

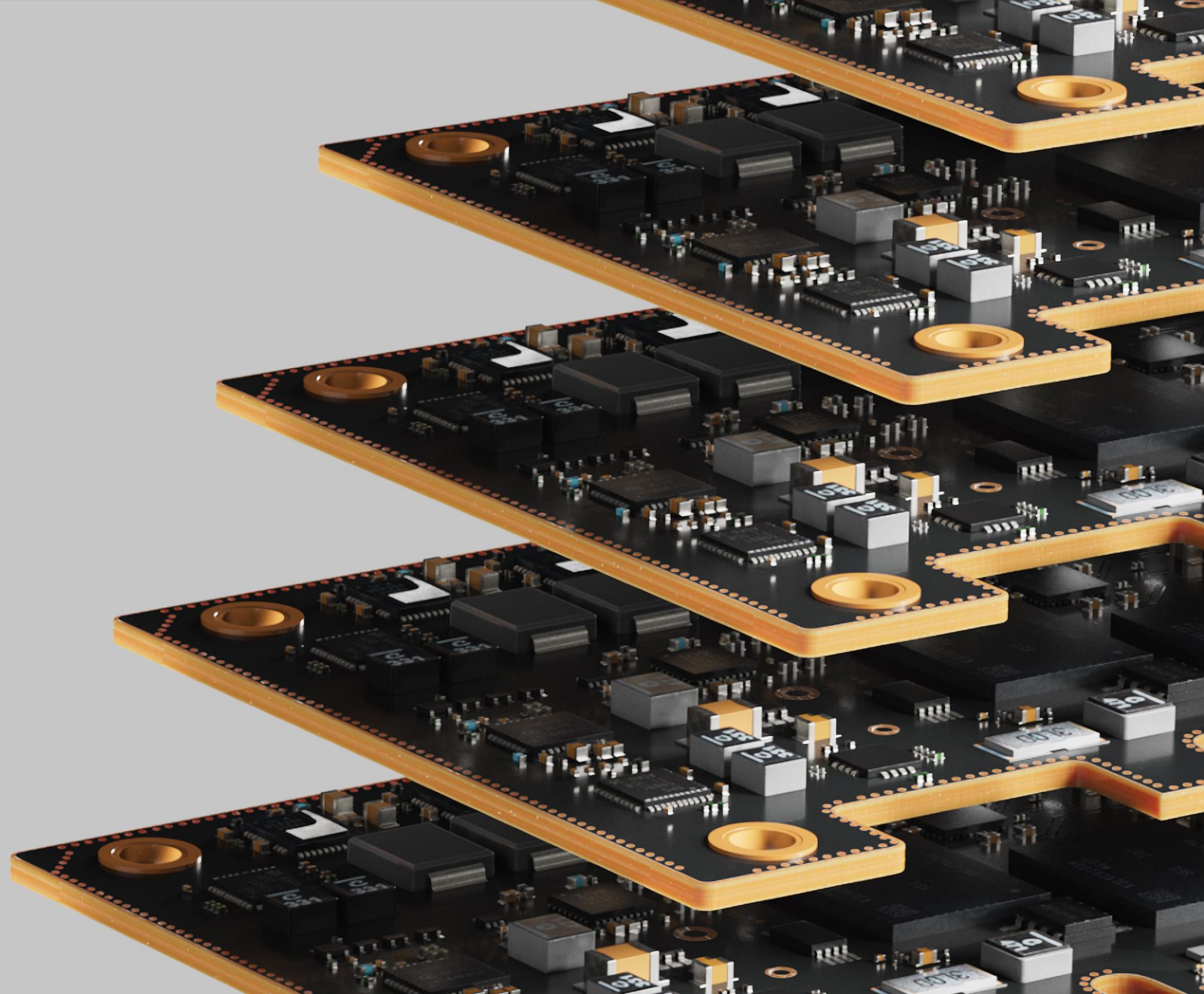
AI Acceleration for Power with Spyre



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

Introducing
IBM Spyre™
for Power...

Turnkey AI
for enterprise
workflows.



Go for days, not months.



“AI is a core enabler of our financial software solutions, accelerating innovation, enabling smarter decisions, and driving sustainable growth for our customers.”
– Jessica Kreuzmann, CEO, Portolan Commerce Solutions



LOB ISV Enterprise Architect



“The IBM Power team excels at rapidly showcasing innovation—from NLP-to-SQL to forecasting and agentic workflows—helping us identify the next generation of capabilities for our financial software solutions.”
– Alexander Nürk, CTO, Portolan Commerce Solutions



LOB ISV Enterprise Architect



“We installed the software stack within a *single morning*, and by the afternoon we were already playing with digital assistants. As IT administrator, I found the *deployment via the AI services catalog incredibly straightforward*.”
– Andre Lutz, Technical Leader Linux on Power, SVA



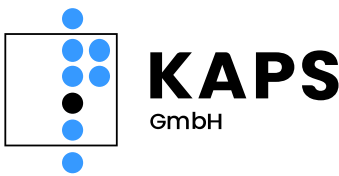
“I like using AI to build and execute plans for specific projects. *Bob's planning [...] provided a better understanding* of the steps through the project from beginning to end.”
– Bob Richardson, ERP and AI Analyst, Wynne Systems



“With supported, standardized APIs, we can deploy workloads across *cloud and on-premises environments*, aligned with our customers’ preferences.”
– Alexander Nürk, CTO, Portolan Commerce Solutions



ISV Enterprise ITOps Architect Manager



“In our Spyre environment, we’ve integrated a *digital assistant* directly into our ERP interface and use *intelligent information extraction* to convert customer inquiries into actual orders. We have also integrated our sales software with Spyre, which proactively supports *B2B sales with data-driven recommendations*.”
– Michael Klinger, CEO, KAPS GmbH



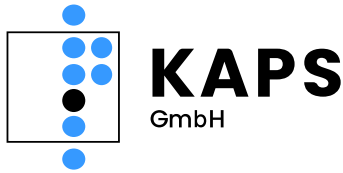
“This [...] allows us to expand our KOBE offering to include local LLM deployment, *bypassing the need for cloud-based services*.”
– Giancarlo Lui, CTO, Horsa Power S.r.l.



