

LES NOUVEAUTÉS
DU Z17 ET
LINUXONE V
NOVIPRO 2025



PRÉSENTATION

Eric Cothenet

Concepteur de
solutions d'entreprise

IBM Champion pour
LinuxONE et Linux on z

Proud to be an IBM Champion 2025!



NOVIPRO INC.

Équipe de vente IBM z



Luc Archambault

**Spécialiste vente -
System Z Hardware**

Responsable de
la plateforme
IBM z et
stockage.



Gilbert Fortin

**Spécialiste vente -
System Z Software**

Responsable des
logiciels IBM z,
conseiller sur les
ELA et assurer le
suivi des
renouvellements

NOVIPRO INC.

Concepteurs de solution et spécialistes IBM z



Eric Cothenet

Concepteur de solutions – IBM System Z & LinuxONE

["IBM LinuxONE Champion"](#)

Conseiller à l'interne et externe pour IBM z et LinuxONE



Karoline Pierre

Administrateur IBM z et stockage

Advocating internally and externally for System Z customers and prospects.



Ugo Perri

Concepteur de solutions mainframe

Conception d'architecture et livraison des services pour IBM z



Pierre Hébert

IBM High-End Storage Architect

Conception d'architecture et livraison des services pour Stockage System Z (DS8K, VTS, Tapes).

OBJECTIFS



Toujours plus vite et plus fort



Encore plus loin pour l'Intelligence Artificielle

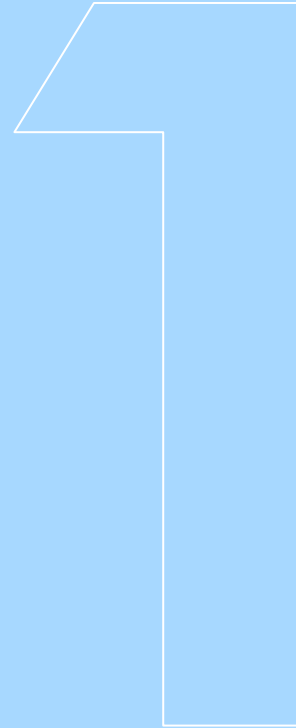


Un vrai système Quantum Safe

AGENDA

1. Quoi de neuf
2. On gagne quoi?
3. AI Inference, IBM Spyre
4. Sécurité
5. Démo

QUOI DE NEUF



IBM Z17 ET LINUXONE

Date d'annonce: 8 avril 2025

Toujours **99.999999 %** de disponibilité ce qui équivaut à 0,32s d'indisponibilité par année

Capable d'exécuter **35 milliards** de transaction OLTP **encrypté** par jour

Jusqu'à **3 000 000** NGINX containers dans un système sous RedHat Linux avec KVM

IBM Spyre Accelerator pour exécuter du code « Generative AI »

Jusqu'à **450 milliards** opérations d'inférence par jour dans un système (en conjonction du processeur Telum II et de la carte d'accélération IBM Spyre)

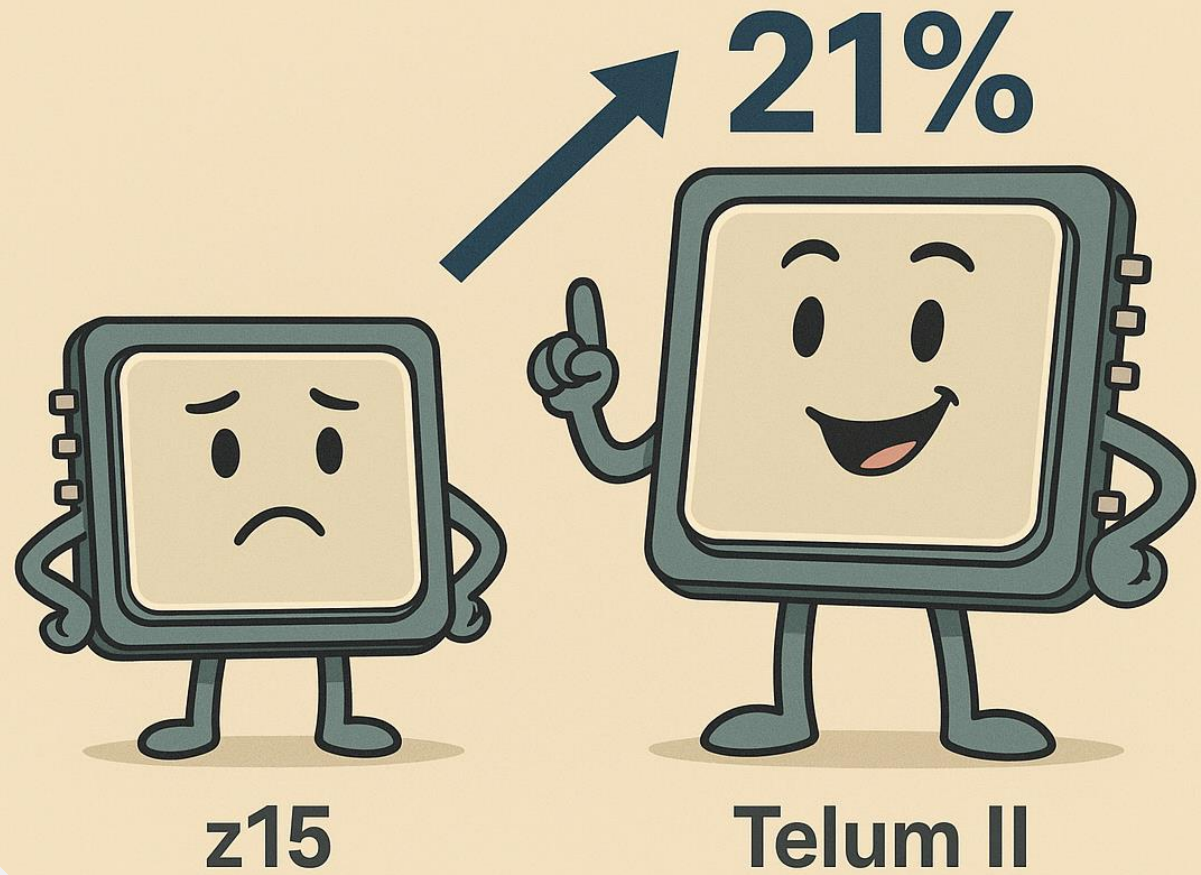
On gagne quoi?

