also a lot faster in terms of the access speed compared to flash and somewhat comparable to DRAM in some respects but still has the density and power advantage, then we will certainly see them appearing in various levels of the memory hierarchy.

Currently flash is more on the storage side. But we should start seeing novel memory devices in the upper layers of the memory hierarchy, maybe at the same level as DRAM or memory cache, or just underneath, in byte addressable memory space. So the jury is still out. There is a lot of research going on to investigate the most effective way of utilizing these devices, especially for HPC applications where memory access is very demanding. But these are very exciting times because, as we have been discussing, this is a very disruptive technology. Being disruptive, this may result in a tremendous increase in capability.

For a long time, people have said that as a rule of thumb for one flop of performance you need one byte of memory; which means that for a petaflop/s machine it is desirable to have a petabyte of memory. This is true across many applications. There are some applications that only need minuscule memory, but there a lot of applications that require substantial memory. Nowadays, even on machines like ours, people cannot use all the cores, because we simply do not give them enough memory. For example, it is not so rare to see people using only a fraction of the CPUs within a node in the TSUBAME supercomputer, as that is how the applications are programmed and require. But new memory technologies may allow tremendous increase in usable memory of the machine, and expand the scope of the applicability of these machines to non-traditional areas, like big data. So this disruptive technology is not just used to expand the solving of existing problems, but also to expand the scope of the applications, and expand our capabilities.

*Primeur Magazine:*Thomas, do you have something to add to that?

Thomas Sterling: I have a, perhaps, more slightly aggressive view. I think we are going back to an old idea that never found a market niche, let alone an architecture niche, because we really do not have a choice. Where I disagree is with such rule of thumb that one needs one byte per flops and so forth, because that is a processor-centric perspective. It still says the processor is more important. Our colleagues say I need two Gbytes per processor core. The concept of processor core is anachronistic. It is a concept that has to go away, as the principal driver.

The idea of a multi-core chip in the classical sense now, is archaic. It cannot work: we now understand the memory bottleneck, the Von Neumann bottleneck. And it only gets worse. I am sure Satoshi's people have reduced the number of cores, thereby getting improved performance. I think what we have to move towards, is the idea of active memory, or some other appropriate term, where there is a pervasive availability of simple processing elements, really tightly coupled. You mentioned Peter Kogge who was one of the pioneers of Processor-In-Memory (or PIM) as was Ken Iobst, who actually coined the term, back around 1990-1992. So they probably did not use the right structures, but the underlying concepts - that the processor is no longer the critical path but the memory bandwidth is - are correct.

I am not sure about the precise technology, but I am a little bit less critical about DRAM. I could be wrong about that. But I do know the following: the capacitance that you need to store a bit, is in part determined by how long you need to store that bit. It turns out, and this I do know, that in a period between a microsecond and a millisecond you do not need the capacitance. It turns out that people that build these things, not just the gate, but the transistor level, use the capacitance on the internal wire on the chip. It is more than enough to avoid the noise and so forth. As transistors for processing are big, even on the die, where they are combined with DRAM, they take up a disproportional large amount compared to the tightly dense and small transistors used in DRAM. I am not making a projection. I am simply saying I am less certain about technology changes.

Satoshi Matsuoka: If you look at DRAMs today, they are like skyscrapers. They are very, very narrow and tall. We are getting to the point where the density in order to hold the capacitance, needs some volume. But if you think of the scaling, the only direction they can go is the z-direction, although there are some innovations. Actually at Tokyo Tech, my institution, they found some ways to make these DRAMs really, really thin, like only a few micrometers. They have successfully shaved the silicon waver so thin you can actually see through them, and this will allow you to stack these DRAMs effectively. There are still problems with power, if you stack these hundreds of DRAMs in the future. They kind of shave the waver. Still, there are all these innovations coming.

