IBM Spyre™ for Power

Dr. Sebastian Lehrig
World-Wide AI with IBM Power
Team Leader, IBM

Alilas

Context matters

Clients align AI use cases with their core workloads on IBM Power.

Cross-industry

ITOps | Development

- IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

...

System house DACH

MR WILLIAMS More Than A Convenience Store Distributor

...

Enterprise Resource Planning

- BI & HR assistant
- Supply chain forecasting
- Order processing assistant
- Product sales assistant

...

Geis Global Logistics

Large retailer US

...

Industry-specific

Banking & Finance

- Analyst assistant (frauds, NPAs, ...)
- Predict NPAs
- Open account agent
- Anti-money laundering

...

CREST — Ideate — Innovate — Implement —

Infosys Finacle

...

Healthcare

- Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

...

SHIBUYA BY PULSEN

SiPH

...

...and more

- Claims & policy management agent
- Private documents assistant
- Agriculture assistant
- Predict risk & underwrite assistant
- 360-degree view assistant
- Real estate assistant

...

...

EDSVÄRD

MIDRANGE

BanFast FÖRVALTNING

Gov. client DACH

...

AI drives transformational outcomes for IBM Power clients today.

5x Increase in business process rate with AI integration into existing enterprise workflows[a]

Uwe Remppel, head of ZSI department, Geis Group

18x Increase in developer productivity with AI-powered code assistant[b]

Jasmine Kaczmarek, VP Technology, MR Williams

> *"IBM Power servers are superscalar, multithreaded, multi-core servers with embedded AI acceleration technology, ideal for high-performance workloads."*

Brock Alston – CEO Rocketgraph

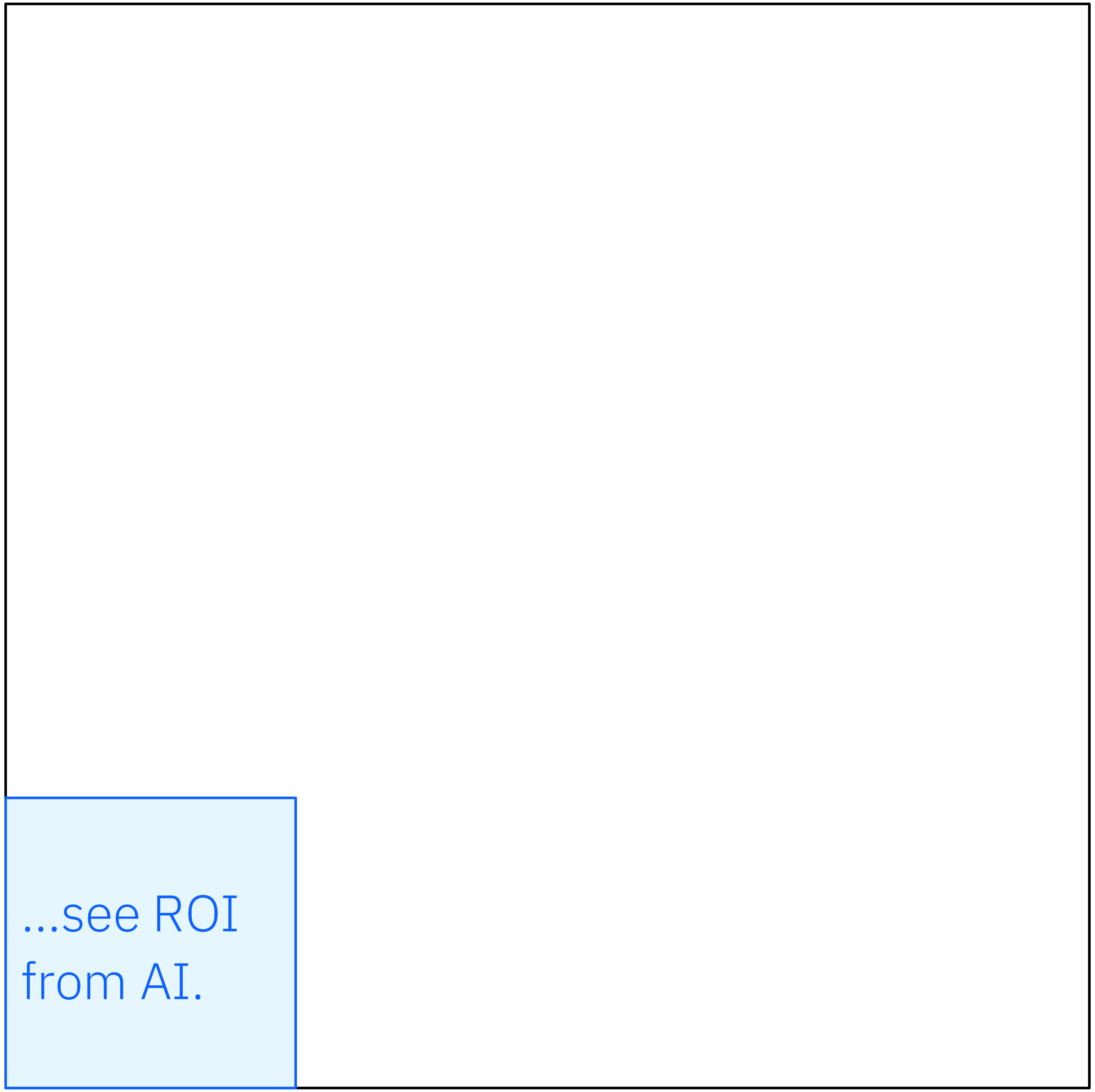
[a] Geis-group.eu: "We achieve a 5x faster business process throughput by integrating AI into our existing processes on IBM Power leveraging on-chip accelerator (MMA and SIMD) technology."* - Uwe Remppel, head of ZSI department, Geis Group *: Measured in staging system.

[b] MR Williams:"I only had 20 minutes to spare, and with the code assistant, I was able to investigate a report, trace how the field was populated, understand the calculation, and fix the issue. A senior developer had spent six hours on it the day before without finding a solution. It was super easy." Jasmine Kaczmarek, VP Technology, MR Williams

Nearly all enterprise pilots are stuck at the starting line

with only…

5%



…see ROI from AI.

Source: MIT, “The GenAI Divide, STATE OF AI INBUSINESS 2025”, Jul 2025.

