IBM Power

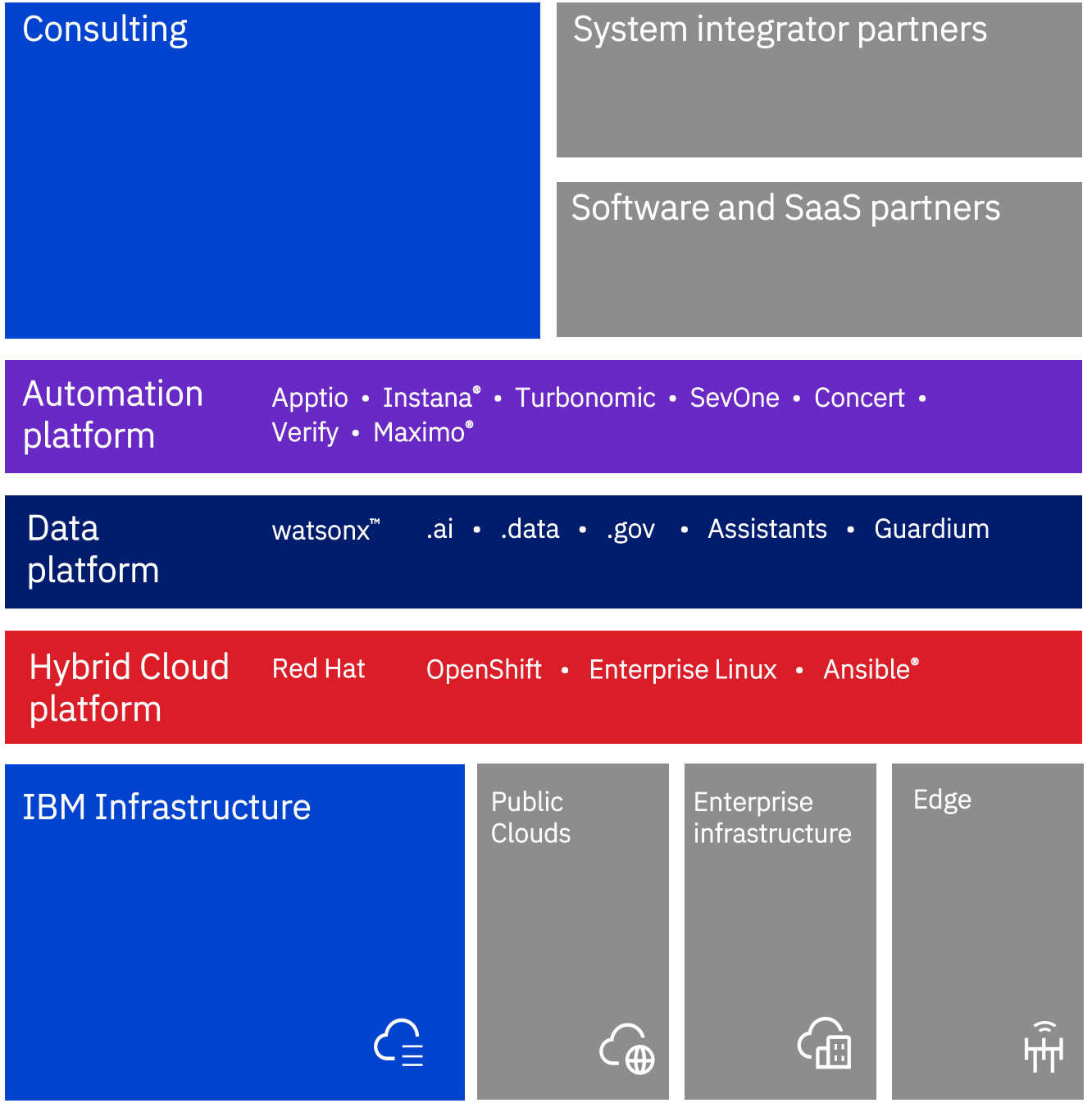
IBM Client Center, Oslo

IBM fagdag med Common

05.05.2026

Espen Skarpsno
Technical Specialist – Power/Cloud
espen.skarpsno@ibm.com

Our hybrid cloud and AI use cases deliver business value and integrate across IBM's strategy



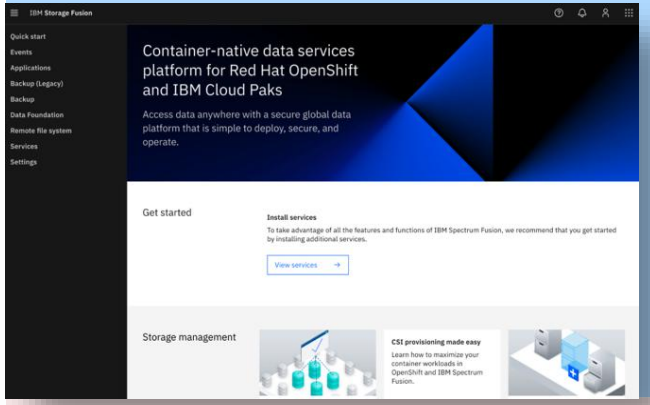
IBM Fusion is the modern platform to simplify providing and consuming modern cloud and VM applications consistently and dynamically

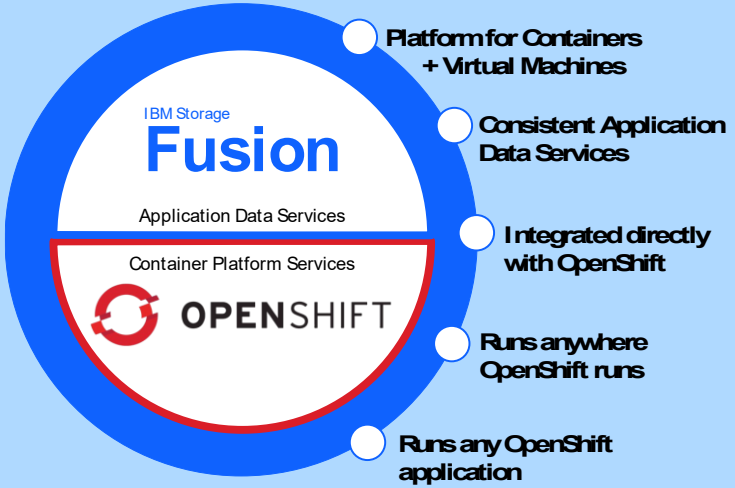
If it runs on OpenShift, it runs on Fusion

All IBM Cloud Paks | Databases | Off the shelf | Custom apps | **watsonx** Generative AI

