Thomas Sterling: 'I Think We Will Never Reach Zettaflops'

Nicole Hemsoth May 7, 2012 © HPCwire

As supercomputing makes its way through the petascale era, the future of the technology has never seemed so uncertain. HPC veteran Thomas Sterling, Professor of Informatics & Computing at Indiana University, takes us through some of the most critical developments in high performance computing, explaining why the transition to exascale is going to be very different than the ones in the past and how the United States is losing its leadership in HPC innovation.

HPCwire: Do you have a sense that other regions, China and Europe in particular, are closing the HPC leadership gap that the US has enjoyed for so long? If so, do you think this is more the result of technology democratization rather than government policy choices?

Thomas Sterling: It is clear that other regions are not closing the HPC leadership gap; they are widening it. Through a series of actions in both the EU and Asia the momentum is shifting overseas where once it was entrenched within the US. The Europeans through their EESI and forthcoming EESI2 efforts are making dramatic strides in planning towards a EU-dominant exascale trajectory. The Russians as well are putting in to place a tightly coordinated exascale program combining Moscow State University, T-Platforms, and government, not simply to duplicate prior US methods but to innovate beyond them.

Asia today represents the biggest surge in top performing machines with the 10 petaflops Kei machine in Kobe Japan and the 2.6 petaflops Tianhe-1A system in Tianjin. Perhaps more defining, if somewhat less powerful, is the Sunway BlueLight, a petaflops-class machine built entirely of Chinese microprocessors. Less clear is the degree to which all of these machines are being applied effectively to end-purpose applications, but it is only a matter of time until these strengths push other science and industry objectives beyond the US sphere of influence.

The reason is a mix of both technology democratization and government policy. Neither is working in favor of the US. While the US will be deploying 10 and 20 petaflops machines over the next couple of years, it is clear that the momentum in innovation is off-shore. What may reverse this distressing trend is the new — no pun intended — energy at DOE in driving towards exascale through innovative advancements in software, programming methods, application parallel algorithms, and eventually at least to some degree in hardware.

HPCwire: How is the transition to exascale different from the other HPC milestone transitions — terascale and petascale — that we've passed through?

Sterling: The transition to exascale is different from the two previous tri-decade transitions through which we have passed, and in two fundamental ways: one related to the past, and the other the future. By the early 1990's, the "killer micro," cheap DRAM, and the emergence of system area networks manifest as MPPs (for example, Intel Touchstone Delta) and commodity clusters (for example, my own Beowulf project). These combined with the foundational Communicating Sequential Processes execution model reflected by the message-passing programming model established a formula to match weak-scaling workloads to VLSI component technologies.

At about 11 year intervals this delivered teraflops-scale computing, ASCI Red in 1999, and petaflops-scale, Roadrunner in 2008. However, this highly successful strategy is unlikely to facilitate the realization of exascale computing, except perhaps for some specialized and carefully crafted workloads. This is because the means adopted by this approach to address key factors of performance degradation will no longer prove adequate.

For example, the fine-grained instruction level parallelism and coarse-grained concurrent processes will not provide sufficient efficient parallelism to meet the billion-plus-way parallelism requirement of exascale. Static resource allocation

and task scheduling is insufficient to provide the necessary efficiency or scalability as well as the introspective techniques required for reliability and power management. I expect the need for new programming models, which may include but not be limited to variations of previous techniques, will be essential for communicating between user applications and underlying execution systems.

As I have asserted in the past, a new execution model as an embodiment of a paradigm shift will drive this transition from old systems to new. We have done this before in the case of scalar to vector and SIMD, and again from these to message passing, MPPs, and clusters. We are now simply — or not so simply — facing another phase shift in HPC system programming, structure, and operation.

Exascale is also different because unlike previous milestones, it is unlikely that we will face yet another one in the future. These words may be thrown back in my face, but I think we will never reach zettaflops, at least not by doing discrete floating point operations. We are reaching the anvil of the technology S-curve and will be approaching an asymptote of single program performance due to a combination of factors including atomic granularity at nanoscale.

Of course I anticipate something else will be devised that is beyond my imagination, perhaps something akin to quantum computing, metaphoric computing, or biological computing. But whatever it is, it won't be what we've been doing for the last seven decades. That is another unique aspect of the exascale milestone and activity. For a number, I'm guessing about 64 exaflops to be the limit, depending on the amount of pain we are prepared to tolerate.

HPCwire: What's the biggest hardware challenge to attain exascale computing?

Sterling: The usual response to this question is either "power" or "resilience" and these are certainly critical challenges to attaining exascale. Depending on the analysis of choice, without innovative ways of managing vertical and lateral data movement power estimates based on anticipated technology trends suggest an order of magnitude greater power demand than is considered practical. Single point failure modes of systems comprising hundreds of millions of cores will exhibit mean-time-to-interrupt on the order of minutes or many seconds, much less than the expected time to service a checkpoint or restart cycle using conventional methods.

