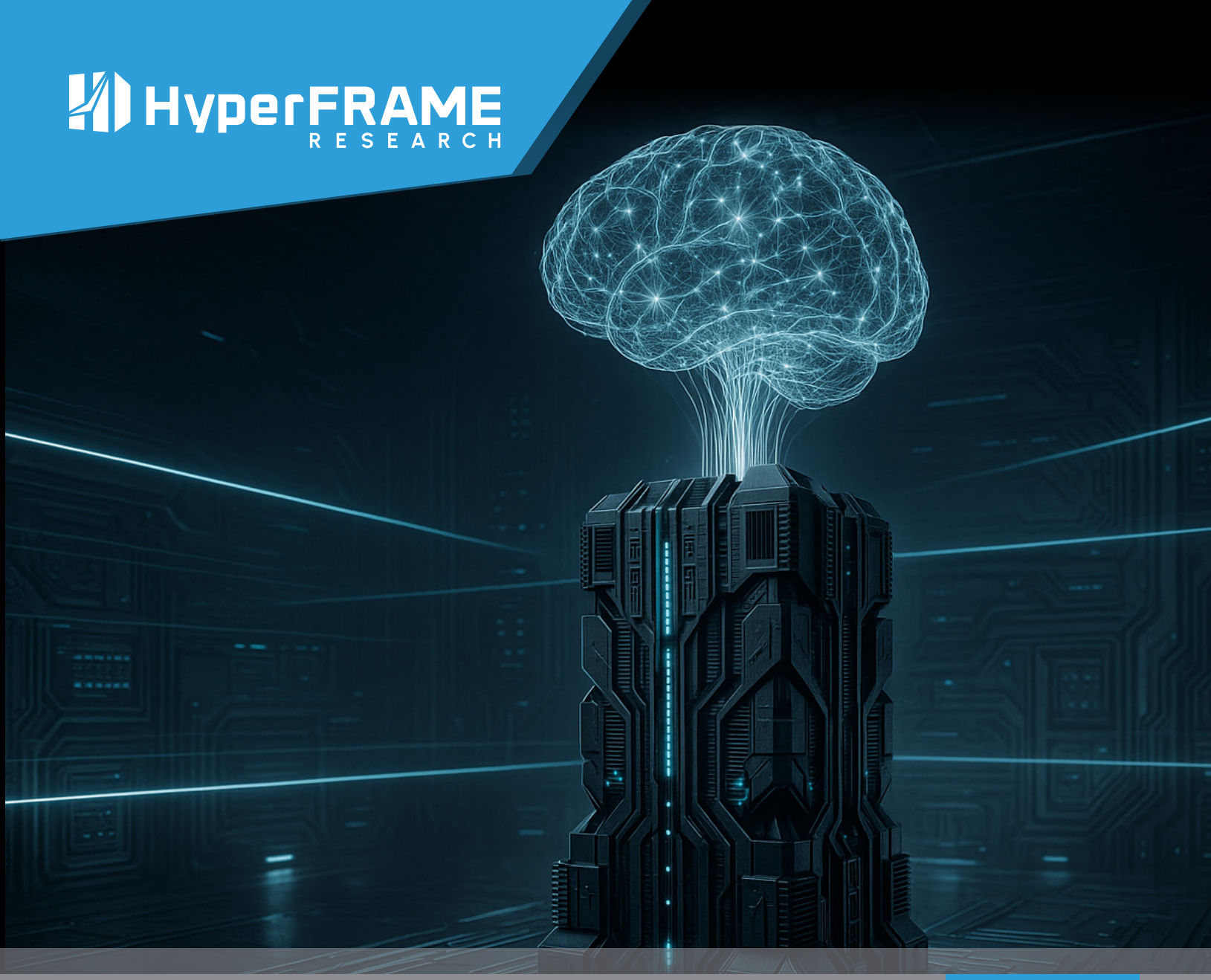
HyperFRAME RESEARCH



WHITE PAPER

z17 - The First AI-Centric Mainframe

IBM launches a new mainframe, designed from the ground up to be a an AI powerhouse.