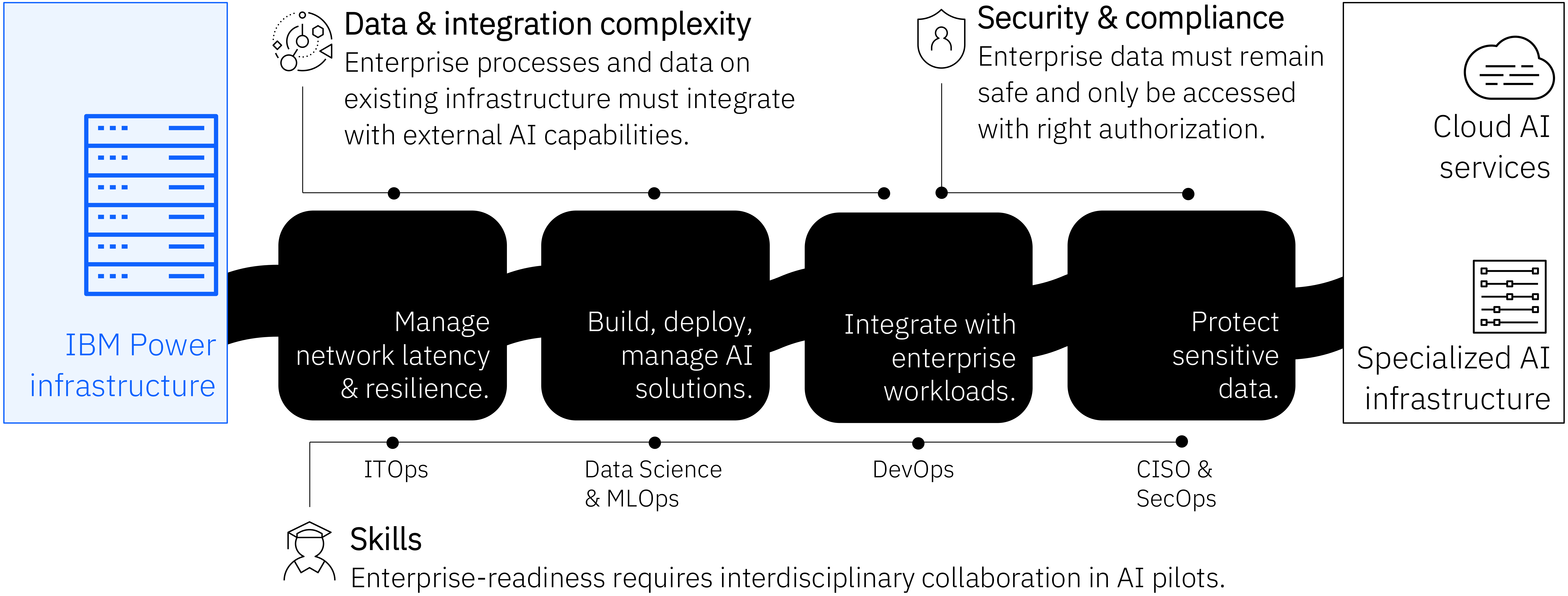
The time for scaling AI is now but enterprises face obstacles.

TOP 3 OBSTACLES

- Data & integration complexity
- Skills
- Security & compliance

Source: Deloitte, “*The State of Generative AI in the Enterprise*”, 2024.

Core enterprise workflows get the largest ROI from AI, if obstacles are addressed.



IBM Power infrastructure

Data & integration complexity
Enterprise processes and data on existing infrastructure must integrate with external AI capabilities.

Security & compliance
Enterprise data must remain safe and only be accessed with right authorization.

Manage network latency & resilience.

Build, deploy, manage AI solutions.

Integrate with enterprise workloads.

Protect sensitive data.

ITOps

Data Science & MLOps

DevOps

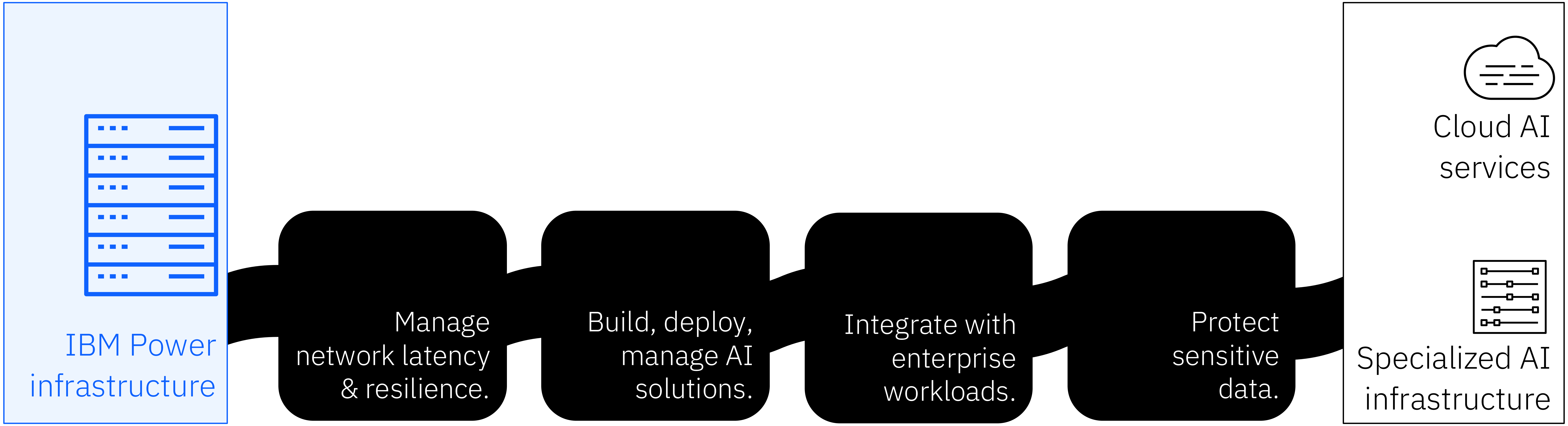
CISO & SecOps

Cloud AI services

Specialized AI infrastructure

Skills
Enterprise-readiness requires interdisciplinary collaboration in AI pilots.

…so, what if all this now comes out-of-the-box?



IBM Power infrastructure

Manage network latency & resilience.

Build, deploy, manage AI solutions.

Integrate with enterprise workloads.

