IBM z16 – A Modern Mainframe for a New Business Era

By Cal Braunstein and Jean S. Bozman

IBM's z16 system provides a new path for customers to "meet the moment" in this New Normal business environment.

In a world that is clearly showing the strain of supply-chain delays affecting shipments of server systems, the IBM z16 is a highly scalable platform that can consolidate and run – and secure – the rapidly growing workloads being generated across an enterprise's business units.

Much has changed since the IBM z15 was announced and shipped starting in 2019 – and customer priorities have shifted, along with a changed business environment due to the COVID-19 pandemic, supply chain challenges, and war in Europe. Business outcomes and speed-to-decisions have risen in priority for customers facing rapidly growing data across their organizations – along with the need to protect the entire organization from cyberattacks and ransomware that can disrupt operations – and take them off-line.

And, while the IBM z16 was not designed as a consolidation platform for x86 workloads, IBM z16's specialty processors are expected to accelerate processing of Linux, open-source, and analytics workloads across an organization's hybrid cloud. Without faster engines, the aggregation of many localized x86 servers running those same workloads would not be able to keep up with customer-driven demand to scale out IT resources. The IBM z16 can be used to scale up Linux, open-source computing – and to accelerate AI-based processing. That can be done on-premises, via the IBM Cloud – or delivered via a third-party firm's cloud services.

Specifically, the IBM z16 system will help customers to gain a unified view of corporate data – an important requirement in a highly distributed world. The new system's enhanced AI capabilities, coupled with integrated on-chip accelerators, are designed to make analytics much faster, and more comprehensive.

According to IBM, the new IBM z16 can process up to 300 billion inference requests per day with just one millisecond (ms) of latency – enabling companies to perform fraud checks on the full volume of high-volume transactions – in contrast to an estimated 10 percent of all potential fraud-checks being handled today. The IBM z16 speeds and scales the processing for artificial

intelligence (AI) workloads, often combined with machine learning (ML), and supports a wide range of Linux, open-source, and containerized applications – the technology foundation that is powering much of modern cloud computing.

There is a hardware element to this high-performance workload acceleration in the IBM z16, as well. Significantly, the IBM z16 is built using up to 200 cores per rack-based system with a 40 percent per-socket performance improvement, leveraging the 7nm IBM Telum processors designed by IBM and fabricated by Samsung.

IBM executives said that the company – long known for its IBM mainframe designs -- is continuing to invest heavily in the zSystems processors – which are due to move from a 7nm process towards a 2nm design sometime in the 2020s. That will further enhance this system, as it speeds and scales the processing for artificial intelligence (AI), machine learning (ML) and Linux, open-source, and containerized applications. There are few competitors in the traditional mainframe space – but the new IBM z16 platform could support AI processing that today is running on supercomputers and clusters of advanced GPU devices for AI-based workloads.

Addressing key demands

IBM has designed the IBM z16 to satisfy its customers' latest requirements and be more than a replacement platform for older IBM zSystems, including the IBM z15. As the IBM z15 focused on hybrid cloud and containerized workloads in 2019, the IBM z16 is "doubling down" on making those customer workloads run faster and more efficiently and securely.

To this, it is adding new objectives: faster business analytics, handling much larger data volumes; supporting next-generation quantum-safe security; and greater processor consolidation, which supports improved energy consumption for enhanced sustainability.

Taking a customer-centric approach to large-scale computing, IBM is addressing these key areas for New Normal-era business demands:

- Gaining actionable data. Data is growing rapidly, but the ability to analyze it quickly and accurately is not keeping up. For two-thirds of Fortune 500 companies and many others, most mission critical data is kept on the mainframe. Yet scalable AI systems in the data center lacked the ability to support real-time data-based insights via fast analytics,

where the compute was not on the mainframe. The new IBM z16 can deliver low-latency to these local datasets and satisfy customer workloads that require detection of credit-card fraud, make AI-based recommendations, and visualize high-resolution images.

- **Ransomware and cyberattacks.** These cyber challenges are top concerns for CIOs, who view finding gaps in security as vital to continuous operations. The IBM z16 has inherited the IBM z15's pervasive encryption of data, and support for end-to-end security via customer-only access keys all along the path to the Edge. The IBM z16 is an engine for pervasive encryption that can be extended to the Edge, for end-to-end security throughout the hybrid cloud. Now, the IBM z16 added two new security features: a new, uniquely-designed quantum-safe system, which is underpinned by lattice-based cryptography; and a secure boot capability, which prevents bad actors from injecting malware into the boot-up process. These security enhancements will help protect data and systems against current and future threats. Combined with the system's LPAR-enforced isolation, and air-gapping to protected storage devices, the IBM z16 enables enterprises to adapt to the changing threat environment that could disrupt enterprise-wide operations.

