Sustainability with AI

Rosalind Radcliffe
IBM Fellow
CTO for Z Ecosystem

Integrated Governance Program - key components

Foundational Data Transformation

Org. Governance; Responsible Tech Board

AI System and Model Build

Data Governance

Model Lifecycle Management

Regulatory Compliance

Organizational Governance AI Ethics Board

- Perform Use Case Risk Assessments to identify potential issues.
- Put in place guardrails to manage risks.

AI System and Model Build

Design use cases in accordance with IBM's Ethics by Design Framework.

Data Governance

- Data pipeline has a direct bearing on the risk profile.
- Facts about the data are collected.

Model Lifecycle Management

- The model is trained on cleared datasets regardless of origin.
- The model is tuned, tested and evaluated throughout its lifecycle.

Regulatory Compliance

A single inventory enables us to filter items in scope for a given regulation and launch targeted compliance campaigns.

Chief Privacy Officer (CPO)

Key concern:
How are we ensuring compliance with the growing number of AI regulations, across the jurisdictions we operate in?

Chief Data Officer (CDO)

Key concern:
Do we have trusted data pipelines to ensure quality and rights of use for model training data?

Compliance/Risk Officer

Key concern:
Do we have a comprehensive inventory of AI models and Systems used in the enterprise? Can we track these across their lifecycle?

IBM z17

Fully engineered stack for AI where it matters most

Transaction processing platform

Operating systems & firmware

IBM Z infrastructure

Built on a foundation of security, resiliency, and high availability.

450 billion

Inference operations per day with 1ms response time vs. 300B on IBM z16[8]

7.5x AI throughput

Utilizing 8 AI processing units vs. one on IBM z16[9]

83% less power

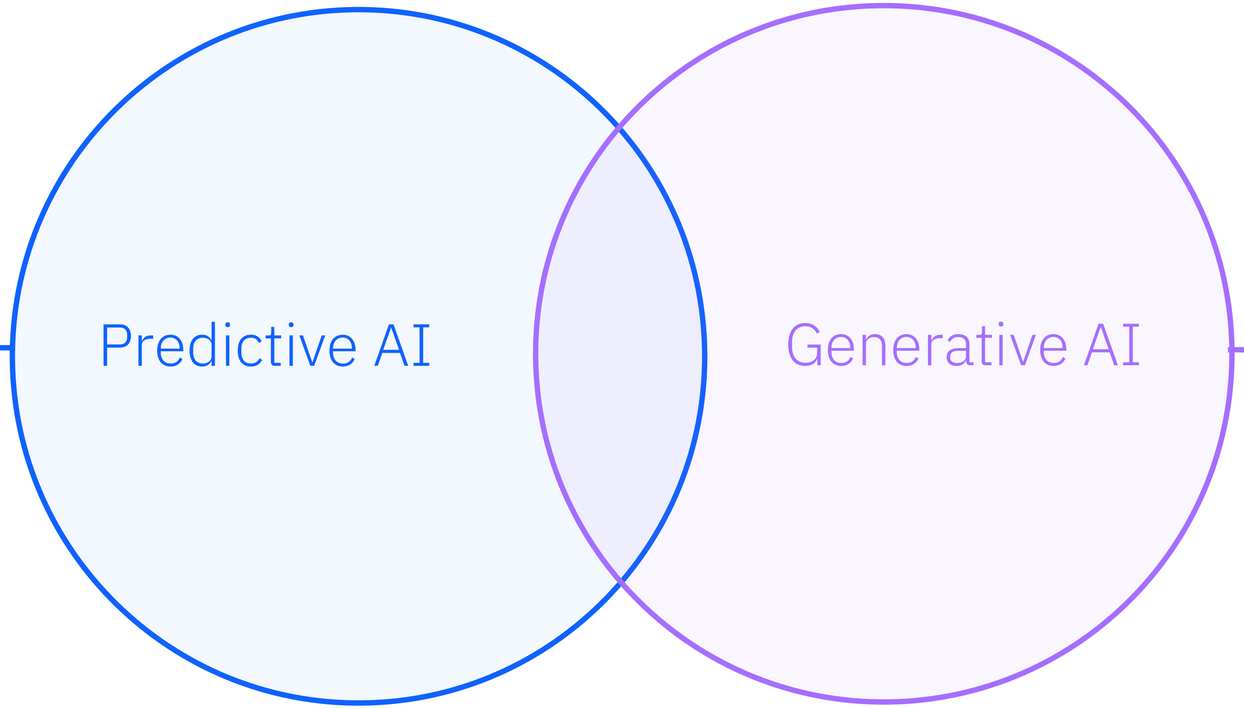
Moving AI-infused OLTP workloads from compared 2 year old x86 servers[10]

Predictive AI and generative AI on IBM z17 and LinuxONE deliver unique capabilities