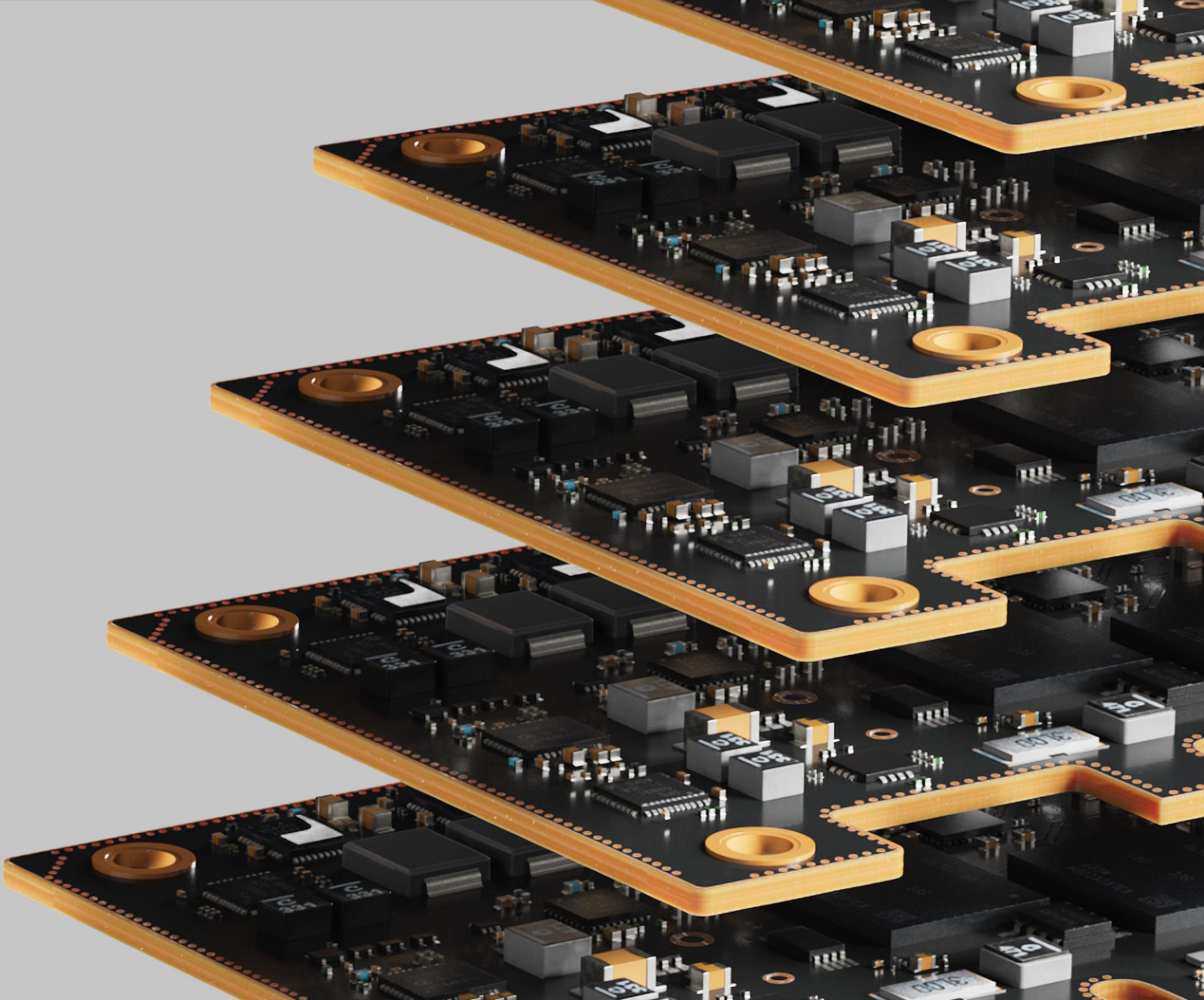
Protect sensitive data.

Cloud AI services

Specialized AI infrastructure

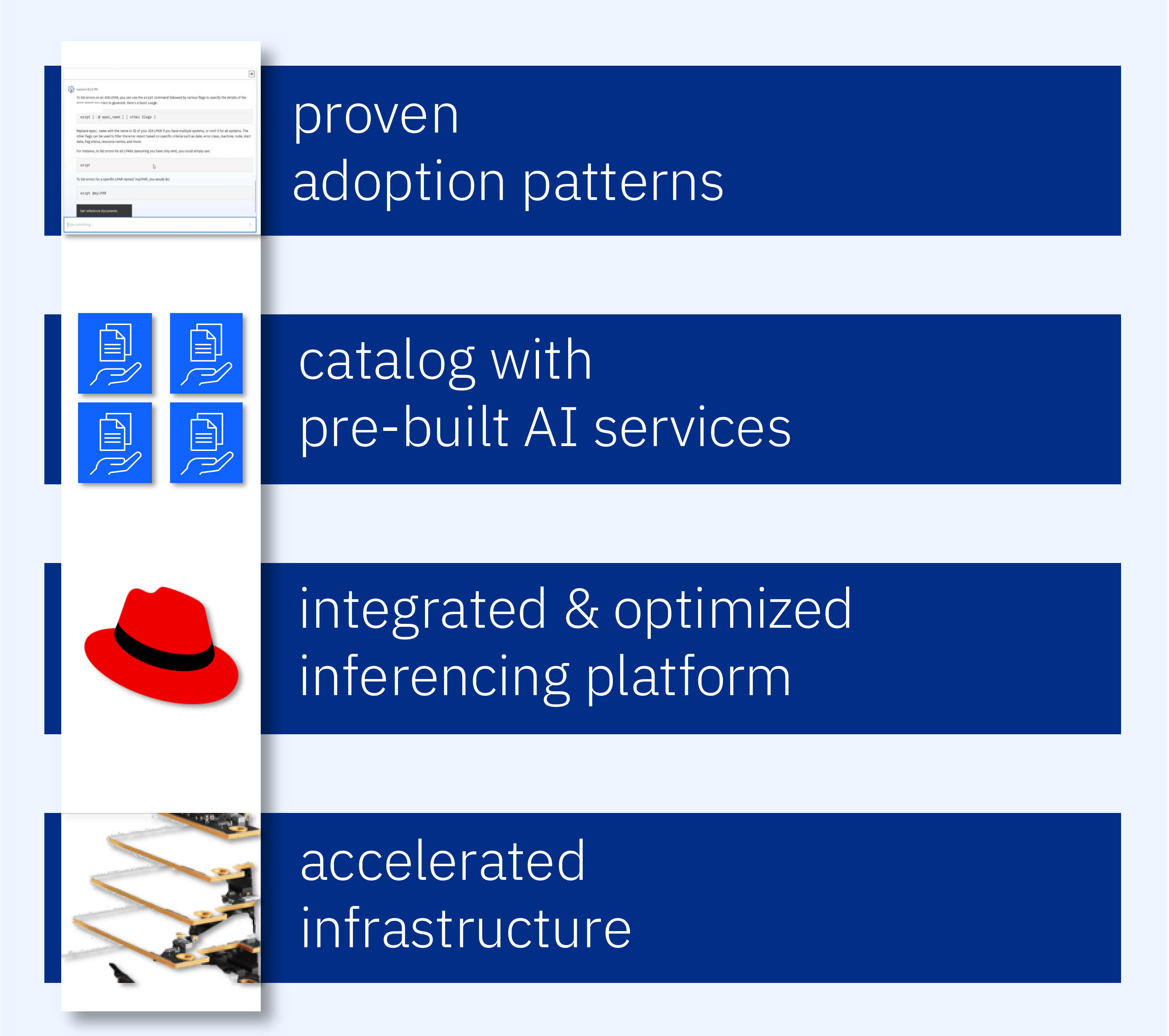
Introducing IBM Spyre™ for Power...

Turnkey AI for enterprise workflows.



IBM Spyre™ for Power

Turnkey AI for enterprise workloads.



proven adoption patterns

catalog with pre-built AI services

integrated & optimized inferencing platform

accelerated infrastructure

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.

