IBM LinuxONE 5 Gets Huge AI Boost

A Cambrian-AI Research Paper
Karl Freund, Founder and Analyst
Sponsored by IBM
July 21, 2025

Introduction

IT organizations have been reaping the benefits of lower TCO from server consolidation for decades. Multi-core x86 servers, VMware, open hypervisors, and container management platforms such as Kubernetes create multiple virtual servers from a single physical machine. While the TCO spreadsheets accurately reflect the potential savings, they do not typically account for the costs of a server outage or a security breach, which would now impact scores of applications running on the server.

IBM® LinuxONE 5 is an enterprise Linux platform that merges the extensive system expertise of IBM with open-source innovations to address modern IT infrastructure demands. LinuxONE excels by providing significant scale, power efficiency, and security for mission-critical Linux workloads. With the inclusion of AI inferencing capabilities, LinuxONE is a powerful platform for datacenter operators seeking to reduce the cost of workload consolidation and sustainably scale AI workloads in a secure environment.

The new IBM Telum II processor combines high-performance hardware, advanced security, and Linux flexibility, targeting organizations that need scalable, secure, and sustainable solutions for mission-critical workloads. Practically speaking, bringing AI inside the LinuxONE 5 enterprise-class server has the potential to be a significant game-changer for enterprise IT.

With the IBM Spyre Accelerator, planned to be available in 4Q 2025, IBM aims to transform the user experience on the platform by running generative AI encoder and decoder models that enable use cases such as document summarization and image analysis. Clients can achieve business goals, efficiencies, and productivity through automation and advanced AI, leveraging Gen AI chat, such as IBM Watson Assistant.

The approach taken by LinuxONE - of consolidating data, models, and applications all onto one scalable system – has the potential to reduce latency, improve performance, maximize security, and minimize costs. Potential use cases include private or open-source AI platforms, co-location of data and AI applications, and lower-energy usage.

This research paper will explore the technology incorporated into LinuxONE 5, and the AI use cases we expect will be a good fit for this new enterprise-class server

IBM LinuxONE Background

IBM launched the LinuxONE system in August 2015, and with IBM's acquisition of Red Hat® in 2019, the LinuxONE platform gained support for additional foundational components, including Red Hat OpenShift®. LinuxONE is a family of enterprise-grade Linux servers that support all major versions of Linux, powered by the IBM Telum and Telum II processors. It brings together decades of IBM expertise in building enterprise systems with the openness of the Linux operating system.

IBM LinuxONE offers an alternative Linux platform that scales far beyond the reach of an x86 server, providing the most secure, resilient, and scalable consolidation platform in the industry. According to IBM performance tests, LinuxONE 5 can support 35 billion transactions per day at 99.999999% availability, which translates to approximately 315 milliseconds of downtime per year. Organizations can save up to 94% in software costs over five years by migrating cloud-native, containerized workloads from an x86 solution to IBM LinuxONE 5, which runs the same software products. Additionally, a single system can do the work of up to 2,944 cores of an x86 solution, making it a highly efficient option for data centers.

Figure 1: The IBM LinuxONE 5

And now, the new LinuxONE 5 features on-chip AI acceleration for inferencing and secure PCIe-based dedicated AI accelerators with IBM Spyre (available in Q4 2025) for multi-model, including generative AI. This enables AI enterprise applications and transaction processing to happen efficiently while eliminating the need to offload enterprise AI workloads to less secure and less reliable hardware.

IBM LinuxONE is a family of enterprise-grade servers designed exclusively for Linux, featuring:

- **High performance:** Optimized for data-intensive and AI workloads, with vertical/horizontal scaling capabilities
- **AI optimizations:** Includes integrated AI accelerators on the Telum II processor, and supports scalable PCIe-based Spyre Accelerator/s for multiple model AI, including generative AI use cases
- **Security:** Hardware-based encryption, confidential computing, and quantum-safe cryptography
- **Sustainability:** Energy-efficient design enabling workload consolidation
- **Hybrid cloud integration:** Supports cloud-native tools like OpenShift, Kubernetes, and Ansible

All these capabilities are then inherited by the software running on LinuxONE, providing a highly secure, performant, and reliable server.

