



## **Accelerating the Future of High Performance Computing**

### **The Need for Speed**

Most users of high performance computing (HPC) systems are driven people. Some are driven to achieve something for the first time, to gain new insight through better accuracy or detail of analysis. Others are driven to produce results sufficiently fast that it enables people to make better decisions, substantially improve their productivity or to predict events that can change or save lives.

The very nature of HPC is that the solution to one set of questions or problems generates yet another set of questions. These next-step problems are generally more complex than their predecessors. Subsequently, more accurate and powerful tools are required to move ahead which generates an insatiable thirst for compute cycles and a compulsion to achieve constantly increasing levels of performance and precision.

Recent advances in technology by the leading processor companies have made HPC capabilities available to many more individuals and organizations by combining excellent performance with the cost benefits of volume markets. Unfortunately the very availability and affordability of these new systems has created a new class of challenges.

### **The Limit of Power Dissipation**

Suppose you dream of building the world's first system capable of achieving a petaflop on LINPACK (one quadrillion floating-point operations per second solving dense systems of linear equations). The excitement builds as you sketch the possibility on the restaurant placemat... a few thousand nodes, each at a sustained speed of a few teraflops... and then you note what that kind of cluster will require in the way of power dissipation and real estate (square footage), and the "gulp" sound can be heard all over the restaurant.

Walk into any Wal-Mart and you will have no problem finding a Windows-based PC (that you can reconfigure to run Linux in a matter of hours) for a few hundred dollars. If you put it in a cluster, how much will you spend on electricity over its useful life of about three years?

Over *a thousand dollars*. Hmm... can that be right? If a server consumes just 300 watts and you keep it on for a year, how much does the electricity cost? At 12 cents per kilowatt-hour, it will use \$375 per year. Google's biggest single line item expense is the power bill for their enormous server farms.

Electricity ranges from about 5 cents per kilowatt-hour in places like the Department of Energy's Pacific Northwest National Laboratory (which is right next to a surplus of nuclear power as well as plenty of hydroelectric power) to about 23 cents per kilowatt-hour at the Maui High-Performance Computing Center. As the price of hardware falls with Moore's law, the price of the energy to flip all those ever-denser bits keeps rising with inflation and the price of a barrel of oil.

Semiconductor and system suppliers have definitely gotten the message. Intel and AMD have moved away from advertising their gigahertz rates and made quite a point about performance-per-watt. Having spent years convincing the typical consumer that speed and clock frequency are the same thing, their marketing departments now face the task of preaching an entirely new figure of merit. It might not be a tough sell to anyone who's run out of laptop battery power on a long plane trip, or has noticed how laptop computers have almost gotten too hot to keep on your lap for very long.

For the physicists among us, there is an interesting thing about making performance-per-watt a figure of merit. Performance is *computational* work per second. A watt is a unit of *physical* work (one joule) per second. (If you do not have a gut feel for what a "joule" is, 3.6 million of them is one kilowatt-hour.) Thus, performance in flops/sec divided by watts is simply flops per joule! As we approach physical limits to computation, there may be a convergence in the metrics used by computing people and those used by physicists.

For the largest supercomputing clusters, the procurement planners now must entangle a power budget with the hardware budget. It is a bit simplistic to equate the power budget with money. It usually does not work that way. The facility has some maximum ability to supply power to the computer room, and increasing it by several megawatts is a major engineering project. The money for doing that usually comes from a different pot, if the money is even obtainable for that purpose.

## **The Limit of Floor Space**

When it was first proposed to Los Alamos National Laboratory (LANL) that they fill a building with racks of microprocessor-based units to reach new supercomputing performance, the suggestion was greeted with uproarious laughter. That was in 1984, and supercomputing floor space was whatever it took for a love seat-shaped mainframe from Cray Research and some boxes full of proprietary disk drives, not an entire building.

Now it is generally accepted that if you want to do supercomputing, you will be filling a building with racks. The issue is how many you can accommodate.

Even if your facility has enough floor space, the communication fabric might limit the maximum distance between racks. And even if the electrical specification does not limit the distance, the speed of light will start to add to the message-passing latency between cabinets if they get too far apart. With MPI latency at about one microsecond lately, that has so far managed to mask plenty of speed-of-light delay; light in a vacuum travels about a thousand feet in a microsecond. But for a system that occupies over 10,000

square feet, connecting the opposite corners of a 100 by 100 foot square might (at best) be done with an optical cable in which signals transmit at only 70% the speed of light. Moreover, the cable is not line-of-sight straight between the corners, so allow perhaps 180 feet of cable between points. That delay adds about 250 nanoseconds to the MPI latency, reducing performance significantly for latency-sensitive applications.

The first thing to do, of course, is to pack each cabinet with as much computing capability as possible to reduce the number of racks. It may be easy to use a 25-inch rack instead of a 19-inch rack, but increasing the height introduces an objection from an unexpected source: those concerned with safety. Taller racks have a greater chance of injuring people if they tip over, and any system that asks administrative staff to stand on stools or ladders incurs a nasty reaction from the insurance company.

Packing more general-purpose computers into a rack, however, intensifies the heat generated. That forces the ugly choice between using more floor space and introducing liquid cooling. Air cooling hits its limit at about 70 watts per liter... near sea level. For a place like LANL, at 7500 feet above sea level, forced-air cooling is only about half as effective. Hence, air-cooled racks hit a limit of about 20 to 30 kilowatts. Besides blowing the heat away from the processors, the requirement to blow the heat out of the computing room is getting arduous as well. The Thunder system at Lawrence Livermore National Laboratory requires an air current under the raised floor at 60 miles per hour. A typical guideline is that the power to remove the heat adds 20% to the total power required for the system.

