IBM z16

—

Reasons to upgrade from your z15 or z14

Edward L. Linde II
Principal Product Manager- IBM z16
ell11@us.ibm.com

IBM z16 is built to build – Why upgrade

Leverage innovative on-chip AI inferencing and quantum-safe technologies to accelerate insights in every customers interaction and protect against the threat of the coming "quantum apocalypse", where bad actors harvest data now to decrypt later. With a z and cloud experience you can modernize applications and accelerate time to value in your digital transformation. We built the world's most powerful and secure platform for business. That's why 67 of the Fortune 100 companies rely on IBM zSystems.

Capability	Value	Proof points and use cases	IBM z16™	IBM z15™	IBM z14®
AI Acceleration	Embed AI into every customer interaction and business process at scale to deliver increased value, reduce risk and improve business productivity.	Leverage IBM z16 on-chip AI acceleration to integrate inferencing easily into transactions with low latency and high performance to improve business results, reduce risks and deliver customer value with every interaction. IBM z16 with z/OS delivers 20x lower response time and up to 19x higher throughput when co-locating applications and inferencing versus sending the same inferencing operations to a compared x86 cloud server with 60ms avg. latency.[1] With IBM z16, process up to 300 billion inference operations per day with 1ms response time using a Credit Card Fraud Detection model.[2]	X		
Quantum-safe Security	Leverage the industry-first quantum safe system[3] to protect against the looming future threat where data is collected NOW and encrypted later using quantum computers. Prevent "quantum- apocalypse"	Mitigate the risk of a cyber attack now or in the future, by using advanced Quantum-safe cryptography that makes your data more resilient to decrypting. IBM Z secure boot technology helps protect IBM z16 firmware from quantum attacks through a built-in dual signature scheme with no changes required. IBM z16 quantum-safe APIs will enable clients to begin using quantum-safe cryptography along with classical cryptography as they begin modernizing existing applications and building new applications. IBM z16 quantum-safe APIs will enable clients to build hybrid quantum-safe key exchange systems with Crypto Express 8S protection of the keys.[3]	X		
Flexible Capacity for Cyber Resiliency	Easily transfer capacity from one data center to another in seconds for disaster recovery, testing, maintenance and compliance at an attractive price point and in a straightforward manner.	Flexible Capacity Transfer optimizes resources between IBM z16 systems in different data centers for disaster recovery, compliance and testing. IBM Z Flexible Capacity for Cyber Resiliency is designed to help organizations proactively reduce the impact of downtime by dynamically shifting their critical workloads to an alternate site for business continuity.[4]	X		
Continuous Compliance	The IBM Z Security and Compliance Center and IBM z16 make regulatory compliance easier and faster, while reducing human errors and risk	IBM Z Security and Compliance Center on the IBM z16 mitigates regulatory risk and reduces audit preparation time. Sponsored user client reporting of projected savings after implementing the solution was a reduction in preparation time of 55%[5] and reducing the number of skilled resources needed for audit preparation functions by 40%[6].	**X**		