Fusion Software

Simple - runs anywhere and everywhere





Simplified support

OPENSHIFT

aws | Microsoft Azure | IBM Cloud | vmware | IBM Z | IBM Power | Bare metal

Fusion HCI

Simplifies deployment & management

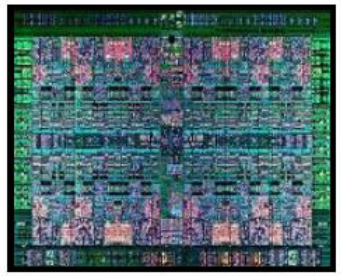
Hyperconverged Infrastructure for OpenShift

Opinionated infrastructure switches, servers, & storage

IBM Power Processor Roadmap

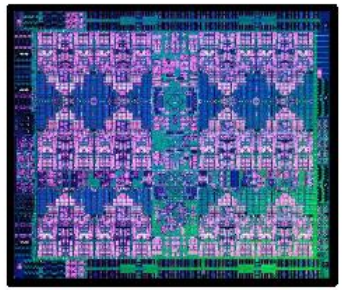
Continuous Platform Innovation and Leadership

POWER8



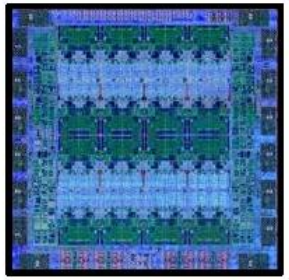
Powerful SMT8 Core
Enterprise Scaling
Big Data Optimized
Agnostic Memory

POWER9



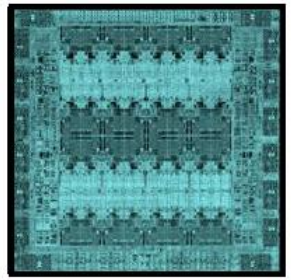
Modular Core Design
Accelerator Attach
(NVlink, OpenCAPI)
Data Plane Bandwidth
DDR & CDIMM Memory

Power10



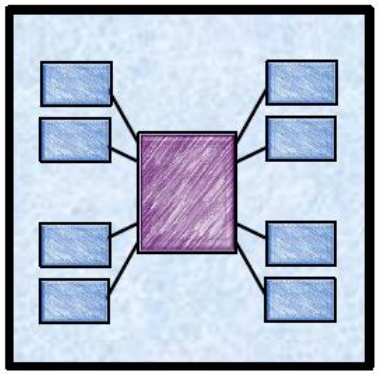
Core / Thread Strength
Socket Performance
In-Core AI Acceleration
Efficiency / Sustainability
HW Accelerated Security

Power11

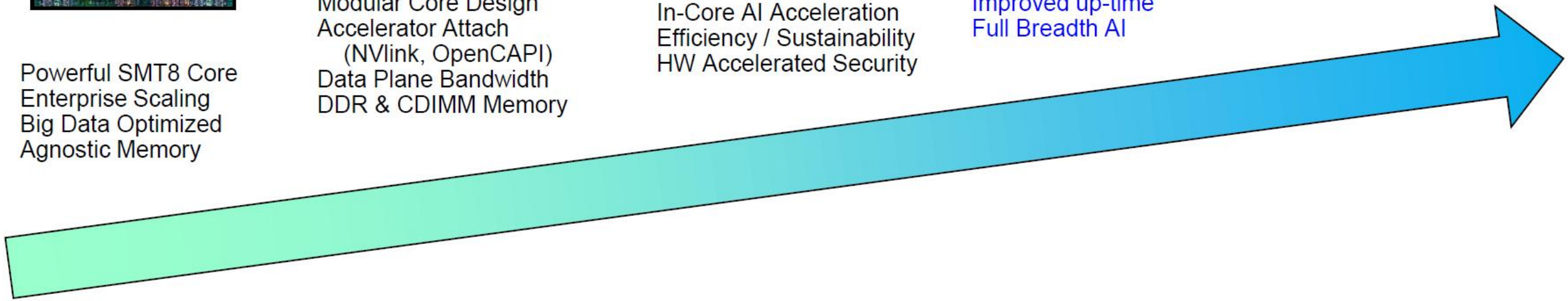


Core / Thread Strength
Socket Performance
Enterprise Scaling
Energy Optimization
Improved up-time
Full Breadth AI

Power Future



(Under development)



Information is forward looking and plans are subject to change based upon IBM's discretion