As for PIM, I am in complete agreement with Thomas. HPC is not just about computing but also about processing large data, and in fact in modern applications data-centric view is more important than a compute-centric view. If you believe in that view, it is important to bring data close to computation, so PIM is a very natural incarnation of this idea. As Thomas said, in the nineties maybe there was not such a need, because at that time we could afford to have a more classic way of Von Neumann architectures. But now, because of the restrictions of physics, this is becoming a problem. We need it because data movement is so expensive.

We now see the rebirth of the idea of PIM. So a lot of people, including myself, are thinking about it. The issue is - and this also relates to the earlier question of the design of novel memory architectures - this is another dimension of what processing we can do in a lightweight fashion in these memory integrated processors. There are tasks that may need more classic processing, very heavy, very fast threads of execution. So how do we design a system where you may have very lightweight ultra parallel PIM memory coupled with DRAMs or SRAMs or novel memory? This coupled with very powerful processors kind of reminds you of a very old architecture, the HTMT (which Thomas proposed many years ago). How do we design that for today's technology? That could be a very interesting endeavour.

Thomas Sterling: Well, I learned two things from that project; pitfalls of PIM like structures. I believe that is a path to go, but I am also very sensitive to the problems. While you can reduce the communication by merging simple logic near the data, or part of it, most operations require two operands per operation, not one. No matter how you line it up or arrange it, this means you still have to move one item.

I like to make a prediction that is bizarre, not one that you find in common conversation. I find when people figure out how they can get over the hump, it is usually because they never go back and reconsider their assumptions. Our field has had an intrinsic assumption, not only the Von Neumann type of structures, but there is another one: that all of our logic and storage, as well as communication have been to use the single bit Shannon model. There are reasons for this. There are fundamental theoretical reasons. But first of all boolean logic is only one of a set of logics that were defined by George Boole and others. Secondly, the classic use of saturation logic, which by the way is expensive in time and energy, has been one which has naturally forced us in this assumption. Technology will probably end up going back and looking at the question: Can we have more states per unit area and switching? Can there be multiple voltage rails, multiple levels of voltage design and can it be multiple state boolean logic?

Quantum computing is another marvelous thing above that. I am not proposing that. I am simply saying that the underlying assumption that we have always presumed is that it is two-state logic. We should go back and we should address that. Because if we do, we get much more power in terms of semantic productivity with the same amount of energy and time overhead to do the lowest level of computation. I predict that somebody, not that many years from now, will have the courage to revise it. Having said that, in a discussion with Bill Dally a couple of years ago, he assured that two-state logic is optimal for energy usage. So I could be completely wrong.

Satoshi Matsuoka: Maybe we should go back to our original topics. This shows that at the hardware level, there are lots of innovations happening: on the device level; and on the architectural level. So the question is: people have conceptions about what computers are; how do we deal with that? It is not so much how do we support legacy applications of course. There are lots and lots of new codes. Things can be ported to new execution models and so forth. We still have to make these systems easy to use. We can be very wild about combining all this and all these wonderful technologies, but then we end up in these enormous complex systems. People have to deal with data locations, with a variety of processors with completely different characteristics. You have to move data around everywhere, and precisely time it, or deal with zillions of asynchronies, some having to synchronise, in complex ways. This is impossible stuff for people to comprehend. The deal is not just the hardware, but how to manage all of these in software, because we are at an inflection point.

*Primeur Magazine:*Before we go to the software, there is also work done on ARM and the like. So taking processors out of the embedded area, and try to see whether they fit somehow in HPC, is that an interesting development?

Satoshi Matsuoka: This is exactly what I was getting to. Software must be put into context, with different types of processors and instruction sets. It is always the software in the ecosystem that is very important. The reason people are able to program, and use the machine, is not just because you can write code. You have libraries, you have tools, debuggers, compilers, and all these supportive software infrastructure. People may not even program, they may use ISV software. When you use your PC, your laptop, or your smartphone, it is supported by a tremendously gigantic IT ecosystem. So the way an instruction set becomes more dominant, is the fact that they carry the ecosystem with it. For HPC this is for instance, the x86. It took a tremendous effort by NVIDIA to engineer not only the GPU execution model, but also the programming, to be an important part of HPC. However, they did not have to do that just for HPC alone; rather they had this whole graphics and gaming market behind them and they leveraged all the efforts. Also their GPU was attached to x86: that ecosystem helped them to some extent.

