Securities Industry News

By Janice Fioravante
March 26, 2007

Since at least the mid 1990s, Wall Street has used a new generation of high-performance computing (HPC) configurations to run tasks such as Monte Carlo simulations. But until very recently, the physical limitations of semiconductor technology stood in the way of bringing true HPC onto trading floors. The basic architecture of the Intel Corp. x86 family used too much power and took up too much space, and applying it to a high-volume application such as equities trading wasn't deemed practical.

In line with Moore's law, the power of the x86 line, and high-end semiconductor chips as a whole, has been doubling roughly every other year, but with each boost in power, the machinery throws off more heat, which in turn requires more air conditioning for cooling, which also approaches physical limits.

"If the chip vendors had stayed on the path of increasing clock speed, we would have had to submerge computers in liquid nitrogen to cool them," quips Ed Turkel, manager of high-performance product marketing at Hewlett-Packard Co. (HP).


Ziegler

Heat and power are not the only issues. The server farms that power the electronic trading systems at the heart of Wall Street operations require physical space that is getting more expensive and scarce by the day. Kurt Ziegler, EVP of product development at New York-based Aspeed Software, says that Manhattan itself isn't big enough to house all the servers Wall Street wants to add. The company's software, Accellerant, adapts applications such as Microsoft Corp.'s Excel to take advantage of parallel processing, multiplying the CPUs (central processing units) available to apply to a complex, compute-intensive task.

Fortunately, the most recent generation of chip and server technology appears to be addressing the heat, power consumption and space issues. Rather than just making chips run faster, Santa Clara, Calif.-based Intel, and its chief rival, Advanced Micro Devices of Sunnyvale, Calif., are increasing the memory of their processors and coming out with semiconductors with two, four and eight cores.

Hardware manufacturers led by HP, IBM Corp. and Sun Microsystems have made space and power considerations a major part of their product development and sales pitches. Specialized vendors including Egenera of Marlboro, Mass., Rackable Systems and Appro, both of Milpitas, Calif., Verari Systems of San Diego and Angstrom Microsystems of Boston have condensed servers into smaller, so-called "blade" footprints to squeeze extra CPUs into data centers. Having the easy-to-install and -replace blades clustered in compact "farms" is also simplifying the task of managing the hardware.

Power of Sharing

"We're just a few years from the very big, large-scale, very expensive proprietary technology," HP's Turkel says. "Now we have clusters of industry-standard, fairly economical multicore processors packed as densely as possible." Being able to share power supplies across several blades also contributes to the reduction in power requirements.

"Clusters are very prominent on Wall Street," says David Turek, IBM's VP of deep computing. "With network connections and software, you can string the racks together. You can do a whole rack of low-end servers for a few hundred thousand dollars."

Over the past six years, blades have become steadily more compressed and easier to manage. "Servers in clusters and grids had cables all over the place and became dreadlock' nightmares," says Turek. "The enclosures for blades have lots of intelligence for better hardware and software control. With clusters of blades, systems can grow organically by adding another blade as the needs of the business grow."

The deployment of blades is also facilitating parallel processing for computationally intensive tasks. "The trick in this transition," says Turek, "is to make applications that used to run on six processors now run on thousands of them."

But, "for maximum performance they have to think in terms of separate tasks, dividing them among processors--that's how to take advantage of multicore, multi-threaded environments," says Nathan Brookwood, research fellow at Insight 64, a Saratoga, Calif.-based research firm.

The computer industry can design chips or platforms that can handle 8, 16, 32, 64 threads, but the software to work on it is not there yet," contends Jim McGregor, research director and principal analyst at Scottsdale, Ariz. research firm In-Stat. "Specific single applications can do that, but to program everything to do 8, 16 or 32 things simultaneously will need a whole new programming paradigm." Today's apps are so customized, says McGregor, that it is difficult to "port" them from one platform to another.

One potential solution for adapting software to take advantage of multicore processors is the Cell Broadband Engine Architecture microprocessor design being jointly developed by IBM, Sony Corp. and Toshiba Corp. The first commercial implementation of it was in Sony's PlayStation 3 game console.

"There's a lot of math in vectors for running [electronic] games," explains Aspeed's Ziegler. It turns out that systems that are optimized for games run many data elements simultaneously, use lots of algorithms and perform many numeric calculations. Since Wall Street does a lot of algorithms and numeric calculations, some financial systems vendors realized that the chip design could be used for markets other than video games.

"It's still in the experimentation stages," says Ilya Mirman, VP of marketing for Interactive Supercomputing of Waltham, Mass. "It's a question mark that it will end up being useful in the real world." Interactive's Star-P software adapts applications for parallel processing on server farms, Mirman says, adding, "There's a shortage of programmers versed in parallel computing on Wall Street. Financial analysts are typically not programming experts, but with Star-P, you don't need to involve programmers to run a model. If the quants can offer an answer faster, then traders can react and make moves in the market much faster."

"The ability to run transactions will accelerate significantly in the next five to ten years," says In-Stat's McGregor. "To be able to break down multiple transactions to run simultaneously on the circuit and system architecture--this is happening and being used by specific apps today, but the wave of the future is being able to do it effectively for all applications."