Major Eras in Computing

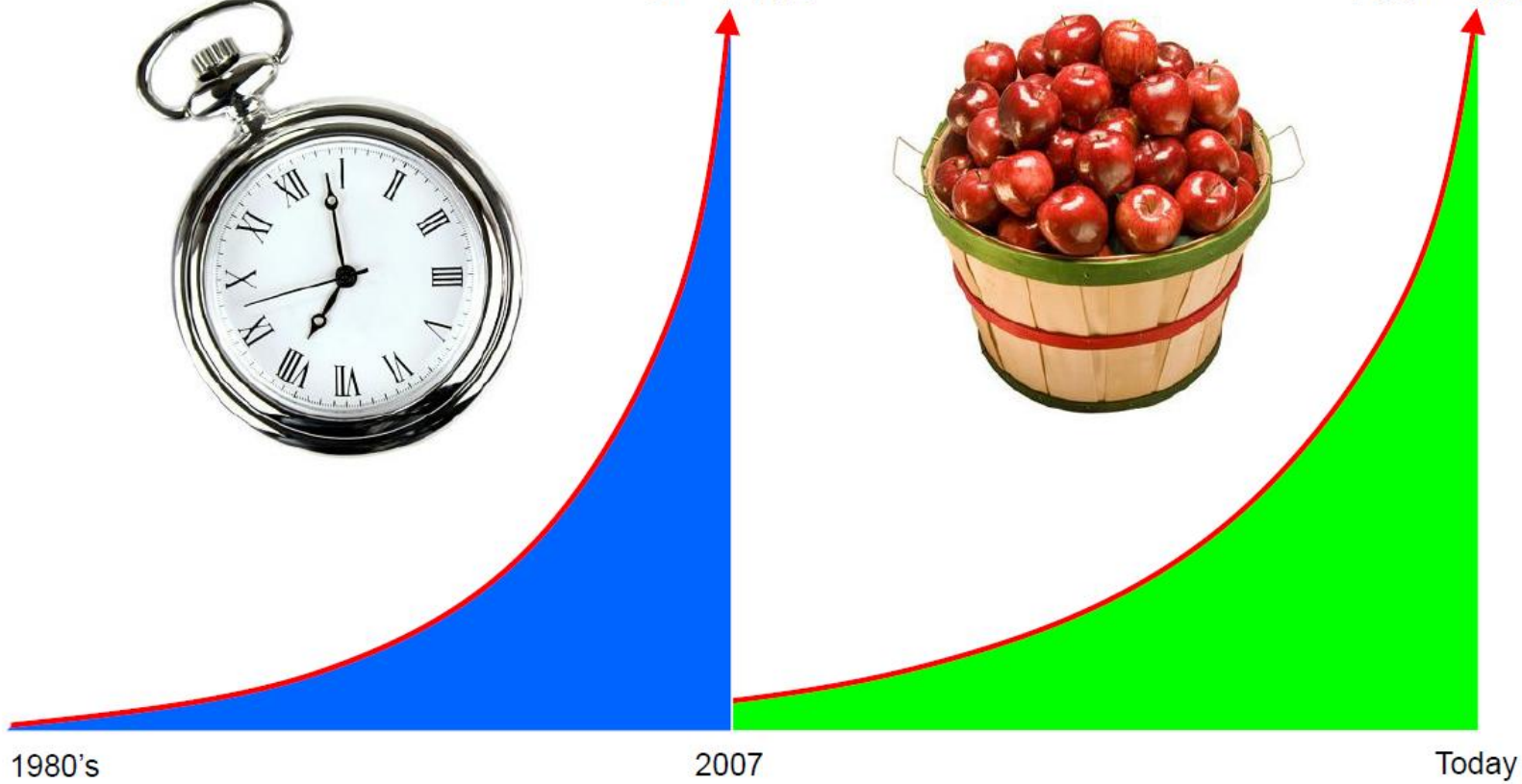
The Future

The Clock Speed Era

5+ GHz

The Multi-core Era

100+ cores



Phases of Processor / System History

The Acceleration Era?

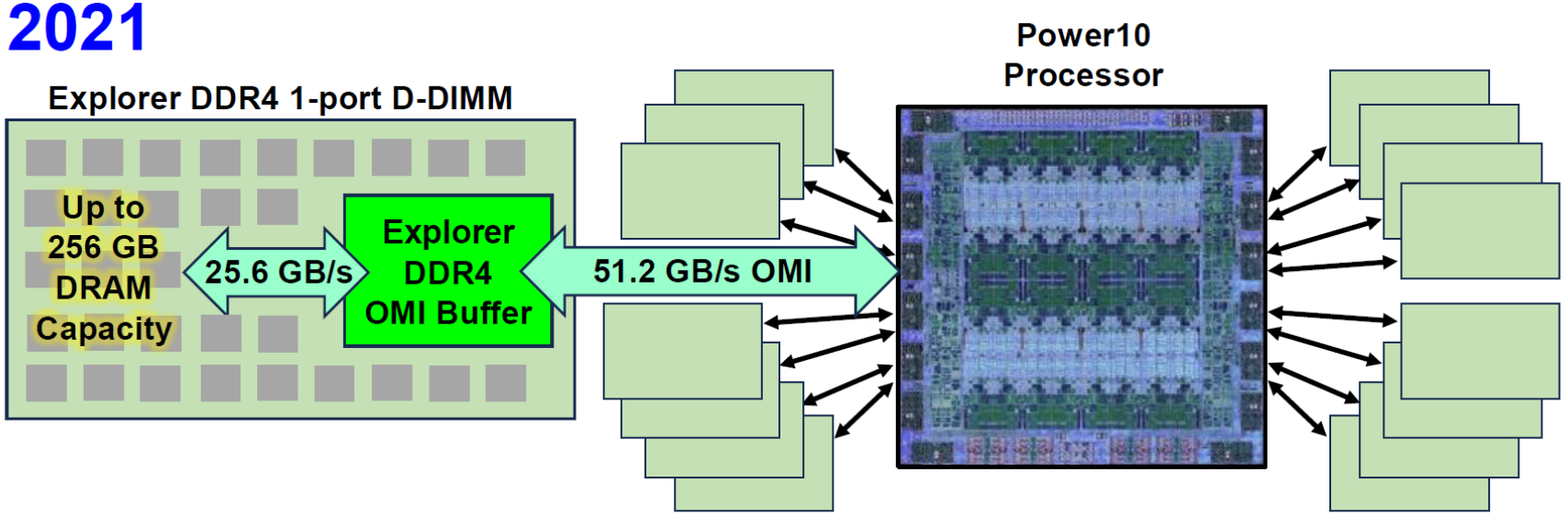
The Memory Era?

How will the Classic Von-Neumann bottleneck Manifest in the AI Era?

© Copyright IBM Corporation 2024

Memory Technology: OMI DDR5 Performance, Scaling, Economics

2021

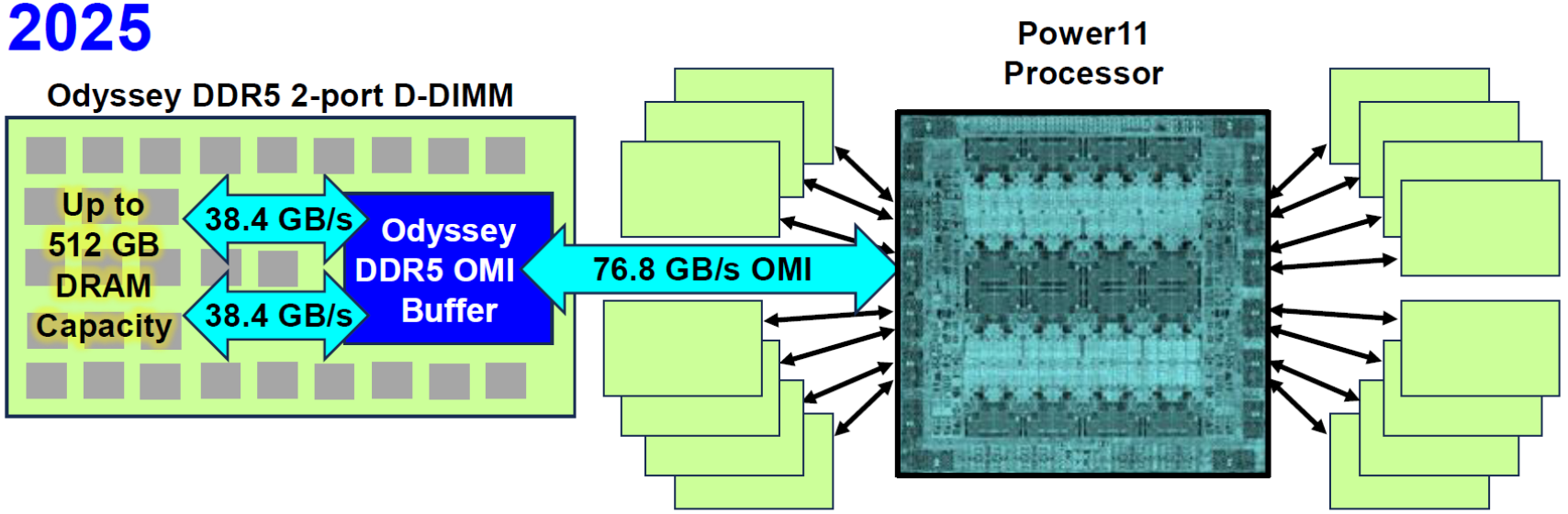


400 GB/s DRAM Bandwidth / Socket

4 Terabyte DRAM Capacity / Socket

700-800 GB/s System Coherence Flow

2025



1200 GB/s DRAM Bandwidth / Socket **3x!!!**

8 Terabyte DRAM Capacity / Socket **2x!!**

1000 GB/s System Coherence Flow **1.3x!**

Power11 - Do more, *with less*

More Performance = less hardware needed, which results in less licenses and lower operational costs

P10 -> P11	P9 -> P11	x86 -> P11
<1.5x	<2x	<4x+
CPU/Server Performance +0-10% per core	CPU/Server Performance	CPU/Server Performance
<3x	<6x	<8x+
RAM Performanc 400 -> 1200GB/s /socket	RAM Performance 200 -> 1200GB/s /socket	RAM Performance

IBM Spyre Accelerator

AI acceleration card in addition to the in-chip AI acceleration

½ TCO

Or even better ..

IBM Power E1180

The industry's most scalable server for consolidating database and mission critical workloads

IBM Power E1150

Built for large data & memory intensive workloads.

IBM Power Entry & Edge

Enterprise-class reliability, security

S/L1124, S/L1122

S1112 → GA July 2026

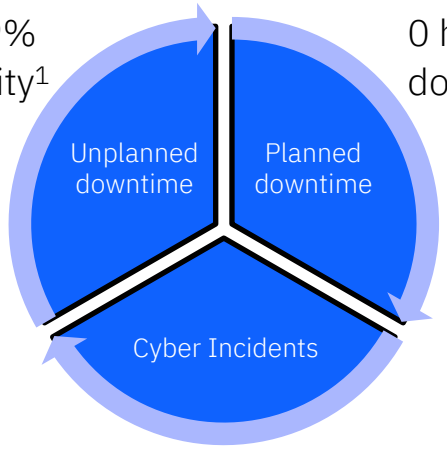
Power11 - Do more, *with less*

More Performance = less hardware needed, which results in less licenses and lower operational costs

x86 → Power - Rule of Thumb

..and why it matters..

99.9999% availability[1]



0 hours of planned downtime[2]

Secure by design
Quantum-enabled IT
<1 min guaranteed ransomware detection with automated response and recovery within minutes[3,4]

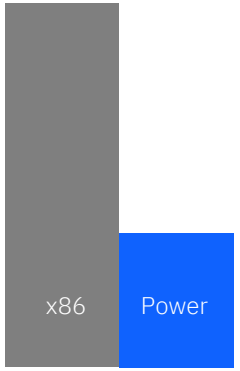
4:1

Core to Core easy sizing

½ TCO

Real world experience
Rule of Thumb put to the test

NAVIAIR



Approximately 1/3 of cost of x86 or Cloud
While achieving 3-4x improved user experience, and reducing downtime to 0 since going into production
https://www.ibm.com/case-studies/naviair

Or even better... More Performance = less hardware needed and in turn less licenses needed and lower operational costs.

1. Based upon estimated average entire-system unplanned downtime for Power E1180 systems as calculated in the Power11 Processor-Based Systems RAS (see section: Enterprise Single Server Availability Standard)
2. Based upon IBM internal testing of system upgrade scenarios; many (i.e. VIOS, hot plug adapters, I/O adapter FW, and concurrent system firmware updates) can be done in-place while some (i.e. non-concurrent system FW and HW maintenance) may require Live Partition Mobility (LPM) support.
3. This guarantee of an alert in less then one minute. Remediation is in the form of drive replacement up to the cost of the Covered Product. Terms and conditions apply; full details can be found here. (Coming to IBM i in 4Q 2025, Roadmap for PowerVS)
4. Recovery is defined as the time it takes to for a Safeguarded Copy to be available in the same Flash System or moved to another defined host server attached to that Flash System. Does not include host server recovery actions that are application dependent. Terms and conditions apply; full details can be found here

Simplify Licensing:

Streamline purchase decision, simplify license compliance, and increase flexibility to deploy capabilities

Accelerate Innovation:

Gain access to more capabilities right out of the gate to modernize your infrastructure

Reduce Total Investment:

Bundles are priced lower than piecing together individual licenses, reducing your upfront and ongoing investment.

Introducing ...

IBM i P30 Enterprise Edition

A strategic bundle that offers comprehensive IBM i enterprise capabilities

IBM i P20 Standard Edition

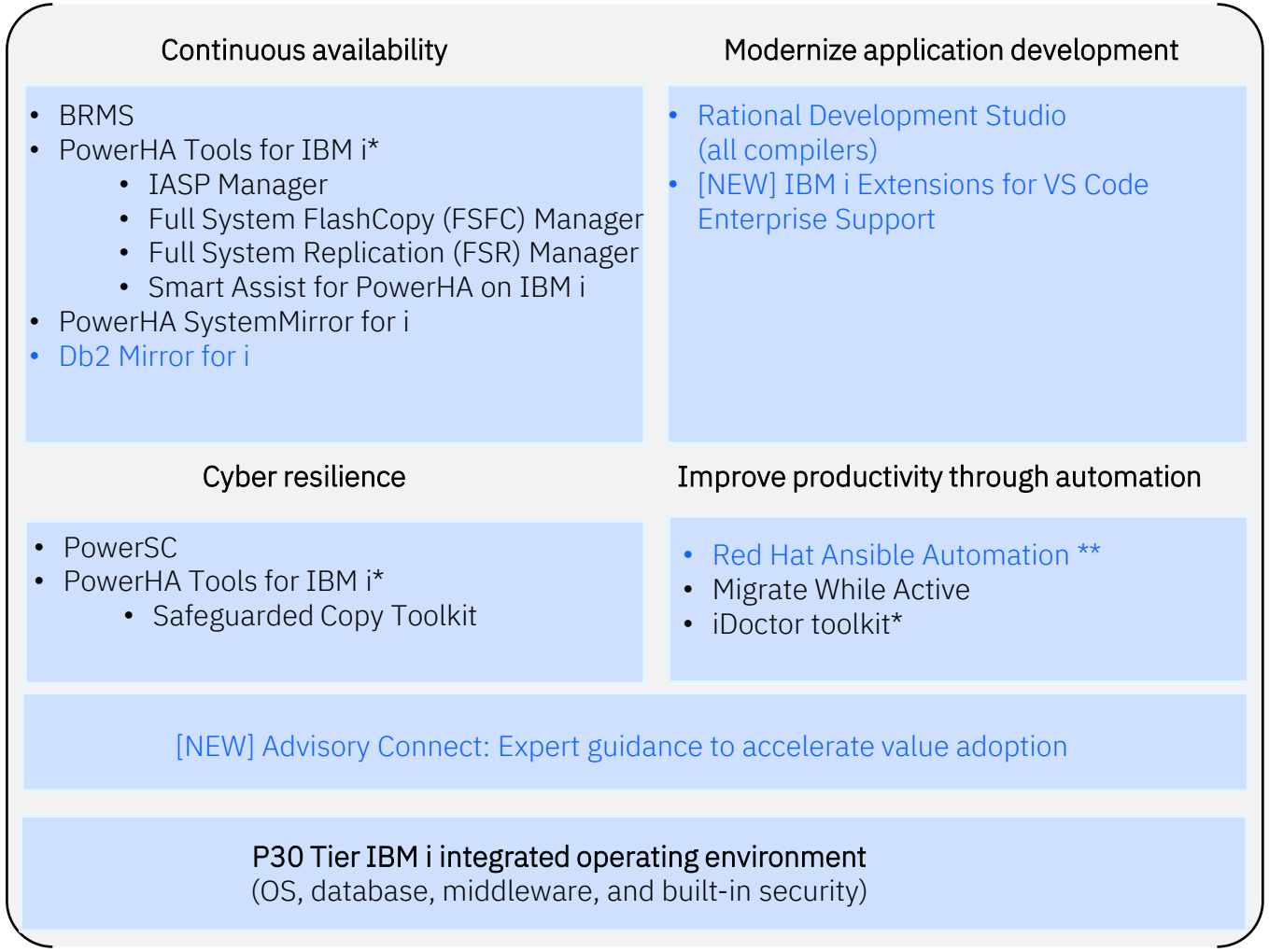
A simplified bundle with core capabilities

@2025 IBM Corporation

IBM i P30 Enterprise Edition (5770-30E)

A ***comprehensive*** bundle that helps Enterprise clients unlock more value of IBM i capabilities:

- Advanced HA/DR and cyber security capabilities to maximize uptime for mission-critical workloads
- Automation-driven operational management to increase IT productivity and minimize disruptions
- Open and modern application development and modernization to accelerate innovation
- Expert guidance to accelerate value adoption



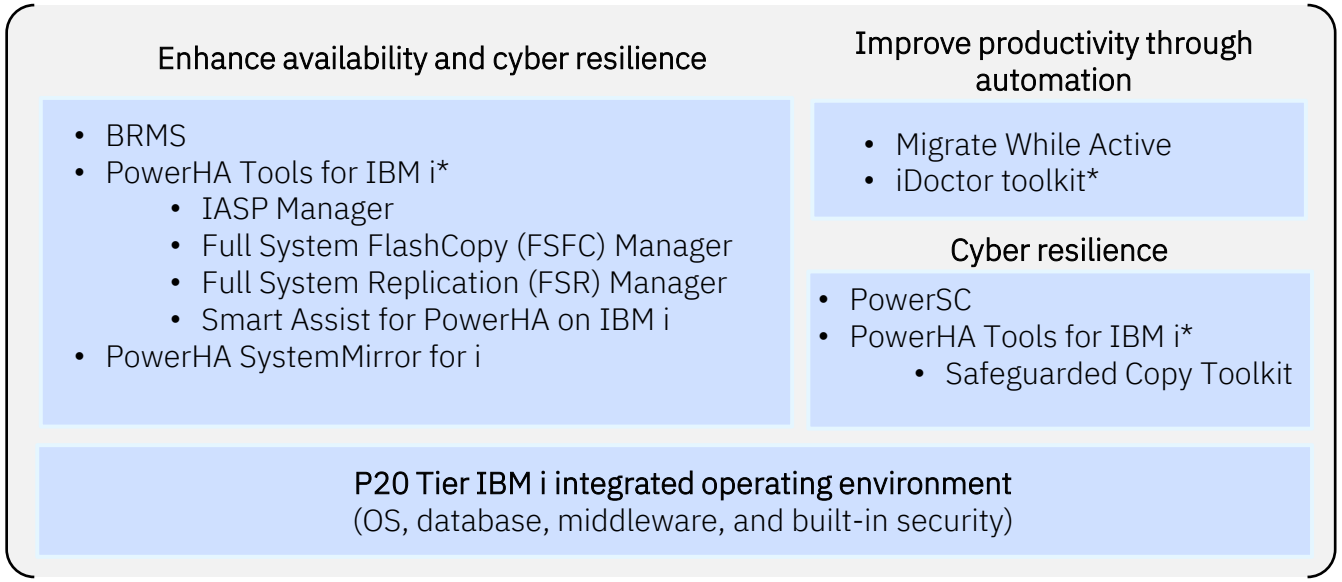
- License and maintenance for Expert Labs assets and tool kits only

** RedHat Ansible Automation: 1 year support for 100 managed nodes with a minimum purchase of 10 P30 Enterprise Edition cores

IBM i P20 Standard Edition

A strategic bundle that helps large clients unlock essential IBM i capabilities:

- Advanced HA/DR and cyber security capabilities to maximize uptime for mission-critical workloads
- Automation-driven operational management to increase IT productivity and minimize disruptions