IBM Spyre® for Power

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

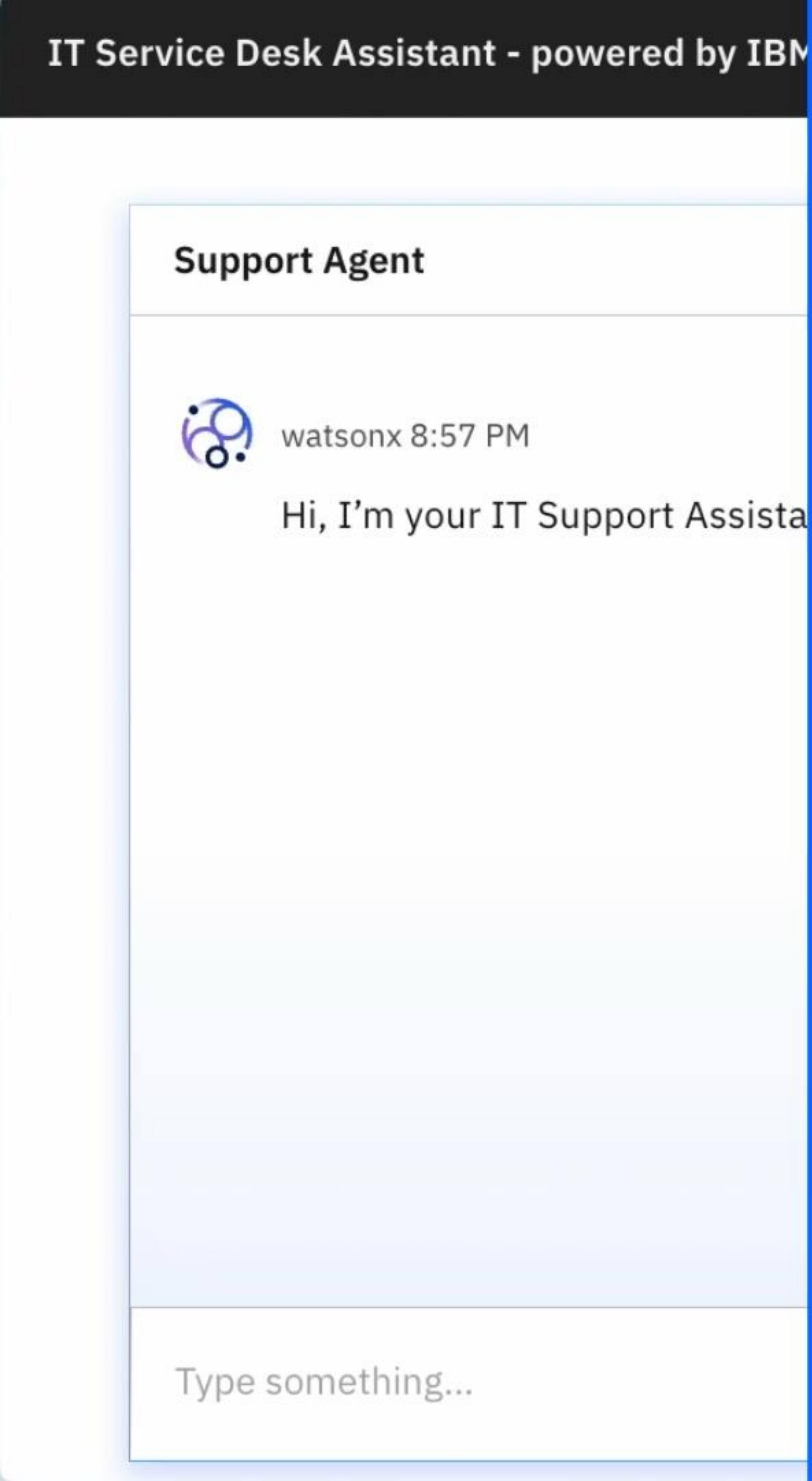
Adoption patterns

- 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



IT Service Desk Assistant - powered by IBM

Support Agent

watsonx 8:57 PM

Hi, I’m your IT Support Assista

“By integrating our digital assistant, Charlie, with IBM Spyre for Power, we’ll be delivering a truly plug-and-play AI solution integrated with proprietary data on the IBM Power platform that maintains strict data compliance and, ultimately, accelerates our business forward.”

– Alexander Edsvärd, CEO, Edsvärd Hållbarhet AB

“With IBM Spyre for Power, we can host our digital AI assistant locally in Sweden. This enables us to serve a wide range of clients, including the ones who demand the highest data protection standards and compliance with European and national regulations.”

– Lars Nylund, CEO, BanFast Förvaltning AB

Disclaimer: BanFast is a real estates company and a client of Edsvärd, using their digital AI assistant Charlie. Edsvärd and BanFast have tested IBM Spyre™ for Power as part of a Tech Preview program. Edsvärd uses the OpenAI API standard to connect their digital assistant Charlie to AI inferencing services. The IBM Spyre™ for Power endpoint also provides OpenAI APIs (via Red Hat AI Inferencing Server). The IBM Spyre™ for Power stack allowed Edsvärd to move to IBM Spyre™ for Power in a plug-and-play fashion. Strict data compliance covers protecting sensitive information, such as personal data and credit card information, from unauthorized access or data breaches according to the most common data compliance regulations and standards such as HIPAA, GDPR, and CCPA and is achieved through an on-premises deployment in a protected IBM Power environment. The IBM Spyre™ for Power stack does not require additional networks to be setup that potentially increase attack vectors because all IBM Spyre™ for Power Accelerators are connected through PCIe cables (and not via network), ensuring protection against network attacks.

Digital assistants

being validating in IBM Spyre™ for Power Tech Preview program.

Cross-industry

ITOps | Development

- System house DACH: IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Supply chain forecasting
- Order processing assistant
- Product sales assistant

Industry-specific

Banking & Finance

- Analyst assistant (frauds, NPAs, ...)
- Predict NPAs
- Open account agent
- Anti-money laundering

Healthcare

- Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

Insurance

- Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Gov. client DACH: Private documents assistant
- 360-degree view assistant

...and more

- SEMICON INDIA: Agriculture assistant
- BanFast FÖRVALTNING: Real estate assistant
- ...

proven Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics
- ...

pre-built AI services

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)
- ...

Deep process integrations

being validating in IBM Spyre™ for Power Tech Preview program.

Cross-industry

ITOps | Development

- IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Supply chain forecasting
- Gels Global Logistics
  Order processing assistant
- Product sales assistant

Industry-specific

Banking & Finance

- Analyst assistant (frauds, NPAs, ...)
- Predict NPAs
- CREST
  Open account agent
- Anti-money laundering

Healthcare

- Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

Insurance

- Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Private documents assistant
- 360-degree view assistant

...and more

- Agriculture assistant
- Real estate assistant
- ...

proven
Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics
- ...

pre-built
AI services

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)
- ...

Real-life example:

Extract ordering details in the order process automatically

Hans Geis, a logistics provider running their core ERP system on IBM i, integrates AI for a…

“5x increase in business process rate with AI integration into existing enterprise workflows.”