Authors:

Steven Dickens
CEO and Principal Analyst

APRIL 2025

Executive Overview

We have had to wait longer than the typical mainframe refresh cycle (8-10 quarters) for the launch of the IBM z17. The IBM z16 was launched on the 5th April 2022 and has, according to the details shared at an Analyst pre-briefing a couple of weeks ago, been the most successful of the post-CMOS mainframe cycles. IBM only shares capacity growth by Millions of Instructions Per Second (MIPS) rather than systems or processors shipped. The current Telum II has 2,500 MIPS per processor for some context.

The current generation marks the 17th generation of CMOS systems since IBM moved from the previous bipolar systems in 1994. Given the innovation in the Telum II processor and accompanying Sypre PCIe-attached accelerator this extended cycle time is to be fully expected, and will no doubt lead to increased revenue in the early quarters of the new z17 cycle, more on that later.

The IBM z17 announcement signals a significant push towards integrating AI deeper into enterprise infrastructure. IBM's claim that the z17 is designed to process 50% more AI inference operations per day compared to the z16 is ambitious, but given my knowledge of IBM's approach to the system claims, is to be believed. The introduction of multi-model AI capabilities suggests a move towards handling more complex AI workloads, potentially opening up new use cases for mainframe technology.

The IBM z17 is architected to deliver improved AI inferencing capabilities, powered by a second-generation on-chip AI accelerator built into the IBM Telum II processor. Steven Dickens, our CEO and Principal Analyst wrote about the Telum II at its announcement back at Hot Chips last year. Suffice to say it's a single-thread powerhouse, perfectly suited to transactional workloads. If AMD, Broadcom or Intel were to launch a 3Nm processor boasting a clock speed of 5.5Ghz, the internet would break with excitement and wonderment, but this is IBM and the mainframe, so analysts and a few hardcore fans like me will give it the kudos the IBM in-house, largely US based chip design deserves.

The system aims to enable more than 450 billion inferencing operations in a day with one millisecond response time. The IBM Spyre Accelerator, expected in 4Q 2025, aims to provide additional AI compute capabilities. At the analyst briefing a few weeks back, IBM was at pains to stress it has over 200+ use cases with early ship clients who are leveraging the AI capabilities of the box, and IBM went further to stress these use cases extend way beyond the obvious fraud mitigation example.

The system also includes new security features and tools that leverage AI for improved system usability and management. While continuing on a long-established trend, the continued focus by IBM on post-quantum encryption is to be applauded. HyperFRAME Research has published on IBM Z Operations Unite so check out that whitepaper here for detailed perspective, suffice to say there is much to like with the combination of IBM's rich heritage with the OMEGAMON range of tools, combined with Open Telemetry and AIOps functionality.

The IBM z17: A New Era of Mainframe Innovation

The IBM z17, powered by the Telum II processor and complemented by the Spyre AI Accelerator, represents a significant leap forward for the IBM Z family. Unveiled at the Hotchips 2024 Conference, the Telum II processor is a marvel of modern engineering, boasting 43 billion transistors across 24 miles of wire, fabricated using Samsung's 5nm High-Performance Process (HPP). With eight 5.5 GHz cores per processor and a robust 36 MB L2 cache (a 40% increase over the z16), the Telum II delivers exceptional single-thread performance, up 11% compared to its predecessor, while supporting up to 2,500 MIPS per processor.

The Spyre Accelerator, a PCIe-connected powerhouse designed for enterprise-grade AI workloads, enhances this further. Due to start shipping in Q4 2025, Spyre offers 26 billion transistors, 32 accelerator cores (distinct from processor cores), and 128 GB of LPDDR5 memory per card, scaling up to 1 TB across eight cards in a logical cluster. With an aggregate memory bandwidth of 1.6 TB/s, Spyre will be purpose-built for multiple-model AI, offering versatile processing with support for int4, int8, fp8, and fp16 datatypes. For technical buyers, Spyre's projected 75W per card power consumption underscores its efficiency, balancing performance with sustainability.