Intel's Xeon Phi is also an attempt to exploit the x86 ecosystem in the HPC space. OpenPOWER is an IBM ecosystem, but Power always used to be more on the enterprise side. Although there are some embedded processors, like PowerPC in the embedded space. Eventually it did quite well in the BlueGene series and was an important endeavour. That was again because IBM had a motivation to push it: because it was their own thing.

So, given all those observations, ARM is a very interesting phenomenon, what is the motivation for using ARM? This is different from OpenPOWER backed by IBM's motivation to exploit the ecosystem, both from the commercial side and the business side, and recently with OpenPOWER the engineering IP side. For GPUs, they have a different interest and position in the ecosystem and they will not throw away the x86; that is they may support alternative processors other than x86, but they will not throw away the existing majority x86 ecosystem.

But ARM is saying: we have this processor for the embedded market, and somehow counters x86 because it initially was low power and now becoming increasingly powerful. When it comes to the HPC space, what is the inherent advantage of ARM over x86 and POWER? Is it low power? Is it cost? Is it the fact that there is an embedded space ecosystem? The answer, although I endorse the project, is that I do not think most people get the idea. People are excited, but I do not think people have a clear idea about what is the inherent advantage of ARM. Except for the fact that it will not end x86 dominance, but increase AMD's influence perhaps. If it is just the intellectual property issue, from an engineering standpoint it is very boring, it is just an alternative choice. Commercial software versus open source.

Thomas Sterling: Let me address the question that Satoshi points to which is, what I think, in line with the question you asked; because it is exactly the right question. If I may allude to my own experience of two decades ago, the answer to the question - and I like to position it a little more concretely, is one word: it is liberating. When I and my colleagues first started exploring the area of commodity clusters using open software, we did not know where we were going to, and certainly did not have any conviction that it would be a useful and successful enterprise. What we found - not because we had great insight, but because we tripped upon it accidentally - was that it took away the vendors strangleholds on the configurations both in hardware and software, which prior to that had been the definition of their IP.

In an analogous fashion we find that ARM, as an Intellectual Property (IP) does not impose the complete capability of the socket and therefore limits the protocols and the functionality of the circuit as a whole. The reason why this is important, and why I think it is differentiating - although I am not a proselytizer type, I do smell blood here - is because none of the processors, including the design of ARM is designed around the use of HPC, designed around its responsibilities in a billion core system. They at most recognize there is a number of additional cores, maybe in an SMP line, maybe slightly more, but they are not defined to operate effectively in a 10^9core environment or under such variability and asynchrony that is needed in storage. I believe that eventually Satoshi's team, and others, will recognize the importance of a billion cores that are designed for the purpose of working in a totality machine. Satoshi used the term ecosphere here. But the ecosphere is not just software, it is all the way down to some of the additional functionalities of the hardware. We understand, and now I use my mantra, if we understand the execution model, we understand what the cores need to do in addition to their local computation. ARM gives you the IP you need, and the flexibility to add more capability onto the same chip integrated with the ARM in order to build a full complete HPC system. That is where I see the opportunity.

Satoshio Matsuoka: Let me give you a concrete description of what Thomas was mentioning. This actually goes back to the original question of the inflection point. The difficulty being the discovered challenges to achieving Exascale. In the old days, when people like Thomas discovered commodity clusters that actually worked, the physical constraints were less challenging. They slapped together server boxes, and that is what we did too. You can slap together these 1U servers and you have a parallel machine. Wow! And we did that. Then we started scaling, and we discovered something: the room was actually getting hot, so we needed to turn up the air-conditioning. We discovered the importance of thermal management, what the vendors already had already learned, the hard way.