- **Connections to distributed computing resources.** IBM has deepened its software-stack support for Linux open-source workloads, including containerized applications, highly distributed data, and Kubernetes orchestration to manage containers in the hybrid cloud. This process began before IBM acquired Red Hat in 2019 – but has accelerated since then, with an expanded software stack made available in 2021. The IBM z16 supports rapid access to scalability for Linux and open-source applications that are already running in a highly distributed hybrid cloud environment.

- **Addressing the New Normal Environment.** The fallout from the global COVID-19 pandemic has shifted IT policies, to address remote workers and remote, IoT devices; and to rapidly growing data resources at the Edge and in the Cloud; and the immediate need to tighten security across-the-board, to preserve business continuity. In this case, new z16 features are supporting these important business objectives.

- **Scalability, computing power and acceleration.** The IBM z16 system's on-chip accelerators, built into its 7nm IBM Telum processors, support speed and efficient

business analytics for an enterprise's rapidly growing datasets. The IBM Telum processors that are at the core of the IBM z16 systems will speed transactional processing – the mainstay of IBM zSystems. At the same time, built-in integrated acceleration in the IBM z16's Telum processors, will speed AI and business analytics, processing higher data volumes in less time – and running business analytics faster to reach actionable results more quickly than before. IBM expects strong demand for the IBM z16's "specialty" engines, which run Linux and data-based analytics.

Customer-Centric Workloads

The portfolio of targeted applications for mainframe computing have expanded, reaching beyond OLTP alone, and adding AI supercomputing features that have been running on specialized HPC systems made by other vendors. The driver behind the Red Hat acquisition is also coming into sharper focus: the new IBM z16 is targeting faster processing for a shifting software landscape that has become the *de facto* environment for all cloud computing: Linux, open-source, containers, and Kubernetes.

This customer landscape will include the IBM z16 as one of the key components in a hybrid cloud environment that has been shaped since the IBM z16 design process began in 2017. This means the IBM z16 is a modern enterprise server that is designed to handle core applications and databases. The design process itself was informed by a "design thinking" process to which more than 70 IBM enterprise customers contributed.

A few of these workloads are:

- **AI for Business Analytics** – The IBM z16 Telum Processor includes an industry-first, integrated on-chip AI accelerator that provides millisecond response times for analytics. This integrated capability enables enterprises to perform real-time analytics on workloads that could not be done before – whether it be fraud analysis, geospatial analysis, medical or other predictive analyses. The IBM z16 supports IBM Watson Machine Learning for IBM z/OS, also announced with the IBM z16, is designed to help clients make faster predictive decisions using insights from operational data, including native z/OS applications.

- **Data mining** – With the new AI accelerator, the IBM z16 becomes the most efficient, least cost solution to mining data kept on mainframe datastores. DB2 13 for z/OS, launched along with the IBM z16, includes a new feature — SQL Data Insights — that is designed to take advantage of the IBM Telum Processor and more quickly and easily identify similarities and clusters in data, while also simplifying the data science process. The system also runs DB2 13 for z/OS, the latest version of the DB2 database, which is widely deployed in IBM zSystem installations. This latest version of DB2 is being leveraged here to perform business analytics on large data warehouses and data lakes.

- **IBM zSystems hybrid cloud modernization stack**. IBM is shipping an expanded set of software tools that complement the IBM z16 platform, and that are designed to help customers' businesses become more flexible, and to innovate faster, with greater productivity.

IBM's objective now is to have customers innovate with the zSystems engines in the cloud or data center – as it has with the new 7nm IBM Telum processors designed by IBM and fabricated by Samsung – and to push forward on the IBM zSystems roadmap for years to come.

Key Takeaways

Analyst Point of View (POV): IBM is making good on its promise of ensuring that AI and hybrid cloud will be primary "movers" of the company's overall strategy in the 2020s. It's worth noting that customers need end-to-end solutions that will allow them to satisfy customer demands, while protecting the customer, sustainable growth, and the environment.

Customer demand for more accelerated AI, applied directly to business analytics, is being supplied on this platform, along with enough compute capacity to accomplish that goal. The customer-centric design thinking process that supports more real-time decision-making (e.g., credit-fraud detection; quantum-safe multi-factor security; and New Normal-era governmental and regulatory compliance, worldwide) was well-played, precisely because it directly addressed business-critical customer issues.

Given IBM's strategy to focus on AI and hybrid cloud, it looks like there is more to come during the IBM z16's product lifecycle, based on a decade-long roadmap: On Day 1, IBM said its

roadmap will extend throughout 2020s, further speeding and scaling customers' business workloads – leveraging the platform for artificial intelligence, machine learning, and a broad portfolio of Linux, open-source and containerized applications. Surely, given the depth and breadth of the IBM z16's on-board capabilities, there is potential to do even more with the system's on-board AI capabilities and its integrated accelerators, to widen the scope of business analytics for applications running in the hybrid cloud. As a result, we expect the IBM z16 to gain significant attraction now – and deep into the decade.

---

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; Contact: inquiry@rfgonline.com