2

IBM Z15 VS Z17

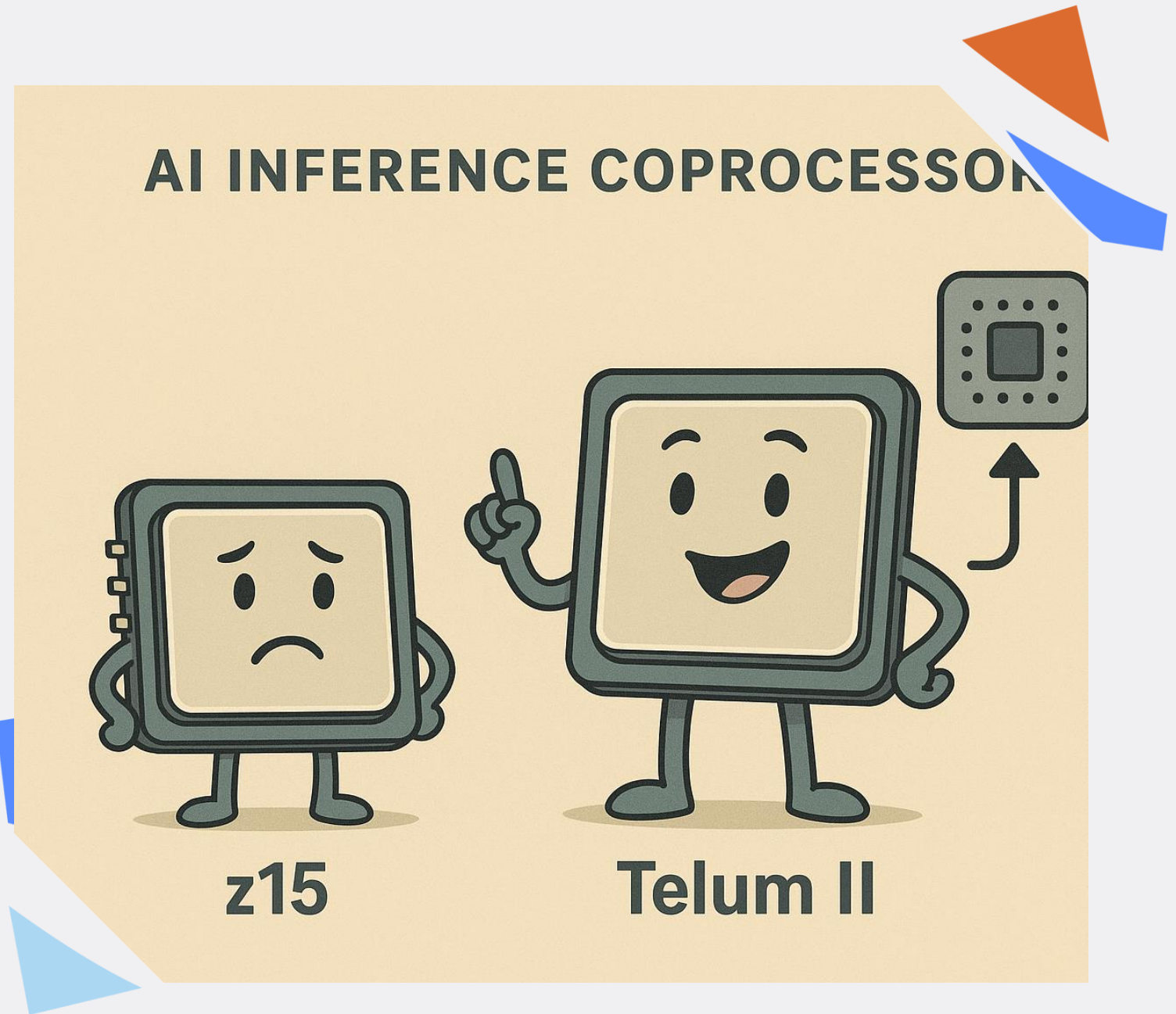
Toujours plus fort

PERFORMANCE PER THREAD



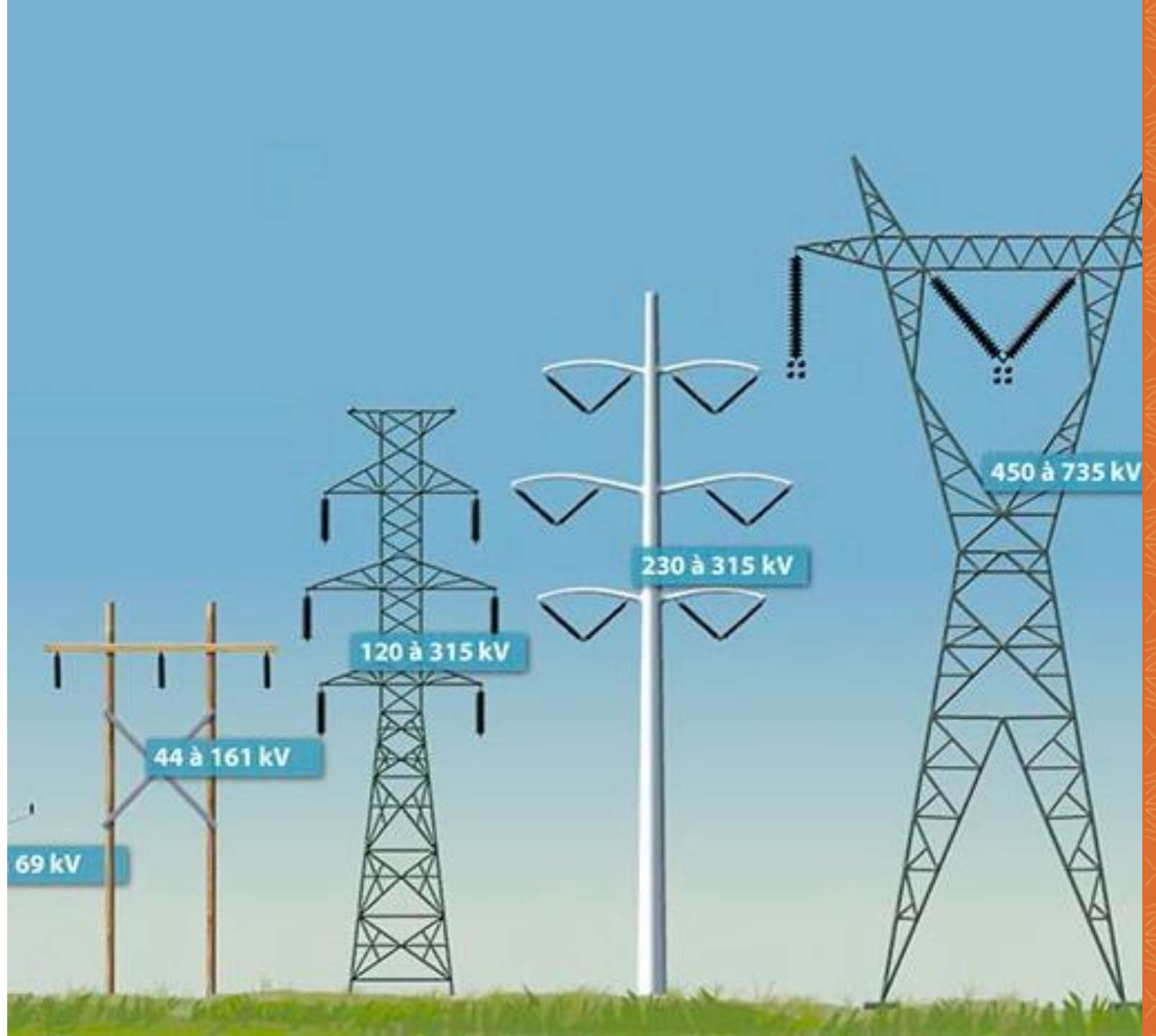