– Uwe Remppel, head of ZSI department, Geis Group[1]

Global Logistics

- 190 branches
- 13 countries
- 8k+ employees

Disclaimer: 1: “We achieve a 5x increase in business process rate with AI integration into our existing processes on IBM Power leveraging on-chip accelerator (MMA and SIMD) technology.” - Uwe Remppel, head of ZSI department, Geis Group. Measured in staging system of Hans Geis. Without AI it takes a senior employee 2 minutes / mail = 120 seconds / mail; with AI, they can completely automate 80% with 0 minutes processing time: (80%*0 minutes+20%* 2 minutes) / mail = 24 seconds / mail; leading to a factor of 5x: 120/24 = 5.

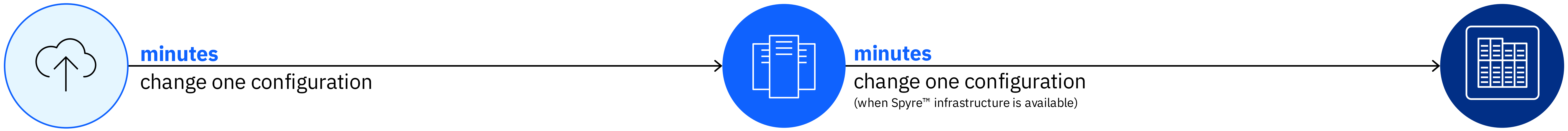
Simplify AI on-boarding

by moving between cloud & on-premises seamlessly.

Explore.

Test.

Launch.



minutes

change one configuration

minutes

change one configuration

(when Spyre™ infrastructure is available)

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

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

Promote to production
IBM Power11
+ IBM Spyre™ for Power

https://alfresco-demo.crestsolution.com:8080/digiworks-enterprise/#/documents/browse
InPrivate
digiworks
Dashboard
Tasks
Documents
Records
Reports
Apps
Hi, Suganesh
SR
Document Library
Upload Documents
Templates
Recent Documents
My Favourites
Current Edits
More
New File Plan
Folders
Contract Management
Crest Infosolutions Sdn B...
Demo Test Site
001- Folder
0123
023-09-25
04-09-2024 test
07-05-2025 folder
08-09-2025
1
1- investment contract
11111
11112
11222
12111
122111
12222
123
18-08-2025
Demo Test Site
001- Folder
Search Files and Folders
New Document
New Folder
Upload From PC/Laptop
Select Template
PDF
Drag & Drop
Or Browse Your Files

“Personally Identifiable Information (PII) redaction helps to protect against data breach & ransom scenarios, which can cause fines on banks, insurance and other customer facing organizations, leading to loss of trust and reputation.

We are planning to leverage IBM Spyre for Power for PII redaction, which we believe will decrease the amount of manual work by an organization, improving their ability to onboard customers faster and more efficiently.”

– Hemant Prasad, Chief Executive Officer, Crest Infosolutions

CREST

— Ideate — Innovate — Implement —

Deep process integrations

being validating in IBM Spyre™ for Power Tech Preview program.

Cross-industry

ITOps | Development

- IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Supply chain forecasting
- Gels Global Logistics — Order processing assistant
- Product sales assistant

Industry-specific

Banking & Finance

- Analyst assistant (frauds, NPAs, ...)
- Predict NPAs
- CREST — Open account agent
- Anti-money laundering

Healthcare

- Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

Insurance

- Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Private documents assistant
- 360-degree view assistant

...and more

- Agriculture assistant
- Real estate assistant

...

proven

Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics

...

pre-built

AI services

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)

...

Client examples & demos

aligned with core workloads; not restricted to Spyre™.

Cross-industry

ITOps | Development

- System house DACH: IT service desk assistant
- MR WILLIAMS: Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Large retailer US: Supply chain forecasting
- Gels Global Logistics: Order processing assistant
- Large retailer US: Product sales assistant

Industry-specific

Banking & Finance

- Demo TechXchange 2024: Analyst assistant (frauds, NPAs, ...)
- Infosys Finacle: Predict NPAs
- CREST: Open account agent
- Anti-money laundering

Healthcare

- SHIBUYA BY PULSEN: Medical assistant
- SiPH: Medical transcription assistant
- Claims & EHR matching agent
- SiPH: Medical image analysis assistant

Insurance

- Demo TechXchange 2025: Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Gov. client DACH: Private documents assistant
- 360-degree view assistant

...and more

- SEMICON INDIA: Agriculture assistant
- MIDRANGE
- BanFast FÖRVALTNING: Real estate assistant
- ...

proven Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics
- ...

pre-built AI services

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)
- ...