While both are clearly important, I think the biggest hardware challenge is architecture. This may surprise many of our colleagues because there is a general expectation that the system architecture is likely to be an evolutionary extension of the current mix of multicore sockets and GPU accelerators. This view is driven by the number one concern, which is parallelism and the need to expose and exploit it. Not only will the system architecture have to provide sufficient hardware concurrency of on the order of a billion or more simultaneous actions for the throughput requirement, it will have to use more of it as a latency mitigating method requiring additional architecture change.

Further, it will have to incorporate mechanisms to reduce overheads in order to make effective use of finer granularity tasks (e.g., lightweight user threads) such as the instantiation of remote actions. Support for advanced forms of global address spaces, their management, and address translation will be required in support of randomly distributed global data (e.g., dynamic graphs). New mechanisms for efficient semantically rich synchronization and continuation (control object) migration to manage locality of control will be part of future designs if they are to succeed at unprecedented scale.

Additional hardware mechanisms will be required for fault tolerance including error detection, isolation, in-memory checkpointing, and recovery through reconfiguration. Power reduction will demand active sensor and control hardware mechanisms to continuously adjust energy usage based on application demands. New processor cores and their relationship to memory (for example, processor in memory) for superior bandwidth, reduced latency, and lower power will further drive hardware innovation needs.

HPCwire: How about software challenges?

Sterling: Every advance in hardware will require corresponding changes in software. But the software challenge extends beyond this supporting role. Perhaps most critical is the development of performance-oriented runtime system software for scalable computing. Such software will include dynamic scheduling for lightweight user threads, message-driven computation for moving the work to the data, global address space management, and again efficient support for powerful

synchronization objects like the futures construct to eliminate the use of global barriers and enable asynchrony control through dynamic adaptation.

The ParalleX execution model, as well as the HPX-3 prototype runtime system and the ETI SWARM that embody many of its principles, are two examples that support these goals, even on today's conventional parallel distributed system architectures. But they are only a beginning as the needs for a new generation of fault tolerance and energy management control will be required, too. With a billion cores, their memory hierarchy, and layered communications, a new scalable and robust operating system will be needed. A new software architecture is required to provide a context in which both runtime and operating system need to be mutually designed.

One major challenge is a new interface and protocol definition between the runtime and OS that enables a unique dynamic for a symbiotic relationship of mutual and interactive support. The presence of a performance runtime system also imposes new demands and class of functionality on future compilers that now play a very different role given the existence of a runtime and the exploitation of introspective techniques. These changes percolate up to affect the need for new application programming interfaces. In combination this suggests possibly an entirely new software stack for exascale computing implying that it is not too early to be investing in and conducting research in these areas already.

HPCwire: Do you think the industry will provide suitable manycore hardware and software products that can be applied to high performance computing — for exascale, but also for HPC in general?

Sterling: This is a complicated question with the answer depending on what is meant by "the industry," "suitable," and "products." I am not sanguine about the current path and offerings as incrementally extended to exascale, and industry roadmaps that assume this approach will have shrinking impact on the total range of problems that will eventually apply exascale capability to their solution space. I don't think we as a community know enough at this point to establish what the right hardware/software machine is for general purpose exascale or even if such a system is possible within the constraints of parallelism, energy, and reliability.

Therefore claims that particular vendors have it under control are of limited value at best. The design space is just too complicated, prior methods for scaling to Moore's Law apply to a decreasing degree, whole new modalities demanding advanced runtime components are yet to be derived but are essential, and generality is already diminishing to a worrisome degree for such assertions to have meaningful validity.

Nonetheless, industry will deliver the systems that will be used in the next decade. There is no other choice. It is clear that vendors would prefer not to have to retool and this is true for users as well. To do so will involve a degree of disruption that would be best avoided if it were possible. And for a portion of the overall workload, even at exascale, this may prove to be possible. But such systems are a placebo to an ailing HPC community that if not in triage, is already showing symptoms of underlying conditions that require attention.

The big hurdle is when industry fully embraces the need to address the system wide parallel computing challenge at the processor core level, refactors the physical and logical relationship between cores and memory banks for minimum latency and maximum bandwidth, and transitions from static to dynamic execution models and system software. I do expect this to happen but not without strong push from the user mission-critical agencies.

Thomas Sterling will be delivering the Wednesday keynote at this year's International Supercomputing Conference (ISC'12), which will take place in Hamburg, Germany from June 17-21. His presentation will examine the achievements over the past 12 months in high performance computing.

Related Articles

The Power to Flop

The Bumpy Road to Exascale: A Q&A with Thomas Sterling

Taking a Disruptive Approach to Exascale