IBM z15 vs z17

Intelligence Artificielle



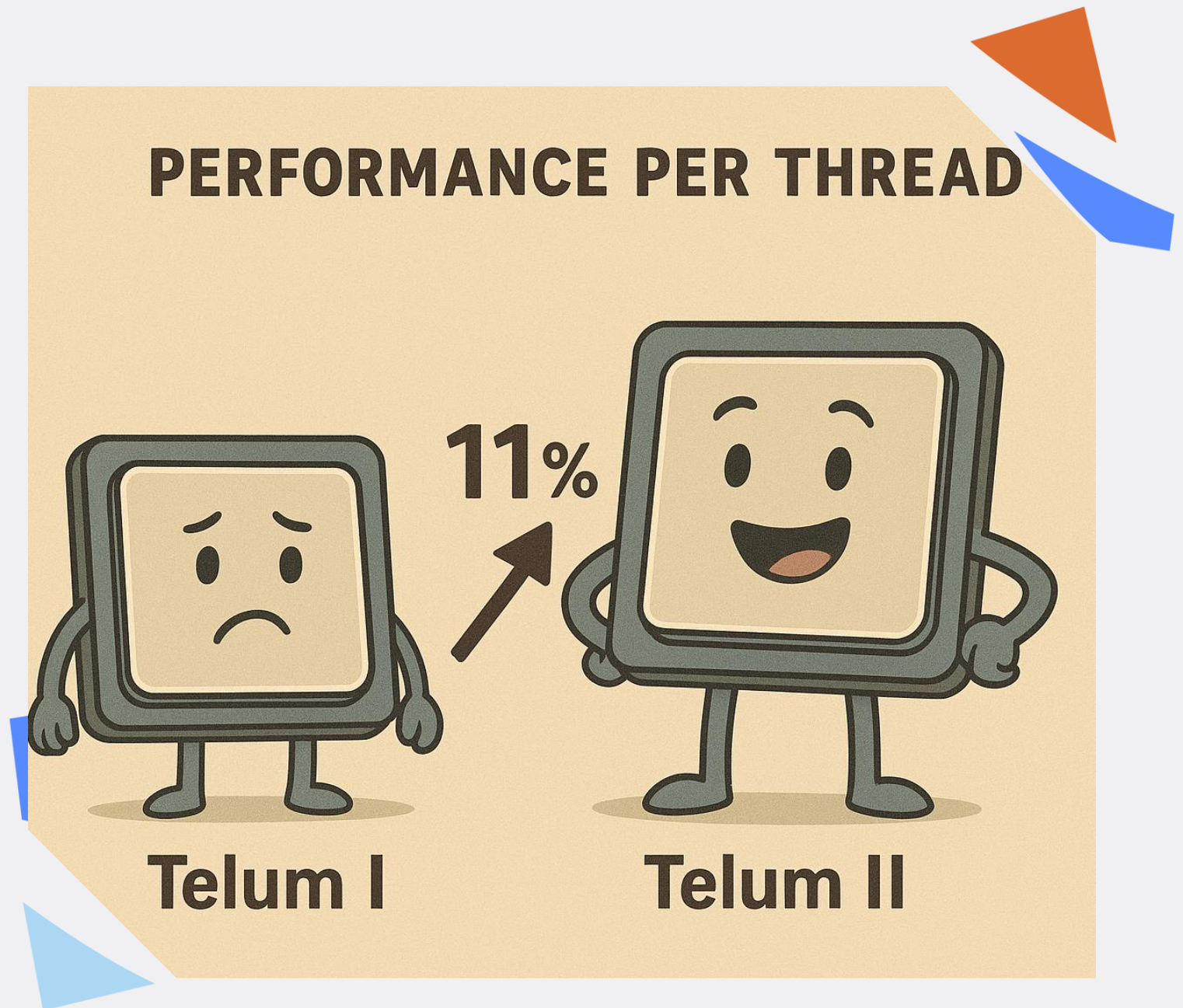
Réduction
d'électricité
I/O Subsystem

-70%



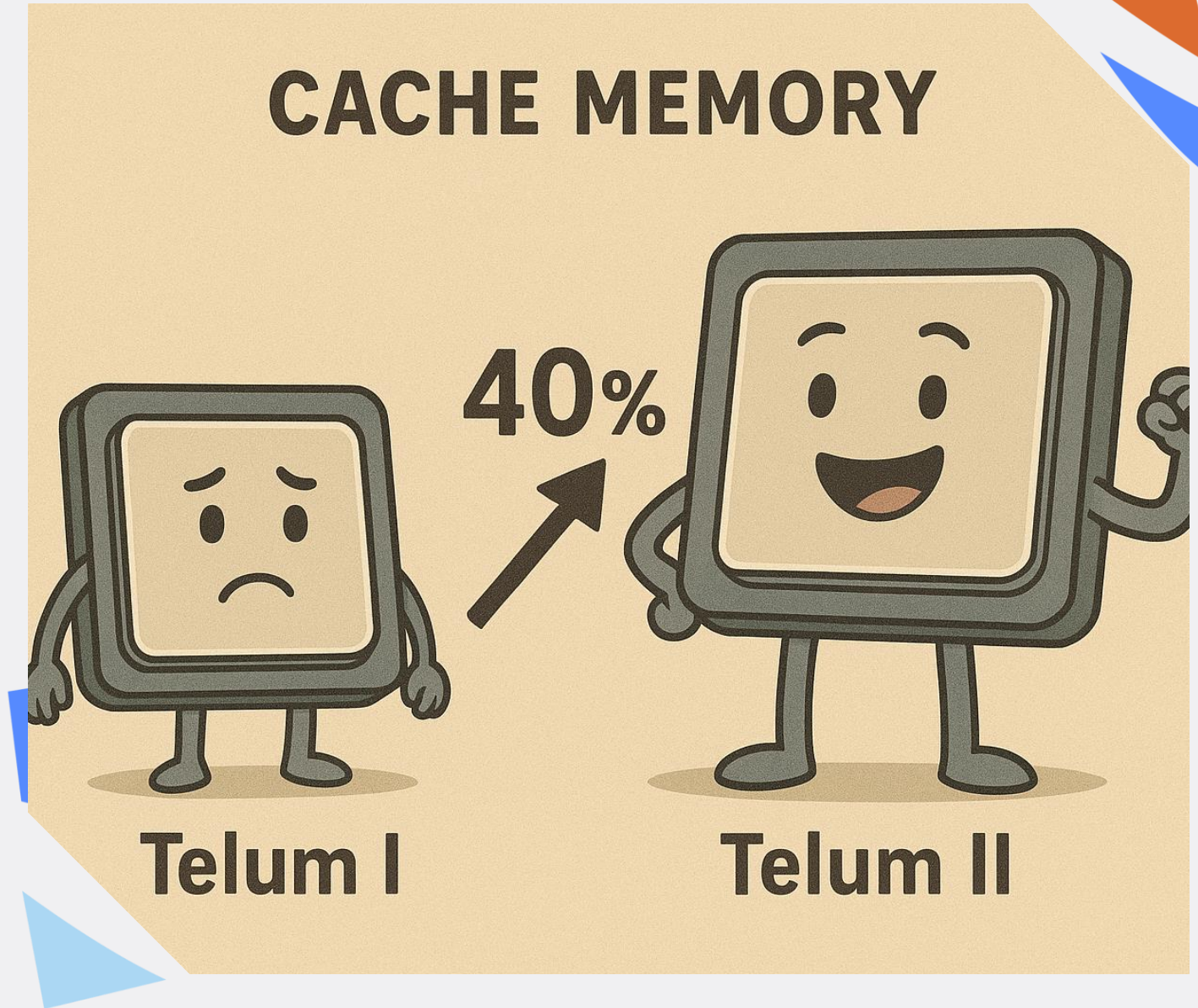