Machine learning

Advantage

- Real-time AI inference
- In-transaction AI at scale
- Inference every transaction[1]



Larger, more complex models

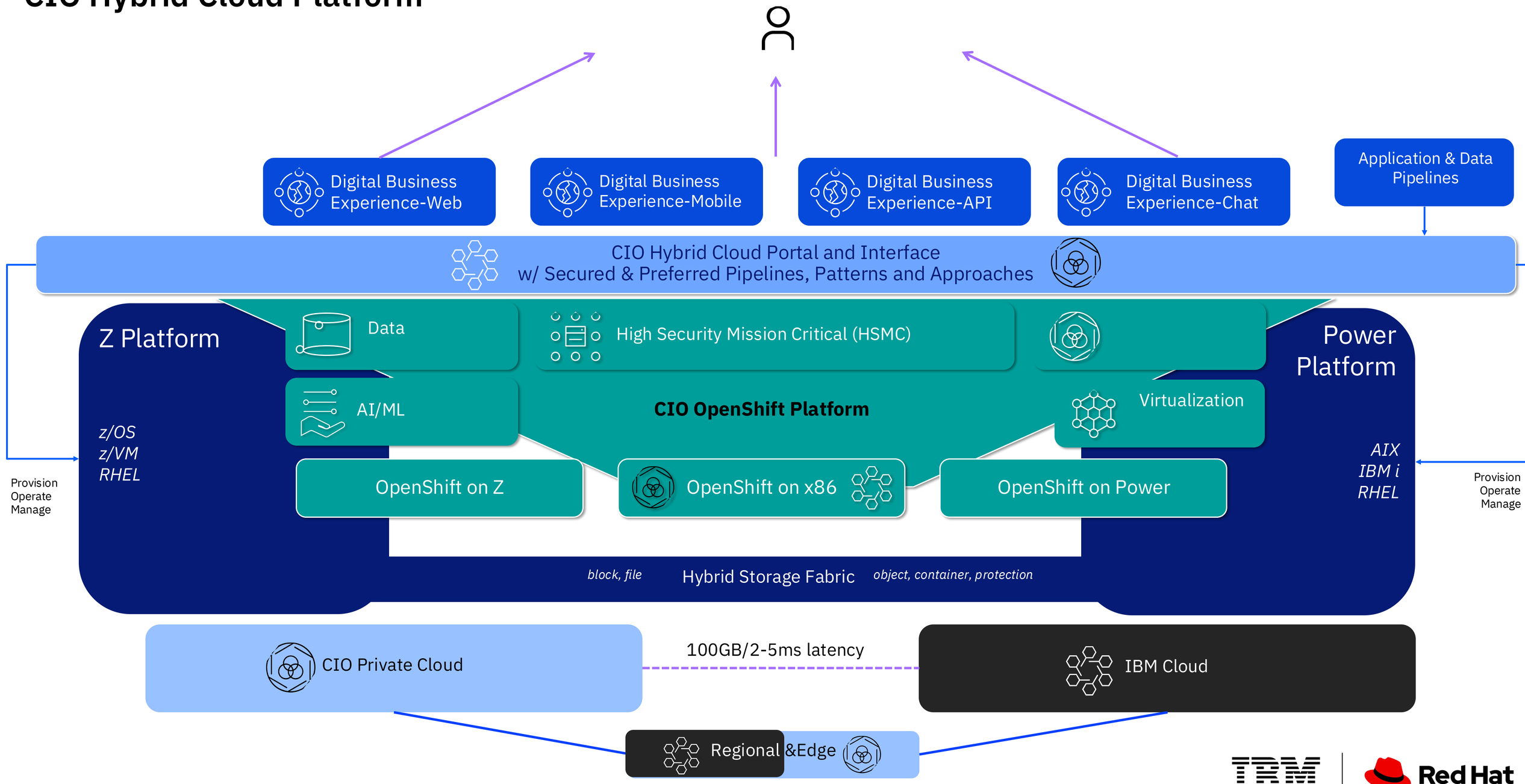
Advantage

- Gen AI with IBM Z RAS
- Security and data protection
- Energy efficiency

Multiple-model AI at speed and scale

Combine machine learning
and LLMs to achieve higher prediction accuracy
and expanded insights

CIO Hybrid Cloud Platform



Digital IT Support transformation powered by AskIT

Outcomes (2023 to 2024)

~79%

- IT Support labor reduced
- IT Support Advisor to employee ratio increased from 1:891 to 1:4248

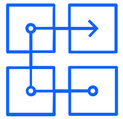
~$18M

- IT Support cost reduced

Eliminate

top support call drivers

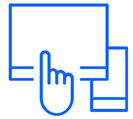
- 514 intents now in AskIT
- IT Support tickets down 56% from 2023-2024