Such ad-hoc configuration and cooling was the case for a long time until the commodity clusters became very dominant. Now, it is completely different - if you look at clusters today, they have become very specialized. No longer can you build scalable and reliable clusters out of 1U server building blocks, but they are now rather very specialized purpose-built machines like the HP Apollo or the SGI ICE-X, those that only these highly skilled vendors with professional system architects can build. Projecting this to the future, system architecture design is again becoming the important differentiation factor. For the future, people realize they need to go to chip level integration in order to save power, and build scalable, reliable systems further raising the hurdle.

People are talking, for example, about designing a special purpose SOC (System-On-a-Chip) for scalable computing. But then you need IPs of the processors themselves. So unfortunately for x86 this is not available (yet). The only two companies in the world that can build x86 are Intel and AMD, nobody else. Others can emulate it, but not build a hardware x86 machine. But you need these IPs in order to build a SOC, something that innovates networks, stacked memory, and all these things going into a processor die. So ARM, if you pay something like 50 million - still a huge amount of money - you can get a license. There are various licenses at various levels around which you can innovate with new features, but still be part of the ecosystem as the basic software will still run. The same software that runs on some other ARM based supercomputer will also run on your machine, although you have added all those innovations. So that is the big motivating factor. If you just take a simple ARM chip, building a cluster is not interesting. Rather, I think people are really seeing the next steps. They start seeing these highly integrated SOCs very customized for HPC. And this is also the strategy, although I do not know whether it will be successful, that IBM is trying to do with OpenPOWER. They are actually making the IP of the processor open so that people can integrate this into SOCs or even make it part of their full ASIC. That then could lead to a big innovation that could come in a few years.

Thomas Sterling: I think Satoshi covered the issue. Fundamentally I only would want to make one little footnote. There is a logical aspect to it as well as the practical, physical and engineering aspects. That is that in a distributed system there is the variability and uncertainty of latency and so forth. We cannot think of a machine that is only moving data from one part, but also moving control stacks, and this means there are different aspects of the semantics at the lowest level of the operation of these machines. So when these associates that Satoshi was referring to, design, there has to be an understanding of control as well as migration sometimes in a completely subcontrol manner.

Satoshi Matsuoka: Of course someone who does something based on ARM will understand the ARM instruction set, but adding these dynamic control migration features as an add-on to lower the latency, that will be very difficult to do with x86. So to put it in a very different way, people that are selling ARM processors in HPC space just because they are low power, just do not understand the issue. Intel is perfectly capable of building a similarly low power processor, and they have done that in the embedded space. The low power offerings of x86 are just as power efficient as ARM. That is not the point. It is the open innovation aspect, exactly as Thomas and his people discovered for clusters 20 years ago, and it is now becoming an innovation in its own right at the chip level.

Thomas Sterling: This is a potential challenge for the MontBlanc project. I do not see in there a programme today that understands these points. I am pleased and excited about European advances in this area, but if MontBlanc is as conservative as it appears to be currently, they are going to miss the opportunity Satoshi is referring to.

Satoshi Matsuoka: I think they are looking at it, but their funding level needs to be higher, because building a SOC is very expensive. Also, you need a real commercial commitment. But having said that, yes, looking at having something production-ripe coming out of it that is a very different thing. I would not blame them too much, because their funding level is very basic.

Next week: Part 2. Exascale software developments.

Ad Emmen

Back to Table of contents

Primeur weekly 2015-02-23

Focus

Special

The Cloud

EuroFlash

USFlash

Technology leaders unite around 'Open Data Platform' to increase enterprise adoption of Apache Hadoop and Big Data

Latest five supercomputer projects have direct application to Wyoming issues

Yahoo selects Splunk's Hunk for Hadoop analytics