IBM Z16 VS Z17

Toujours plus fort



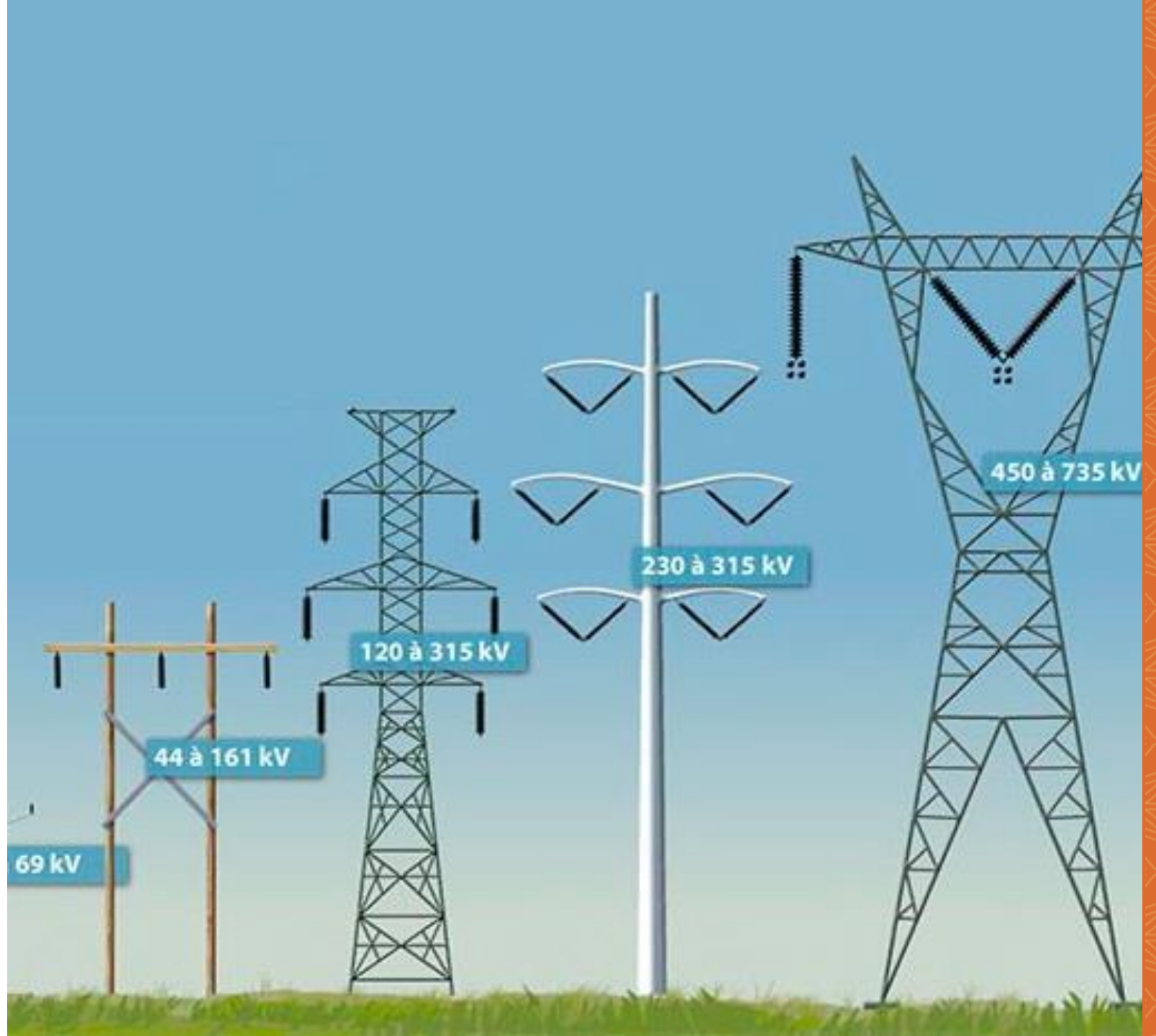
IBM Z16 VS Z17

Toujours plus vite



Réduction
d'électricité
I/O subsystem

-70%



10 - 36MB L2 caches

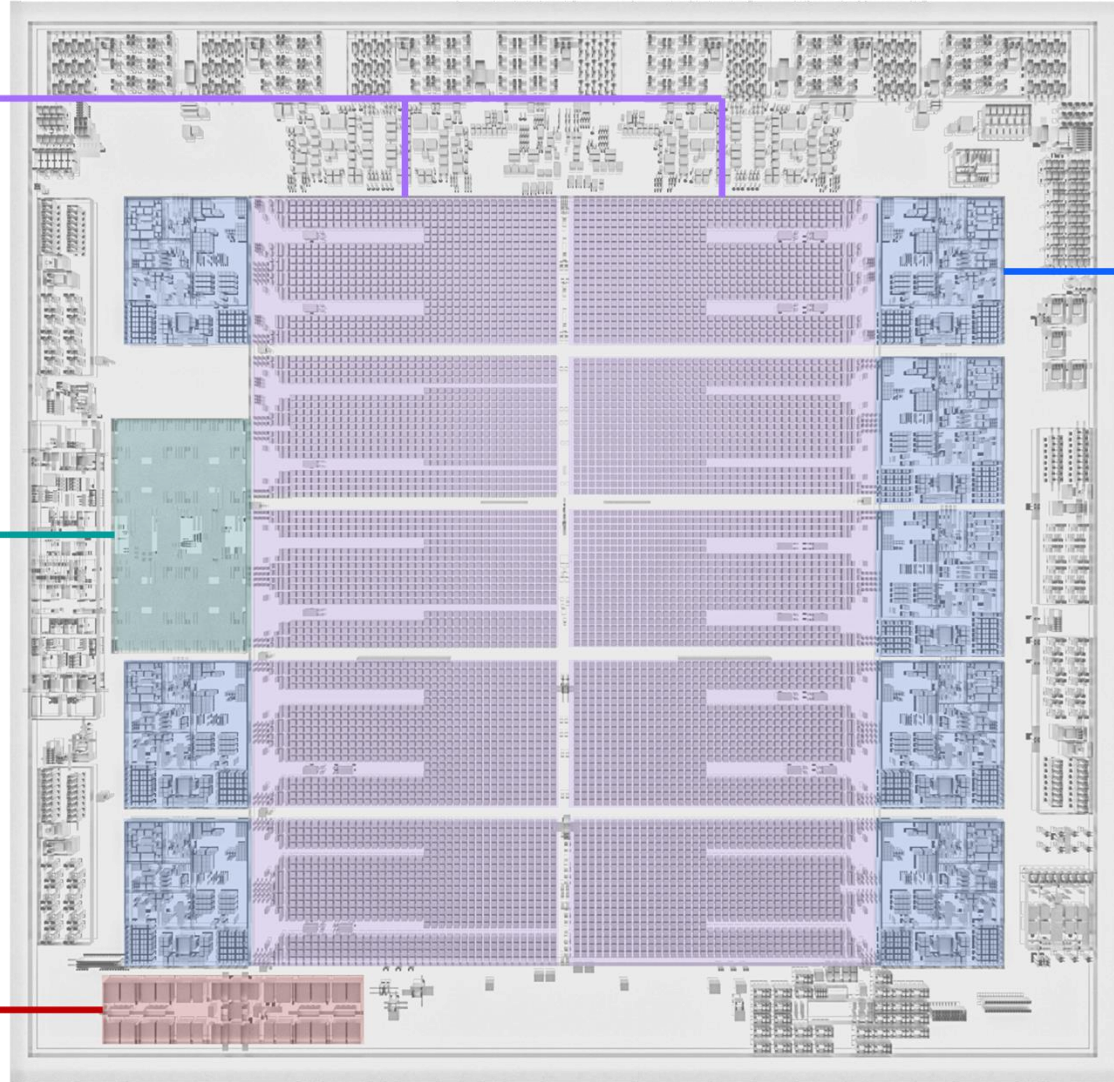
- 40% more virtual L3 and virtual L4 cache per core compared to IBM z16

I/O Data Processing Unit

- Redesigned I/O subsystem resulting in power and data center footprint reduction

2nd gen AI accelerator

- Improved quantization and matrix operations
- 8x accelerators available per core



8 - 5.5GHz cores

- 11% increase in single thread performance
- 20% area reduction
- 15% power reduction

Process up to 450 billion inference operations per day with 1 ms response time – a 50% increase over IBM z16

AI Inference IBM Spyre



L'écosystème de l'Intelligence Artificielle dans IBM z et LinuxONE



Business Insights

Infuse AI into every transaction *in real time*



Machine Learning for IBM z/OS

Deliver AI solutions at an unprecedented speed



IBM Db2 for z/OS with SQL Data Insights

Uncover hidden patterns from data locked in z/OS



IBM Cloud Pak for Data on IBM Z

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



AI Toolkit for IBM Z and IBM 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-generation protection for the most crucial data



IBM Db2 AI for z/OS

Enhance database performance with ML



IBM Concert for Z

Proactively identify and mitigate Ops issues



AI - Infused IBM z/OS

Enable intelligent administration and automation



IBM watsonx Assistant for Z

Use conversational AI to automate tasks and build skills



IBM watsonx Code Assistant for Z

Accelerate mainframe application modernization



Snap ML



spaCy



TensorFlow



Apache Spark



Keras



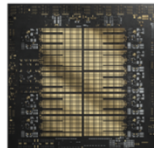
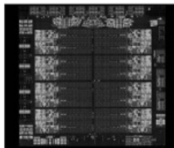
MAGE



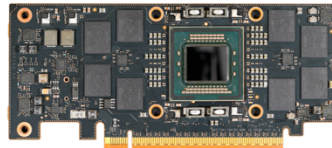
RED HAT OPENS SHIFT

mlflow

Enable market-leading open-source AI tools & frameworks on IBM Z



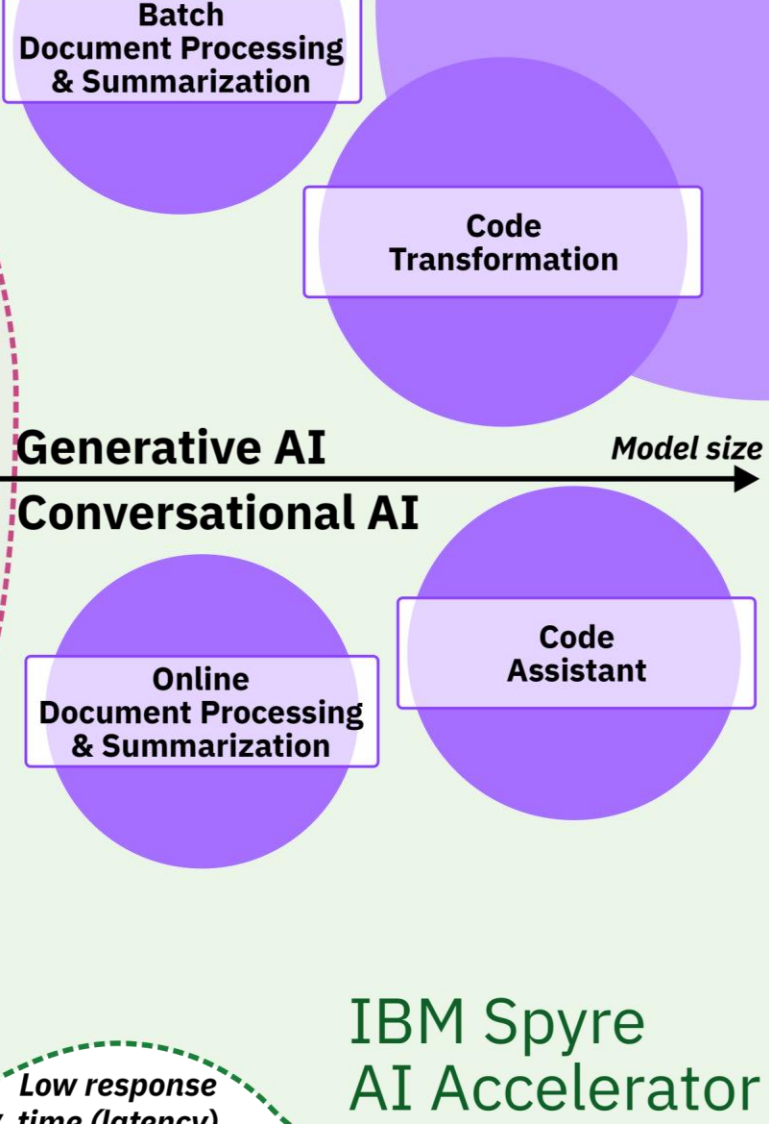
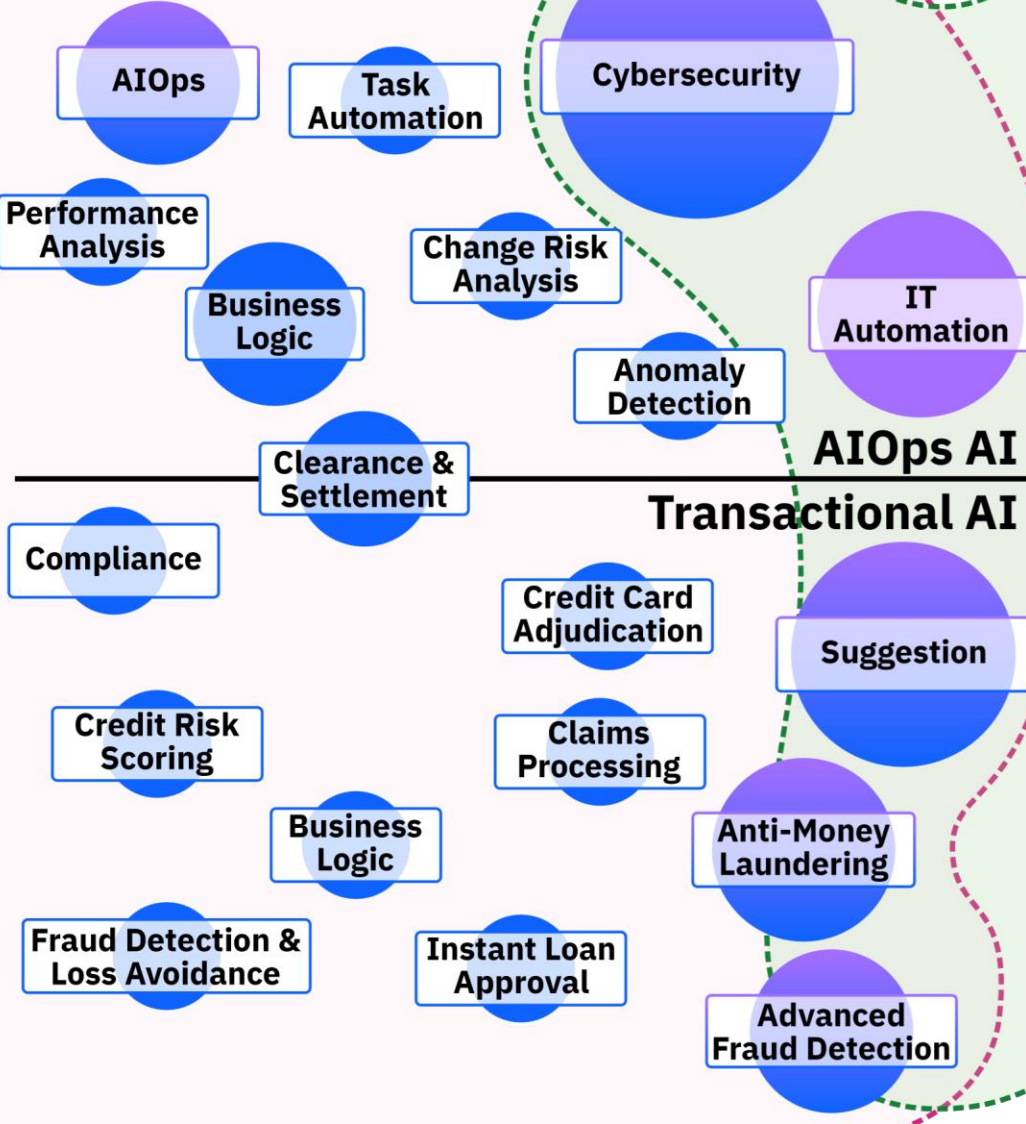
Telum I and Telum II
On-chip AI accelerator



Spyre Accelerator (available starting in 4Q 2025)
Host attached AI accelerator

Process billions of inference requests per day, in real time

IBM Telum II Processor



Legend

- Compute & bandwidth needed → Large
 Small
- Predictive AI →
- Multiple Model AI →
- Generative AI →

Sécurité



SECURITÉ

QUANTUM SAFE

The advent of quantum computing poses a significant threat to some of the classical cryptographic methods protecting our current systems and data.

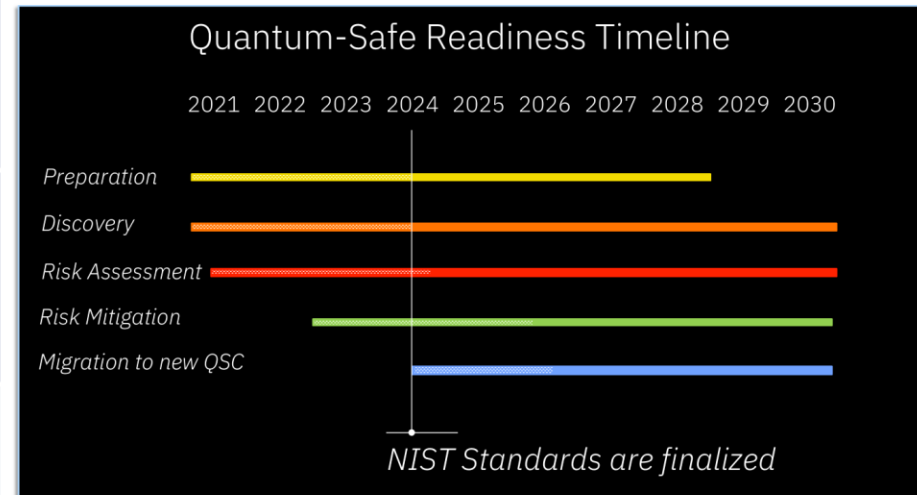
Organizations rely heavily on secure communication for transactions, client information, and regulatory compliance.