Simplify

end-to-end support

- AskIT triage and resolution
- Phone lines no longer needed
- Human chat support 24 x 7 for complex issues



Automate

manual tasks

- Mac Recovery Key automated for end user
- Device compliance check
- Device upgrade eligibility
- Support ticket creation

Phase 1 | Phase 2 | Phase 3 - Today | Our Journey Forward

CSAT - 90%
Chat Quality - 96%

Consolidate, Standardize

- Support content silos consolidated into w3 IT Support
- ServiceNow migration for strategic content and ticketing

AskIT IT Support Front Door

- Strategic tool implementation
- Automate top call drivers preventing help desk contacts
- Sunset phone lines to drive AskIT front end with 24 x 7 human chat support as back-up
- Executive IT Support reimagined

80% Automation for Support

- Eliminated 80% of support queries being handled through to the Help Desk
- Automate further with Watsonx Orchestrate conversations
- Eliminate simple tasks like password resets and certificate simplify resolution steps like Mac recovery Key

Enhancements & Reimagined ThinkDesk

- Guided resolution beyond keyword intents
- Device telemetry for proactive support
- IT Support data for more productivity insights
- Reimagined ThinkDesk experience powered by AskIT to triage and offer on site appointment if IT Advisor needed (piloting at selected sites)
- Emergency device loaner via on site lockers & emergency IT peripherals at pilot sites

Use Case: Digital Labor - HR

Conversational AI Transformation of HR Support

AskHR

Challenge

HR Support has multi-channels and multi-tiers resulting in poor employee experience

Solution

Single Digital channel for all employee engagement and 2-tier support model (digital/human)

Interactive
- Accessible via Mobile, Intranet and HR pages
- Accessible via SLACK / including proactive ”push” notifications related to different HR events

Integrated
- Search over 4700 policy pages
- 2700+ FAQ’s
- Integrations with Workday, WF360, Concur. Weather Channel, Org Risk Insights, Zendesk

Personalized
- Key HR Links, News & Updates added
- Country-specific responses for multiple persona’s
- 79 HR Task Automations e.g. job transfer, time-off, compensation planning,

Personas
- Employee
- Manager
- Executive
- HR Business Partner
- Assignee
- Alumni*
- Candidate**

10.1M interactions

243K Unique users (81 Countries)

75% CSAT

94% Containment Rate

756K HR Transactions Automated

75% quicker transactions execution ($5m productivity)

6M Slack Notifications

97% Manager Usage

94% Executive Usage

61% Ticket Reduction (7 years)

Solution includes: watsonx Assistant

Amplifying developer productivity with AI-powered automation

The Hidden Complexity

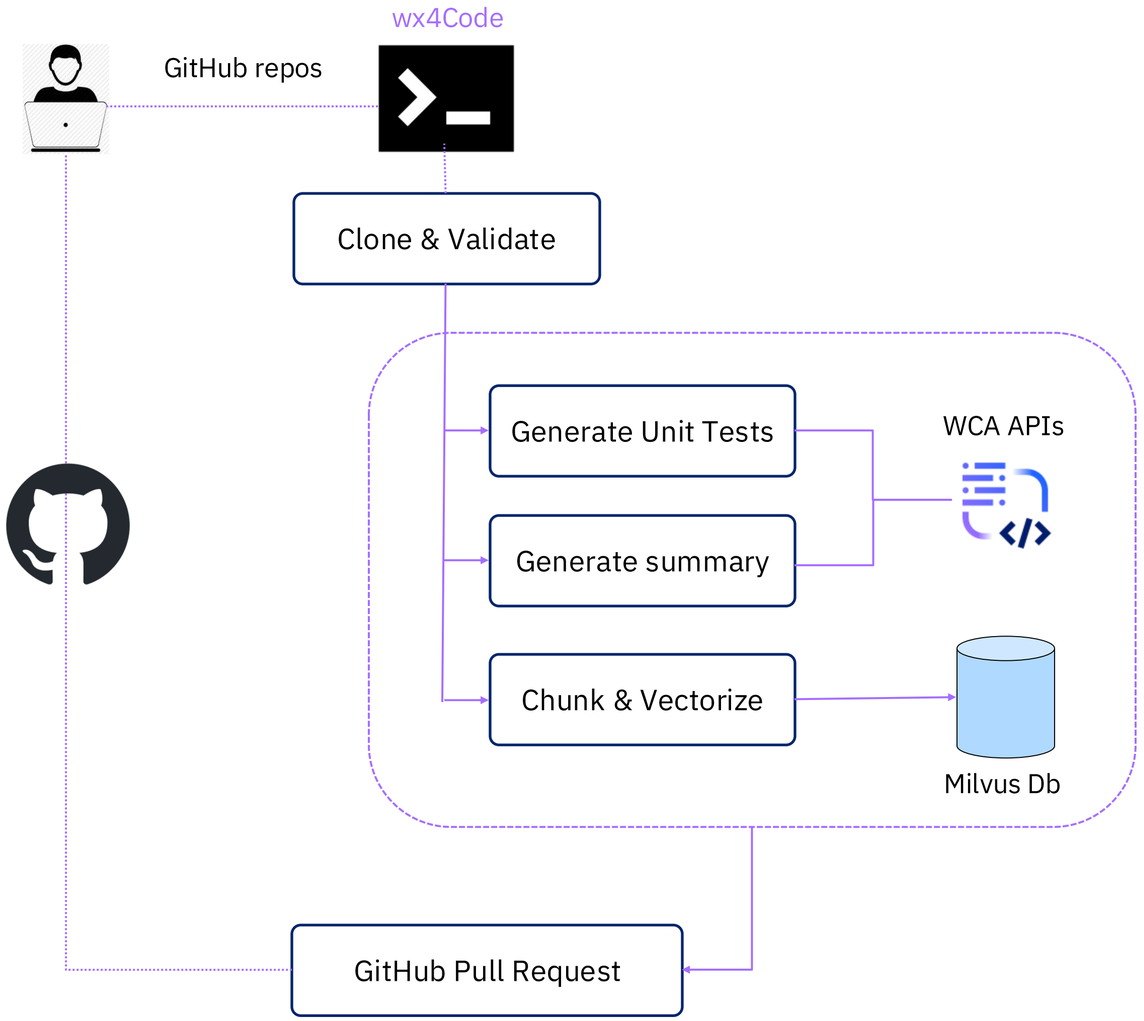
As we strive to stay ahead in today's fast-paced digital transformation, many enterprises are hindered by the weight of their legacy codebases.