IBM is committed to open

Models

Granite

Tools

InstructLab

Docling

data-prep-kit

Data Prep Kit

Agents

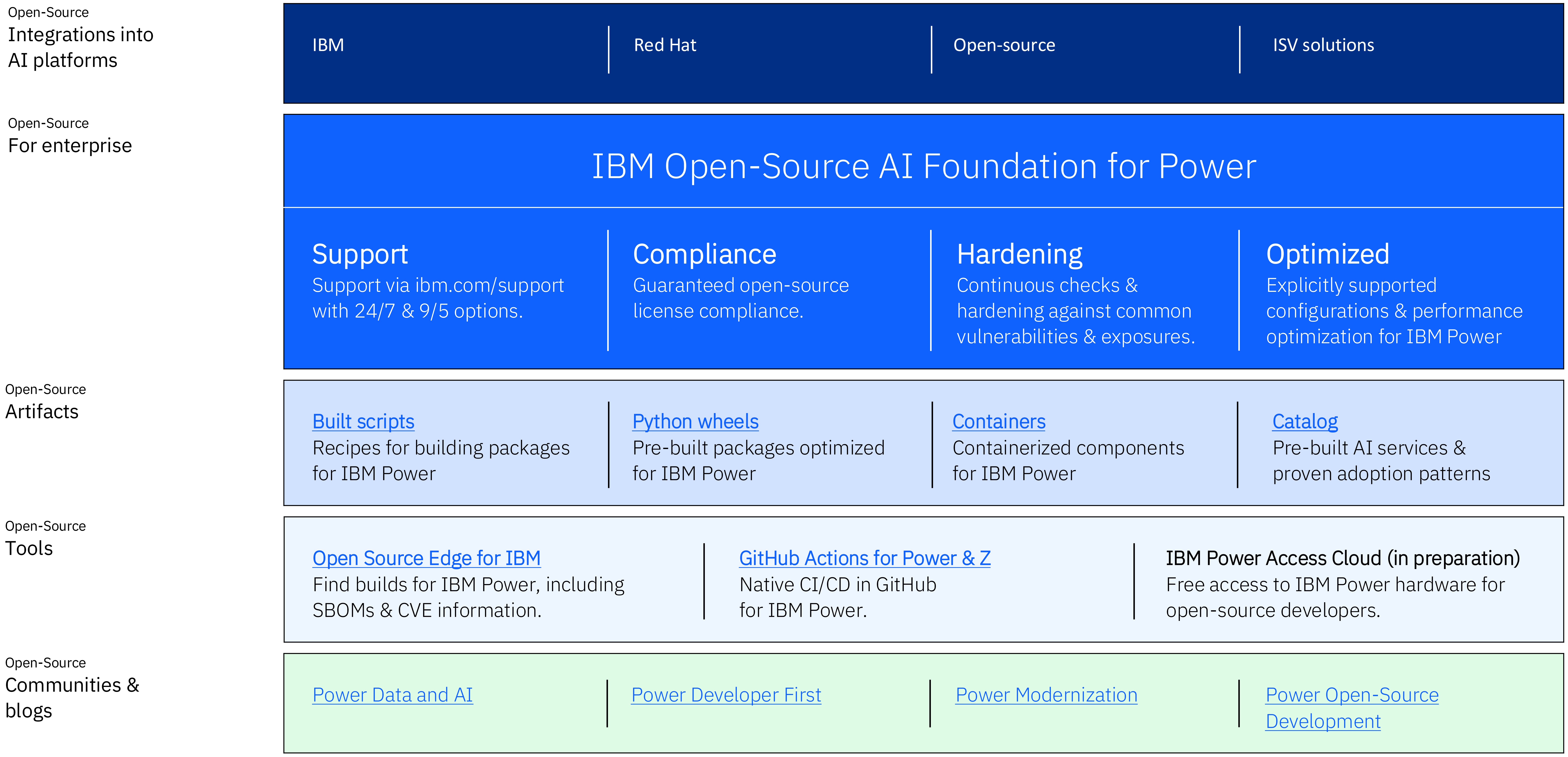
BeeAI

Open platform

Hugging Face kubernetes Kubeflow PyTorch vLLM OpenAI Triton

Based on: “Artificial Intelligence Needs Community Intelligence” presented at All Things Open AI 2025, Sriram Raghavan, VP IBM Research AI, March 2025.

Enable developers with a holistic open-source ecosystem for IBM Power



Open-Source
Integrations into AI platforms

IBM | Red Hat | Open-source | ISV solutions

Open-Source
For enterprise

IBM Open-Source AI Foundation for Power

Support
Support via ibm.com/support with 24/7 & 9/5 options.

Compliance
Guaranteed open-source license compliance.

Hardening
Continuous checks & hardening against common vulnerabilities & exposures.

Optimized
Explicitly supported configurations & performance optimization for IBM Power

Open-Source
Artifacts

Built scripts
Recipes for building packages for IBM Power

Python wheels
Pre-built packages optimized for IBM Power

Containers
Containerized components for IBM Power

Catalog
Pre-built AI services & proven adoption patterns

Open-Source
Tools

Open Source Edge for IBM
Find builds for IBM Power, including SBOMs & CVE information.

GitHub Actions for Power & Z
Native CI/CD in GitHub for IBM Power.

IBM Power Access Cloud (in preparation)
Free access to IBM Power hardware for open-source developers.

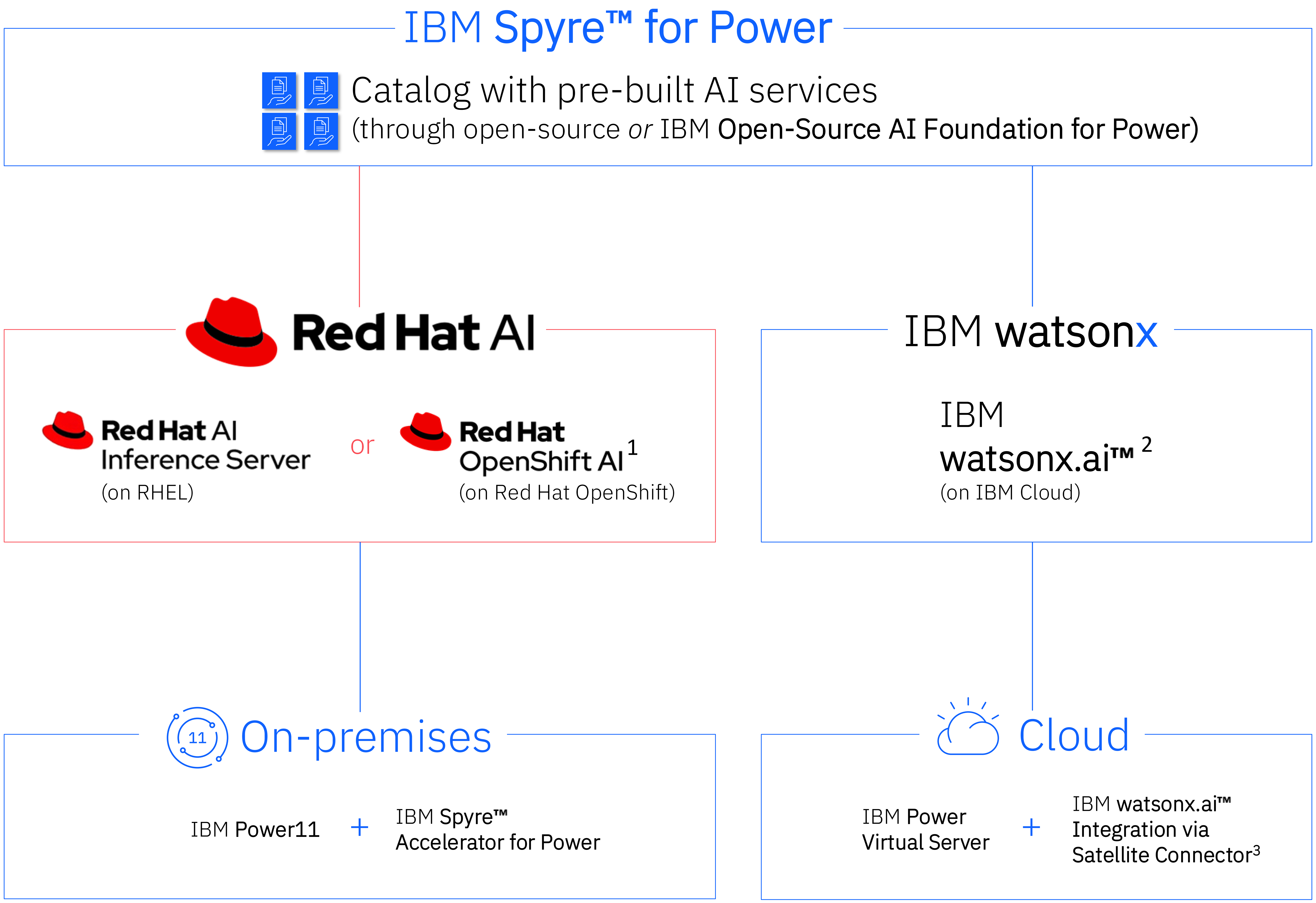
Open-Source
Communities & blogs

Power Data and AI | Power Developer First | Power Modernization | Power Open-Source Development

Simplify with pre-built open-source AI services...