“MMA *reduced production planning optimization time by ~46%* for Becom, enabling them to securely integrate optimization workflows with their IBM i data and ERP system.”
– Clemens Zauchner, Senior Data Scientist, IT-PS



LOB ISV Enterprise ITOps Architect Manager



“Early customer previews have generated strong excitement, and we are now *preparing for general availability*.”
– Michael Klinger, CEO, KAPS GmbH



“Given these results, *we are officially incorporating Spyre hardware into our commercial proposal for KOBE*.”
– Giancarlo Lui, CTO, Horsa Power S.r.l.



“Forecasting and optimization are reshaping our entire production chain: Cutting volatility, optimizing batch sizes, and enhancing coordination between materials and manufacturing capacity. We have successfully *rolled out the solution across our ERP systems in Europe*.”
– Johann Katona, Head of Organizational Development & IT, BECOM Electronics GmbH



LOB ITOps Manager

Identify AI needs and objectives

- Clarify needs and pains
- Outline opportunities
- Align compliance context

Align and prioritize use cases

- Assess ROI potential
- Assess feasibility
- Specify success criteria

Plan and engineer solutions

- Scope and run POC/MVP
- Design and test solution
- Review continuously

Evaluate ROI and day 1 ops

- Assess success criteria and incorporate feedback
- Deploy and initialize
- Implement change management

Use and day 2 ops

- Employ innovation within everyday business tasks
- Validate in-production ROI
- Ensure day-2 operations

Disclaimers: (1) SVA has externally shared experiences with Spyre on LinkedIn (https://www.linkedin.com/posts/udo-sachs_ibmchampion-ai-ibmpower-ugcPost-7442243262792343552-xWHX/7utm_source=share&utm_medium=member_desktop&rcm=ACoAABafyM0B3VuT78e0AUTiufiZT0ADhZmimB4) and within a digital (German) live event, published at YouTube: (1) “KI-Strategie: Warum sie scheitert – und wie Erwartungsmanagement gepaart mit schlüsselfertiger KI sie retten kann” (https://youtu.be/vqksbP0_uCQ) and (2) “Schlüsselfertige KI mit IBM Spyre im technischen Fokus: Architektur und KI-Workloads.” (<https://youtu.be/creYkigango>). The quotes on this slide come from these venues, were freely translated into English, and then approved by SVA. (2) Client Becom and their partner IT-PS have internally measured the average total execution times when running their production planning optimization (based on linear programming) over a given input data set (static data sample with all assembly units that involve several hundred individual optimizations). Their test setup resembles the production environment with a Linux LPAR on IBM Power with 3 physical IBM Power cores (SMT8, 24 Linux virtual cores, nproc=24), 23.8GB memory, and an external storage (IBM Storwize Flash Storage). On IBM Power9-based servers (IBM S922), they have measured 9.95 minutes and on IBM Power10-based servers (IBM S1022), they have measured 5.3 minutes, leading to a reduction of approximately 46%. More detailed client stories are available at <https://www.ibm.com/downloads/documents/us-en/15db45f02d520e7a> and <https://www.ibm.com/it-infrastructure/us-en/resources/power/ibm-i-customer-stories/> (BECOM Electronics GmbH).

IBM Spyre™ Accelerator PCIe attached card

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





Front



Back

IBM Power11

Change version

9856-42H

☒ Show full table of contents

Filter on titles

PCIe5 x16 AI Accelerator adapter (FC ECSE; CCIN 2E1F)

PCIe4 4-port NVMe JBOF adapter (FC EJ1X and EJ1Y; CCIN 6B87)

PCIe4 cable adapter (FC EJ24; CCIN 6B92)

PCIe4 cable adapter (FC EJ2A; CCIN 6B99)

PCIe3 x8 SAS quad-port 12 Gb tape adapter (FC EJ2B and FC EJ2C; CCIN 57F2)

4769 Cryptographic Coprocessor (FC EJ35 and EJ37 for BSC; CCIN C0AF)

PCIe3 x8 2-port Fibre Channel (32 Gb/s); (FC EN1A and EN1B; CCIN 578F)

[All products](#) / [IBM Power11](#) / [9856-42H](#) /Was this topic helpful?  

PCIe5 x16 AI Accelerator adapter (FC ECSE; CCIN 2E1F)

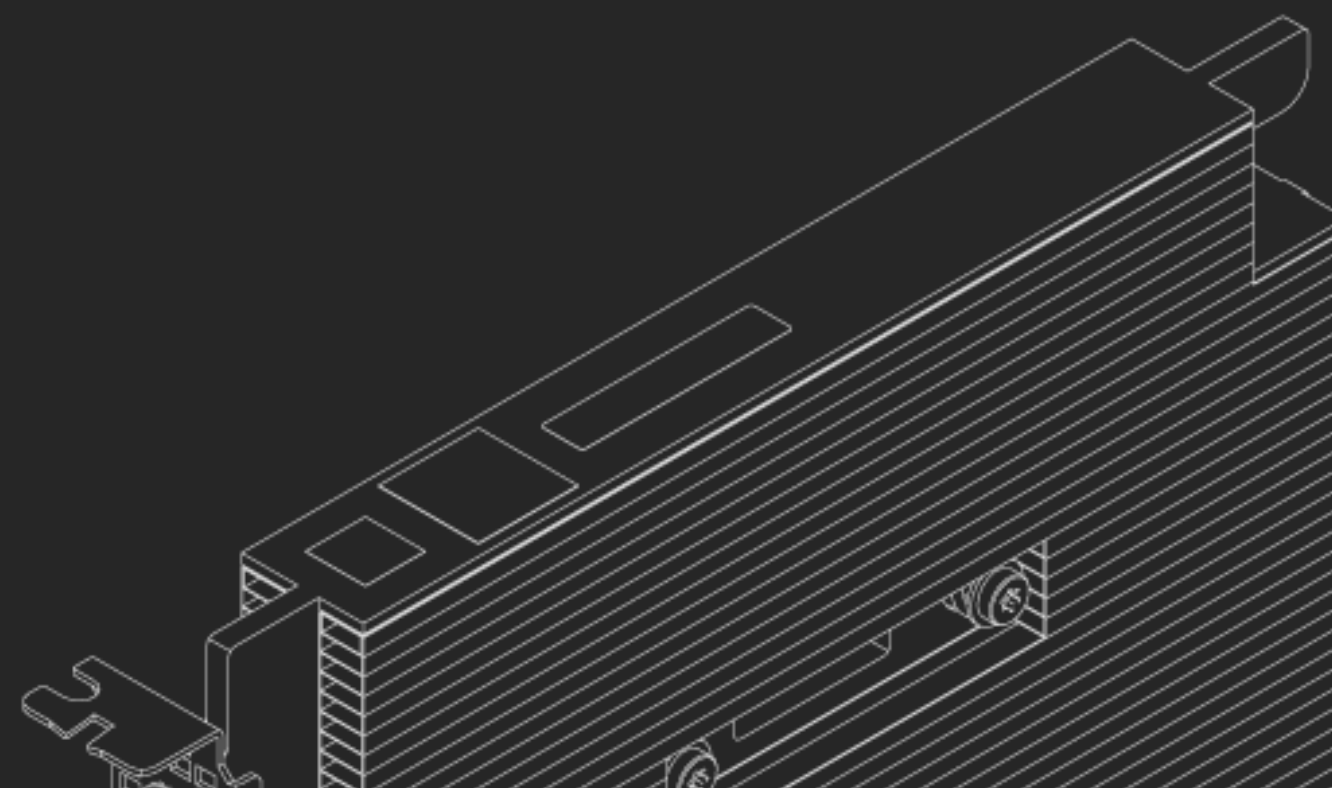
Last Updated: 2026-01-28

Learn about the specifications and operating system requirements for the feature code (FC) ECSE adapter.

Overview

The PCIe5 x16 Artificial Intelligence (AI) Accelerator adapter is a PCI Express (PCIe) generation 5 (Gen5) x16 adapter. The adapter utilizes a Samsung 5LPE chip with 32 accelerator cores and 128 GB of LPDDR5 memory. The adapter is only supported in ENZ0 PCIe4 expansion drawers with unit type U5B67 and feature code ENZA fanout modules.

Figure 1. PCIe5 x16 AI Accelerator adapter

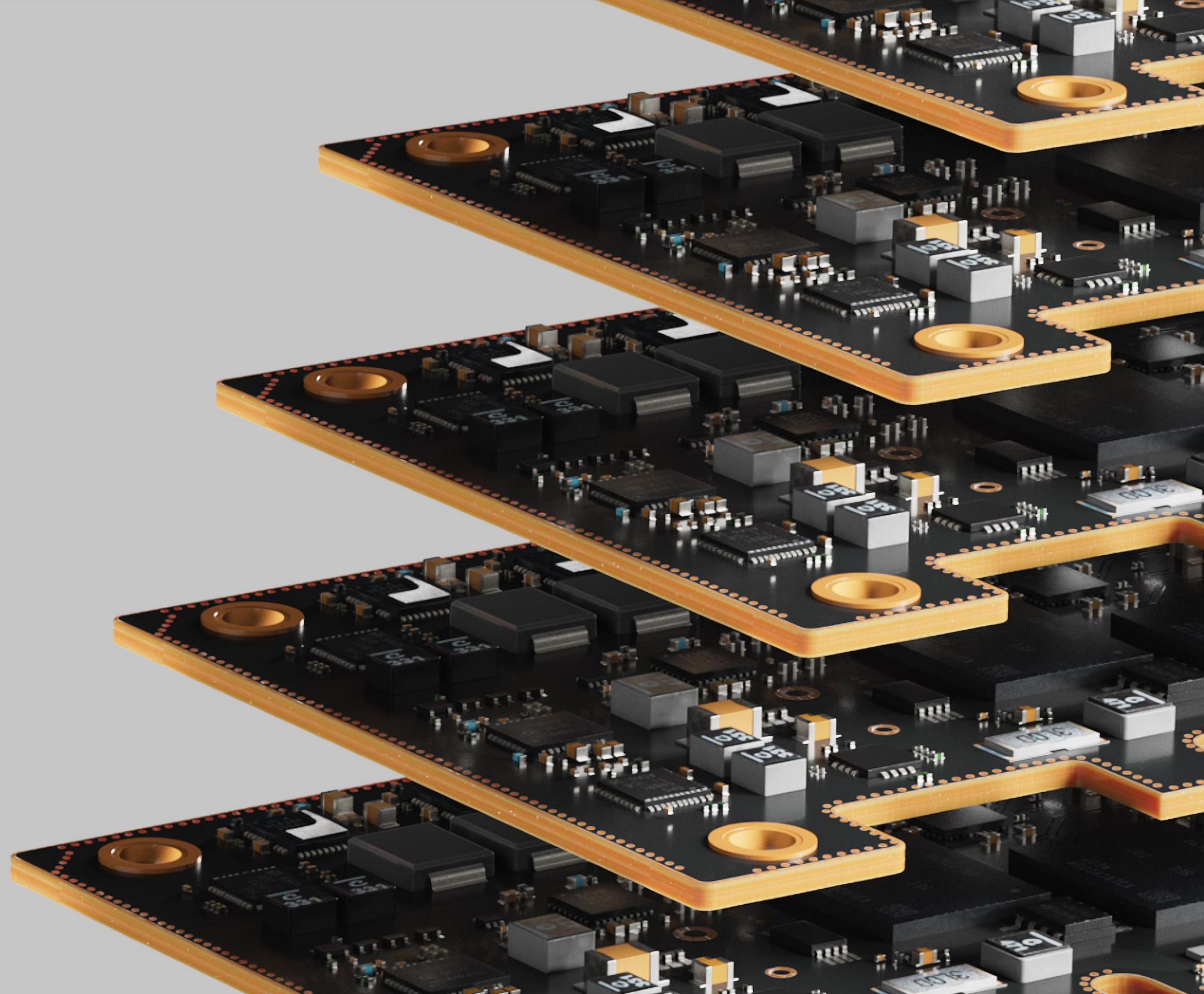
**On this page**[Overview](#)[Specifications](#)[Operating system or partition requirements](#)

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



ENZ0 PCIe Gen4 I/O expansion drawer

Last Updated: 2025-07-08

This topic provides users and service providers with installation and maintenance information for the ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer).

– Overview of the ENZ0 PCIe Gen4 I/O expansion drawer

The ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer) is a 482.6 mm (19 in), 4U PCIe Gen4 I/O drawer. Each ENZ0 PCIe4 expansion drawer provides up to 12 PCIe Gen4 adapter slots.

– Installing an ENZ0 PCIe Gen4 I/O expansion drawer

Follow the steps provided here to install an ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer).

– Removing an ENZ0 PCIe Gen4 I/O expansion drawer from an existing configuration

Learn how to remove an ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer) from an existing configuration when the system power is turned off.

– Connector locations for the ENZ0 PCIe4 expansion drawer

Learn about connector locations for the ENZ0 PCIe4 expansion drawer.

– Installing parts

Use the installing parts procedures when you want to add parts to your ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer).

– Removing and replacing parts

Use the removing and replacing parts procedures to exchange, maintain, or repair parts in your ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer).

– Finding parts, locations, and addresses

Locate physical part locations and identify parts with system diagrams.

– Common procedures for the ENZ0 PCIe Gen4 I/O expansion drawer

This section contains all the common procedures related to managing the ENZ0 PCIe Gen4 I/O expansion drawer (ENZ0 PCIe4 expansion drawer).

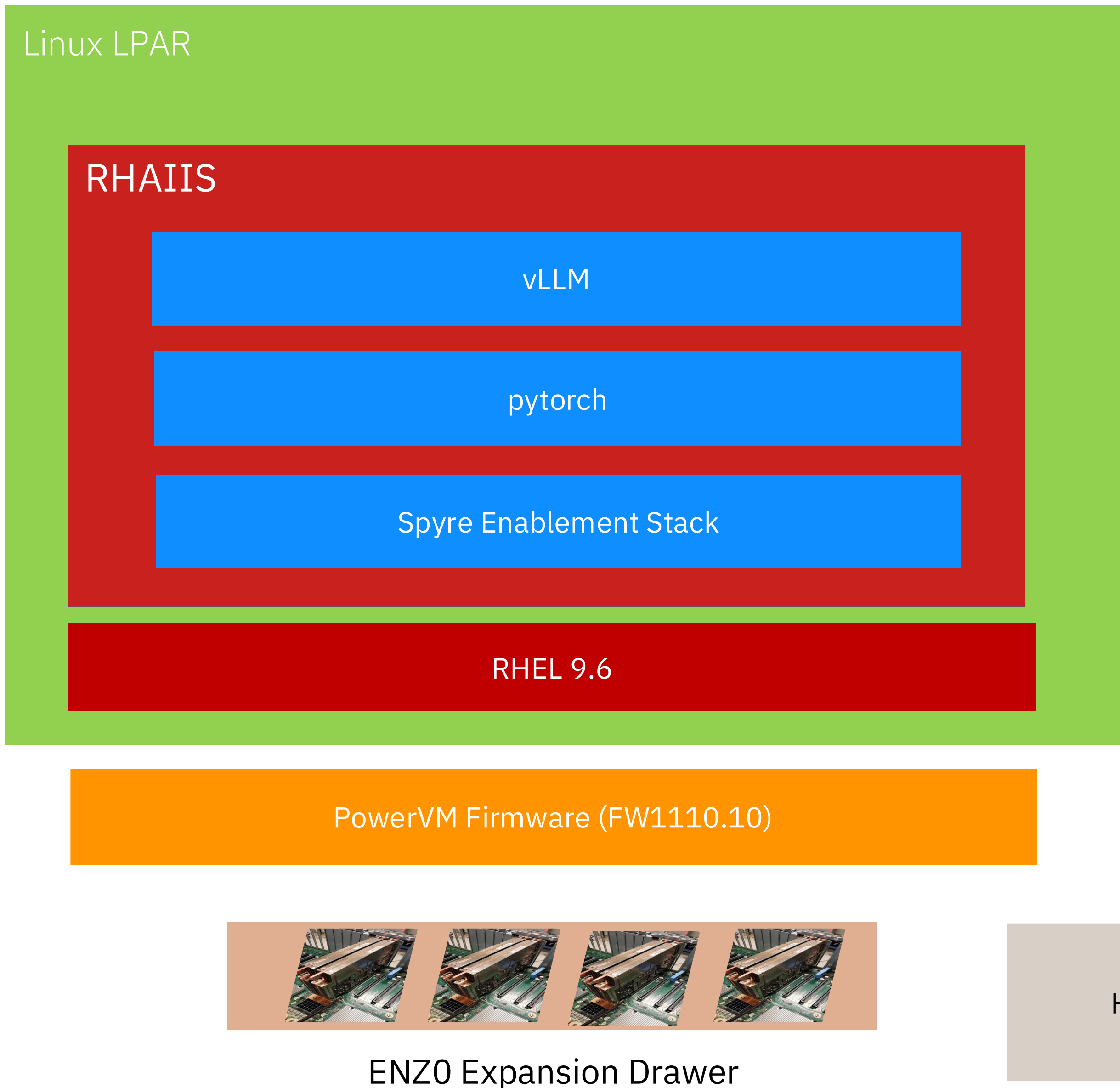
[Home](#) > [Products](#) > [Red Hat AI](#) > [Red ...](#)

Red Hat AI Inference Server

Red Hat® AI Inference Server optimizes inference across the hybrid cloud for faster, cost-effective model deployments.

[Try It](#)[Talk to a Red Hatter](#)[Explore the latest release →](#)

Spyre on Power Software Stack



- Vllm is an open souce library that provides inference serving at scale
- Pytorch is an open source deep learning library
- Spyre enablement stack includes the low level software to interface with the Spyre hardware
- RHAIIS is delivered by Red Hat as a container image which can be deployed on RHEL using podman
- The Spyre enablement stack is part of the RHAIIS container image
- The Spyre enablement stack is released quarterly with new features, functions, and model support



☒ Show full table of contents

Filter on titles

Welcome

IBM Spyre Accelerator for Power



Introduction to IBM Spyre Accelerator



Overview of use cases



Configuration and installation



Serviceability



Known issues and limitations

Container environment variables for Spyre

Announcements & sales manuals



Download PDF

[All products](#) / [IBM Spyre Enablement Stack for Power](#) /

Was this topic helpful?



IBM Spyre Accelerator for Power

Last Updated: 2025-12-11

IBM Spyre™ Accelerator for Power is an energy-efficient, high-performance PCIe-attached AI inference card that is available for Power11 systems. Built on 5 nm SoC technology, the Spyre Accelerator for Power supports FP16 execution and continuous batching within a 75 W power envelope.



Note: This publication contains information about deploying and using the IBM Spyre Accelerator for Power that is integrated in the Red Hat® AI Inference Server (RHAIS) release for IBM Power Systems.

The PDF version of this publication, IBM Spyre Accelerator for Power, is available for download [here](#).

– Introduction to IBM Spyre Accelerator

IBM Spyre™ Accelerator for Power is available for Power11 systems in the ENZO PCIe4 expansion drawer and requires Red Hat® AI Inference Server (RHAIS).

– Overview of use cases

IBM Spyre™ Accelerator for Power supports various use cases and models.

– Configuration and installation

Configure the IBM Spyre Enablement Stack for Power.

– Serviceability

You can improve system reliability and troubleshoot with serviceability features, including hardware information, vital product data (VPD), and data collection tools such as aiu-smi for performance analysis.

– Known issues and limitations

Known issues and limitations contain important information that can help you avoid problems on your system.

– Container environment variables for Spyre

Environment variables are used to configure settings in the container.

Getting started

1. [Red Hat® Enterprise Linux® \(RHEL\) 9.6](#) installed and registered with Red Hat subscription
2. [Red Hat AI Inference Server \(RHAIS\) repository](#) enabled
3. [System firmware](#) updated as per platform requirements
4. [Root or sudo](#) access
5. [Outbound network](#) access for model download
6. [IBM Power Tool Repository](#) enabled

Installation

(1)

1. Install prerequisite software
 - `dnf install -y container-tools git git-lfs wget`
 - `rpm -ivh`
<https://public.dhe.ibm.com/software/server/POWER/Linux/yum/download/ibm-power-repo-latest.noarch.rpm>
 - `dnf install -y ServiceReport ibm-power-managed`
2. Create non root user for spyre
 - `adduser senuser`
 - `passwd senuser`
3. Validate system for Spyre
 - `servicereport -v -p spyre`
4. Configure the system for Spyre
 - `servicereport -r -p spyre`
5. Add user to sentient group
 - `usermod -aG sentient senuser`

Installation

(2)

1. Create the model source directory for downloading models
 - `install -d -m 0775 -o root -g sentient /opt/ibm/spyre/models/src`
2. Enable systemd linger for non-root user
 - `loginctl enable-linger senuser`
3. Enable persistent storage for container logs
 - `install -d -m 2755 -o root -g systemd-journal /var/log/journal`
 - `sed -ri 's/^#?Storage=.* /Storage=persistent/' /etc/systemd/journald.conf`
 - `systemctl restart systemd-journald`

Installation

(3)

1. Login as senuser
2. Setup user directory
 1. `mkdir -p ~/.config/containers/systemd`
3. Download AI models
 1. `git lfs install`
 2. `cd /opt/ibm/spyre/models/src`
 3. `git clone https://huggingface.co/ibm-granite/granite-3.3-8b-instruct`
 4. `git clone https://huggingface.co/ibm-granite/granite-embedding-278m-multilingual`
4. Install Red Hat AI Inference Server
 1. `podman login registry.redhat.io`
 2. `podman pull registry.redhat.io/rhaiis/vllm-spyre:3.2.5`
 3. `sudo setsebool -P container_use_devices 1`



☒ Show full table of contents

 Filter on titles

Welcome

AI Services



Introduction

Configuration and Installation



RAG Chatbot



CLI Reference



Troubleshooting



FAQ

Limitations

Developer Guide



Glossary

Announcements & sales manuals



Download PDF 

[All products](#) / [IBM Open-Source AI Foundation for Power](#) /

Was this topic helpful?



AI Services

Last Updated: 2026-02-10

- [Introduction](#)
Overview of AI Services
- [Configuration and Installation](#)
Configure and install AI Services
- [RAG Chatbot](#)
Retrieval-Augmented Generation (RAG) Chatbot Overview
- [CLI Reference](#)
Comprehensive guide to AI Services Command-Line Interface
- [Troubleshooting](#)
Guidance to diagnose and resolve common issues
- [FAQ](#)
Frequently asked questions about AI Services
- [Limitations](#)
Details on design constraints and documented challenges
- [Developer Guide](#)
Developer guide for integrating with the AI Services
- [Glossary](#)

References & demonstrations help clients aligning AI use cases with their core workloads quickly.

Cross-industry

ITOps | Development

System house
DACH

IT service desk assistant

Detect & fix incidents agent

MR WILLIAMS
More Than A Convenience Store Distributor

Code assistants
(RPG, Ansible, ...)

Forecast & plan capacity assistant

Industry-specific

Enterprise Resource Planning

BI & HR assistant

Order processing assistant

Large retailer
US

BECOM
it's possible.

Production planning

Large retailer
US

KAPS
omni

Product sales assistant

Banking & Finance

Demos
TechXchange
2024/2026

PORTOLAN
Commerce Solutions

Analyst assistant
(frauds, NPAs, ...)

Infosys
Finacle

Predict NPAs

Anti-money laundering

CREST

Open account agent

Healthcare

SHIBUYA
BY PULSEN

Medical assistant

Claims & EHR matching agent

SiPH

Medical transcription assistant

Medical image analysis assistant

Insurance

Demo
TechXchange 2025

Claims & policy management agent

Predict risk & underwrite assistant

Public | Real Est.

Gov. client
DACH

Private documents assistant

Real estate assistant

...and more

SEMICON
INDIA

Agriculture assistant

Slide search assistant

...

integrated intelligent apps

Assisted workflows
(deep integration of AI services into enterprise workflows)

Digital assistants & advisors
(retrieval augmented generation, chat interactions, ...)

Agentic collaborators
(tools & skills, autonomous orchestration, ...)

pre-built services & tools

Digitalize documents
(manual, invoice,...)

Find similar items
(product, client, ...)

Q&A from knowledge

Extract information
(orders, PII, ...)

Redact documents
(invoices, forms, ...)

Summarize

Translate

Transcribe
(meetings, phone calls, dictation, ...)

NLP to SQL
(Db2, Oracle, SAP HANA, ...)

Forecast
(expenses, supply, demand, revenue)

...

Digital assistants

Services

Use case references

Use case references

Ready-to-explore use cases based on real-world AI solutions to help you envision how AI can solve common business problems—and accelerate your AI journey.

🔍 Search

Filters

Provider

- ☐ IBM (9)
- ☐ Private
- ☐ Public (third-party)
- ☐ IBM certified (any provider) (9)

Domains

- ☐ Agriculture (1)
- ☐ Banking and finance (1)
- ☐ Enterprise resource plan (1)
- ☐ Insurance (1)
- ☐ Professional services (1)

Assets

- ☐ Video demos (3)
- ☐ Interactive prototypes
- ☐ Code sources (2)
- ☐ Reference stories (5)



Agriculture

Agriculture assistant

Provides farmers with accurate, validated crop-growing guidance tailored to their region and preferred language.

Digital assistant



Banking and finance

Analyst assistant

Enables finance teams to investigate fraud, identify high-value clients, and uncover non-performing assets using their core data—without needing specialized AI or IT skills such as SQL.

Digital assistant



Enterprise resource planning

BI assistant

Enables teams to ask a question to analyze their data, understand business logic, and uncover what's happening, why it matters, and what to do next.

Digital assistant



Professional services

Conference slide search

Innovates conference experiences by enabling fast access to presentation slides from current and past events through a search-driven digital assistant.

Digital assistant



Insurance

Claims and policy management



IT operations

IT service desk assistant

Digital assistant

Services

Use case references

Digital assistant

Production-ready AI solutions that combine multiple services into intelligent, integrated systems for complex use cases using Retrieval-Augmented Generation (RAG).

DeploymentsAbout







Deploy


Deployment name	Status	Uptime	Type	Messages
Incident troubleshooting	Error	Mar 4, 2026	Digital assistant	Error message goes ...
Process FAQs	Running	2 days	Summarize	
Permissions ops	Running	Mar 4, 2026	Digital assistant	
Digitize documents	Running		Service	
Find similar item	Running		Service	
Question and answer	Running		Service	
Summarize	Running		Service	
Deals tracker	Running	Jan 15, 2026	Find similar item	
Contract analysis agent	Running	Jan 15, 2026	Question and answer	
Case routing	Running	Jan 15, 2026	Digitize documents	

Items per page: 20

1–20 of 100 items

1
of 5 pages

Digital assistant

Services

Use case references

Digital assistant

Production-ready AI solutions that combine multiple services into intelligent, integrated systems for complex use cases using Retrieval-Augmented Generation (RAG).

Deployments

About

Deploy



Card title

Services

- Digitize documents
- Find similar items
- Question and answer
- Summarize

Use case references

- Agriculture assistant
- Analyst assistant
- BI assistant
- Claims & policy management agent
- Conference slide search
- Financial documents assistant
- Invoice matching assistant
- IT service desk assistant
- Open account agent
- Order processing assistant
- Private documents assistant
- Product sales assistant
- Real estate assistant

Minimum resource allocation

Required cores

0.5 - 2.0

Required memory

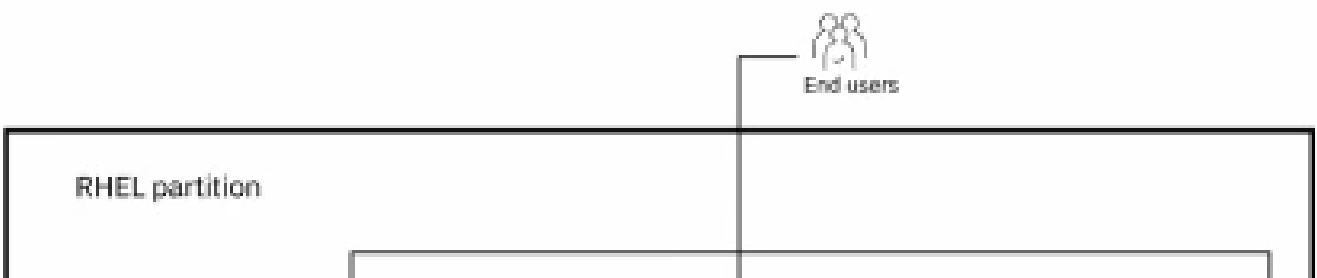
15GB -25GB

Required Spyre cards

4 cards

Code and architecture

View code



Demos and prototypes



Digital assistant

Services

Use case references

Digital assistant

Production-ready AI solutions that combine multiple services into intelligent, integrated systems for complex use cases using Retrieval-Augmented Generation (RAG).