It will take from 5 to 15 years to migrate to quantum safe options, so start now!



Quantum-Safe Readiness

Preparation	<ul style="list-style-type: none"> Educate teams and/or Security Stakeholders Follow standards community and quantum-safe computing Research migration best practices Redbook : Transitioning to Quantum-Safe Cryptography on IBM Z
Discovery	<ul style="list-style-type: none"> Create a crypto inventory (reusable security asset) <ul style="list-style-type: none"> Inventory cryptographic assets and cryptography use Inventory data handled by the organization Inventory suppliers of cryptographic assets Aid crypto discovery with existing and new tools for IBM z17 <ul style="list-style-type: none"> Crypto Discovery and Inventory Tool
Risk Assessment	<ul style="list-style-type: none"> Perform Gap Analysis to discover security risks Understand internal and external dependencies Leverage risk assessment reports to prioritize your execution roadmap Quantum Safety Assessment with IBM Expert Labs
Risk Mitigation	<ul style="list-style-type: none"> Reconsider and/or possibly redesign how crypto is consumed in the environment Determine best mitigation action: retire it, accept it, or fix it Quantum Safety Assessment with IBM Expert Labs
Migration to Post-Quantum Crypto (PQC) / Quantum-Safe	<ul style="list-style-type: none"> Update old and build new applications, leverage application transparent technology z17, Crypto Express 8S, ICSF, Java on z/OS Security, and more Protection of data at rest with Pervasive Encryption Protection of data in flight with AT-TLS



Démo



Démo interactive

<https://www.ibm.com/demos/it-infrastructure/index.html#C178>

Groupe zMeetup

<https://www.meetup.com/groupe-meetup-montreal-mainframe-et-linuxone/>

Liste des projets opensources

<https://community.ibm.com/zsystems/oss/>

OBJECTIFS



Toujours plus vite et plus fort



Encore plus loin pour l'Intelligence Artificielle



Un vrai système Quantum Safe

Différence matérielle

	z15	z17	z17 differentiation
Processor Chip	14nm technology 5.2GHz 12 cores per chip 14.5 miles of wire per chip 9.2B transistors	5nm technology 5.5 GHz 8 cores per chip 24.1 Miles of wire per chip 43B transistors	Core processor instruction set and performance improvements +21% single thread performance
Capacity	190 cores	208 cores	+35% capacity
Cache	4MB L2, 256MB L3, 960MB L4	36MB L2, 360MB Virtual L3, 2.8GB virtual L4	+200% total cache growth
Total system memory	40TB max memory	64TB max memory 4U DDR5 DDIMM	+60% system memory growth
On-chip AI acceleration	No AI acceleration	Telum II 2 nd -gen on-chip AI accelerator	On-processor ultra-low latency AI inference engine Up to 8x on-chip AI processors available per core Support for LLM compute primitives Improved quantization: Int8, FP16 datatypes Multi-model real-time AI on 100% of transactions
Off-chip PCIe Accelerator card	Not available	Spyre AI accelerator card 32 Gen AI-ready cores Up to 48 PCIe gen5 x16 adapter	Available only on IBM z17 On-prem watsonx Code Assistant and watsonx Assistant Supported LLMs, generative, and agentic AI with Spyre
System I/O	Off-chip IO Processing	On-processor chip IO DPU	70% Reduced power for IO management Double density (Up to -21% system power for I/O rich configs) <ul style="list-style-type: none"> • FICON Express 32G (4-port) • Converged network adapter (RoCE and OSA)
Other features	Not available	Quantum-Safe Secure Boot Memory Encryption Flexible Capacity for Cyber Resiliency	Significantly improved security and compliance Infrastructure and compute agility Consumption-based model for flexible compute

Différence matérielle

	z16	z17	z17 differentiation
Processor Chip	7nm technology 5.2G Hz 8 cores per chip 18.8 Miles of wire per chip 22.5B transistors	5nm technology 5.5 GHz 8 cores per chip 24.1 Miles of wire per chip 43B transistors	20% core processor area reduction 15% power reduction +11% single thread performance
Capacity	200 cores	208 cores	+12-20% capacity
Cache	32MB L2 256MB virtual L3 2GB virtual L4	36MB L2 360MB Virtual L3 2.8GB virtual L4	+40% cache growth
Total system memory	4U DDR4 DDIMM 40TB max memory	4U DDR5 DDIMM 64TB max memory	+60% system memory growth
On-chip AI acceleration	Telum I	Telum II	More AI processing per chip Up to 8x on-chip AI processors available per core Support for LLM compute primitives Improved quantization: Int8, FP16 datatypes
Off-chip PCIe Accelerator card	N/A	Spyre accelerator 32 Gen AI-ready cores on extended adapters 75W PCIe gen5 x16 adapter Up to 48 adapters per system	Available only on IBM z17
System I/O	Off-chip IO processing	On-processor chip IO DPU	70% Reduced power for IO management Double density with new FICON Express 32G (4-port) Double density with converged network adapter (RoCE and OSA)

Un monstre de IO

- Une latence réduite grâce à la cohérence mémoire.
- Une meilleure efficacité énergétique.
- Une meilleure performance globale du système, surtout pour les workloads transactionnels à haut débit.
- Workloads IA + transactionnels en simultané.

Aspect	SAP (ancien modèle)	DPU (Telum II)
Localisation	Externe au CPU	Intégré dans le die du CPU
Rôle	Gestion des I/O	Accélération I/O + cohérence mémoire
Architecture	Processeur dédié	4 clusters × 4 cœurs + cache L2
Communication	Via bus système	Directement sur le tissu SMP du CPU
Avantage principal	Décharge CPU	Latence réduite, performance accrue
Énergie	Consommation élevée	Réduction de 70 % de la consommation I/O