...integrated with trusted, consistent, and comprehensive inferencing platforms...

...empowered by reliable & safe infrastructure options.



IBM Spyre™ for Power

Catalog with pre-built AI services
(through open-source *or* IBM **Open-Source AI Foundation for Power**)

Red Hat AI

Red Hat AI Inference Server
(on RHEL)

or

Red Hat OpenShift AI[1]
(on Red Hat OpenShift)

IBM watsonx

IBM **watsonx.ai™**[2]
(on IBM Cloud)

On-premises

IBM **Power11** + IBM **Spyre™ Accelerator for Power**

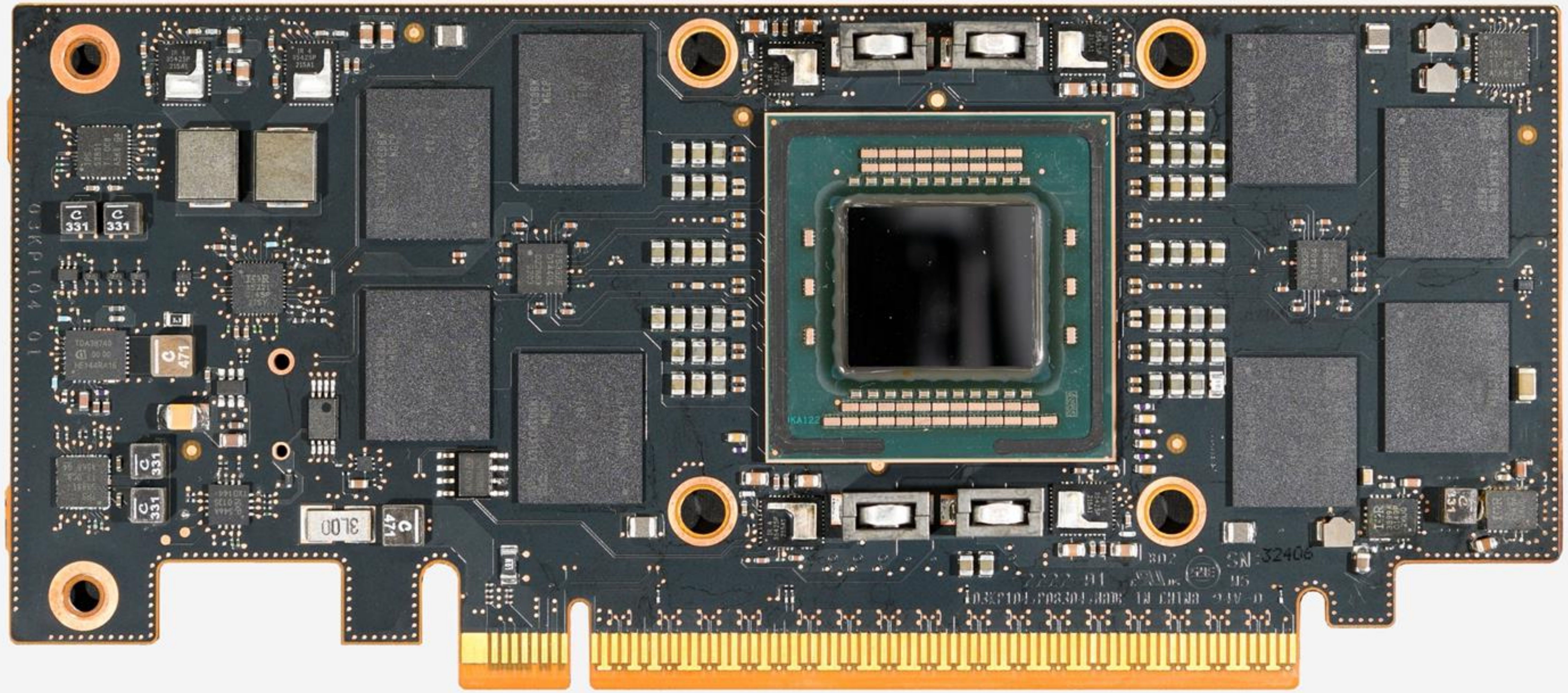
Cloud

IBM **Power Virtual Server** + IBM **watsonx.ai™ Integration via Satellite Connector**[3]

[1] GA with Spyre 1Q'26
[2] Integration of IBM Spyre™ for Power catalog of AI services targeted for 2026
[3] https://cloud.ibm.com/docs/powervs-watsonx-toolkit?topic=powervs-watsonx-toolkit-powervs-watsonx-ra