Deployments

About

Search input text

Deploy

	Deployment name	Status	Uptime	Type	Messages
▼	Incident troubleshooting	Error	Mar 4, 2026	Digital assistant	Error message goes ...
	Process FAQs	Running	2 days	Summarize	
^	Permissions ops	Running	Mar 4, 2026	Digital assistant	
	Digitize documents	Running		Service	
	Find similar item	Running		Service	
	Question and answer	Running		Service	
	Summarize	Running		Service	
	Deals tracker	Running	Jan 15, 2026	Find similar item	
	Contract analysis agent	Running	Jan 15, 2026	Question and answer	
	Case routing	Running	Jan 15, 2026	Digitize documents	

Items per page: 20

1–20 of 100 items

1 of 5 pages

Q

Search input text

Deploy

Deployment name	Status	Uptime	Type	Messages
Incident troubleshooting	Error	Mar 4, 2026	Digital assistant	Error message goes ...
Process FAQs	Running	2 days	Deep process integration	
Permission requests	Running	Mar 4, 2026	Digital assistant	
Contract analysis agent	Running	Mar 4, 2026	Summary	
Case routing	Running	Mar 4, 2026	Translation	
Deals tracker	Running	12 hours	Digital assistant	
Privacy, redaction, audit	Running	Jan 2, 2026	Question and an...	
IT support triage	Running	25 minutes	Digital assistant	
Sales deck generator	Running	Nov 9, 2025	Digital assistant	

Items per page: 100

1–100 of 100 items

1 of 10 pages



Support Agent



watsonx 8:57 PM

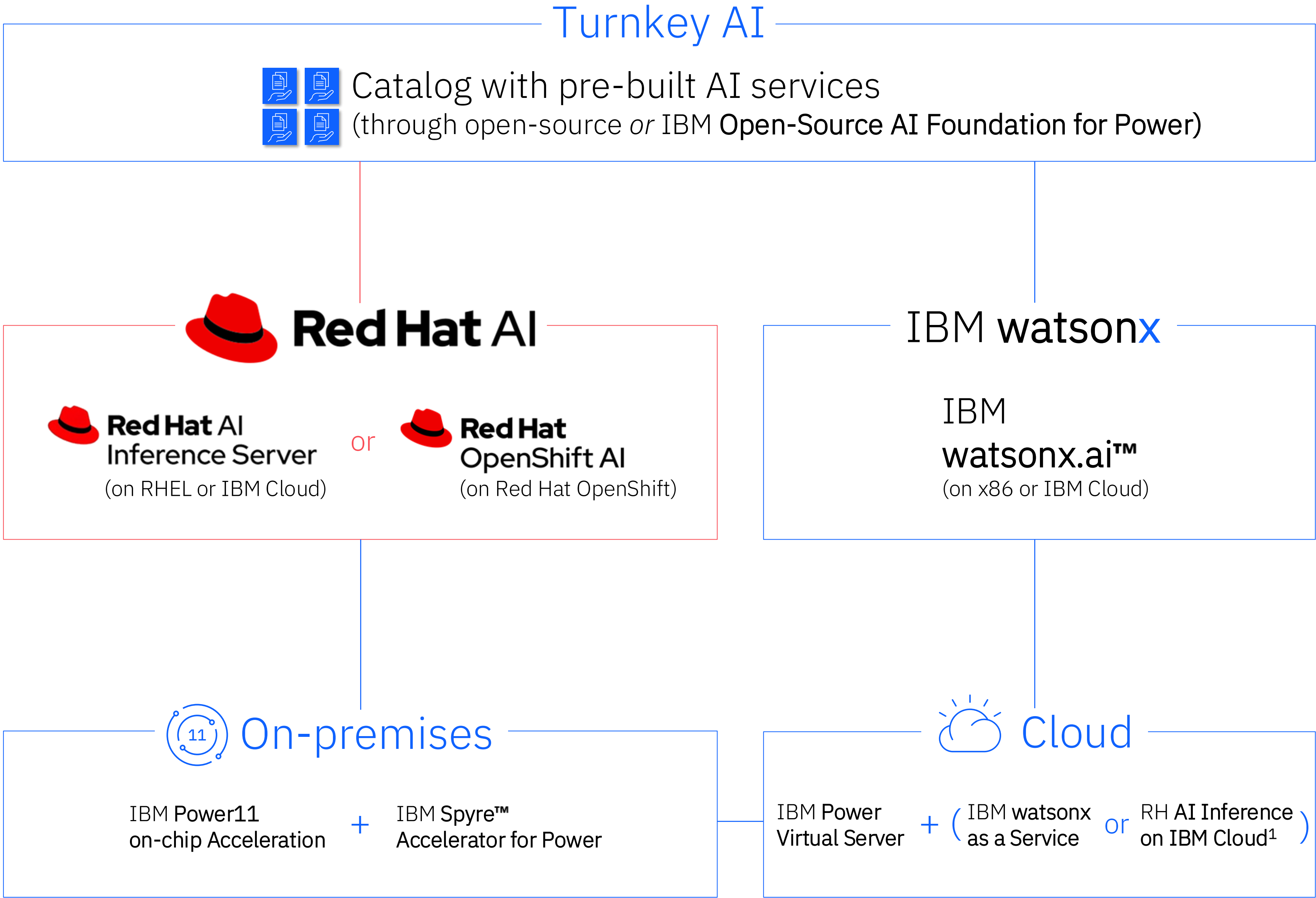
Hi, I'm your IT Support Assista

Type something...

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

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

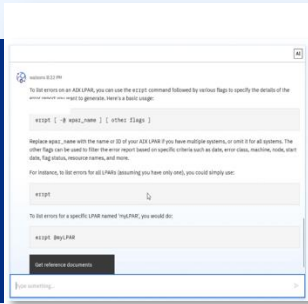
...empowered by reliable & safe infrastructure options.