The Telum II Processor

When IBM launched the first Telum processor at the Hot Chips conference in 2021, it was the first processor to incorporate on-chip acceleration for artificial intelligence inferencing during transaction processing. This accelerator enabled applications to perform credit card detection within every transaction, rather than sampling a subset, thereby improving business outcomes and customer service.

The new LinuxONE 5 supports up to 208 secure Telum II cores, featuring 32 on-chip AI accelerators and 64 TB of memory, providing 4 times more compute power than its predecessor. The on-chip AI accelerator can process up to 5 million inference operations per second with INT8 precision at a response time of less than 1 ms, using a Credit Card Fraud Detection Deep Learning Model. The larger accelerator enables more complex AI workloads and the integration of larger models, including encoder Large Language Models (LLMs). Telum II also boasts a 40% increase in on-chip cache memory. It features 10 L2 caches, each with 36MB, and larger virtual L3 (360MB) and L4 (2.88GB) caches, which reduce latency and improve performance for data-intensive applications.

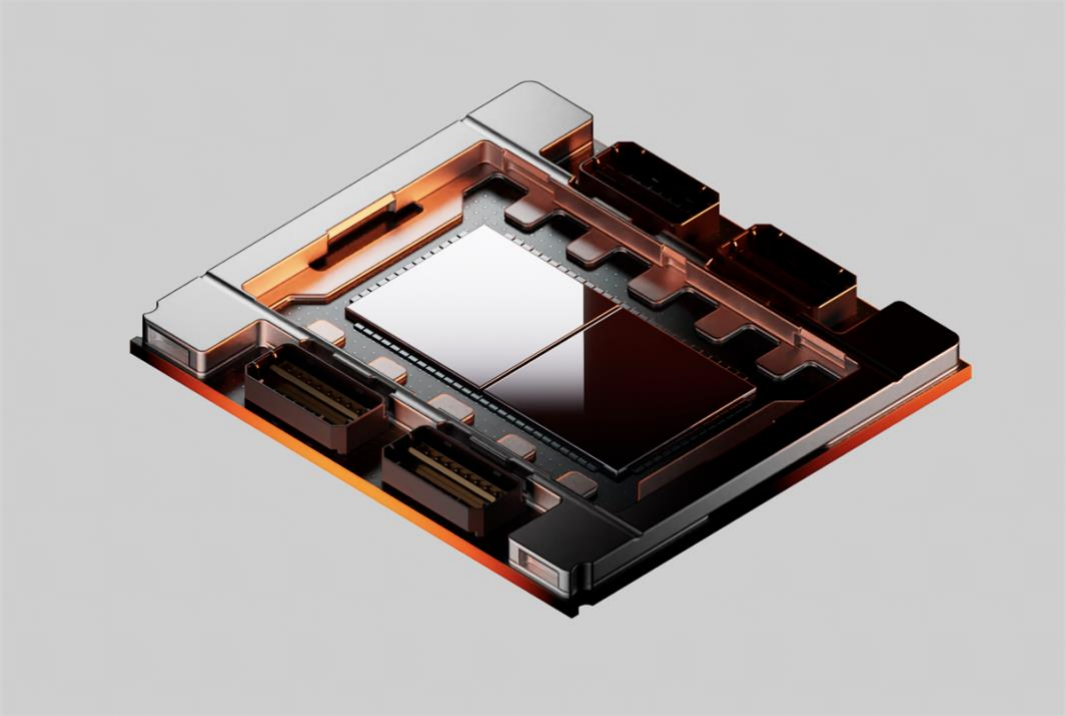

Figure 2: The IBM Telum II Processor.

The on-chip accelerator is ideal for traditional machine learning algorithms and smaller generative AI models of less than eight billion parameters. Using a single Integrated Accelerator for AI on an OLTP workload on IBM LinuxONE, 5 matches the throughput of running inference on a compared remote x86 server with 13 cores.[i] Replacing a compared x86 solution comprised of two-year-old servers running AI-infused OLTP workloads with an IBM LinuxONE 5 can save up to $24 million on the total cost of ownership over five years.

With IBM LinuxONE 5, the energy consumption of an OLTP workload can be reduced by 81% when running inference operations with multiple AI models on the platform compared to running them remotely on a comparable x86 server with a GPU.

The IBM Telum II also introduces a new data processing unit (DPU) that can reduce the power required for input/output (I/O) management in an extensive IBM LinuxONE 5 system by over 90% compared to a similarly configured IBM LinuxONE 4 system.

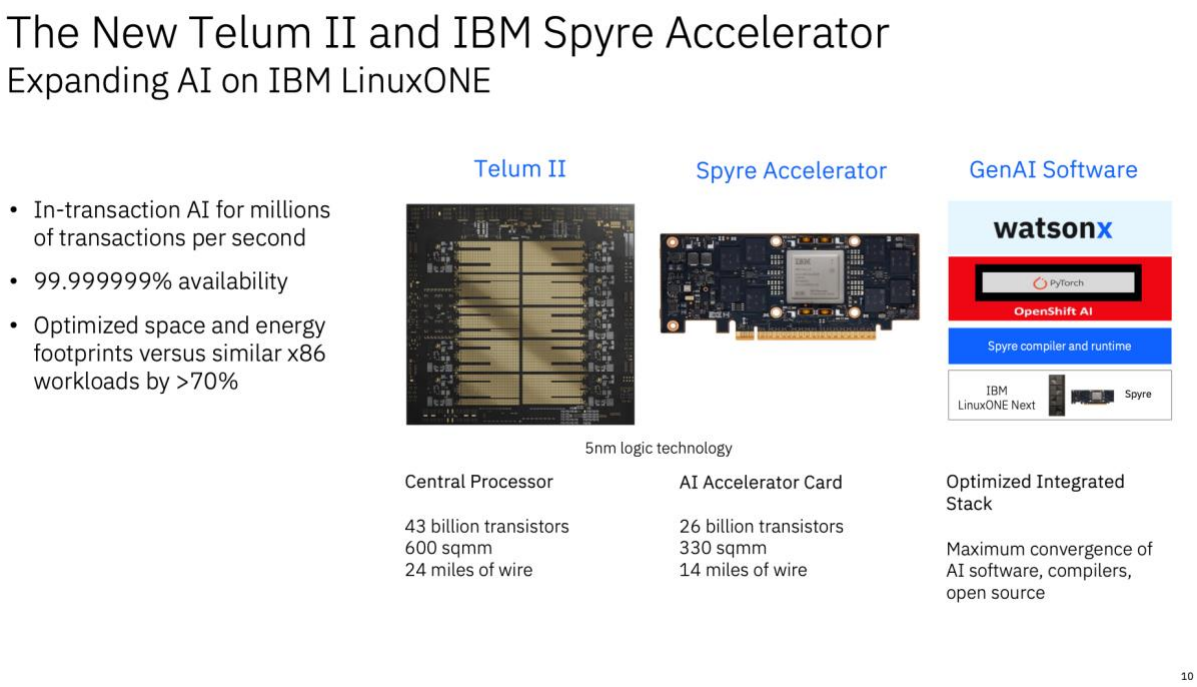

Figure 3: AI on LinuxONE supports the on-chip Telum II accelerators and PCIe-based Spyre accelerator.

The IBM Spyre Accelerator

Recognizing the need to support larger generative AI models in enterprise applications, IBM has engineered the Spyre Accelerator, delivered by PCIe card, which IBM plans to make available in Q4 2025. With up to 256 accelerator cores and 1 TB of memory, the platform's scalable architecture and optimized energy consumption can contribute to improved business efficiency and reduced operational costs. LinuxONE can support up to 48 Spyre cards to handle larger AI models and multiple model applications, such as complex document summarization and image analysis, as well as Gen AI, including IBM® watsonx Assistant. More details about the capacity and performance of Spyre will become available when IBM launches it for general availability (GA) later this year.

We will explore example use cases below to examine which accelerator is best for specific workloads.

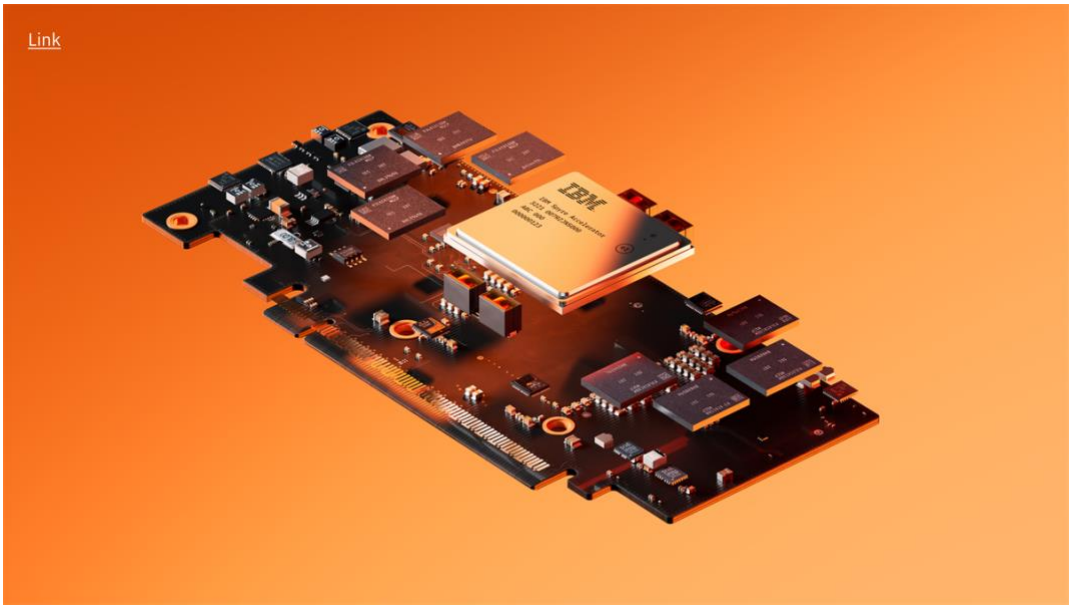

Figure 4: The IBM Spyre AI accelerator

AI Software

IBM offers a comprehensive suite of AI inferencing software stacks to leverage hardware capabilities, including popular frameworks in the AI Toolkit for LinuxONE and through support for Red Hat OpenShift AI. IBM supports the Nvidia Triton Inference Server, which Nvidia released as open source, enabling decreased latency and increased throughput, thereby providing high-performance inference capabilities across various industries. Its ability to handle multiple concurrent requests efficiently makes it particularly useful in real-time applications. It will be interesting to see whether IBM takes the same approach and supports the Nvidia Dynamo open software, which acts as an OS for running enterprise AI.

The availability of open-source AI software that leverages AI hardware accelerators is crucial to the success of LinuxONE in the Linux marketplace.

- **AI Toolkit for LinuxONE**: Includes popular open-source AI frameworks and tooling optimized to take full advantage of Telum II, paired with IBM Elite Support. It features PyTorch, TensorFlow, Triton Inference Server, SnapML, and the IBM Deep Learning Compiler.
- **Red Hat OpenShift AI for IBM LinuxONE:** Combines Red Hat's open-source expertise with the powerful capabilities of IBM LinuxONE, providing organizations with a robust foundation for building, training, deploying, and monitoring AI models across hybrid cloud environments. It's currently available as a tech preview and is expected to GA later in 4Q25.

- **IBM watsonx:** Under development for LinuxONE, this will provide an end-to-end solution framework when it becomes available. Whether running foundation models, building models, utilizing IBM models, or a combination of these, watsonx is designed to help fine-tune these models and run real-time inference with the trust, scalability, stability, and accuracy needed by many enterprises today.

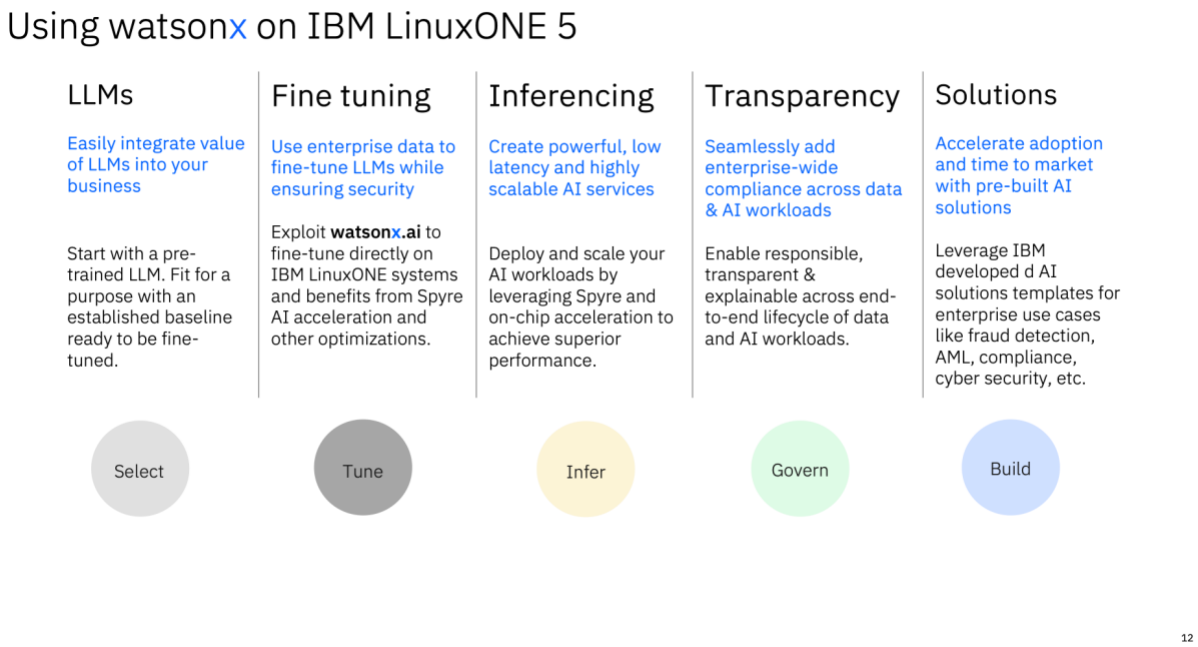

Figure 5: IBM's vision for watsonx support for LinuxONE.

Advanced LinuxONE Security Features for AI

IBM LinuxONE 5 can protect against threats and attacks on sensitive data, models, and AI workloads. It can leverage confidential computing techniques and quantum-safe encryption models to protect sensitive data, along with other solutions to strengthen the security posture of enterprise AI. Security features include:

- **Confidential computing: L**everages a confidential computing environment to protect AI workloads, data, and creates secure boundaries around a model in development and use. Protects data in memory with encryption, preventing unauthorized access from within the system.
- **Hardware Security Module (HSM):** Provides a secure environment for storing and managing cryptographic keys. The Crypto Express 8S is the latest generation of IBM HSM. It performs top-level security processing and high-speed cryptographic functions with a high throughput rate and reduced latency. It

supports advanced cryptography by making available to applications the primary post-quantum cryptography algorithms standardized by NIST, ML-KEM, which enables key establishment schemes, and ML-DSA, which generates digital signatures.
- **Quantum-Safe Foundation:** Defends against potential future threats from quantum computers with quantum-safe cryptographic algorithms.

To support clients applying and taking advantage of these capabilities, several software components are available that can strengthen the security posture of enterprise AI:

- **IBM Synthetic Data Sets:** Protects sensitive private data by training with pre-built artificial data, excluding real Personally Identifiable Information (PII), curated for AI use cases on IBM LinuxONE.
- **IBM Hyper Protect for Red Hat ecosystem:** Integrated part of the Red Hat ecosystem with two products, IBM Hyper Protect Container Runtime (HPCR) for Red Hat Virtualization Solutions and Hyper Protect Confidential Containers (HPCC) for Red Hat OpenShift, tailored to a wide range of use cases. These two products secure sensitive data from development to deployment and throughout its usage in an application. Together, they form a hardware-based foundation for confidential computing in an AI solution stack.
- **IBM Secure Engineering & IBM Elite Support for Open-Source AI Frameworks:** Vetted and scanned open-source AI-serving frameworks delivered via IBM-certified containers for security vulnerabilities. These can be paired with IBM Elite Support to offer additional support for deploying open-source AI in a production environment.
- **AI Governance:** Drive transparency with AI lifecycle governance and risk management across the entire AI model lifecycle with IBM Cloud Pak for Data or watsonx-based governance on IBM LinuxONE.

Running AI Applications on IBM LinuxONE 5

IBM LinuxONE 5 has the potential to reshape the enterprise Linux IT landscape, enabling businesses to deploy both traditional and generative AI models directly where their data resides, minimizing latency and increasing performance, and all within a highly scalable and secure environment.

The LinuxONE ecosystem includes a range of AI solutions from ISVs, IBM, and the open-source community. Predictive AI can run encoder LLMs in real-time for every transaction. The on-chip Telum II AI accelerator handles this workload at low latency and high throughput.

Multiple-model AI applications combine traditional AI models with generative AI on every transaction by offloading the more computationally intensive generative AI to the Spyre Accelerator. And, of course, larger LLMs demanding dedicated decoder hardware can run on one or more Spyre Accelerators.

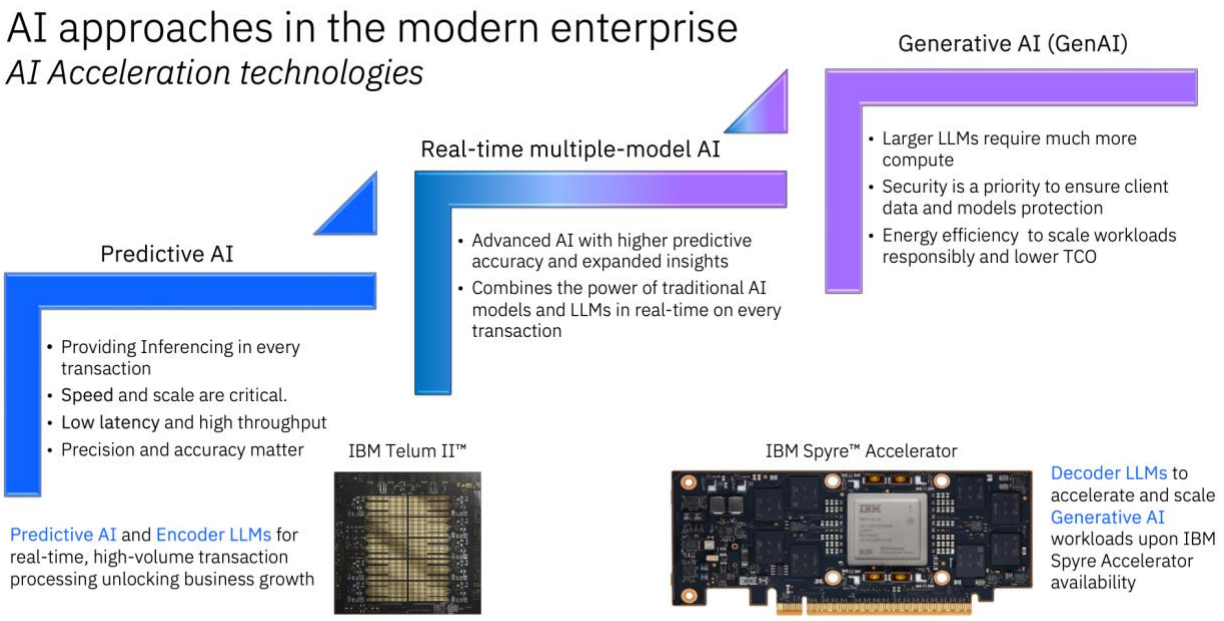

Figure 6: The range of AI applications.

