

# Meeting the Need for Speed

Firms are struggling to overcome the five obstacles to scalable throughput and fast response times in today's financial services applications.

By **RAM APPALARAJU**

**I**n financial services, IT infrastructure is all about high-speed transaction processing and the ability to maintain that speed throughout wild variations in load and unpredictable surges in transaction volumes. On Tuesday, Feb. 27, the Dow Jones Industrial Average suffered its worst performance since the aftermath of the September 2001 terrorist attacks; it plummeted in a matter of minutes and lost 416 points that day. What some have called the "Crash of 2007" was compounded by a computer system that couldn't handle an unexpected surge in the volume of trades—about 4.5 billion, double the daily average.

IT industry experts are not surprised that even the most efficient IT systems, including those that we rely on for our most critical business processes, can be overwhelmed by unexpected volume. While a handful of trading systems have proprietary technology aimed at solving this challenge—though not guaranteed to do so—most trading systems around the world, including systems for foreign exchange, derivatives, and commodities, are at risk. Furthermore, many financial services firms are grappling with the challenge of implementing real-time risk analysis concurrent with transaction processing, which further burden their systems.

While the minor meltdown that occurred in February affected Wall Street and the financial services industry, no firm is immune to the threat of unexpectedly high transaction processing volumes. Those at risk include online retailers, auction houses, insurance companies, betting houses, and travel booking agencies. Even RFID tracking systems and many telecom companies are vulnerable. For most of these organizations, transaction processing speed translates into revenue, and the failure to maintain speed during peak periods means lost business.

## Five Obstacles to a Faster Infrastructure

Analyst firm Gartner estimates that by the end of 2008, more than 80 percent of all new e-business application development will be based on virtual machine technology, Java or other managed runtime environments. But while these platforms offer a host of benefits, the hardware systems they run on simply weren't designed for virtual machine technologies and suffer from other limitations as well.

Here are the five key obstacles to achieving optimal speed and reliability for high-volume transaction processing applications on the Java platform:

**Dedicated Compute Resources:** The standard today in most datacenters is still a fixed number of servers, and therefore a fixed amount of compute resources, dedicated to each application. While datacenter managers attempt to plan for maximum demand, if they are wrong and an unexpected surge in demand occurs, nothing can be done immediately to bring more computing resources to the application. The result is application performance degradation and unmet service levels.

**Garbage Collection Pauses:** The Java platform uses memory in the form of a "heap." Objects are created in this heap as needed by the application, and once the objects are no longer needed, the application pauses briefly so a process called "garbage collection" can reclaim the memory for reuse. The trend in investment banking is that ever-increasing trading volumes are driving dataset sizes to 10 to 100 times larger than what traditional Java virtual machines are designed for, which causes the applications to pause at unpredictable times. In such large-scale transaction-intensive applications, these pauses can run from several seconds to minutes. This causes service levels to plummet and creates unacceptable delays for clients.

**Limited Memory Heap Size:** To avoid the long garbage collection pauses, vendors of conventional application servers recommend limiting memory heap size to 2 GB, but

“What some have called the ‘Crash of 2007’ was compounded by a computer system that couldn't handle an unexpected surge in the volume of trades—about 4.5 billion, double the daily average.”

such a small heap size restricts the application's ability to grow and deliver enhanced functionality. Organizations face an unacceptable choice: Limit the power of their applications, resulting in user dissatisfaction and lost competitiveness, or expand their applications and risk processing delays resulting in an equal level of user dissatisfaction.

**Lock Contentions Limit Scalable Performance:** This may seem like a small and very technical issue, but the impact on application processing is significant. Although today's applications may be multi-threaded, keeping the threads from interfering with one another is a problem that grows in complexity as the number of threads increase. This is especially true when many threads need to access memory. To preserve the correct order in which threads access individual memory locations, access is usually serialized via "memory locks." In large-scale applications, this locking behavior can create unnecessary contention for shared objects in memory, dramatically reducing parallel performance and causing unacceptable delays in application processing.

**Costs: Utilization, Footprint, Power and Cooling:** In an attempt to ensure reliable speed, IT managers tend to throw more servers at the problem. But each server must be over-specified to meet peak demand. With utilization at 10 to 20 percent, companies are paying huge sums for a lot of idle computing power, and the number of servers continues to proliferate.

Server proliferation results in costly datacenter expansion and additional real estate costs, while attempts at consolidation bring equally painful results. Costs for power and cooling are rising dramatically relative to server costs—in excess of 70 percent in 2007—and this trend will only continue.

### **The Answer: Extreme Transaction Processing**

Firms need "extreme transaction processing," and the goal for Java applications is to scale and respond in real time to meet business needs.

Gartner defines "extreme transaction processing" (XTP) as "an emerging application style aimed at enabling the implementation of large-scale, business-critical, transactional applications on the basis of distributed architectures implemented by leveraging commodity hardware and standards-based software."

According to a Gartner report, XTP solutions must integrate with pre-existing applications in advanced service-oriented architecture (SOA) scenarios and should scale horizontally on commodity hardware and standards-based software to minimize initial deployment costs and optimize the cost of scaling up. The report also suggests that today's XTP solutions should be based on a mix of incremental and disruptive technologies, and organizations must be willing to take some risks with these



technologies to gain competitive advantage. One of the technologies listed in the report is the Azul Compute Appliance.

The Azul Compute Appliance offloads Java processing from application servers and eliminates obstacles to a faster, more reliable infrastructure with a new computing model and optimized hardware. The appliance creates a shared pool of compute resources available to any application that needs them. This eliminates capacity planning at the application level and can ensure available resources even in the most unexpected situations. Hardware-assisted garbage collection and optimistic thread concurrency are new technologies that eliminate garbage collection pauses and scalability bottlenecks caused by lock contention.

These technologies enable the appliance to take full advantage of Azul's Java computing appliances that contain up to 768 processing cores and support up to 768 GB of memory, ensuring that applications can continue processing transactions reliably and at maximum speed, no matter what the load. By offloading Java processing, the Azul Compute Appliance dramatically reduces the number of servers required for an application, reducing the datacenter footprint and dramatically reducing power and cooling requirements. For performance-critical applications in investment banking, such as in trading and risk analysis, Azul Compute Appliances have delivered dramatically increased application throughput by five to 10 times and at the same time improved application response times by a factor of five to 10.

Unless financial services firms immediately begin to upgrade their systems to handle unpredictable transaction volumes, we can expect many more days when system delays cause financial upheaval, whether on a worldwide or corporate scale. The Azul Compute Appliance, available today, should be among those technologies carefully considered by the financial services industry as a compelling part of solving this application transaction processing problem. ●

Ram Appalaraju is vice president of marketing product technology for Azul Systems.