- **Undocumented codebases:** Substantial codebases without documentation.
- **Outdated unit tests**: Inadequate tests, posing challenges for new developers.
- **Knowledge gap**: Developers struggling to understand code they didn't write.
- **Security risks**: Vulnerabilities and weaknesses in outdated code

Solution

We created a wx4Code that leverages Watsonx Granite code models(Granite-34b-code-instruct) to automatically update documentation and unit tests at scale, revolutionizing the way we maintain our codebase.

- AI-enabled capability to process multiple GitHub Repos for Code Summary and Unit Tests, improving maintainability (MVP Scope)
- Automated Pull Request Reviews with Code fixes
- Identify patterns for future standardization and code optimizations at scale
- Data driven decision making for prioritizing code fixes



watsonx Challenge

Here are the problems that still remain:

What’s the right way to show the cost of AI?

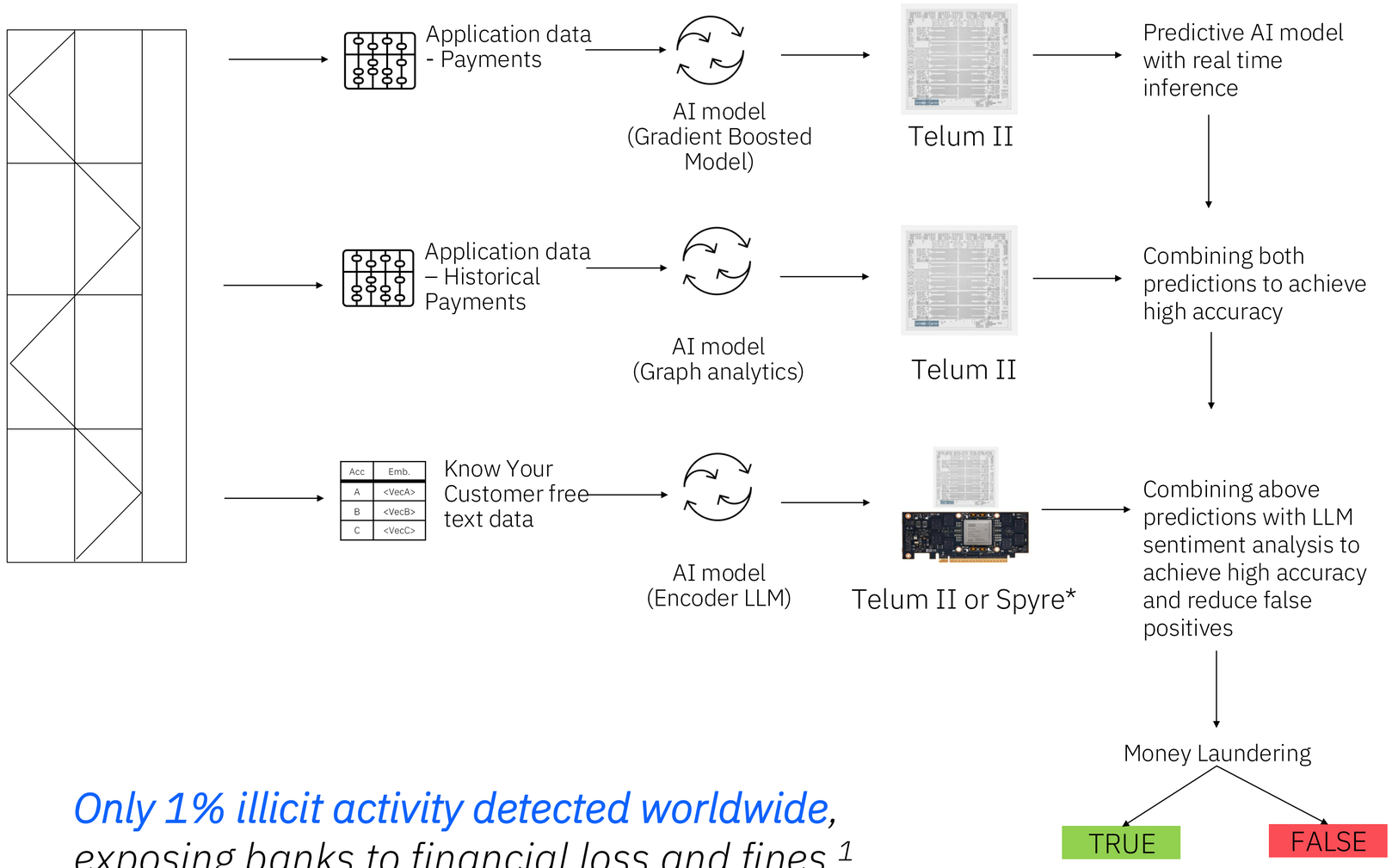
Including Power and water utilization.

Deliver business growth with deep insights

IBM z17 with Telum II

Advanced anti-money laundering

- Improved accuracy with reduced false positives
- Greater ROI achieved by reducing risk of fines with AI
- Improved client loyalty
- Improved efficiency and reduced cost by removing manual analyses and processing



Only 1% illicit activity detected worldwide, *exposing banks to financial loss and fines* [1]

*Available starting in 4Q25

12

AI on IBM Z Ecosystem Stack

Designed for Business Insights and Intelligent Infrastructure

BUSINESS INSIGHTS

Infuse AI in real time into every transaction

MACHINE LEARNING FOR Z/OS

Deliver AI solutions at an unprecedented speed

CLOUD PAK FOR DATA ON IBM Z

IBM's market leading, cloud native Data and AI platform

DB2 FOR Z/OS WITH SQL DATA INSIGHTS

Uncover hidden patterns from data locked in z/OS

AI TOOLKIT FOR IBM Z AND LINUXONE

Support popular open-source AI tools on IBM Z

INTELLIGENT INFRASTRUCTURE

Improve automation, security, privacy and ITOps with AI

AI-POWERED IBM SECURITY

Next gen protection for your most crucial data

IBM Z ANOMALY ANALYTICS

Proactively identify and mitigate Ops issues

watsonx Assistant for IBM Z

Enable conversational AI, automate tasks and build

IBM DB2 AI FOR Z/OS

Enhance database performance with ML

AI-INFUSED IBM Z/OS V3.1

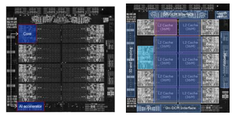
Enable intelligent admin, ops and automation

watsonx Code Assistant for IBM Z

Accelerate mainframe application modernization

ONNX, scikit-learn, Snap ML, python, spaCy, RASA, TensorFlow, Kubeflow, OpenCV, ANACONDA, Keras, PyTorch, Hugging Face, MAGE, NVIDIA Triton Inference Server, Red Hat OpenShift AI, Apache Spark, mlflow

Enable market leading AI / ML ecosystem on IBM Z



Telum I & Telum II
On-chip AI accelerator

Spyre Accelerator
Host attached AI accelerator

Deliver billions of inference requests per day in real time

More AI acceleration on IBM z17 and LinuxONE

In-transaction AI with encoder LLMs and multiple AI model techniques

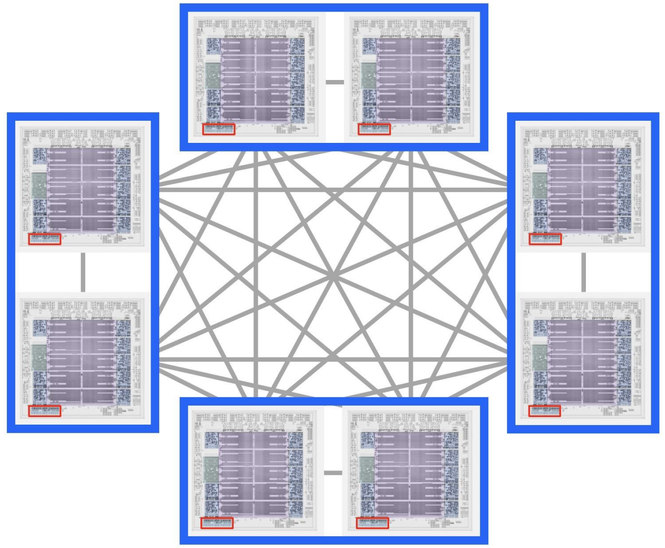
2nd Gen on-chip AI accelerator in Telum II

- Support for LLM compute primitives
- Improved quantization and matrix operations
- Improved AI processing over IBM z16[9]

AI workload balancing during peak usage

In-drawer intelligent routing

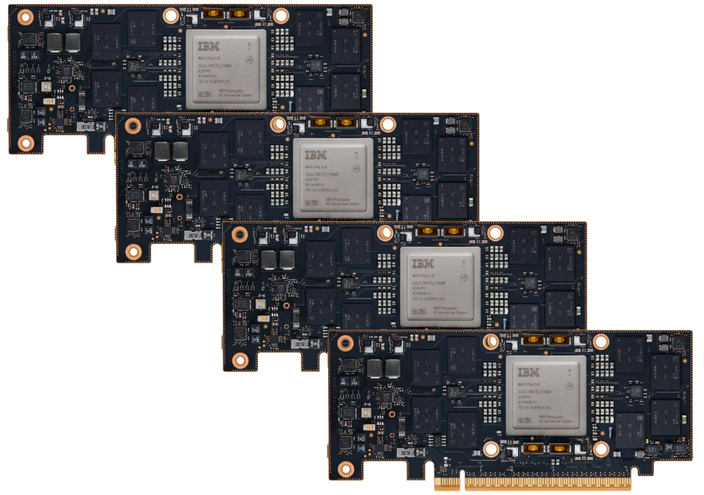
- Remote AI processing
- Up to 8x AI processing available



Optimize generative AI and LLM use cases

IBM Spyre Accelerator cards*

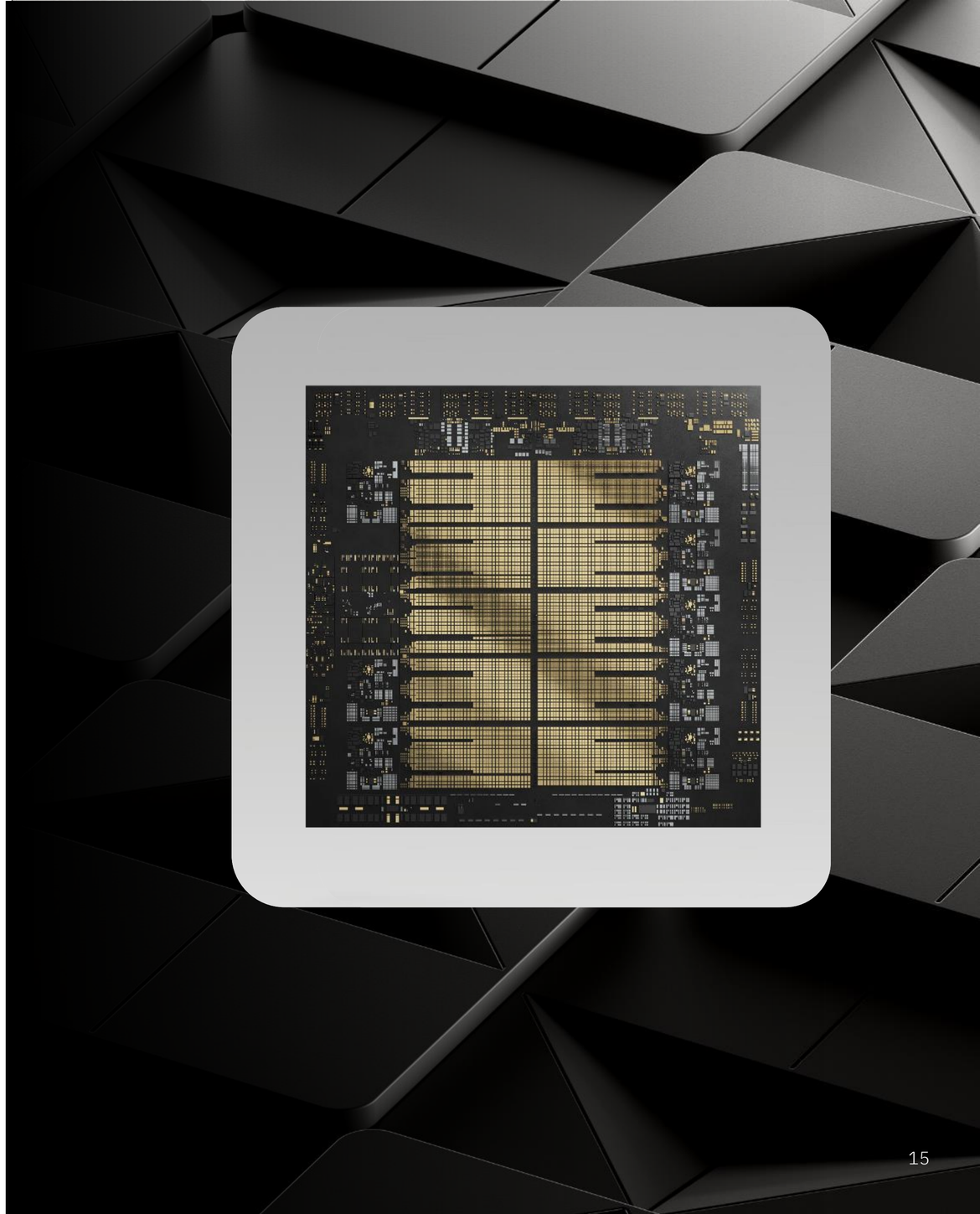
- 32 Gen AI-ready cores per adapter card
- Up to 48 adapter cards per system



*Available starting in 4Q25

IBM Telum II Processor

- 5nm technology, 5.5GHz
- 8 cores with 20% area reduction and improved microprocessor power management
- 40% more cache per core
- **NEW:** On-chip Data Processing Unit (DPU): Increased I/O performance with 70% reduction in power for I/O management, RAS, reduced latency
- 2nd-gen AI Accelerator for high-speed inferencing with fine tuning
- 8x dedicated AI processing per core



Thank you

Disclaimer for IBM Spyre™ Accelerator

The IBM Spyre™ AI Accelerator will not be available with IBM z17 general availability. The IBM Spyre AI Accelerator is currently expected to be available in 4Q of 2025. Any capabilities discussed in this presentation with respect to the IBM Spyre AI Accelerator will not be enabled by IBM z17 until the accelerator cards are installed in the system. See the IBM z17 announcement letter for the statement of direction for the IBM Spyre Accelerator.

Citations/Claims/Disclaimers

1. Katov PhD, N. (April 2025), Mitigating Fraud in The AI Age: Supporting Transaction Fraud Detection at Scale on IBM z17, Celent
2. the-economics-of-technology-investments-downloadTop Ten Payments Companies Processed $9 Trillion in 2022 Payment Card Volume, GlobeNewswire
3. For clients running z/OS v3.1 or higher with a configured high availability IBM software stack on IBM z16 or IBM z17, users can expect up to 99.999999% availability or 315.58 milliseconds of downtime per year when using a GDPS 4.7 Continuous Availability (CA) configuration and workloads.

   DISCLAIMER: The claim is based on IBM internal data and a GDPS CA three-site configuration, 2 active Sysplex sites and 1 Disaster Recovery (DR) site, consisting of z/OS 3.1 or higher with a Recovery Time objective (RTO) of 2 minutes or less, one of the required GDPS CA IBM middleware stack workloads and replication products running on IBM z16 or IBM z17.
   GDPS CA includes resiliency features such as Parallel Sysplex enabled data sharing applications, GDPS Metro Mirror replication (Hyperswap), software replication, and other CA configuration documented high availability features. A supported GDPS CA middleware stack could include CICS v6.2, IMS v15.5, MQ v9.4, and Db2 v13 or at later releases. Clients must follow maintenance, configuration, capacity planning and testing best practices for the entire software stack and hardware configuration. This includes enabling all the resiliency technology for their workloads as defined by GDPS CA, z/OS, and workload related software products. Other configurations may have different availability characteristics.
4. The AI boom could use a shocking amount of electricity, Scientific American, 13 October 2023

5 Computational power and AI, AI Now Institute, 27 September 2023

IBM Spyre™ Accelerator is
in tech preview until Q42025

Citations/Claims/Disclaimers

8.. With IBM z17, process up to 450 billion inference operations per day with 1 ms response time using a Credit Card Fraud Detection Deep Learning model .DISCLAIMER: Performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. The benchmark was executed with 1 thread performing local inference operations using a LSTM based synthetic Credit Card Fraud Detection model (https://github.com/IBM/ai-on-z-fraud-detection) to exploit the Integrated Accelerator for AI. A batch size of 160 was used. IBM Systems Hardware configuration: 1 LPAR running Red Hat® Enterprise Linux® 9.4 with 6 IFLs (SMT), 128 GB memory. 1 LPAR with 2 CPs, 4 zIIPs and 256 GB memory running IBM z/OS® 3.1 with IBM z/OS Container Extensions (zCX) feature. Results may vary.