IBM is also very active in helping ISVs develop their solutions on LinuxONE. A good example is Hazelcast, an ISV focused on enterprise AI applications, which has recently partnered with IBM to enable real-time data processing and automation for the AI age.

"Hazelcast on IBM LinuxONE brings together two enterprise-grade IT leaders to deliver extreme performance, ironclad resilience, and consistent security from core to cloud," said Eric Bochner, CEO of Hazelcast. *"The joint stack inherits LinuxONE's fault-tolerant hardware and Hazelcast's 'write once, run everywhere' portability, enabling seamless deployment across mixed environments without code changes. The stack helps developers modernize and scale AI pipelines and agentic workloads on a secure enterprise platform while preserving governance and uptime."*

AI Use Cases for LinuxONE

Here are some example workloads where LinuxONE can deliver the AI performance and total cost of ownership (TCO) that enterprises demand. The first few are where multiple AI models collaborate to achieve the desired results.

Predictive and Multiple AI Model Examples

Advanced Fraud Detection (AFD)

AFD is perhaps the most frequently cited example of an enterprise-class server utilizing AI to detect potential fraud. IBM LinuxONE 5 can help prevent financial fraud with an advanced multiple AI model technique, a challenge that cost banks $448 billion globally in 2024[ii]. With AI on LinuxONE, it is possible to review every transaction in real-time. IBM LinuxONE 5 delivers enhanced AI inferencing capacity, enabling the processing of 50% more transactions within a single day using a Credit Card Fraud Detection model compared to IBM LinuxONE 4.[iii] The AI accelerator solutions with Telum II can identify sophisticated fraud attempts using traditional AI and augment them with generative AI from Spyre to mitigate significant financial losses associated with high-risk transactions. A reduction of a few percentage points can potentially save a financial institution millions in losses.

Anti-Money Laundering

AI can enhance AML screening to reduce false positives and deliver substantial cost savings for financial institutions, resulting in reduced AML compliance and manual investigation expenses. With Telum II, institutions can mitigate regulatory risks by leveraging multiple AI models in a combined approach, resulting in a more accurate AML screening process.

Insurance Claims Processing

IBM LinuxONE 5 can help prevent insurance fraud with advanced multiple AI model techniques, a challenge that costs the insurance industry $83 billion globally.[iv] With multi-model AI, LinuxONE 5, using Telum II and Spyre together can optimize and accelerate claim processing by prioritizing urgent claims and leveraging additional insights from unstructured data to reduce losses. AI can reduce processing times from days to hours or minutes, thereby reducing manual effort and accelerating claims processing. Increased speed and throughput with LinuxONE 5 can also address scalability requirements, enabling the processing of a high volume of claims and resulting in enhanced customer satisfaction and retention.

Procurement and Supply Chain Risk Management

Contract compliance and risk management for procurement teams can enhance traditional AI methods with large language models (LLMs). LinuxONE 5 can identify potential risks by analyzing contract text that contains ambiguous terms and regulatory inconsistencies, such as pricing discrepancies or missing clauses, to ensure compliance. This enables organizations to reduce legal exposure and the risk of loss while ensuring stronger governance, supplier adherence to contract terms, and compliance across the supply chain.

Generative AI Examples

Here are a few workloads that can benefit from running generative AI on the Spyre accelerator, keeping enterprise data and analytics secure.

Tax Document Processing

GenAI can summarize submitted tax documents and highlight key findings, such as mismatched income reporting and unusually high deductions. In the US alone, this is estimated to enable potential annual savings of more than $5 billion from fraud.

Financial Documents Processing

GenAI can also summarize financial documents and business reports to extract key data points, such as financial metrics and performance indicators, and identify essential information needed for compliance processes and financial audits, freeing up significant time to handle priority details, with estimates suggesting a 30% to 70% reduction in time spent on these tasks.

Information Search and Data Extraction

AI can extract key points from content stored as unstructured text and identify essential information needed for compliance processes, such as financial audits and regulatory compliance.

Customs Screening for Suspicious Cargo

Customs organizations can utilize generative AI to identify potentially suspicious cargo through sophisticated image processing techniques and analysis of textual descriptions associated with each shipment.

Conclusions

IBM LinuxONE 5 offers a compelling alternative to traditional Linux servers for organizations that demand maximum security, reliability, and efficiency at scale. While the upfront investment may be higher, the platform's ability to consolidate workloads,

reduce operational costs, and deliver unmatched uptime makes it a strong choice for mission-critical environments. The total cost of ownership can be lower than alternatives, especially when one considers that LinuxONE can mitigate the costly impacts of outages and security breaches.

The addition of shared on-processor and PCIe-based AI accelerators, enabled with IBM and open-source AI software, creates an enterprise platform that brings AI to where organizations run their businesses, and does so securely with 99.999% availability at scale.

When security, scale, availability, and TCO are paramount requirements, the IBM LinuxONE 5 with new AI capabilities offers a compelling alternative to traditional server consolidation approaches for mission-critical AI applications and models.

[i] Performance results are based on IBM® internal tests running on IBM Systems Hardware of machine type 9175. The OLTP application (https://github.com/IBM/megacard-standalone) and PostgreSQL was deployed on the IBM Systems Hardware. The Credit Card Fraud Detection (CCFD) ensemble AI setup consists of two models (LSTM: https://github.com/IBM/ai-on-z-fraud-detection, TabFormer: https://github.com/IBM/TabFormer). On IBM Systems Hardware, running the OLTP application with IBM Z Deep Learning Compiler (zDLC) compiled jar and IBM Z Accelerated for NVIDIA® Triton™ Inference Server locally and processing the AI inference operations on cores and the Integrated Accelerator for AI versus running the OLTP application locally and processing remote AI inference operations on a x86 server running NVIDIA Triton Inference Server with OpenVINO™ runtime backend on CPU (with AMX). Each scenario was driven from Apache JMeter™ 5.6.3 with 64 parallel users. IBM Systems Hardware configuration: 1 LPAR running Ubuntu 24.04 with 7 dedicated cores (SMT), 256 GB memory, and IBM FlashSystem® 9500 storage. The Network adapters were dedicated for NETH on Linux. x86 server configuration: 1 x86 server running Ubuntu 24.04 with 28 Emerald Rapids Intel® Xeon® Gold CPUs @ 2.20 GHz with Hyper-Threading turned on, 1 TB memory, local SSDs, UEFI with maximum performance profile enabled, CPU P-State Control and C-States disabled. Results may vary.

[ii] Banking industry fraud numbers are from the Celent paper 'Mitigating fraud in the AI age' which was commissioned by IBM.

[iii] Both IBM LinuxONE 5 and IBM LinuxONE 4 feature eight cores and one Integrated Accelerator for AI. On both systems, 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.

- With IBM LinuxONE 5, process up to 450 billion inference operations per day with 1 ms response time. Performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. 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. Results may vary.
- With IBM LinuxONE 4, process up to 300 billion inference requests per day with 1ms response time. Performance result is extrapolated from IBM internal tests running on IBM LinuxONE 4 LPAR with 48 IFLs and 128 GB memory on Ubuntu 20.04 (SMT mode). The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The lscpu command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results may vary.

[iv] Insurance industry fraud numbers are from the Celent paper 'Mitigating fraud in the AI age' which was commissioned by IBM.

Disclosures: This white paper was sponsored by IBM, expresses the opinions of the author, and is not to be taken as advice to purchase from or invest in the companies mentioned. My firm, Cambrian-AI Research, is fortunate to have many semiconductor firms as our clients, including Baya Systems BrainChip, Cadence, Cerebras Systems, D-Matrix, Esperanto, Flex, Groq, IBM, Intel, Micron, NVIDIA, Qualcomm, Graphcore, SImA.ai, Synopsys, Tenstorrent, Ventana Microsystems, and scores of investors. I have no investment positions in any of the companies mentioned in this article. For more information, please visit our website at https://cambrian-AI.com.