Finally, the limit of floor space is like the limit of power dissipation, in that it does not simply translate into cost. The floor space may not be available, at any price. At a national laboratory, creating new building space is literally a line item for Congress to approve. In the financial centers like Manhattan and London that use HPC for financial modeling, the space is not only expensive but unlikely to be for sale right where it is needed.

### **So, What Should We Do?**

The new frontier in chip design is finding clever ways to compute with less energy. Perhaps the most visible ray of hope is in the emergence of multi-core processor chips running at lower clock speeds. Every time a microprocessor vendor doubles the number of processing elements but lowers the clock speed, it drops the power consumption yet raises the effective performance.

Despite these impressive advances in microprocessor technology, a tradeoff still persists between designing a computer for the full range of applications and designing it for technical applications that make heavy use of floating-point arithmetic. The extra hardware demanded by high-performance computing (HPC) raises the price of a processor too much to incorporate it into a general-purpose chip.

The rise of graphics accelerators such as those from ATI and nVIDIA® illustrates this situation. A microprocessor can certainly perform graphics functions, albeit at a lower

speed than the graphics card. Designers could allocate some of the transistors on a chip that currently go to caches and other general performance enhancements to graphics processing; however, many microprocessor applications involve little or no graphics, whereas almost all applications benefit from caches, look-ahead features, etc. relegating both graphics and HPC functions to separate hardware moves the choice from the chip engineer to the user, avoiding the tradeoff issues.

The ClearSpeed CSX600 chip is currently the most radical example of energy-efficient, multi-core processor designs with 96 processing elements running at only 250 megahertz. The result is a chip that is simultaneously the fastest at 64-bit floating point speed (about 25 GFLOPS/s sustained) yet one of the lowest for power consumption (under 10 watts). The ClearSpeed chip is a true coprocessor, and depends on a general-purpose processor for a host. If you need double-precision floating-point performance, you can achieve it without adding to the facilities burdens of power demand or floor space. Using chips optimized for an application regime in combination with standard, generic processors is something we can do right now to mitigate facilities costs.

This “hybrid” computing approach is very reminiscent of hybrid cars... two kinds of motors for two kinds of driving, resulting in much higher efficiency than if one tries to use one kind of engine for everything. The resurgence of interest in the coprocessor approach may remind people of the old days of attached processors from Floating Point Systems, but the FPS approach never had anything to do with saving electricity or space in the computing room. Coprocessors are no longer separate cabinets, but instead are available via plug-in boards that consume an otherwise empty expansion slot on a server or workstation. The ClearSpeed Advance™ accelerator board combines two CSX600 processors that deliver 50 GFlops of sustained double-precision matrix multiply (DGEMM) performance for while averaging 25Watts power consumption.

The Advance board works in conjunction with the host processor to manage the most computationally intensive portions of the application. When a call is made by an application to a ClearSpeed supported standard math library, it is intercepted by ClearSpeed’s accelerated math library, CSXL, which calculates if the function call can be accelerated. When it can be accelerated, the required data is transferred to the Advance board, the answer is calculated, and the results passed back to the host CPU. Throughout this process the only perceivable difference between a function running on the host system and a function running on the Advance board is the speed. The acceleration is transparent to the end user and the application.

### **Measured Linpack performance**

The Linpack Benchmark is used to solve a dense system of linear equations. The Top500 which tracks the 500 most powerful commercially available computer systems known on a semi-annual basis uses a version of the benchmark that allows the user to scale the size of the problem and to optimize the software in order to achieve the best performance for a given machine.

Benchmarks performed by ClearSpeed Technology in August 2006 on a single server with two 3.0 GHz Intel® Xeon® 5160 (Woodcrest) dual core processors system delivered 34 GFLOPS without acceleration. A cluster of four such nodes delivered an impressive 136 GFLOPS from its 8 Intel® Xeon® 5160 (Woodcrest) dual core processors while consuming 1,940 Watts of power.

With two Advance accelerator boards in each server, a single node delivered 90 GFLOPS and the cluster performance was increased to over 364 GFLOPS while adding only 200 Watts to the overall power levels, representing over 1GFLOP Linpack per Watt of additional performance. The ClearSpeed accelerated cluster completed the Linpack benchmark run in just 18.4 minutes while using only 40% of the energy required by the non-accelerated cluster which took 48.4 minutes to finish.

To put these results in context, the performance delivered by the four node ClearSpeed accelerated cluster, (a total of 16 CPU cores,) is equivalent to the world's most powerful supercomputer according to the Top500 results from November 1996. That supercomputer was a massive 2048 CPU Hitachi system at the Center For Computational Science at the University Of Tsukuba in Japan that delivered 368.2 GFLOPS.

### **Accelerating into the Future**

While satisfying an insatiable demand for compute cycles is by definition impossible, it is also clear that shifting from traditional supercomputing architectures to new hybrid approaches offers a way for the HPC community to overcome the challenges presented by floor space and energy consumption limits.

By collaborating on complementary solutions, mainstream vendors can deliver increasingly more efficient processors designed for the needs of mass market, while specialist suppliers like ClearSpeed Technology can focus on the needs of the HPC community. Together they can deliver the precision, the performance and the reduced power consumption required for the HPC community to continue pushing the boundaries of knowledge.