Performance and Scalability: Breaking Down the Numbers

The IBM z17 is engineered to deliver measurable improvements over the z16, catering to the demands of modern workloads. Let's dive into the specifics:

Box Details

- Frames: 1-4, offering flexibility for varying data center footprints.
- Customer Cores: Up to 208, providing 15-20% capacity growth over the z16.
- Single-Thread Performance: 11% improvement over z16, driven by the Telum II's advanced architecture.
- Memory: Up to 64 TB (60% more than z16), ensuring ample headroom for memory-intensive applications.
- Power Efficiency: 17-27% less power consumption than z16, with a 35 kW envelope, balancing performance with sustainability.

Telum II Processor

- Samsung HPP process
- 43 billion transistors
- 24 miles of wire
- 18 layer wafer
- 8x 5.5Ghz cores per processor
- On-chip Data Processing Unit (DPU)
- 10x 36MB L2 cache
- 40% L2 Cache growth over z16

Spyre - PCIe connected

- 26 billion transistors
- 14 miles of wire
- 32 cores
- 2MB scratchpad
- PCIe gen5 x16 adapter
- 128GB of LPDDR5 memory
- 8 cards in I/O drawer form a logical cluster
- 1TB of memory
- 1.6TB per second aggregate memory bandwidth

Configuration Flexibility

The z17's configurability is a boon for technical buyers seeking tailored solutions. Here's a breakdown of its feature options:

Feature	CPC Drawers	CPs	IFLs	Unassigned IFLs	zIIPs	ICFs	IFPs	Std SAPs	Spare PUs
Max43	1	0 - 43	0 - 43	0 - 42	0 - 42	0 - 43	2	5	2
Max90	2	0 - 90	0 - 90	0 - 89	0 - 89	0 - 90	10	(N/A)	2
Max136	3	0 - 136	0 - 136	0 - 135	0 - 135	0 - 136	16	(N/A)	2
Max183	4	0 - 183	0 - 183	0 - 182	0 - 182	0 - 183	21	(N/A)	2
Max208	4	0 - 208	0 - 208	0 - 207	0 - 207	0 - 208	24	(N/A)	2

These configurations support a range of processor types—CPs, IFLs, zIIPs, ICFs, and IFPs—allowing buyers to optimize for specific workloads, such as Linux-based applications (IFLs) or specialty tasks (zIIPs).

AI at Scale: The Telum II and Spyre Advantage

The standout feature of the z17 lies in its AI capabilities, driven by the Telum II processor and the forthcoming Spyre Accelerator. According to the analyst briefing prior to the launch, IBM is working with clients on over 250 AI use cases, leveraging Telum II's ability to process complex models with low latency. Telum II delivers up to 450 billion inference operations per day with a 1ms response time, enabling real-time AI scoring on 100% of transactions—a capability also present in the z16. However, unlike the z16, the z17 introduces large language models (LLMs) alongside deep learning (DL) and machine learning (ML) models while maintaining this low latency, marking a significant advancement.

Spyre, when available, will further enhance this by integrating directly into the mainframe via PCIe, offering dedicated accelerator cores for AI workloads. While specific energy efficiency claims are under review, Spyre's projected efficiency and the Telum II's on-chip AI accelerator make the z17 a compelling platform for latency-sensitive applications like fraud detection or real-time analytics.

The z17's future GenAI capabilities, underpinned by IBM's Reliability, Availability, and Serviceability (RAS) framework, will help ensure that AI-driven insights are delivered with the uptime and security enterprises demand. Whether optimizing decision-making or integrating AI with mission-critical transactions, the z17 offers a robust foundation for digital transformation.

Security and Resilience: Quantum-Safe and Beyond

Security remains a hallmark of the IBM Z platform, and the z17 takes it to new heights with quantum-safe cryptography. Technical buyers will appreciate the immediate availability of quantum-safe APIs and crypto discovery tools, which protect data against current

and future threats, including “harvest now, decrypt later” attacks. The z17’s quantum-safe capabilities include:

- **Key generation and encryption.**
- **Key encapsulation mechanisms.**
- **Hybrid key exchange and dual digital signature schemes.**

These features, combined with tools like IBM Application Discovery and Delivery Intelligence (ADDI) and Integrated Cryptographic Service Facility (ICSF), enable early threat detection and simplified compliance—crucial for industries like finance and healthcare. On the resilience front, the z17 introduces IBM Z Flexible Capacity for Cyber Resiliency, allowing capacity shifts between systems without onsite intervention, ensuring business continuity during unplanned events or maintenance.

Power and Efficiency: A Greener Mainframe

With a 35 kW power envelope and 17-27% less power consumption than the z16, the z17 aligns with sustainability goals—a growing priority for eco-conscious buyers. Spyre’s projected 75W per card efficiency will further enhance this, making the z17 a compelling choice for organizations balancing performance with environmental impact.

Beyond the Box – The Software Surrounding The Mainframe

The mainframe is experiencing a significant transformation through modern software practices and AI-fueled tooling. Open-source tools, containerization, and API integrations are actively reshaping its ecosystem, promoting agility and innovation. Software development on the mainframe is evolving from an isolated practice to an integral part of interconnected enterprise architectures. The increasing availability of cloud-native technologies for z/OS is empowering developers to build and deploy applications with unprecedented speed and flexibility. This software-driven revolution is critical for the mainframe to maintain its position as a vital component of modern enterprise computing, ensuring its continued relevance in a dynamic technological environment.

While the focus of the z17 launch was largely on the box, and its undoubtedly AI focused credentials, IBM did also take the opportunity of the Z launch to drop some new software capabilities.

IBM Z Operations Unite: Enhancing Mainframe Management with AI and OpenTelemetry

IBM’s unveiling of IBM Z Operations Unite marks a notable evolution in mainframe operations for its IBM Z ecosystem. This solution aggregates performance metrics and logs from disparate IBM Z sources into a standardized OpenTelemetry format, employing AI to improve operational efficiency. Z operations Unite targets faster anomaly detection, precise incident isolation, and reduced resolution times, all of which are key priorities for enterprises managing high-stakes workloads.

The use of OpenTelemetry, an open-source observability framework, underpins this offering by unifying telemetry data—metrics, logs, and traces—into a single stream. This enables AI-driven analysis, allowing operations teams to identify and mitigate irregularities proactively. Such capabilities could lower downtime risks, a pressing concern for sectors like finance and healthcare reliant on IBM Z for transactional processing. When integrated with IBM Concert, an enterprise-wide platform, Z Operations Unite extends its scope, correlating mainframe data with hybrid environments to provide a comprehensive view of system performance. This reflects IBM’s strategy to infuse AI into IT operations, complementing the z17 mainframe’s hardware advancements, including the Telum II processor and forthcoming Spyre Accelerator.

While the solution emphasizes efficiency, concrete performance metrics or case studies remain undisclosed, leaving its real-world impact unquantified. Adoption of OpenTelemetry aligns IBM with open standards, potentially easing integration, though smaller enterprises may face resource constraints. Z Operations Unite represents a calculated effort to modernize mainframe management, merging legacy strengths with AI-driven insights to address contemporary operational demands.

IBM also announced plans to enhance Z Backup Resiliency through integrations with DFSMS Cloud Data Manager and Threat Detection for z/OS. The DFSMS integration will enable the visualization and reporting of dataset copies stored in cloud object storage via SMF record ingestion. Functionality will include JCL generation for restoring cloud data to z/OS. The Threat Detection integration aims to

display and filter security incident records logged by TDz within the Backup Resiliency interface. User activity visualization, associated with detected anomalies, will be provided for investigative purposes. The system will also facilitate recovery point analysis based on TDz records to expedite data recovery processes.

The Mainframe Software Ecosystem

BMC Highlights Momentum with AMI Solutions

To coincide with the z17 launch BMC Software published a blog post dissecting IBM's latest mainframe milestone. BMC Software's response to IBM's z17 launch, articulated in a blog by John McKenny, Senior Vice President of Intelligent Z Optimization and Transformation, frames the mainframe as a pivotal enterprise asset. The z17 integrates AI for real-time transactional insights, targeting over 250 use cases across industries like finance and retail. This positions it as an evolution of mainframe capabilities, blending reliability with modern AI requirements.

BMC underscores its Automated Mainframe Intelligence (AMI) portfolio—tools like AMI DevX and AMI Ops—designed to enhance developer productivity and operational performance in tandem with z17's architecture. Early collaboration with IBM ensured compatibility from launch, reflecting strategic alignment. The blog highlights the mainframe's relevance amid hybrid cloud and generative AI trends, supported by BMC's 19th annual Mainframe Survey, though it lacks detailed findings to substantiate this optimism. Challenges such as implementation costs or technical complexity are glossed over, favoring a narrative of seamless progress. BMC's analysis serves dual purposes: affirming IBM's innovation and positioning its tools as integral to z17's ecosystem, targeting enterprises committed to mainframe longevity.

Broadcom Champions IBM z17 with Day One Support and Strategic Alignment

On launch day of the z17 Broadcom released a detailed statement celebrating IBM's z17 mainframe launch, emphasizing its proactive preparation and software readiness. Broadcom's statement on the z17 launch emphasizes its proactive preparation, with its mainframe software portfolio fully tested for compatibility prior to release. This "Day One" readiness aims to accelerate modernization, resilience, compliance, and ROI for enterprises leveraging z17's AI and hybrid cloud features. Solutions spanning automation, cybersecurity, data management, and DevOps integrate with the platform, validated through testing in Broadcom's z17-equipped labs and close IBM collaboration.

The z17 is presented as a transformative system, with AI enabling real-time insights and operational streamlining. Broadcom complements this with customer-focused offerings—expert guidance, flexible licensing, and resource optimization—while addressing the skills gap through initiatives like the Vitality Residency and Mainframe Skills Council. This dual emphasis on technology and talent development positions Broadcom as a comprehensive partner. However, the statement avoids discussing potential adoption barriers, such as cost or complexity, projecting an optimistic view of mainframe evolution. Broadcom's commitment extends beyond technical integration, aiming to sustain enterprise reliance on IBM Z infrastructure.

Rocket Software Embraces IBM z17 for Mainframe Modernization

Rocket Software penned a succinct yet impactful blog post celebrating IBM's z17 mainframe launch, framing it as a catalyst for a new era in enterprise computing. Rocket Software's blog post on the z17 launch highlights the mainframe's critical role, managing 75% of enterprise data and over $3 trillion in daily commerce across systems like payroll and banking. The z17's AI and hybrid cloud capabilities, powered by the Telum II processor, reinforce this foundation. Rocket positions itself as a key enabler, offering tools to enhance security, hybrid IT, and automation without operational upheaval.

Its security solutions provide advanced threat detection and data recovery, while DataEdge facilitates real-time data access and Gen AI-driven interfaces for cross-platform utility. Automation tools incorporate predictive analytics to optimize workloads and preempt issues. Rocket's 18 IBM Champions lend technical credibility to its claims of seamless z17 integration. The narrative prioritizes continuity and innovation, though it omits discussion of costs or integration challenges, maintaining a streamlined modernization focus. For enterprises tied to IBM Z, Rocket offers a practical, if underspecified, path forward, aligning legacy systems with future demands.

What Does the New z17 Mean for IBM and the Industry as a Whole

This announcement needs to be seen in the context of the evolving AI landscape, the mainframe installed base, and IBM as a whole. Firstly, while cloud-based AI solutions have gained significant traction, IBM is rightly doubling down on on-premises AI processing, aiming to deliver low latency and high throughput for critical applications. I fully expect to see big banks, telcos, retailers and government departments leveraging the AI capabilities of the Telum II and Spyre DPU (when it ships later in the year) to infuse transactional flows with AI. Secondly, the mainframe installed base. With this system coming 12 quarters after the z16, the upgrade financials will be more compelling for many of the installed base as they will have gone all the way through their leases, so IBM's sales teams won't need to perform the usual gymnastics to entice customers to upgrade.

This leads to the final point, the impact on IBM of the z17. The z17 hardware will drag 3-4x the revenue for the system, which HyperFRAME Research estimates to be in the $4bn range annually. The 3-4x drag will manifest in the likes of DS8000 storage, high-margin transactional software such as CICS and Db2 and then support services and IBM Consulting. All told the long-stack revenue associated, will by my estimates, be in the region of $15bn, with the vast majority of it being at high margins. The IBM z17 matters to IBM revenue and margins.

With IBM already posting robust growth in the low single digit range, the next four quarters will have the tailwind of the mainframe refresh cycle, with the majority of clients most likely to upgrade in the first year of the cycle. Put simply, expect Arvind Krishna and Jim Kavanaugh to be talking about the mainframe refresh cycle in the coming quarters.

Looking Ahead

The launch of the IBM z17 on April 8, 2025, represents a defining moment for the mainframe's role in enterprise computing, reinforcing its position as a linchpin for mission-critical workloads in an era dominated by artificial intelligence (AI), hybrid cloud architectures, and quantum-safe security imperatives. This whitepaper has explored the z17's technical advancements, strategic ecosystem enhancements, and the broader market dynamics shaping its trajectory. As HyperFRAME Research's analysis underscores, the z17's success will hinge on its ability to deliver measurable outcomes, bridge legacy and modern technological paradigms, and address practical challenges in adoption, integration, and skill development. This conclusion synthesizes these insights, offering a forward-looking perspective on the z17's potential to sustain the mainframe's relevance in a rapidly evolving digital landscape.

A critical trend to monitor is the adoption rate of the z17, particularly the Spyre Accelerator's performance beyond its established fraud detection use case. Set to ship in Q4 2025, the Spyre, with its 26 billion transistors, 32 accelerator cores, and 1.6 TB/s aggregate memory bandwidth, is poised to redefine enterprise-grade AI. Its support for diverse datatypes (int4, int8, fp8, fp16) and projected 75W per card efficiency make it a versatile platform for real-time analytics, decision optimization, and generative AI applications. However, its real-world impact remains untested. HyperFRAME Research will track adoption metrics, focusing on deployments in non-fraud use cases such as supply chain optimization, customer personalization, or predictive maintenance across industries like retail, telecommunications, and government. IBM's claim of over 250 AI use cases provides a promising starting point, but the depth and scalability of these implementations will determine whether Spyre can transcend niche applications to become a cornerstone of mainframe-based AI innovation.

The z17's broader market significance lies in its reaffirmation of the mainframe's enduring relevance, countering outdated narratives that dismiss it as a legacy platform. With up to 208 customer cores, 64 TB of memory, and a 15-20% capacity increase over the z16, the z17 is engineered for the most demanding workloads, from high-transaction financial systems to hybrid cloud integrations. The Telum II processor, boasting 43 billion transistors and 8x 5.5 GHz cores, delivers an 11% improvement in single-thread performance, positioning it as a powerhouse for transactional AI. IBM's assertion of a 50% increase in AI inference operations per day (up to 450 billion with 1ms latency) compared to the z16 is ambitious, yet credible given IBM's track record. However, achieving widespread adoption will require robust proof points—detailed case studies, performance benchmarks, and transparent ROI metrics—to convince risk-averse enterprises, particularly in high-stakes sectors like finance and healthcare.