Enhanced System Recovery Boost	Leverage middleware boost, SVC dump processing boost and HyperSwap configuration load boost on the IBM z16 to extend the System Recovery Boost benefits introduced on the IBM z15 to further reduce the recovery time from planned and unplanned outages.	System Recovery Boost on z16 provides additional capacity to your z/OS system at middleware startup to process your transaction backlog up to 35% faster than on z15 by delivering up to 25% more throughput and up to 30% better response time during the boost period.[7] During middleware startup, System Recovery Boost on IBM z16 provides additional processor parallelism and up to 40% increased general CP processing capacity during the boost period.[8] On the IBM z16, the use of System Recovery Boost when capturing large SVC Dumps to catch up on paused work is up to 30% faster during the boost period.[9]	**X+**	**X**	
Sustainability	Lower your carbon footprint with the IBM z16 to help you meet your sustainability objectives by achieving energy efficiency vs comparable workloads on x86 or on the IBM z14. Sustainability is built into every aspect of the product lifecycle from design to product packaging to disposal.[6]	Each IBM z16 single frame saves an average of $16,250 in data center floor space and power consumption costs per year vs. compared x86 2U servers running the same workloads and throughput.[10] Accessing your database while running an OLTP workload on OpenShift Container Platform, requires up to 3.6x fewer cores running your workload when co-located on IBM z16 versus running the workload on compared x86 platform connecting remotely to the IBM z16.[11] An IBM z16 single frame system saves up to $6,373 in data center floor space cost per year vs. IBM z14.[12]	**X**	**X**	
Scalability	Achieve massive scale with the IBM z16 to address your toughest workload demands and position you for non-disruptive growth.	The largest IBM z16 has approximately 17% more capacity than the largest z15[13] The largest IBM z16 has approximately 47% more capacity than the largest IBM z14[13] The IBM z16 is capable of processing up to 25 billion encrypted z/OS OLTP transactions per day[14] The IBM z16 can execute up to 20 billion HTTPS transactions per day with OLTP microservice applications running on the Red Hat OpenShift container platform[15]	**200 Cores 40 TB RAM**	**190 Cores 40 TB RAM**	**170 Cores 32 TB RAM**
Performance	Meet workload growth challenges, implement containerized cloud applications, consolidate Linux workload for sustainability and cost savings. Deliver exceptional customer experiences with blazing fast and predictable performance	11% performance gain for single thread vs IBM z15 or 27 % vs IBM z14.[13] New cache architecture with 1.5X more cache reduces latency and improves performance New 7nm processor chip technology running at 5.2 GHZ with encrypted memory New FICON Express32S has 2X the bandwidth as 16 Gbps adapters for faster data transfer Significantly improved scaling for CF images larger than 9 ICFs On IBM z16, the enhanced ICA-SR coupling link protocol provides up to 10% improvement for read requests and lock requests, and up to 25% for write requests and duplexed write requests, compared to CF service times on IBM z15 systems.[16] Deliver exceptional customer experiences with blazing fast and predictable performance	**1.1X**	X	.73X

Notes and Disclaimers:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

1. DISCLAIMER: Performance results based on IBM internal tests using a CICS OLTP credit card workload with in-transaction fraud detection. A synthetic credit card fraud detection model was used: https://github.com/IBM/ai-on-z-fraud-detection. On z16, inferencing was done with WMLz on zCX. Tensorflow Serving was used on the compared x86 server. A Linux on IBM Z LPAR, located on the same IBM z16, was used to bridge the network connection between the measured z/OS LPAR and the x86 server. Additional network latency was introduced with the Linux "tc-netem" command to simulate a remote cloud environment with 60ms average latency. Results may vary. IBM z16 configuration: Measurements were run using a z/OS (v2R4) LPAR with WMLz (OSCE) and zCX with APAR– oa61559 and APAR - OA62310 applied, 8 CPs, 16 zIIPs, and 8GB of memory. x86 configuration: Tensorflow Serving 2.4 ran on Ubuntu 20.04.3 LTS on 8 Skylake Intel® Xeon® Gold CPUs @ 2.30 GHz with Hyperthreading turned on, 1.5 TB memory, RAID5 local SSD Storage.
2. DISCLAIMER: Performance result is extrapolated from IBM internal tests running local inference operations in an IBM z16 LPAR with 48 IFLs and 128 GB memory on Ubuntu 20.04 (SMT mode) using a synthetic credit card fraud detection model (https://github.com/IBM/ai-on-z-fraud-detection) exploiting the Integrated Accelerator for AI. The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The lscpu command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results were also reproduced using a z/OS V2R4 LPAR with 24 CPs and 256GB memory on IBM z16. The same credit card fraud detection model was used. The benchmark was executed with a single thread performing inference operations. A batch size of 128 inference operations was used. Results may vary.
3. DISCLAIMER: IBM z16 with the Crypto Express 8S card provides quantum- safe APIs providing access to quantum-safe algorithms which have been selected as finalists during the PQC standardization process conducted by NIST. https://csrc.nist.gov/Projects/post-quantum-cryptography/round-3- submissions. Quantum-safe cryptography refers to efforts to identify algorithms that are resistant to attacks by both classical and quantum computers, to keep information assets secure even after a large-scale quantum computer has been built. Source: https://www.etsi.org/technologies/quantum-safe- cryptography." These algorithms are used to help ensure the integrity of a number of the firmware and boot processes. IBM z16 is the Industry-first system protected by quantum-safe technology across multiple layers of firmware.
4. DISCLAIMER: The IBM z16 systems must be installed in different locations, using z/OS Version 2.2 or above.
5. DISCLAIMER: IBM does not ensure regulatory compliance. The intent is to provide a point in time statement of your current posture for a specific group of resources. The responsibility of ensuring systems are configured in accordance with regulatory controls is on the individual businesses who are using the IBM Z security and compliance Center and IBM does not take responsibility for any compliance oversights or penalties associated with data breaches. The survey consisted of 9 responses across 6 unique customers. Sourced from the IBM ZSCC Sponsor User Program and zDC.
6. DISCLAIMER: IBM does not ensure regulatory compliance. The intent is to provide a point in time statement of your current posture for a specific group of resources. The responsibility of ensuring systems are configured in accordance with regulatory controls is on the individual businesses who are using the IBM Z security and compliance Center and IBM does not take responsibility for any compliance oversights or penalties associated with data breaches. The Survey Consisted of 8 responses across 5 unique customers. Sourced from the IBM ZSCC Sponsor User Program and zDC.
7. DISCLAIMER: System Recovery Boost on IBM z16 provides a 5-minute boost on the z/OS system where the middleware is started. WLM Classification Rules must be configured to enable the boost for the selected middleware products. The magnitude of the capacity boost will depend on the speed boost available for your machine model and the availability of zIIP capacity during the boost period. Measurements were done with 4 GCPs and 2 zIIPs on a z15-704 and a pre-release model z16-704. Pre-release IBM z16 results were adjusted to account for final improvements. Your results may vary.
8. DISCLAIMER: System Recovery Boost on IBM z16 provides a 5-minute boost on the z/OS system where the middleware is started. WLM Classification Rules must be configured to enable the boost for the selected middleware products. The magnitude of the capacity boost will depend on the speed boost available for your machine model and the availability of zIIP capacity during the boost period. Measurements were done on a z16-704 with 4 CPs and 2 zIIPs. Your results may vary.
9. DISCLAIMER: System Recovery Boost on IBM z16 provides a 2-minute boost on the z/OS system where the SVC dump is captured. The magnitude of the capacity boost will depend on the speed boost available for your machine model and the availability of zIIP capacity during the boost period. Measurements were done on a z16-704 with 4 CPs and 2 zIIPs. Dumps were 71 GiB in size. For testing with boost, the dump option, RPBMINSZ, was set to 60G. Your results may vary.

Notes and Disclaimers (Continued):

10. Based on comparable IBM z15 study. Disclaimer: The floor space covered by the systems includes doors and covers. The z15 system includes 3 CPC drawers with 108 configurable processor units and one I/O drawer. x86 systems ran at various utilizations according to 15 customer surveys, representing Development, Test, Quality Assurance, and Production levels of utilization and throughput. Workloads tested are a mix of leading databases and application servers, such as WebSphere, Node.js, MongoDB and Db2. Each consolidated workload ran at the same throughput and SLA response time on Z and x86. All x86 systems are 2U form factor, and 21 x86 systems fully populate a standard 42U rack. External storage floor space is not included. z15 performance data was projected from actual z14 performance data by assuming a 10% performance improvement on z15. Compared x86 models are all 2-socket systems containing a mix of the following x86 processor models: 8-core Xeon E5-2667 v4, 12-core Xeon E7-8857 v2, 12-core Xeon E5-2680 v3, 8-core Xeon E5-4650, 8-core Xeon E5-2650, and 14-core Xeon E5-2690 v4. Average annual data center floor space cost is $221.02 per square foot; average US commercial power rate is $0.10 per kWh; and Average Power Usage. Effectiveness ratio in Data Centers is 1.67 (67% additional power is required for cooling the data center), according to IBM IT Economics.
11. DISCLAIMER: This is an IBM internal study designed to replicate banking OLTP workload usage in the marketplace deployed on OpenShift Container Platform (OCP) 4.9 on IBM z16 using z/VM versus on compared x86 platform using KVM accessing the same PostgreSQL 12 database running in a z16 LPAR. IBM z16 configuration: The PostgreSQL database ran in a LPAR with 12 dedicated IFLs, 128 GB memory, 1TB FlashSystem 900 storage, RHEL 7.7 (SMT mode). The Compute nodes ran on z/VM 7.2 in a LPAR with 8 dedicated IFLs, 188 GB memory, DASD storage, and OSA connection to the PostgreSQL LPAR. The OCP Proxy server ran in an LPAR with 1 IFL, 4 GB memory and RHEL 8.5. x86 configuration: The Compute nodes ran on KVM on RHEL 8.5 on 32 Cascade Lake Intel® Xeon® Gold CPU @ 2.30GHz with Hyperthreading turned on, 192 GB memory, RAID5 local SSD storage, and 10Gbit Ethernet connection to the PostgreSQL LPAR. Both systems are delivering equal throughput. Results may vary.
12. Disclaimer: The actual floor space covered by the system includes doors and covers. The "Radiator-cooled with I/O top exit" version of z14 was used for all z14 models. All z14 models have standard front and rear covers. Floor space areas of all z14 models (M01-M05) are identical. Assume annual data center floor space cost is $221.02 per square foot
13. IBM z16 to z15 and z14 performance comparisons are based on internal measurements. Results may vary by customer based on individual workload, configuration and software levels. Visit LSPR website for more details at: https://www-40.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex
14. DISCLAIMER: Performance result is extrapolated from IBM internal tests running a z/OS CICS-VSAM OLTP workload on IBM z16. The measurement environment consisted of 2 z/OS 2.5 LPARs, each with 6 CPs, and an Integrated Coupling Facility with 4 engines. Both data set encryption and Coupling Facility encryption were enabled. Results may vary.
15. DISCLAIMER: Performance result is extrapolated from IBM internal tests running in an IBM z16 LPAR with 24 dedicated IFLs, 560 GB memory and DASD storage the Acme Air microservice benchmark (https://github.com/blueperf/acmeair-mainservice-java) on OpenShift Container Platform (OCP) 4.9 using RHEL 8.4 KVM. On 4 OCP Compute nodes 4 Acme Air instances were running in parallel, each driven remotely from JMeter 5.2.1 with 384 parallel users. The KVM guests with OCP Compute nodes were configured with 12 vCPUs and 64 GB memory each. The KVM guests with OCP Management nodes and OCP Infrastructure nodes were configured with 4 vCPUs and 16 GB memory each. Results may vary.
16. DISCLAIMER: Measurements were done with an IBM internal workload generating a representative mix of coupling facility requests on an IBM z16 running two z/OS partitions with 16 GCPs on each partition and coupling facility image(s) with 4 ICFs each running at about 30% utilization. Measured with shared ICA-SR links. The amount of improvement will vary based on workload and configuration.

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

IBM*	IBM Z*	FICON*	zHyperlink
ibm.com	IBM z16	z14*	z/OS*
IBM (Logo)*	Db2*	z15	

* Registered trademarks of IBM Corporation

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
ITIL is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
Red Hat®, JBoss®, OpenShift®, Fedora®, Hibernate®, Ansible®, CloudForms®, RHCA®, RHCE®, RHCSA®, Ceph®, and Gluster® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.
RStudio®, the RStudio logo and Shiny® are registered trademarks of RStudio, Inc.
UNIX is a registered trademark of The Open Group in the United States and other countries.
VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.
Zowe™, the Zowe™ logo and the Open Mainframe Project™ are trademarks of The Linux Foundation.
Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.
This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.