IDC
Analyze the Future

I D C A N A L Y S T C O N N E C T I O N

Peter Rutten
Research Manager, Server Solutions

High-Value Business Applications on x86: The Need for True Fault-Tolerant Systems

May 2015

IDC defines fault-tolerant servers as having a combination of multiple hardware and software components that allow a near-instantaneous failover to alternate hardware/software resources so that business processing continues as before without interruption. IDC recommends taking inventory of all installed servers to assess the appropriate availability level for those servers, given the workloads they support. For highly demanding mission-critical workloads accessed by large numbers of end users, lengthy periods of downtime — especially unplanned downtime — are costly to the business and should not be tolerated.

The following questions were posed by HP and Intel to Peter Rutten, research manager, Server Solutions with IDC's Enterprise Platforms Group, on behalf of HP's customers.

Q. How has the universal acceptance of x86 and the decline of Unix forced IT departments with high business value workloads to reevaluate their strategies?

A. Datacenters are going through a transformation, driven by what IDC calls the 3rd Platform of compute, which is characterized by sky-rocketing mobile transaction and data volumes, new deployment models in the cloud, and the increasing importance of social networking for businesses to respond to the marketplace. This transition has affected basic workloads such as IT infrastructure, Web infrastructure, application development, and collaborative as well as high-value workloads such as OLTP, batch, ERP, and database. In particular, OLTP and database-related workloads have become critical as they process the surging transactions that bring revenue, manage the terabytes of data that are a firm's greatest asset, and provide the analytical insights that guide business strategies. These workloads are invaluable for any business today and require the highest levels of protection.

At the same time, we have witnessed about a decade and a half of standardization on low-cost x86 servers to manage the capex in the datacenter, mostly for basic workloads. The high-value workloads largely remained on legacy systems, such as Unix, which have traditionally delivered the highest levels of availability, including true fault tolerance at the very high end. While sales of new Unix-based systems declined year after year, many enterprises remained dependent on these large, expensive, and complex systems because they ran critical — sometimes custom — applications; because migration to x86 would be too difficult, disruptive, and risky; and — most importantly — because there simply were no x86-based systems on the market that could match their reliability, availability, and serviceability (RAS) and their scalability. These firms have now reached a point where they can no longer delay updating the systems that run their high-value workloads, and CIOs are facing difficult choices.

The ecosystem for Unix is getting smaller and smaller, and several server vendors have begun offering scale-up or converged x86-based systems that mostly aim to address what we call "server sprawl" and its opex implications. Some of these systems have improved high-availability features through redundant hardware; from Linux, Windows, or a hypervisor; or from a built-in clustering solution. Few of them can match the RAS features of traditional Unix systems, and none of them can provide true end-to-end fault tolerance (unless, in one case, running in parallel with an identical twin). A fault-tolerant system can detect any error in any layer, whether software or hardware, isolate the malfunctioning element, and continue the workload without interruption or data loss. It can support extremely high OLTP and database volumes with 100% guaranteed reliability and data integrity and with near-linear scalability. Fault tolerance is typically achieved through years of intensely integrated system design at every level of the stack, from the processor, the networking, and the storage to the middleware, the operating system, and the applications.

For the CIO, selecting an x86 platform for high-volume, high-value workloads implicitly means potentially exposing those workloads, and hence the enterprise in its entirety, to unacceptable risks of downtime and data loss, with severe implications for the firm's revenue and reputation. IDC research has determined that for a large firm, the mean cost of downtime per hour is nearly $1.7 million across industries, with some specific industries approaching $10 million lost per hour of downtime. The advent of an extremely scalable and truly fault-tolerant system on x86 resolves this conundrum.

Q. How do you characterize the relationship between the workload and the server platform?

A. IT deems OLTP, batch, ERP, and database-related workloads to be of the highest criticality. Nearly 80% of worldwide IT spending on high-end servers is allocated for executing these high-value workloads. In contrast, only 20% of worldwide IT spending on low-end servers is used for the same purpose. What this means is that, despite the standardization on x86, IT has never felt that the high-value workloads could be performed anywhere but on the most capable platforms with the highest RAS features, which — as mentioned earlier — used to be primarily Unix based. Scalability also plays an important role. There's a strong relationship between the number of sockets on a system and the percentage of high-value workloads the system performs: IT deploys 8-, 16-, and 32-socket systems for high-value workloads because scalability is a crucial component of guaranteeing the SLAs on those workloads.

And then there's architecture. Between 60% and 70% of worldwide customer spending on EPIC-, RISC-, and CISC-based servers is invested to support high-value workloads; for x86-based servers, this number is about 30%. If you juxtapose this data with the ongoing drive toward standardization on x86, what you'll notice is that IT invests in x86 using what you might call a "RAS and scalability barometer." Every time the RAS and scalability on x86 shift further upward, IT will be inclined to move workloads with greater criticality onto the x86 platform. Though it's a cliché, an end-to-end fault-tolerant, massively scalable platform on x86 may turn out to be a classic "if you build it, they will come" scenario for the highest-value workloads. At that point, the only things that are still top considerations for the CIO are the total cost of ownership (TCO) and the ease of migration of high-value workloads onto that platform.

Q. Beyond the traditional industries like telecommunications (telecom) and financial services, what other industries and workloads are reaching a point where downtime is unacceptable for any reason?