9. The IBM z17 Telum II processor is designed to seamlessly scale peak AI workloads as each core on the chip can access each of the 8 integrated accelerators for AI. By allowing routing of inference requests to any idle IBM Integrated Accelerators for AI within the same drawer, the IBM Integrated Accelerator for AI can increase inference throughput by up to 7.5x as compared to IBM z16..

   DISCLAIMER: Performance results are based on internal tests exploiting the IBM Integrated Accelerator for AI for inference operations on IBM z16 and z17. On IBM z17, each IBM Integrated Accelerator for AI allows any CPU within a drawer to direct AI inference request to any of the 8 idle AI accelerators on the same drawer. The tests involved running inference operations on 8 parallel threads with batch size of 1. Both IBM z16 and z17 were configured with 2 GCPs, 4 zIIPs with SMT and 256 GB memory on IBM z/OS V3R1 with IBM Z Deep Learning Compiler 4.3.0, using a synthetic credit card fraud detection model (https://github.com/IBM/ai-on-z-fraud-detection). Results may vary.

10. Save up to 83% of power consumption by replacing a compared x86 solution comprised of two-year-old servers running AI-infused OLTP workloads with an IBM z17 DISCLAIMER: The total cost of ownership (TCO) is based on IBM® internal performance tests running on IBM Systems Hardware of machine type 9175 compared to the same tests running on a commercially available enterprise server with 2x 28 Intel Xeon Gold 5420+ CPU @ 2.20 GHz.
    The MegaCard benchmark (https://github.com/IBM/megacard-standalone) is a containerized IBM WebSphere Liberty v24 application deployed on Red Hat® OpenShift® Container Platform (RHOCP) 4.17 on Red Hat Enterprise Linux® (RHEL) 9.4 with KVM. EDB Postgres for Kubernetes v1.25 is used as the database.
    The TCO model extrapolated the test results to a typical, complete customer IT solution that includes isolated from each other production and non-production IT environments. TCO included software, hardware, energy, network, data center space, and labor costs. On the IBM z17 side the complete solution requires one IBM z17 Type 9175 MAX 136, and on x86 side, the complete IT solution requires 72 of the compared servers. Results may vary.

11. “Invest Implications: Forecast Analysis: Artificial Intelligence Software", Gartner, 2023-2027. https://www.gartner.com/en/documents/4925331

12. IBM IBV - Mainframes as mainstays of digital transformation - https://www.ibm.com/downloads/documents/us-en/10c31775c85402a2

13. Forrester Consulting Study conducted by Forrester Consulting on behalf of Deloitte, and IBM Institute of Business Value Mainframes as mainstays of digital transformation, October 2024 (https://ibm.co/mainframe-hybrid-cloud)

14. IBM Cost of a Data Breach Report 2024 https://www.ibm.com/reports/data-breach

15. 82% of IT executives say that leveraging AI for monitoring, analyzing, detecting, and responding to cyber threats is important to their organizations. - IBM Institute of Business Value Mainframes as mainstays of digital transformation, October 2024 (https://ibm.co/mainframe-hybrid-cloud)

IBM Spyre™ Accelerator is
in tech preview until Q42025

Citations/Claims/Disclaimers

16. Single processor capacity of IBM z17 for equal n-way at common client configurations is approximately 11% greater than on IBM z16 with some variation based on workload and configuration. DISCLAIMER: Based on internal measurements. Results may vary by customer based on individual workload, configuration and software levels. Visit LSPR website for more details at: http://www.ibm.com/support/pages/ibm-z-large-systems-performance-reference

17 Within each single drawer, IBM z17 provides 20% greater capacity than IBM z16 for standard models and 15% greater capacity on the max config model, enabling efficient scaling of partitions. DISCLAIMER: Based on internal measurements. Standard models include IBM z17 Max43, Max90, Max136 and Max183 and IBM z16 Max39, Max82, Max125 and Max168. Max config models are IBM z17 Max208 and IBM z16 Max200. Results may vary by customer based on individual workload, configuration and software levels. Visit LSPR website for more details at: http://www.ibm.com/support/pages/ibm-z-large-systems-performance-reference

•18. 14. The anticipated typical IBM z17 system is estimated to reduce system power consumption by approximately 19% compared to a similarly configured IBM z16 system. DISCLAIMER: Based on an expected typical IBM z17 system configuration based on the actual historical average IBM z16 system configuration. IBM z17 is Max90 with 8TB memory, 82 active processors, 5 ICA-SR 2.0, and 3 PCIe+ I/O drawers with 43 I/O adapters. The IBM z16 is configured to provide the same hardware capability. Power consumption is based on the Power Estimation Tool, available at ibm.com/support/resourcelink/api/content/public/PowerEstimationTool-legacy.html. Results may vary.

IBM Spyre™ Accelerator is
in tech preview until Q42025