ACCELERATED COMPUTING: A TIPPING POINT FOR HPC

Michael Feldman **Addison Snell**

Sponsored research report

November 2015

INTERSECT360 RESEARCH ANALYSIS

The High Performance Computing (HPC) industry has gone through several eras marked by distinct architectures, each with its own programming model for application scalability. For decades vector processing was dominant, with codes that were "vectorized" to take full advantage of the model. But soon scalar processors became more economically viable, and a generation of RISC-based symmetric multi-processing (SMP) architectures dominated the market, using shared-memory programming models such as OpenMP. In the late 1990s, x86-based clusters came into the market, following the "Beowulf" model of industry-standard parts into commodity HPC systems, and the application base migrated to message-passing parallelism techniques such as MPI.

Over the last three years, accelerators have become firmly established in high performance computing (HPC), a significant part of the transition into another new era of HPC, the many-core era. Like previous architectural transitions, this migration is being driven by the economics of price/performance. More cores yield more computational power. This newest transition is amplified by a supply-side shift, as the breakdown of Moore's Law and Dennard scaling leaves the market no choice but to evolve.

"Accelerated computing has reached a tipping point in HPC. Within a year or two, the majority of systems will be equipped with accelerators."

The result is a great diversity of architectural options available, or soon to be available, to the HPC user. "Standard" x86 processors are themselves adding more and more cores, and the door is open for the adoption of systems based on POWER or even ARM. But the greatest proliferation of alternatives has been in accelerated computing, based on the premise that accelerators can execute parallel software more efficiently, from a transistor real estate perspective, than scalar processors. In general, accelerator architectures are highly parallel in design, more so than multi-core CPUs, and thus can deliver a lot of computational throughput with a relatively small number of transistors. As a group, they also offer superior memory bandwidth, and in most cases, provide a large amount of floating point capability. The current implementations span a range of architectures and include GPUs (NVIDIA and AMD), the Xeon Phi (Intel)[1], DSPs (Texas Instruments), and FPGAs (Altera and Xilinx).

According to the most recent Intersect360 Research surveys, approximately a third of HPC systems operating today are equipped with accelerators and nearly half of all newly deployed systems have them[2]. Currently, the most widely used accelerator type is the GPU, with about 80% of the market according to our latest site

[1] Although Intel is releasing its upcoming "Knights Landing" Xeon Phi as a standalone microprocessor as well as a co-processor, we treat both versions as accelerators for the purpose of this report due to Xeon Phi's many-core design and its perception in the market.

[2] "HPC User Site Census: Processors," October 2015, http://www.intersect360.com/industry/reports.php?id=129.

surveys. NVIDIA is by far the largest accelerator supplier, with 78% of that 80%, the 2% remainder being from AMD. Intel, with its Xeon Phi offering, has over 10% of the market, and we expect that share to grow appreciably as that product line matures.

A look at the adoption patterns leads us to conclude that accelerated computing has reached a tipping point in HPC. Overall, we expect the accelerator market to expand to the point that within a year or two, the majority of new systems will be equipped with accelerators. The battle over implementations will be fought in software as much as in hardware.

In concert with the hardware advances we've seen over the last several years, there has been a steady accumulation of accelerator-supported HPC software libraries and application packages. Here again, the vast majority of accelerated codes are implemented on GPUs, specifically with NVIDIA's CUDA (Compute Unified Device Architecture) programming tools. OpenCL, an open-standard programming framework for heterogeneous computing, is also being used for application acceleration. OpenCL has the advantage of being able to support a variety of accelerator platforms, including GPUs, APUs, FPGAs, the Intel Xeon Phi, DSPs, and even multi-core CPUs.

There are already many HPC application packages available and supported in GPU-accelerated versions. In a separate sponsored research report[3], Intersect360 Research found that 32 of the 50 most commonly cited HPC applications[4] now include support for GPUs (with two more in development). Nine of the top 10 codes offer GPU acceleration. In certain areas such as chemical research, physics, structural analysis, and visualization, the availability of GPU acceleration is nearly ubiquitous. In other areas, like biosciences and environmental modeling, penetration remains limited.

Even with the wide array of development tools available, writing code for accelerators is challenging. Nevertheless, the performance advantages inherent in this technology have motivated the HPC community to take up that challenge, just as it did with previous migrations (vector to SMP to MPI). The availability of ported application packages reflects that effort and bodes well for the future prospects of accelerated computing.

[3] "HPC Application Support for GPU Computing," November 2015.
[4] Based on Intersect360 Research HPC User Site Census data, 2015.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com