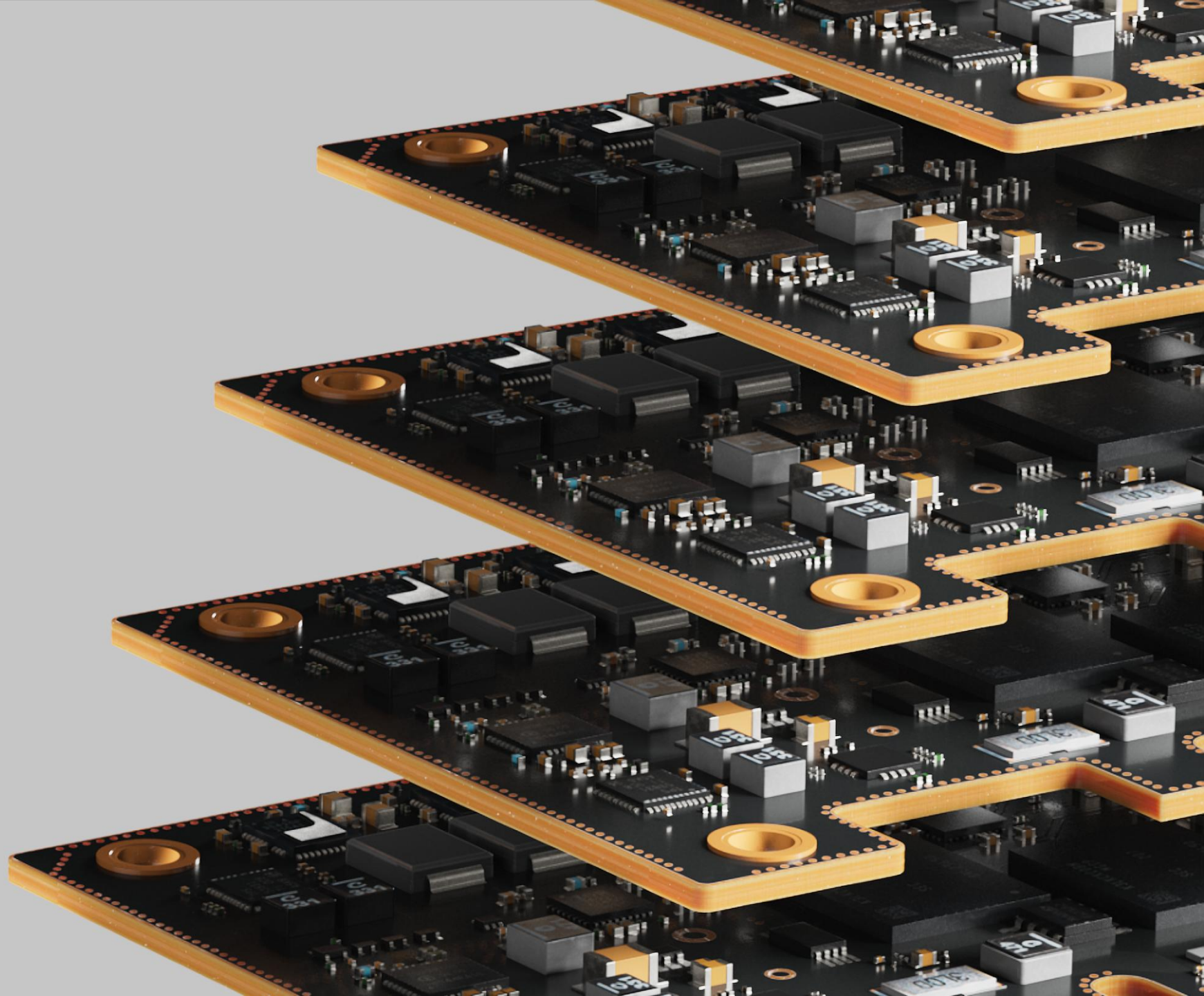


* License and maintenance for assets only

- Additional products may be purchased separately
  - Db2 Mirror, RDS, etc

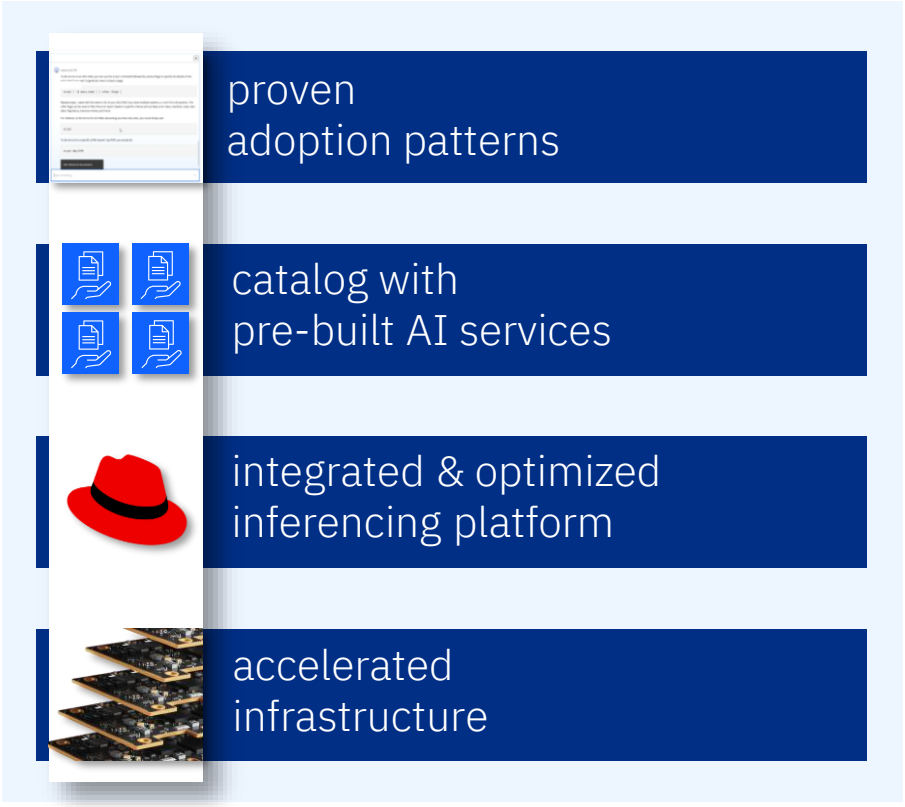
Introducing IBM Spyre™ for Power...

Turnkey AI for enterprise workflows.



IBM Spyre™ for Power

Turnkey AI for enterprise workloads.



1 click ...to install AI services from the IBM-supported catalog.[1]

1 configuration ...to move AI services of the IBM-supported catalog between IBM Power & IBM Power Virtual Server.[2]

>8 million/hour ...document embeddings for knowledge base integration using IBM Spyre™ Accelerator for Power with batch and prompt sizes of 128.[3]

Disclaimer: 1: AI services of the IBM-supported catalog are delivered as one or a set of containers that can be deployed with a single deployment command. The provided UI for the catalog executes such commands in the backend based on a single click within the UI page of the respective AI service. 2: A single configuration is enabled by exposed industry standard APIs to decouple services at the top and the backing inferencing service for all AI services that are part of the IBM-supported catalog. Any service that requires AI inferencing capabilities can connect inferencing services that provide OpenAI API or watsonx.ai API compliant inferencing endpoints (Spyre endpoint, RH AI Inferencing Server, IBM Cloud, OpenAI, Azure, AWS, GCP, ...). Services can run either on IBM Power or on IBM Power Virtual Server. 3: Based upon internal testing running 1M unit data set with prompt size 128, batch size 128 using 1-card container. Individual results may vary based on workload size, use of storage subsystems and other conditions.

Enterprise use cases

IT Ops | Development | Enterprise Resource Planning | Banking and Finance | Healthcare | Insurance | Public | Other

- Code Assistant
- Detect & Fix Agent
- Forecast & Plan Capacity Assistant
- IT Service Desk Assistant

Adoption patterns

- Data & Content Management
- Deep Process Integration
- Digital Assistant
- Forecasting
- Fraud Detection
- Image & Video Analytics
- Recommender System

Pre-built AI services

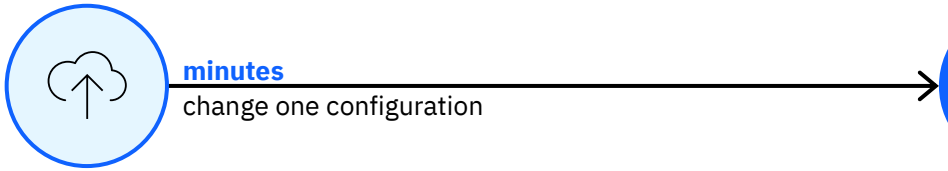
- Digitize Documents
- Extract & Tag Information
- Generate Reports
- Knowledge Management
- NLP to SQL
- Q&A
- Serve Models
- Similarity Search
- Transcribe
- Translate & Summarize

Install services

Simplify AI on-boarding

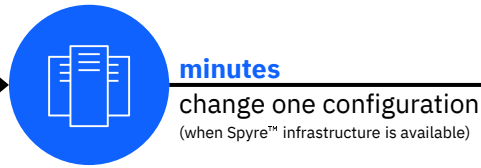
by moving between cloud & on-premises seamlessly.

Explore.



Explore cloud-based AI services
IBM Cloud, AWS, Azure, GCP, OpenAI, ...

Test.

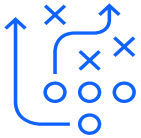


Co-locate AI services with system of records
IBM Power on-chip acceleration

Launch.

Promote to production
IBM Power11
+ IBM Spyre™ for Power

Discover turnkey AI with IBM Spyre™ for Power.



INFORM YOURSELF

Explore AI on IBM Power yourself:

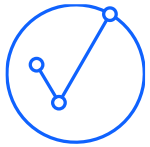
ibm.com/products/power/ai

30 minutes or less

GET BRIEFED

Meet with an IBM AI expert for custom demonstrations of IBM Spyre™. Understand where AI can help boost productivity & drive growth without long pilots.

2 – 4 hours



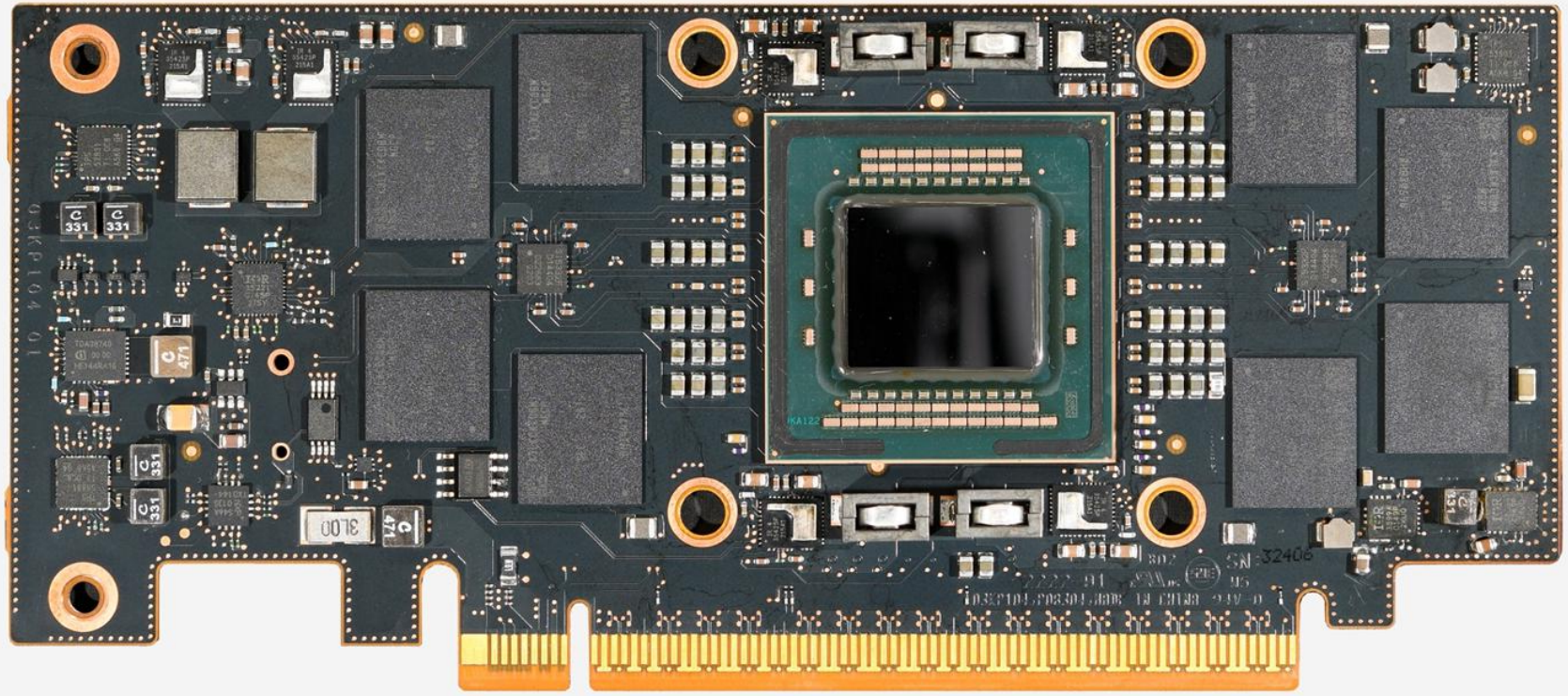
SEE VALUE

Run a use case alignment workshop, try AI services for your use cases at the same day, and plan the enterprise integration.

1 day

IBM Spyre™ Accelerator PCIe attached card

- 360 INT8/FP8 TOPS
- 75W PCIe gen5 x16 adapter
- 128GB of LPDDR5 memory at 205 GB/s

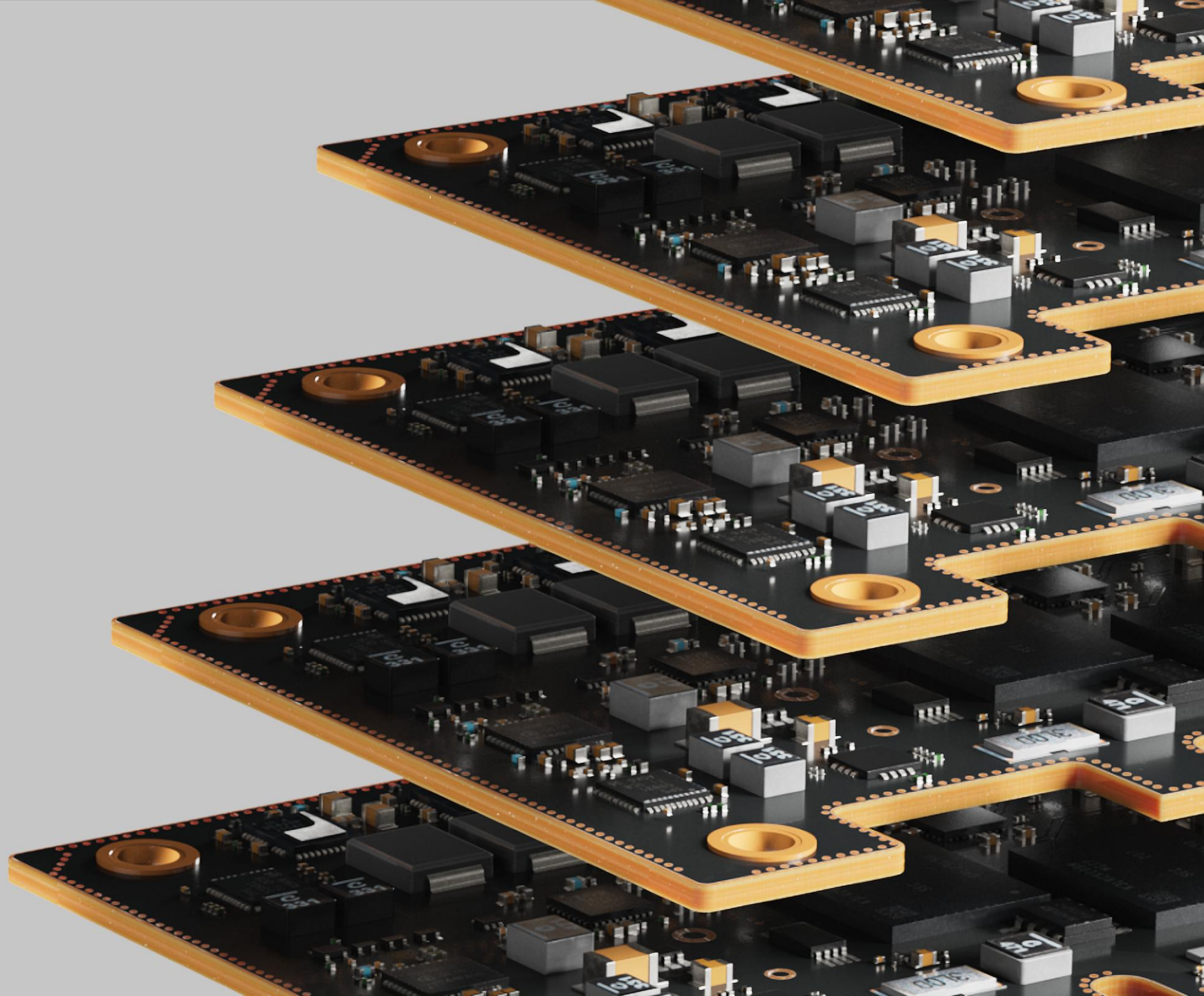


Designed to handle generative AI use cases

8 cards in I/O drawer form a logical cluster

1TB of memory

1.6TB per second aggregate memory bandwidth



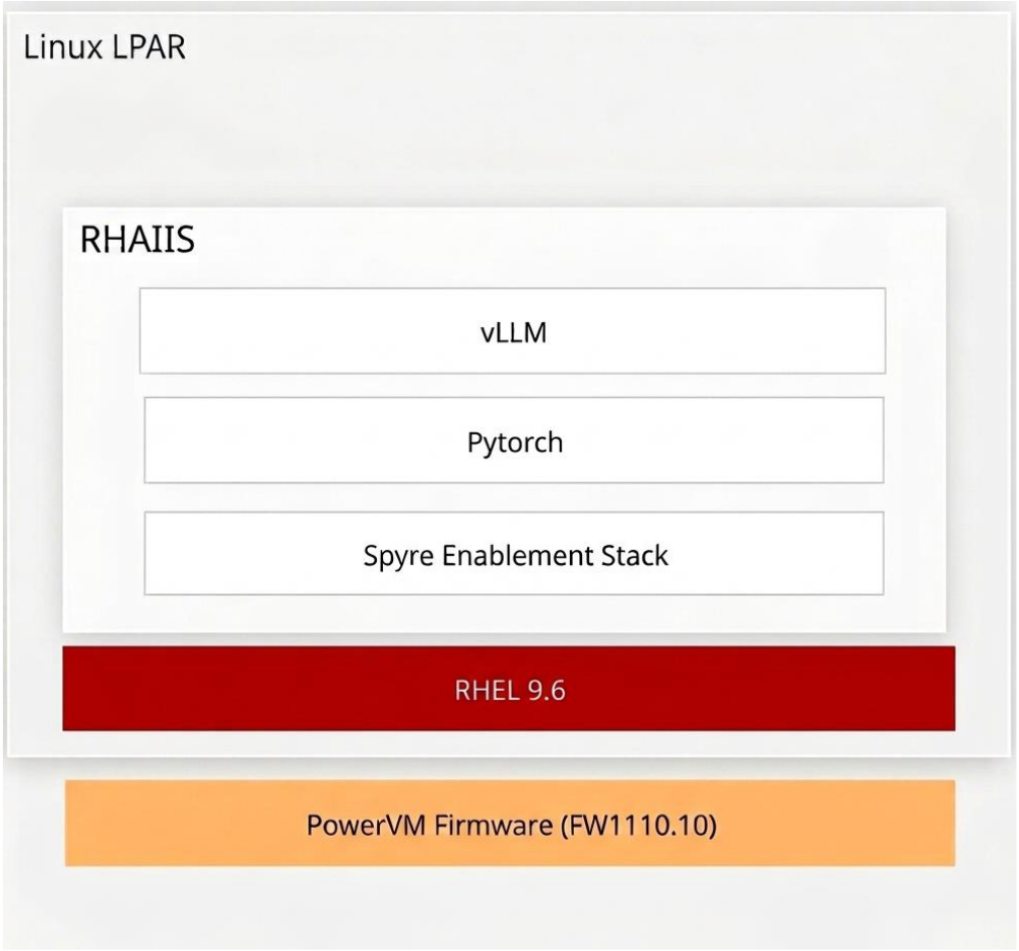
IBM Spyre Accelerator Adapter

- Key features
  - Enables low-latency inference pipelines that runs on-premises
  - Supports FP16/FP8 datatype inference execution
  - Continuous batching
  - Supports vLLM runtime
  - Supports multi-card deployment and scaling
  - Precompiled model caching

- Requires Red Hat® AI Inference Server (RHAIIS) or Red Hat® OpenShift AI (RHOAI)

IBM Power Platform Integration

- IBM Spyre Enablement Stack for Power includes
  - Runtime and drivers built for ppc64le architecture
  - Container-based deployment and orchestration through RHAIIS and Podman quadlets
  - Zero-copy accelerator access (VFIO) from Power CPU to Spyre card. Direct device access to userspace.
  - Tooling for performance monitoring (aiu-smi) and diagnostics
- RHAIIS is delivered as a container image which can be deployed on RHEL using Podman



Software Prerequisities

- System firmware update as per platform requirements (FW1110.10)
- Red Hat Enterprise Linux 9.6 installed and registered with RHEL subscription
- RHAIIS repository enabled

```
$ podman login registry.redhat.io
```

- Root or sudo access
- Outbound network access to registry.redhat.io and huggingface.co for model download
- IBM Power Tools repository enabled

IBM