IBM Spyre™ Accelerator PCIe attached card

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

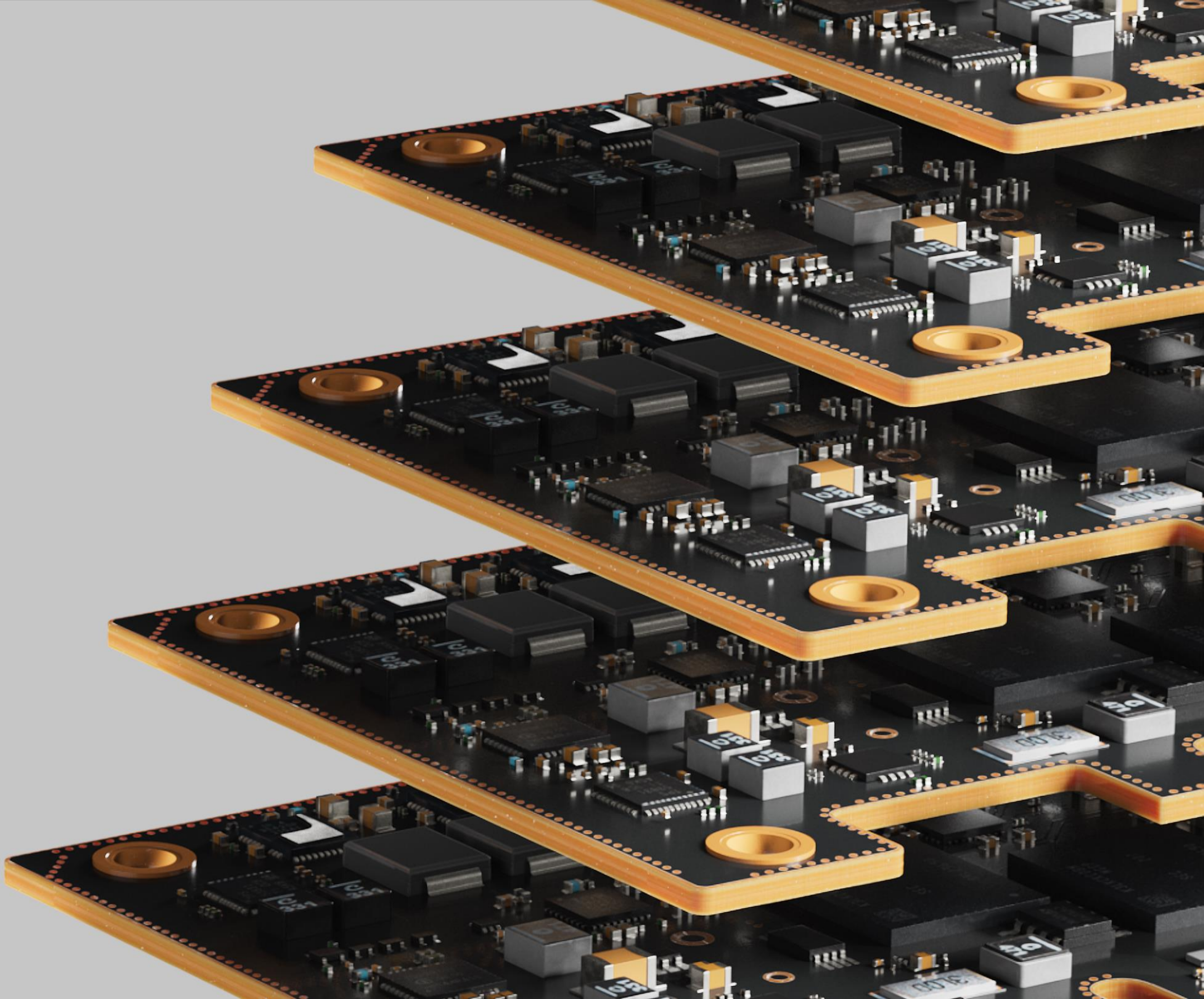


Designed to handle generative AI use cases

8 cards in I/O drawer form a logical cluster

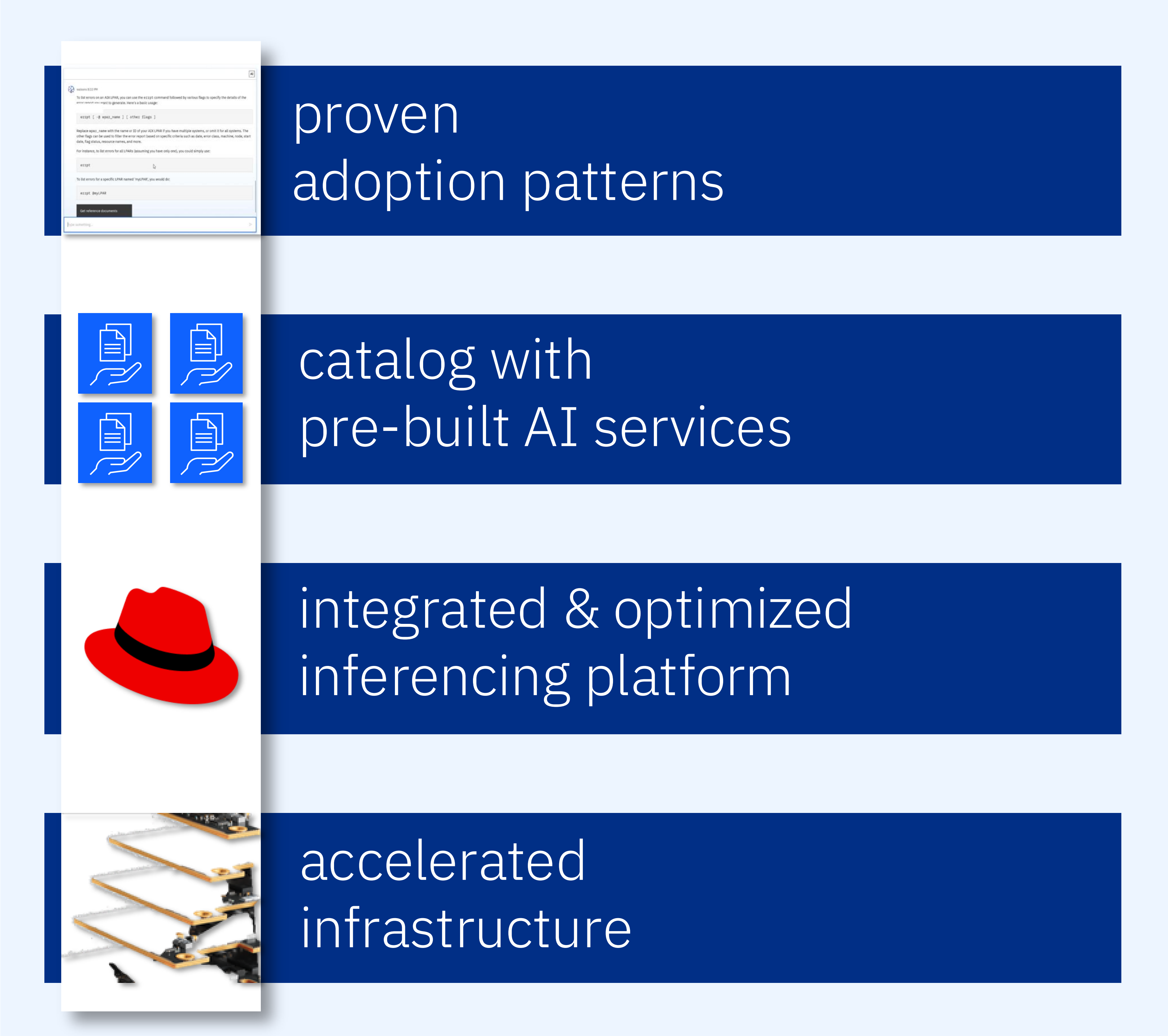
1TB of memory

1.6TB per second aggregate memory bandwidth



IBM Spyre™ for Power

Turnkey AI for enterprise workloads.



proven adoption patterns

catalog with pre-built AI services

integrated & optimized inferencing platform

accelerated infrastructure

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.

Put AI to work with IBM Power.

- ✓ Transform enterprise processes
- ✓ Boost productivity
- ✓ Unlock enterprise data’s value

Blue: covered today

IBM & ISV ecosystem

IBM Project Bob[1] | IBM Concert Integration
Equitus.ai | Rocketgraph | Finacle | Wallaroo.AI | Crest | Edsvärd | ElinarAI

Data fabric

IBM watsonx.data[2] | IBM DataStage | IBM Knowledge Catalog | IBM Orchestration Pipelines

AI foundation

Red Hat OpenShift AI[3] | Red Hat AI Inference Server[2] | IBM Open-Source AI Foundation for Power[2]

AI-ready infrastructure

IBM Spyre™ Accelerator[2] | IBM PowerVS watsonx Toolkit

1. Private Preview
2. GA 4Q ‘25
3. Tech Preview, SoD GA (MMA) 4Q ‘25, GA (Spyre) 1Q’26

Discover turnkey AI with IBM Spyre™ for Power.

INFORM YOURSELF

Explore AI on IBM Power yourself:

ibm.com/products/power/ai

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.

SEE VALUE

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

Thank you

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM