How mainframes are rewriting the AI playbook

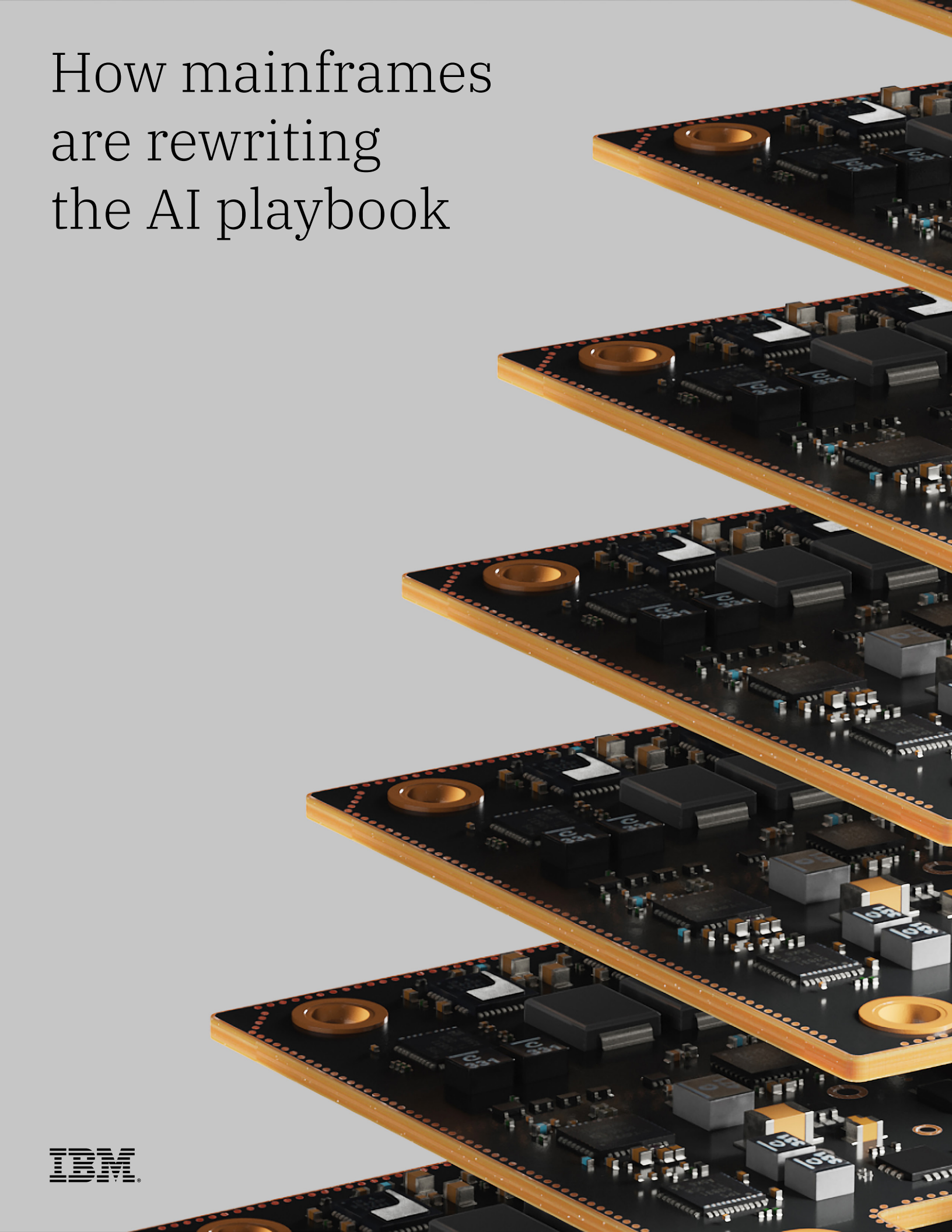
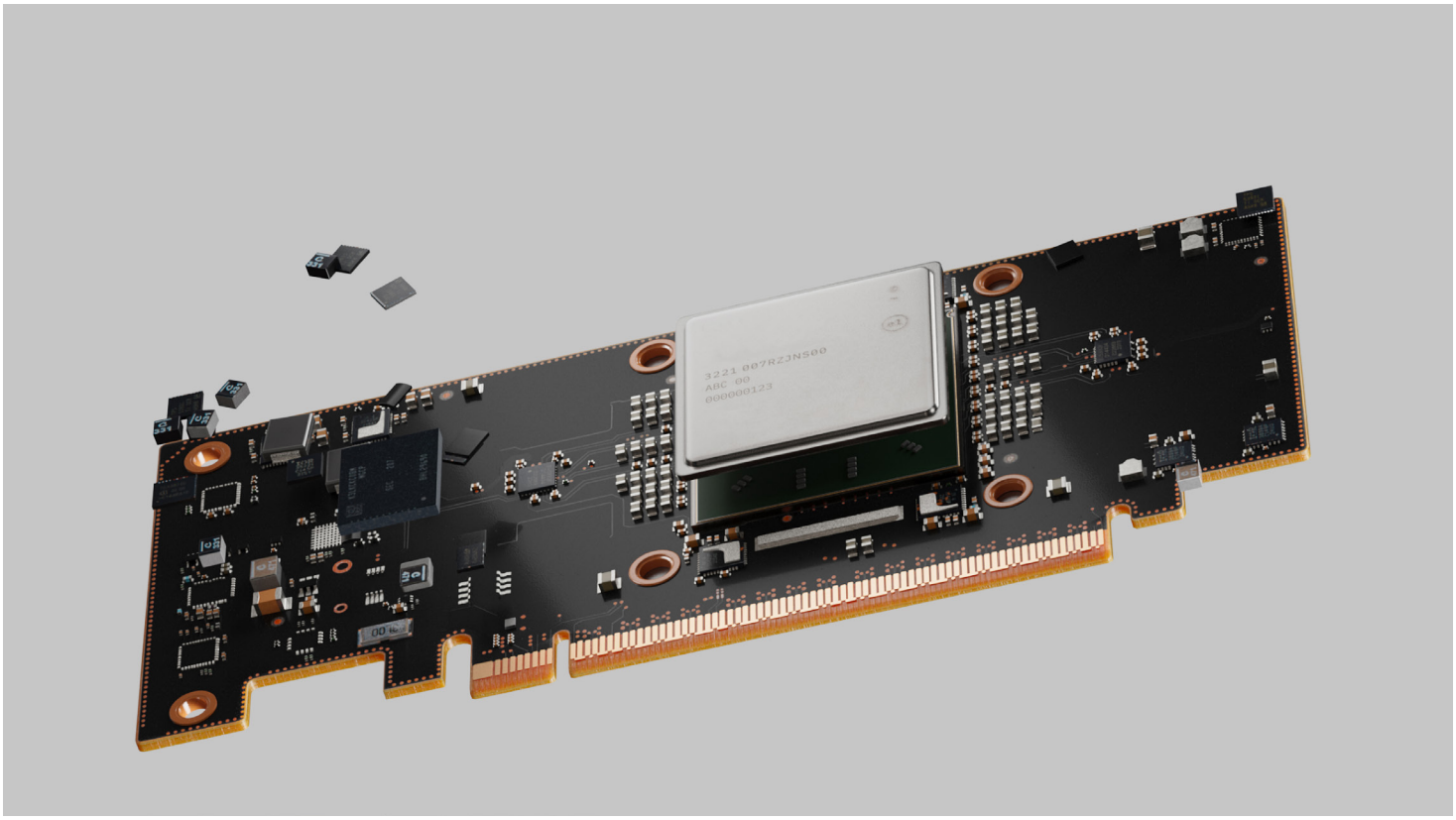


Table of contents



01

The modern mainframe

The modern mainframe

The modern mainframe can process more than 1 million transactions per second. To put that information into context, imagine all of Amazon's daily transactions in the US—around 12 million on an average.[1] The modern mainframe can process all the transactions Amazon handles in a day in just 12 seconds.

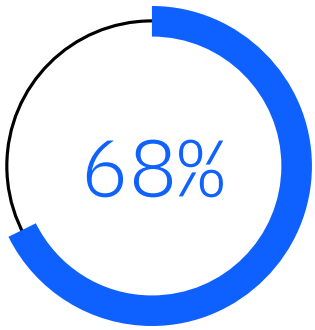
Mainframes have long been a cornerstone of enterprise IT and form the backbone of the global economy, providing critical technology infrastructure to organizations in banking, life sciences, healthcare, public sector and more. Every day, people around the world benefit from the mainframe without even realizing it. When they pay using credit cards, withdraw cash from ATMs, use instant payment apps, make reservations with airline companies, file their taxes or access their medical records, the chances are those transactions are processed by a mainframe.

Mainframes handle 68% of the world's production IT workloads and 72% of the world's transactional workloads and process 90% of all credit card transactions.[2,3,4] 8 of the top 10 payment companies in the US use mainframes. This means, over USD 8.5 trillion in processing volume from over 156 billion transactions by 7.3 million merchants in the US alone are handled by mainframes.[5] 43 of the world's top 50 banks, 63 of the world's top 75 banks and 77 of the world's top 100 banks—as ranked by S&P as of November 2022—also use mainframes.[6] In addition, 8 of the top 10 insurers, 8 of the top 10 telcos and 7 of the top 10 retailers use mainframes for their core workloads.[7] And mainframe use continues to rise—the global mainframe market is projected to reach USD 6.2 billion by 2032, registering a CAGR of 7.9% during the forecast period (2024-2032).[8]

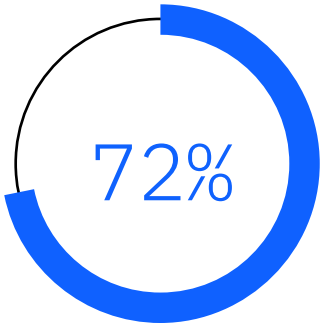
"Mainframes have a very specific role to play in this modern ecosystem. Many enterprise core data assets in financial services, manufacturing, healthcare, and retail rely on mainframes quite extensively."[9]

Chirag Dekate
Analyst at Gartner

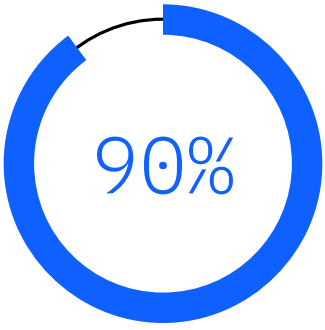
Mainframes handle high-volume workloads at high speeds.



of the world's production IT workloads



of the world's transactional workloads



of the world's credit card transactions

According to the 2024 State of Mainframes report from Forrester Research, 54% of surveyed organizations that currently use a mainframe expect to increase their mainframe use over the next two years.[10] And new research from the IBM Institute for Business Value shows that 75% of surveyed global IT leaders consider mainframes equal to or better than cloud in terms of total cost of ownership (TCO).[11]

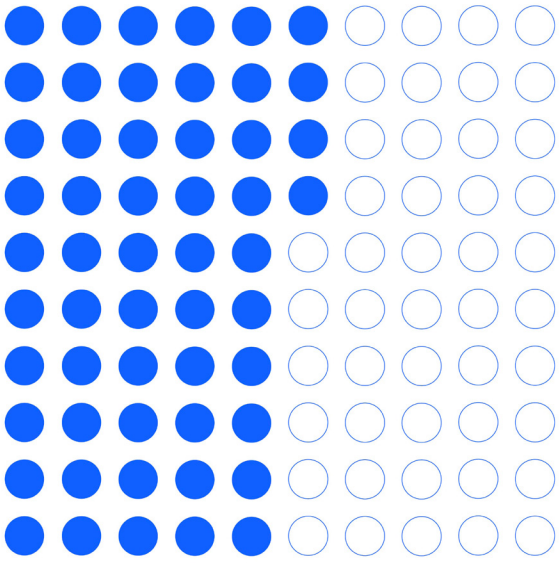
But what makes the mainframe a unique and powerful computing platform and why is it foundational to enterprise IT in the age of cloud, AI and quantum? To understand that we need to understand how the mainframe really works.

How is the mainframe able to process 1 trillion web transactions—comparable to the sales volume from 55 Black Fridays—daily, with the highest levels of security and reliability?[12] Why is it relied upon for high-volume transaction processing? And why do organizations choose to invest in it?

As a result of a technology strategy that has fundamentally evolved with the market, the industry and the users—and a design philosophy that seamlessly integrates and optimizes hardware and software—today's mainframes are workhorses that deliver unparalleled transaction throughput and scalability.

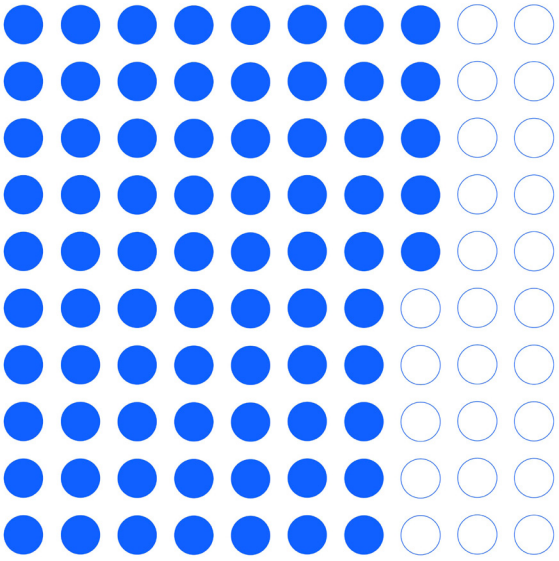
These same qualities are making mainframes' stock rise in the AI era. Mainframes are designed to store and process massive amounts of business-critical data, now organizations can apply AI to their data at the source and achieve AI outcomes faster and more cost-effectively.

In this white paper, we'll examine the different elements that make up the current and next generation of mainframes. As we do that, we'll explore how a decades-long commitment to transformation and the importance of the architecture has resulted in a powerful platform that calculates its mean time between failures in decades.



54%

of surveyed organizations that currently use mainframe expect to increase their mainframe use over the next 2 years.



75%

of surveyed global IT leaders consider mainframes equal to or better than cloud in terms of TCO.

02

Architecture

“The IBM mainframe was built to be resilient, scalable and adaptable as market demand evolves. The technology has stood the test of time.”

Suresh Chandrasekaran
Head of Payments and Fintech Solutions, APAC
Fiserv

Architecture and design: The anatomy of a mainframe

Unparalleled speed, scalability and security. How do mainframes deliver the high performance and reliability that have become their defining characteristics? The answer is simple: they're engineered that way. Their fundamental architecture is designed and built with precision and discipline to complete as many transactions as possible with the highest levels of security and reliability. In comparison, a precision Swiss watch isn't as precise; it may lose or gain 730 seconds a year.

Mainframes are made for workloads that must be available 24x7, need to scale on demand, contain highly sensitive data and are extremely high volume, such as banking transactions, insurance claims, paychecks, Black Friday shopping spikes, credit card transactions, airline reservations, and more.

In sharp contrast to other platforms with *modular* designs, mainframes have been developed with a *total system, full stack view*, with hardware and software tightly coupled. This approach helps prioritize and optimize overall system performance and resilience to support large-scale, mission-critical applications. As a result, the processor units (PUs), memory, input/output (I/O), and network communications in mainframes are uniquely integrated, enabling them to process and scale on demand. Imagine an orchestra with all the instruments working together in perfect harmony and pitch. This is by design. Both an orchestra and the mainframe are examples of complex systems where harmony, precision and coordination are key to success. Their structured approaches allow them to perform intricate tasks seamlessly and efficiently. Today, the mainframe stack is built around digital services, agile application development, connectivity, and system management—resulting in an integrated platform with specialized hardware and advanced computing capabilities.

A unique feature of the mainframe architecture is the ability to isolate resources and define controls by partitioning the system into logical partitions (LPARs). LPARs are subsets of the processor hardware defined to operate independently with their own operating systems. With LPARs, users can design processes to use hardware resources—CPU, memory and I/O—in a way that boosts performance and reduces resource contention. In other words, users can architect a single mainframe to securely run multiple isolated workloads or applications at the same time so that the amount of available compute is comparable to a farm of smaller computers. On that single mainframe, users can connect and integrate many different workloads in a small footprint with significantly less cabling, networking and power consumption.

"One very relevant [mainframe] capability to financial services is the granular ability to manage, control and audit resources so you always know who is accessing what, and what is running under which authorization. This depth and span of granular control is a key benefit."

Suresh Chandrasekaran
Head of Payments and
Fintech Solutions, APAC
Fiserv

As mainframes have evolved, so has the architecture. Like precision-engineered time pieces and luxury cars, the mainframe integrates new technologies to evolve with the market but retains its core architecture. Every new generation of the mainframe has introduced progressively innovative features and capabilities to deliver more speed, security, agility and flexibility. 60 years ago, mainframes had up to 6 processors and 8 MB of memory. Today's mainframes have up to 200 processors and 40 TB of memory. The processors are significantly more advanced, featuring 8 cores with each clocking over 5.2 GHz. and supported by a 32 MB private level 2 cache. The level 2 caches combine to form a 256 MB virtual level 3 cache and a 2 GB level 4 cache. The result is 1.5 times more cache per core than the previous generation and, therefore, better per-thread performance and overall capacity.[13] This means, with their unconventionally large number of processors, mainframes can provide high processing power and low latency to help organizations process more data faster to consistently meet the growing, unpredictable digital demand in today's world. For example, leading global payments technology companies handle hundreds of million transactions every day and the total global daily average is around 1.9 billion transactions.[14] Mainframes—with their ability to process more than a million transactions per second—enable companies to process massive volumes of transactions without any delay to deliver quick and efficient payment services to their customers.

But it isn't just high processor density that enables modern mainframes to support superfast online and batch processes. It's also their ability to fully leverage processor capacities with low-level processor control and management, made possible through an assembler that allows users to have fine-grained

control over the hardware and resources. Mainframes also have better utilization rates than x86 servers. Mainframes typically run at 80-90% utilization compared to x86 that servers run at 20-30% utilization, enabling clients to maximize hardware investment while minimizing floor space and power costs.[15]

The next generation of processors are expected to continue a long history of generation-to-generation improvements. They are projected to clock in at 5.5 GHz. and include ten 36 MB level 2 caches.[16] They'll feature built-in low-latency data processing for accelerated I/O as well as a completely redesigned cache and chip-interconnection infrastructure for more on-chip cache and compute capacity. When multiple cores or processors access and modify shared data, it's crucial to maintain data consistency. The improved chip cache coherency will ensure that each core's cache reflects the most recent data to ease the implementation of highly efficient, strongly consistent transactional workloads.

Today's mainframes also have extensions and accelerators that integrate with the core systems. These specialized add-ons are designed to enable the adoption of technologies such as Java, cloud and AI by accelerating computing paradigms that are essential for high-volume, low-latency transaction processing.

The next crop of AI accelerators are expected to be significantly enhanced—with each accelerator designed to deliver 4 times more compute power, reaching 24 trillion operations per second (TOPS).[16] The I/O and cache improvements will enable even faster processing and analysis of large amounts of data and consolidation of workloads running across multiple servers, for savings in data center space and power costs. And the new accelerators will provide increased capacity to enable additional transaction clock time to perform enhanced in-transaction AI inferencing.

> "There are very few platforms out there that can offer hardware-assisted AI. Everybody thinks that GPUs are the only ways you can run AI, and that's hardly true."[9]
>
> **Ashish Nadkarni**
> Analyst at IDC

But simple statements of scalability, resiliency and security don't fully capture the elegance of the mainframe architecture. It's important to understand that the mainframe and its architecture are shaped by an enduring dedication to engineering to meet the needs of the modern-day enterprise through ongoing collaboration with IT leaders and professionals across industries.

03

Performance and scale

"For many organizations, there's no other platform that can give them what they need with regard to transaction performance, reliability and security."[17]

Peter Rutten
Research Vice President
Performance Intensive Computing
IDC

Performance: A need for speed, high availability and efficiency

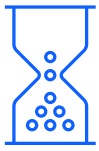
5.2 billion instructions per second.[13] 300 billion inference requests per day with just 1 millisecond of latency.[18] Up to 25 billion encrypted online transaction processing (OLTP) transactions per day.[13] The current generation of mainframes are designed for high-performance and speed beyond comprehension—shaped by organizations' need for high availability and resilience and engineered to maximize uptime. They can process colossal amounts of mission-critical data while maintaining an average of 99.999999% uptime, which translates to around 3.15 milliseconds downtime per year.[19]

Speed, data processing power and uptime are hallmarks of mainframe performance. Every second, millions of bank account transactions happen around the world. To ensure these and other time-sensitive transactions move through the system as quickly as possible, the mainframe architecture is uniquely designed for single thread performance with high-speed cores and pipelines. As data moves through any system, the tables where the data is stored and managed must be kept coherent through locking. For example, if a payment is being processed, the "payments" table with details such as transaction IDs and amounts needs to be locked while being updated to ensure transactional consistency. Long lock times can mean lower throughput and slower transactions, but the mainframe's distinct single thread design minimizes lock times to increase transaction speed.

Caches and the I/O subsystem also play key roles. In today's mainframes, the caches are big—8 to 12 times bigger than the alternatives in the industry—to help data get to the cores as quickly and efficiently as possible. Underneath the caches is a unique I/O subsystem of more than 300 I/O channels and dedicated processors for moving data.[13] This setup of large caches and I/O processors allocated exclusively for data movement is a differentiated feature of the mainframe. It allows the general-purpose processors to focus on running the business workloads—both data-intensive workloads, such as databases, and transactional workloads—without being bogged down by data retrieval or I/O operations, for maximum efficiency and speed. When the business need is to process massive amounts of data as quickly as possible, the mainframe's ability to offload these tasks to achieve high throughput and low latency is a real competitive differentiator.

5.2B

instructions per second



300B

inference requests per day with just 1 millisecond of latency

25B

billion encrypted OLTP transactions per day

Scale: A capacity for infinite scale

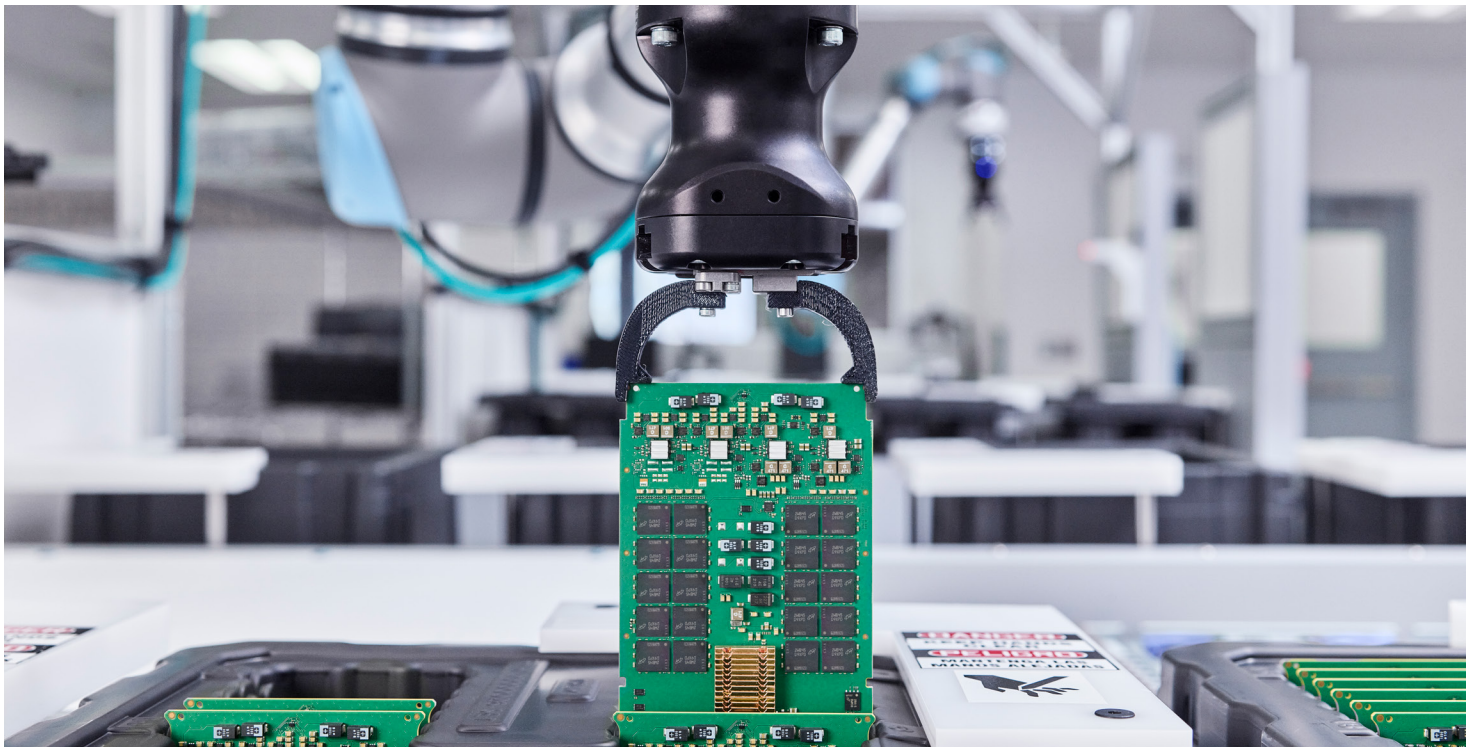
What makes a system scalable? Typically, if it can dynamically manage capacity to meet changing business needs, a system is considered scalable. Mainframes are designed for infinite scale, which means organizations can overcome peak season or unexpected spikes and maintain consistent levels of service with limited impact on reliability. The mainframe architecture delivers the ability to scale both vertically and horizontally and is uniquely engineered to add capacity without any disruption to critical processes.

This ability to manage variances in demand and volumes—and efficiently deliver a consistent quality of service regardless of the volume—is due to the multifaceted design of the mainframe. Dynamic partitioning and advanced virtualization capabilities provide vertical scalability, enabling the platform to scale up to handle more workloads on a single system without the need for additional hardware. At the same time, clustering and parallel processing help increase horizontal scalability, allowing mainframes to scale out to distribute workloads across multiple systems while functioning as a single logical system.

Other systems such as x86 servers add capacity through a scale out model, but mainframes are unique in their ability to scale up and grow in the same footprint. Not only does this capability help improve system performance and agility, it also helps reduce costs by eliminating the need for more hardware.

> "The interesting thing about mainframes from a scale perspective, it's not only the scaling, it's the ability to control the scaling."
>
> **Suresh Chandrasekaran**
> Head of Payments and
> Fintech Solutions, APAC
> Fiserv

With x86 servers, users typically have to reconfigure and then reboot the systems when they add new resources. Similarly, scaling on demand in the cloud means adding discrete servers connected through the network. This process is inefficient, adds complexity and impacts performance and resilience. It is particularly detrimental to workloads that require strong consistency—as opposed to eventual consistency—and transactional integrity. Today's mainframes are designed to let users add resources, be it memory, compute or I/O, without disruption. They help users optimize resource management without ever actually having to reboot the machine. More importantly, they help users scale seamlessly without affecting workload consistency and integrity.



04

Security and resilience

“From the very beginning, mainframes have taken a sustained, disciplined approach to the architecture and have applied it to resiliency and recoverability.”

Mark Anzani
Special Projects Executive, IBM Z
IBM Infrastructure

Security and resilience: What makes a platform securable?

Mainframes have long been associated with security and resilience. Some experts describe them as being fundamentally bulletproof.

Today's mainframes have security, resiliency and redundancy built in, from the next-generation chip to the operating system. They're inherently resistant to hacking and information theft. A key feature that sets mainframes apart is the ability to minimize vulnerabilities—whether they are caused by external factors or internal factors. Mainframes have multiple layers of protection and controls built into their hardware microcode to support process isolation and data integrity. Other features in the underlying architecture, such as virtualization, memory protection and access controls, also add to their unique security and resiliency capabilities.

According to the ITIC 2023 Global Server Hardware Security Survey, the leading mainframe in the market delivered the strongest server security, experiencing the fewest number of successful data breaches, the least amount of downtime due to security-related incidents and the fastest mean time to detection (MTTD). The survey also found that 97% of enterprises using the leading mainframes were able to detect, isolate and shut down attempted data breaches immediately to within the first 10 minutes.[20]

"The [mainframe] architecture is highly resilient. So, in a box, you're able to maintain failovers, you're able to maintain fail-safe mechanisms. You don't have to worry about it breaking down."

Suresh Chandrasekaran
Head of Payments and
Fintech Solutions, APAC
Fiserv

Even if the systems are hacked, mainframes are engineered to provide pervasive encryption capabilities for extensive encryption—of data in flight and at rest—and end-to-end data protection. Pervasive encryption is a hardware-based approach to system-wide encryption, unique to mainframes, that helps significantly reduce the risk of data exposure in the event of breaches. The actual encryption is performed by dedicated cryptographic cards, and therefore requires considerably less effort and is more cost effective than traditional encryption methods.

Modern mainframes also have quantum-safe functions to protect data from future cyberattacks initiated by quantum computers that will be able to break modern encryption. Specifically, they use post-quantum cryptographic algorithms, recently approved by the National Institute of Standards and Technology (NIST), to secure data against any potential quantum attack.[21]

In addition, mainframes feature flexible capacity capabilities to help proactively avoid disruptions from unplanned events and from planned scenarios such as site facility maintenance. They also have enhanced system recovery features for boosted processor capacity and parallelism to expedite CPU recovery. In fact, a large percentage of the mainframe system is dedicated to intelligent error handling and redundancy. It is earthquake tested up to 8.0 on the Richter scale and also put through extreme heat and cold testing which is critical in today's climate extremes.

05

Cost and risk

“The mainframe’s greatest strength is also its greatest weakness—its capital-intensive economies of scale.”

Roger Rogers
Executive IT Economics Consultant
IBM Infrastructure

Cost: The true worth of a mainframe

Mainframes are designed for high volume and the ability to process large workloads at the lowest possible cost. Which is why despite high upfront system costs, over their lifetime, mainframes can be more efficient and cost-effective than similarly configured distributed environments. Even though mainframes handle 72% of transaction workloads worldwide, they account for only 8% of total IT costs.[3]

There are many aspects to TCO, including sustainability and footprint. To understand the economy of mainframes, two points are key. First, we need to put the costs in the context of the business benefits they deliver. Second, we need to compare the costs with all contributing costs and long-term TCO of commodity servers.

Mainframes demonstrate compelling cost advantages when organizations process over 20,000 million transactions per second or exceed 250,000 transactions per second. At this scale, the cost per transaction drops to roughly USD 0.01 for mainframes compared to USD 0.04 for distributed systems. The economics become even more favorable as transaction volumes increase, with every additional 5,000 transactions per second improving ROI by approximately 25%. The mainframe's superior 99.999999% uptime helps prevent costly downtime of more than USD 100,000 per hour and its built-in security and compliance management capabilities can help save USD 500 thousand to USD 1 million annually in reduced risk. It helps achieve additional operational costs savings because the mainframe requires only 3 administrators compared to 15 for distributed systems. When we factor all these gains in, the TCO over a 5-year period can be less than half that of equivalent distributed systems.[22]

Mainframes are designed not only to achieve great economies of scale but also for significant system and data center energy efficiency with differentiated architectural elements.

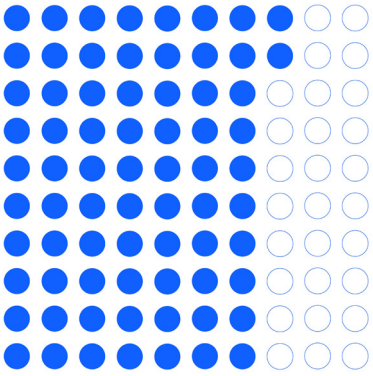
When running certain database workloads, leading mainframes require 16 times fewer cores than the compared x86 servers. If scaled up to a complete IT solution, it means that when running this workload, the mainframe system would be doing the work of about 2000 cores of the compared x86 servers.[23] That's remarkable not just from a cost and footprint perspective, but also from a sustainability and power point of view.

Similarly, consolidating Linux® workloads on mainframes instead of running them on compared x86 servers under similar conditions can reduce energy consumption by 75%. It can also reduce space by 50% and the CO2e footprint by over 850 metric tons annually.[24]

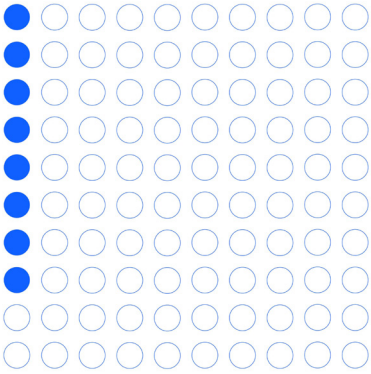
According to the ITIC 2023 Global Server Hardware, Server OS Reliability Report, servers that demonstrate five, six, seven, eight and even nine nines of reliability deliver the lowest TCO and the greatest economies of scale. The study listed today's leading mainframes as the most consistently reliable, secure and robust servers on the market, delivering the lowest TCO and fastest ROI among 18 different server hardware and server OS distributions.[25]

In addition, research from Rubin Worldwide, based on data from approximately 2,400 organizations across more than 20 sectors, shows that the mainframe has the highest level of cost-effective scalability among different technology asset classes. It scales easily and well since it's designed as an extendable platform. A mainframe can deliver 2 times scale growth, which can result in more than 60% unit cost reduction.[3]

Mainframes demonstrably deliver economies of scale with their speed, efficiency, reliability and workload consolidation, and organizations that are dependent on large-scale transaction processing, use data-intensive applications or require robust security find them to be cost-effective.



72% Nearly three-quarters of transaction workloads worldwide are handled by mainframes.



8% Yet, mainframes only account for a small percentage of the total IT costs.

Risk and auditability: No risky business

In conversations about IT and operations risks, IP licensing and code auditability seldom take center stage. But they play a critical role in helping maintain the security, reliability and performance of software and hardware systems and in managing compliance with licensing agreements and regulations.

Mainframe licensing can be very different from licensing for commodity servers and distributed systems, and there are many ways mainframes can increase the ability to be supported and maintained without exposure to risk. Mainframes mostly follow a centralized licensing model that simplifies software license management and reduces the risk of non-compliance.

This type of licensing method allows users to create multiple LPARs without having to buy separate licenses for each LPAR and enables them to run multiple applications and operating systems on a single physical machine.

Code auditability is critically important for organizations in highly regulated industries such as financial services, healthcare and government. Mainframes promote code auditability through several features including centralized control, comprehensive logs, source code management, automated code testing and OpenTelemetry (OTeL). Mainframes' ability to support OTeL—an open-source framework for instrumenting, generating, collecting and exporting telemetry data—is particularly useful during audits. Auditors can use OTeL to analyze software performance and identify potential vulnerabilities and other issues in the code more quickly and efficiently.

But perhaps the easiest—and the most overlooked—way mainframes enable code auditability is by supporting COBOL. Designed specifically for business data processing needs, COBOL or Common Business Oriented Language is a high-level, English-like, compiled programming language. Because the language is self-documenting and easy to understand, auditors can read and understand it and therefore readily verify whether applications match the business process they're implementing.

The built-in, native OS in today's leading mainframes requires no patching. Therefore, there's no risk of harmful code or vulnerabilities being inserted at the OS level, and unknown applications cannot be launched without specific permissions and memory allocations being set up first. COBOL's self-documenting feature complements these benefits and, as a result, helps mainframe users easily understand, maintain, and even secure the code.

Today, COBOL's imperative, procedural and object-oriented configuration serves as the foundation for more than 40% of all online banking systems.[26] COBOL handles decimal numbers efficiently by leveraging hardware-based decimal computing built into the mainframe processors. Organizations that can't compromise on clarity, transparency and auditability have much to gain from mainframes and COBOL.

06

AI, hybrid cloud and people

"The idea we can use AI in a manner that helps the operational characteristics of the machine and simplifies the operational characteristics of the machine. That's absolutely great."

Mark Anzani
Special Projects Executive, IBM Z
IBM Infrastructure

The future of AI: The whole world on an AI chip

Organizations run their most business-critical transactions on the mainframe, why not apply AI to these workloads at the hardware source? With the majority of enterprise data residing on them, mainframes are a perfect fit for integrated AI.

> "You can't do machine learning and cool AI stuff without good data. And if you have a mainframe, what kind of data is on that mainframe? The answer almost 100% of the time is: The most important data to my company — and it's there in spades."[27]

Brian Klingbeil
Chief Strategy Officer
Ensono

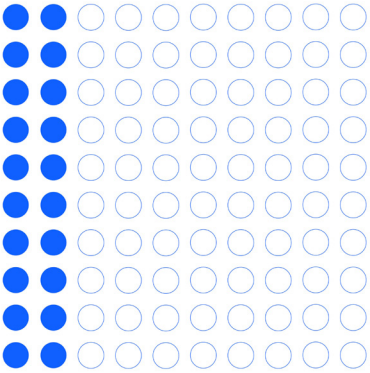
As we move further into the AI era, the importance of a fit-for-purpose architecture becomes clearer. AI on mainframes is more relevant and timely than ever. Today's mainframes are engineered to enable organizations to run AI models along with the most demanding enterprise workloads, without compromising on throughput or inferencing quality. With the focus steadily shifting from training to inferencing, especially in industries that prize real-time decision-making, mainframes can provide AI compute without the need for additional specialized AI hardware or software. Even with organizations increasingly using multimodel AI—which combines traditional AI models and large language models (LLMs)—for better, more accurate results, mainframes can deliver a highly niche AI infrastructure solution unlike commodity servers.

Mainframes today are built to have powerful AI capabilities. They have advanced microprocessors with on-chip AI acceleration that can help run AI inferencing at the speed of a transaction. The on-chip AI scoring logic makes submicrosecond AI inferencing possible, enabling the accelerator to scale up to handle hundreds of billions of inference requests per day. This means organizations can apply AI to transaction workloads on the mainframe to predict product sales and detect payment fraud. When one US bank's credit scoring model was deployed to an on-premises cloud platform, only 20% of the credit card transactions could be scored. When deployed to a mainframe system, 100% of the transactions—15,000 transactions per second—could be scored in real time. Moving to the mainframe resulted in significant fraud reduction and helped the bank save over USD 20 million in annual fraud prevention spend.[11]

The next generation of the accelerator architecture is expected to be more efficient for AI tasks. Unlike standard CPUs, the chip architecture will have a simpler layout, designed to send data directly from one compute engine, and use a range of lower-precision numeric formats. These enhancements are expected to make running AI models more energy efficient and far less memory intensive. As a result, mainframe users can leverage much more complex AI models and perform AI inferencing at a greater scale than is possible today.

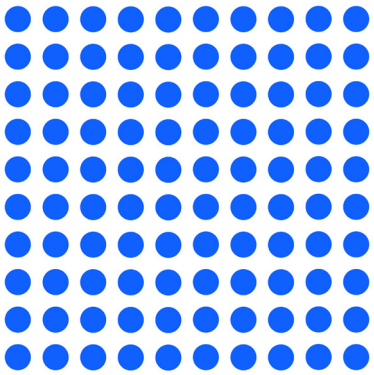
Integrating AI into the mainframe isn't just good for the business, it also helps the platform—transforming the way organizations manage and maintain their systems. It helps mainframe teams enable an intelligent system and make use of integrated AIOps to automate system management. It also helps them take advantage of predictive analytics, self-healing and self-tuning capabilities to optimize their workflows, reduce manual errors and improve overall efficiency. And bringing AI to the data is easier, less risky and less expensive than moving the data and the attendant security, reliability and performance protections to the AI tool for just one type of application workload.

A US bank's credit scoring model deployed on an on-premises cloud platform versus a mainframe system



20%

Only 20% of credit card transactions scored on the on-premises cloud platform.



100%

100% of credit card transactions scored when deployed to the mainframe system.

Hybrid by design: The future of mainframe is hybrid

The mainframe provides hybrid capabilities whether organizations want to manage it as self-contained private cloud or integrate it seamlessly with other cloud resources and services. According to a recent IBM Institute for Business Value report, 91% of surveyed executives agree that their organization promotes a hybrid architecture and 81% consider it important the mainframe easily integrates with other technologies.[11]

A hybrid by design approach, in particular, can help organizations integrate their entire technology estate, including cloud, mainframes and more, to achieve the agility, speed and scale needed to drive business value. Because mainframes house organizations' most critical data, they're central to the hybrid by design approach. This approach allows organizations to take advantage of all the speed, security and reliability benefits of the mainframe platform while also enjoying the flexibility and scalability of public cloud solutions. The IBM Institute for Business Value report shows that organizations that adopt the hybrid by design approach and innovate on the mainframe outperform their competitors. Over the past 3 years, these organizations have experienced higher levels of revenue growth, profitability and innovation than their peers.[11]

Today's mainframes are built for frictionless hybrid cloud integration and offer many capabilities to modernize core applications so they can seamlessly interact with cloud and other modern technologies. A digital integration hub connects core business applications running on the mainframes with hybrid cloud applications. This capability enables flow of real-time data at scale, protects production environments from unpredictable inquiry traffic and provides flexible interactions with APIs—helping organizations accelerate modernization on mainframes.

Mainframes have long stepped out of the shadow of their legacy origins; they are now more ready than ever to help organizations achieve greater agility, efficiency and cost savings through hybrid cloud, AI and other new innovations.

"Mainframes continue to occupy a central role in the hybrid world and are evolving to serve new use cases, with AI and security increasingly influencing modernization plans."[28]

Petra Goude
Global Practice Leader,
Core Enterprise and zCloud
Kyndryl

People: The minds that make the mainframe

When it comes to the mainframe, it's hard to separate the people from the platform. IBM® System/360 wouldn't have been possible without the contributions of many pioneers. Similarly, the later incarnations of the mainframe were possible because the scientists and the engineers continued to drive improvements even as the computing world changed around them.

It's clear that the engineers who oversee the development and creation of the mainframe play a huge role in the continued evolution and development of the platform. And it's this culture of continuous innovation and improvement and the commitment to delivering industry-first solutions to the evolving, unique needs of enterprise clients that has kept the mainframe at the forefront for decades.

Mainframes are the result of a long tradition of innovation, collaboration and community. Few other platforms today can make that claim. And while the community can often be considered closed off, mainframe engineers and developers are unique in the industry. They've been instrumental in shaping the mainframe into the platform it is today, developing new capabilities, improving existing ones and addressing technical challenges to ensure it remains valuable to organizations that require the highest levels of security, resilience and scale. Today, keeping up with the AI zeitgeist, they are working on open-source AI solution templates to help run AI models and LLMs on the mainframe.

A mainframe frame of mind

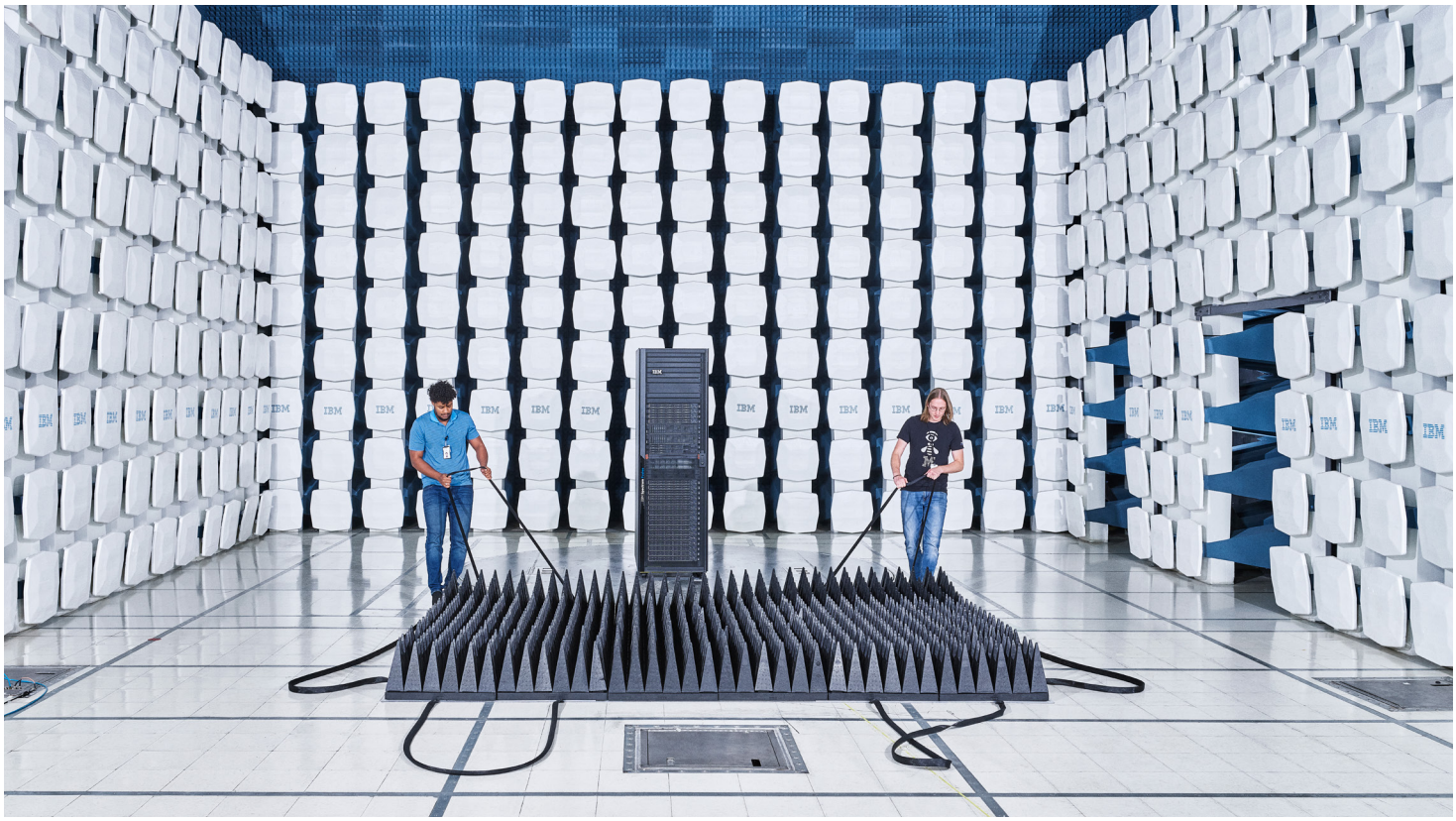
"There are people who are stuck in a picture of the mainframe that hasn't been true for 50 years."

Roger Rogers
Executive IT Economics Consultant
IBM Infrastructure

The story of the mainframe is one of remarkable endurance. Staying relevant in one of the fastest-changing industries in the world for more than 60 years is no mean feat; it's a testament to the power of timely evolution and timeless engineering.

The mainframe today is an open, secure, resilient and adaptable platform. But even as it has changed, it has stayed true to its essential architecture. Whenever new capabilities are implemented, the mainframe architecture is extended rather than replaced.

With the mainframe playing a bigger role in the AI story, we'll see it evolving more to accommodate and accelerate newer innovations. And this architecture-first strategy will continue to help sustain the compatibility, integrity and longevity of the platform in the future.



Take the next step

Connect with an IBM mainframe expert.
If you're ready to talk with an IBM mainframe expert or learn how IBM can guide you to drive AI and hybrid cloud transformation with the mainframe, schedule a consultation.

For more information, contact your IBM Business Partner:

Global IT & Recovery Services, LLC.

1. Amazon Statistics: Key Numbers and Fun Facts
2. Mainframes: The next 60 Years,
The Open Mainframe Project, April 2024.
3. An Economic Model for Optimizing Technology Investments,
Rubin Worldwide, May 2023.
4. Why Is Mainframe Still Relevant and Thriving in 2022?
Planet Mainframe, 20 December 2022.
5. Top Ten Payments Companies Processed $9 Trillion in 2022 Payment Card Volume, GlobeNewswire,
29 March 2023.
6. The world's 100 largest banks, 2022, S&P Global,
11 April 2022.
7. Announcing IBM z16: Real-time AI for Transaction Processing at Scale and Industry's First Quantum-Safe System, IBM Newsroom,
5 April 2022.
8. Mainframe Market Size, Share & Trends Analysis Report By Type (Z Systems, GS Series), By End-User (BFSI, IT and Telecom, Government and Public Sector, Retail and E-commerce, Healthcare, Travel and Transportation, Manufacturing) and By Region (North America, Europe, APAC, Middle East and Africa, LATAM) Forecasts, 2024–2032, Straits Research,
7 June 2024.
9. AI on the mainframe? IBM may be onto something,
CIO.com, 3 October 2024.
10. The State Of Mainframes, Global, 2024,
Forrester Research, 21 March 2024.
11. Mainframes as mainstays of digital transformation,
IBM Institute for Business Value, 8 October 2024.
12. What is a mainframe?
IBM, 1 March 2024.
13. IBM® z16® Technical Introduction,
IBM Redbooks®, April 2023.
14. Number of Credit Card Transactions per Second, Day & Year,
CapitalOne Shopping Research, 30 May 2024.
15. Chapter 20: Data Center IT Efficiency Measures. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures; Period of Performance: September 2011 – December 2014
16. Projections for Telum II, based on pre-release hardware measurements and configurations.
17. Why Mainframes Remain Essential for Many Organizations, BizTech, 12 June 2023.
IBM Newsroom, 5 April 2022.
18. Announcing IBM z16: Real-time AI for Transaction Processing at Scale and Industry's First Quantum-Safe System,
IBM Newsroom, 5 April 2022.
19. IBM z16 and Power10 Deliver Highest Reliability Among Mainstream Servers for 15th Consecutive Year,
TechChannel, 5 July 2023.
20. ITIC 2023 Reliability Survey IBM Z Results, Information Technology Intelligence Consulting,
August 2023.
21. NIST's post-quantum cryptography standards are here,
13 August 2024.
22. Calculations and assumptions based on external as well as internal IBM data.

DISCLAIMER
Assumptions
Initial Investment Comparison (5-year period) Mainframe: Hardware/Software-USD 5M, Implementation: USD 1M, Initial Training: USD 200K Total: 6.2M; Distributed Systems: Hardware (multiple servers): USD 3M, Software licenses: USD 2M, Implementation: USD 500K, Initial Training: USD 300K Total: USD 5.8M Annual Operating Costs Mainframe: Staff (3 FTEs): USD 450K, Power/Cooling: USD 100K, Maintenance: USD 300K, Software maintenance: USD 400K Annual Total: USD 1.25M; Distributed Systems: Staff (15 FTEs): USD 2.25M, Power/Cooling: USD 400K, Maintenance: USD 500K, Software maintenance: USD 800K Annual Total: USD 3.95M

Formulas and calculations
Base formula for Total Cost of Ownership (TCO) over 5 years is TCO = Initial Investment + (Annual Operating Costs × 5) - (Consolidation Savings × 5) - (Reliability Factor × Downtime Costs × 5); Mainframe Calculation: USD 12.45M = USD 6.2M + (USD 1.25M × 5) - (Consolidation Savings) - (0.99999 × Downtime Prevention Value); Distributed Calculation: USD 25.55M = USD 5.8M + (USD 3.95M × 5) - (0.999 × Downtime Prevention Value); Cost Per Transaction (CPT): CPT = TCO ÷ (Transactions per second × Seconds per year × Years), Mainframe Calculation: USD 0.01 = USD 12.45M ÷ (250,000 × 31.5M × 5), Distributed Calculation: USD 0.04 = USD 25.55M ÷ (250,000 × 31.5M × 5); ROI Improvement Formula: ROI Increase = (Additional MIPS Cost Savings ÷ Base TCO) × 100 = (~USD 500K savings per 5,000 MIPS ÷ USD 12.45M) × 100 ≈ 25%.

Additional Notes
When transaction volume exceeds 20,000 MIPS, Cost per transaction drops to USD 0.01 vs USD 0.04 for distributed, ROI increases by approximately 25% per additional 5,000 MIPS, when security requirements demand real-time encryption, pervasive encryption capabilities, compliance monitoring added value: USD 500K-USD 1M annually in reduced risk/compliance costs, when downtime costs exceed, USD 100K per hour, requiring 99.999% uptime.

23. IBM internal tests show that when running IBM WebSphere® and IBM Db2® workloads, IBM z16 requires 16 times fewer cores than the compared x86 servers. If you scale this test up to a complete IT solution, it means that when running this workload, the IBM z16 Max125 would be doing the work of about 2,000 cores of the compared x86 servers.

DISCLAIMER:
This study is an IBM internal study that's designed to replicate a typical IBM client workload use in the marketplace. Results might vary. The core consolidation study targeted a comparison of the following servers: IBM Machine Type 3931 Max125 system that consisted of 3 CPC drawers containing 125 configurable Integrated Facility for Linux (IFL) processor units and 2 I/O drawers to support both network and external storage. Lenovo ThinkSystem SR650 2-rack units (2U) with 2 2nd Gen Intel Xeon Platinum 2.1 GHz processors with 16 cores per CPU. Both solutions had access to the same storage array. The workload consisted of a transactional application running on IBM WebSphere Application Server and IBM Db2 simulating core online banking functions. The actual test results were extrapolated to the stated above x86 servers by using IDC Qualified Performance Indicator (QPI) metrics and IBM sizing methodology by using the following assumptions on a typical IT environment of a banking client using x86 servers. The production IT environment has 16 x86 servers running at 50% average utilization. There are 48 x86 servers in the nonproduction IT environments: development (4 environments with 2 servers each, 8 servers total), development test environment (4 servers), system integration test environment (8 servers), performance test environment (16 servers), user acceptance test environment (4 servers) and production fix test environment (8 servers). A typical average CPU utilization is 7% across all nonproduction environments. An equivalent IBM Machine Type 3931 solution requires a single Max125 server running at 85% average utilization across all IT environments that are separated by using LPAR technology.

24. Consolidating Linux workloads on 5 IBM z16 systems instead of running them on compared x86 servers under similar conditions can reduce energy consumption by 75%, space by 50% and the CO2e footprint by over 850 metric tons annually.

DISCLAIMER:
We compared 5 IBM Machine Type 3931 Max125 models that consisted of 3 CPC drawers containing 125 configurable cores (CPs, zIIPs or IFLs) and two I/O drawers to support both network and external storage versus 192 x86 systems with a total of 10,364 cores. IBM Machine Type 3931 power consumption was based on inputs to the IBM Machine Type 3931 IBM Power® Estimation Tool for a memo configuration. Power consumption of the x86 servers was based on March 2022 IDC QPI power values for 7 Cascade Lake and 5 Ice Lake server models, with 32 to 112 cores per server. All compared x86 servers were 2-socket or 4-socket servers. IBM Z and x86 severs are running 24x7x365 with production and nonproduction workloads. Savings assumes a power usage effectiveness (PUE) ratio of 1.57 to calculate more power for data center cooling. PUE is based on the Uptime Institute Global Data Center Survey 2021. CO2e and other equivalencies that are based on the EPA greenhouse gas (GHG) calculator use US national weighted averages. Results might vary based on client-specific usage and location.

25. ITIC 2023 Global Server Hardware, Server OS Reliability Report,
Information Technology Intelligence Consulting, August/September 2023.
26. What is COBOL?
IBM, 19 April 2024.
27. An unlikely hero is running generative AI workloads: the mainframe,
CIODIVE, 17 September 2024.
28. Kyndryl survey reveals 86% of enterprises are moving fast to adopt AI to accelerate mainframe modernization,
Kyndryl press release, 10 September 2024.

This document is current as of the initial date of publication and may be changed by IBM at any time.

Not all offerings are available in every country in which IBM operates.

Examples presented as illustrative only. Actual results will vary based on client configurations and conditions and, therefore, generally expected results cannot be provided.

It is the user's responsibility to verify the operation of any non-IBM products or programs with IBM products and programs. IBM is not responsible for non-IBM products and programs.

No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.