A. Most industries use a mix of high-availability solutions for various workloads — from load balancing to virtualization to clustering and, ultimately, — for the high-value workloads — fault tolerance. Financial services and telecom are the usual suspects, known to require a high percentage of their systems to be 100% fault tolerant. Just think of ATM or point-of-sale networks that perform real-time debit and credit card processing — these systems cannot go down. Similarly, telecoms need to manage billions of cell phone calls. Where are you when you make a call? Which cell tower are you being linked up to? These systems are processing hundreds of millions of customers a day and cannot tolerate any downtime.

Beyond telecom and financial services, manufacturing has been going through a tremendous automation revolution. Systems on the factory floor are being connected vertically to higher-level enterprise systems, including ERP and database, and horizontally across sophisticated real-time supply chains that do not tolerate downtime at any link; indeed, they will financially penalize a supplier in the chain for any disruptions. And we're increasingly seeing other verticals requiring fault tolerance — industries that you might not have expected just a few years ago.

Healthcare, for example, is one of the fastest-growing markets for high-availability servers. As more and more of our health insurance data, medical records, and treatment plans are being linked and made available to multiple parties within various regulatory limitations, the healthcare industry is becoming very sensitive to reliability, availability, and serviceability of the database workloads they run, and true fault tolerance is mentioned by a high percentage of healthcare providers as the way they describe high availability in their datacenters. The same is true for retail, for which online transaction processing is absolutely essential for online revenue streams while database workloads are the heart of operations. Indeed, retailers currently describe their high-availability requirements as "fault tolerance" more than any other vertical.

Q. What competitive advantage do enterprises gain with continuous availability and data integrity? What other attributes are game changers?

A. Organizations gain the following advantages:

- **Seamless availability.** By avoiding any planned or unplanned downtime, organizations can maximize revenue generation through transaction processing in a world in which consumers and business customers make lightning-fast decisions to purchase a product or service based on a complex set of parameters among which effortless, seamless availability plays a dominant role.
- **Trustworthiness.** For most businesses, reputation makes or breaks their brand. By ensuring that customers have guaranteed uninterrupted and fast access to services and can always expect 100% data integrity, organizations will see increases in customer loyalty and retention, word of mouth, new customer acquisition, and overall reputation. In today's always-on world, nothing is more damaging than rapidly spreading negative customer sentiment about downtime or lost data.
- **Empowering business goals.** A system that is scalable enough to process large, fluctuating volumes of customer transactions and data without interruption, even if they exceed business projections, allows business goals to be met or exceeded without causing infrastructural limitations.

- **Remaining 100% compliant.** Businesses in all industries are increasingly subjected to regulations that demand that their services be available at all times and that no data be lost under any circumstance; failure to be compliant can have serious consequences for continued operations, including large fees, loss of business licenses, and loss of jobs.

Other game changers are:

- **Responsive flexibility to new trends.** As consumers shift how they interact with companies and pay for services, sometimes suddenly and in large numbers, firms must be inherently capable of quickly developing and rolling out new technologies, scaling them and running them without a glitch to remain competitive.

- **Value chain reliability.** Downtime reverberates through an entire value chain. What this means is that your downtime becomes someone else's downtime. In many industries, the implications aren't just financial (e.g., lost revenue from online transactions); they can be legal or contractual, damaging relations with business partners. The ability to maintain continuous availability, whether you are a bank, an auto manufacturer, or an online retailer, has become a basic competitive requirement.

Q. Do you see an increase in adoption of fully integrated, end-to-end, fault-tolerant systems now that a solution is available on x86?

A. When looking at IDC's current market projections for AL4 servers, you can clearly see where the friction lies. The worldwide market for Unix-based AL4 server revenue is projected to decline from nearly $4 billion in 2009 to almost half that amount in 2018 because IT is abandoning Unix. For Windows, this revenue is nearly nonexistent, and there is no anticipated growth because there are very few Windows-based fault-tolerant systems. On the other hand, revenue for AL4 Linux systems is forecast to grow from near zero in 2009 to almost $1 billion in 2018 given that IT is looking for standardized solutions and open source Linux fits well with the rest of the infrastructure. Finally, revenue from other operating systems in AL4 is slowly declining as well.

Next, take a look at the percentage of CIOs and directors of IT who say they aim to increase availability in the next 24 months for business processing workloads such as OLTP (45%) and database-related workloads (36%), and look at the response from IT when asked what it considers to be the most important feature of a future high-availability solution, which is that it utilize standard hardware. Interestingly, until 2011, the market for high-end x86-based servers was divided among four smaller vendors. At that point, one of the top 3 vendors started marketing a high-end, database-oriented x86-based integrated system that can achieve fault tolerance by being clustered. This particular solution has seen very strong growth even though it requires a doubling of the investment to achieve AL4.

We believe that all of this data demonstrates that customers are looking for end-to-end fault tolerance on x86, which should drive significant demand for truly fault-tolerant systems on x86. Indeed, any system that delivers on that need should be able to not only capture a part of the existing AL4 market but also increase the overall AL4 market size by bringing in new customer segments as they evolve with the 3rd Platform trends and are now seeking true fault tolerance for their highest-value workloads.

ABOUT THIS ANALYST

Peter Rutten is a research manager, Server Solutions with IDC's Enterprise Platforms Group. In this role, Mr. Rutten focuses on the high-end server market, including Unix, mainframes, and high-end x86 servers. His research also covers subjects that are of critical importance to high-end server customers such as high availability, big data, and flash storage.

ABOUT THIS PUBLICATION

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

COPYRIGHT AND RESTRICTIONS

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com