The z17's launch also brings into sharp focus the holistic ecosystem that underpins its value proposition. The integration of AI-driven tools, such as IBM Z Operations Unite, marks a significant step toward modernizing mainframe operations. By aggregating performance metrics and logs into a standardized OpenTelemetry format, Z Operations Unite leverages AI to enhance anomaly detection, incident isolation, and resolution times, addressing critical operational priorities for enterprises managing high-stakes workloads. Its integration with IBM Concert extends visibility across hybrid environments, aligning with the industry's shift toward unified observability. Similarly, enhancements to Z Backup Resiliency through integrations with DFSMS Cloud Data Manager and Threat Detection for z/OS strengthen data protection and recovery, reinforcing the mainframe's reliability. These software advancements, coupled with day-one support from ecosystem partners like BMC Software, Broadcom, and Rocket Software, reflect a concerted effort to bridge legacy systems with contemporary demands for observability, security, and automation.

The ecosystem's adoption of open standards, such as OpenTelemetry, and its focus on AI-powered analytics are pivotal for ensuring the mainframe's continued relevance in hybrid cloud environments. BMC's Automated Mainframe Intelligence portfolio, Broadcom's comprehensive software suite, and Rocket Software's tools for security and automation enhance the z17's capabilities, enabling enterprises to modernize without operational disruption. However, the ecosystem's success depends on its ability to deliver tangible benefits—reduced downtime, improved developer productivity, and enhanced security—while addressing practical barriers. Smaller enterprises, in particular, may face resource constraints in adopting these solutions, and the industry must provide clear guidance on implementation costs and integration complexities to broaden market penetration.

Ultimately, the z17's long-term viability hinges on its capacity to demonstrate measurable improvements in operational efficiency, security, and agility. While strategic partnerships and software integrations signal a promising direction, the industry must move beyond optimistic narratives to confront real-world challenges. Implementation complexities, such as integrating AI-driven tools with existing workflows, require robust support frameworks to minimize disruption. The mainframe's skills gap, a persistent concern, demands sustained investment in initiatives like Broadcom's Vitality Residency and Mainframe Skills Council to cultivate the next generation of expertise. Moreover, enterprises navigating complex digital transformations will prioritize solutions that deliver quantifiable ROI, making it imperative for IBM and its partners to provide transparent performance data and customer success stories.

From a financial perspective, the z17 is poised to drive significant revenue for IBM, estimated at $4 billion annually for the system, with a 3-4x multiplier from associated storage, high-margin software (e.g., CICS, Db2), support services, and IBM Consulting, totaling approximately $15 billion in high-margin revenue. Coming 12 quarters after the z16's launch, the z17 benefits from favorable upgrade dynamics, as many customers will have completed their lease cycles, reducing financial barriers to adoption. IBM's focus on on-premises AI processing, targeting low-latency, high-throughput applications, positions the z17 as a strategic asset for industries like banking, telecom, and government, where real-time transactional insights are paramount. Over the next four quarters, IBM's leadership is likely to highlight the mainframe refresh cycle as a key growth driver, with the z17's impact resonating across the company's portfolio.

In conclusion, the IBM z17 and its ecosystem represent a bold step toward redefining enterprise computing in the age of AI and hybrid cloud. Its technical prowess, exemplified by the Telum II processor and Spyre Accelerator, positions it as a platform capable of handling the most demanding workloads. The ecosystem's software innovations and vendor support underscore a commitment to modernization, bridging the gap between legacy infrastructure and contemporary technological imperatives. However, the z17's success will depend on its ability to deliver on its promises—through scalable AI adoption, operational efficiencies, and robust security—while addressing practical challenges in implementation and skill development. As HyperFRAME Research continues to track the z17's performance, customer adoption, and market impact, the mainframe's ability to evolve will determine its role as a cornerstone of enterprise computing in an increasingly AI-driven future. The industry must seize this opportunity to move beyond rhetoric, grounding the z17's potential in measurable outcomes that empower enterprises to navigate the complexities of digital transformation with confidence.

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers indepth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research
Email Address:
steven.dickens@hyperframeresearch.com
Telephone Number:
+1 845 505 1678
X: - @StevenDickens3
LinkedIn: Steven Dickens
BlueSky: Steven Dickens

CONTRIBUTORS

Steven Dickens CEO & Principal Analyst
HyperFRAME Research

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.

LICENSING

DISCLOSURES

HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.


HyperFRAME
RESEARCH