1: Integration of AI services with RH AI Inference on IBM Cloud targeted for 3Q 2027.

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

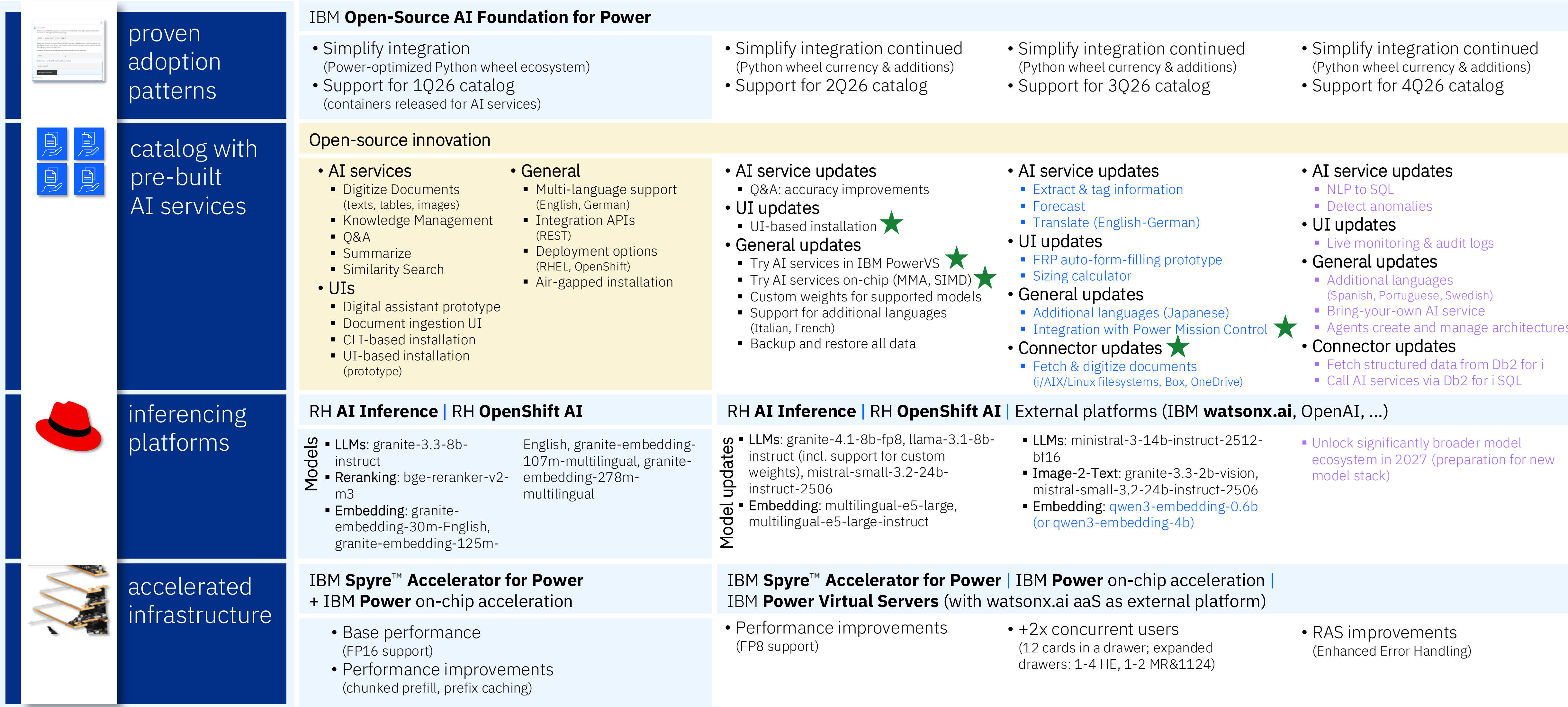
...to move AI services of the IBM-supported catalog between IBM Power & IBM Power Virtual Server.²

> 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.

Delivery Roadmap



Roadmap is subject to change; IBM has the right to change roadmap without prior notification.

What's New

Last Updated: 2026-06-16

[Edit online](#)

- Introduced the **Similarity Search** service, enabling efficient and accurate semantic search using advanced vector-based retrieval techniques.
- Expanded language support to include **French** and **Italian**, allowing broader multilingual adoption of AI Services applications.
- Added support for **multi-turn conversations** with context management, enhancing the Q&A chatbot experience by enabling more natural, context-aware interactions.
- Introduced a **Catalog UI** for complete application lifecycle management, including deployment and management of services. This feature is currently available for the **Podman runtime**.
- Integrated **Caddy reverse proxy** to provide secure HTTPS communication across all services, improving overall security and simplifying configuration.
- Added support for **Watsonx integration** as an alternative to instruct CPU-based vLLM for users who do not require Spyre, enabling inference workloads to run on the IBM Cloud Watsonx service. **LiteLLM** can also be used as a flexible routing layer for model access.
- Expanded Digitize capabilities to support **.docx files**, allowing seamless ingestion and processing of Word documents.
- Introduced support for custom system prompts and **secure vLLM access** through API key–based authentication. These features must be explicitly enabled by users.

Parent topic:

→ [AI Services](#)

Turnkey Connectors (in planning)

Quick Start Data Management

Get your data connected in minutes



Connect SharePoint

Automatic sync keeps your knowledge base up to date

Connect Now

Recommended



Upload Local Files

Quick testing and development with local files

Upload Files

Recent Files

☐	File name	Source	Size	Last updated	Status	Version		
☐	 Budget_report_2025_Q2.pdf	SharePoint	2.5 MB	2 minutes ago	Ready	undefined	:	
☐	 FAQ_Document.docx	Local	1.2 MB	1 hour ago	Ready	undefined	:	
☐	 Product_Guide_2024.pdf	SharePoint	3.1 MB	3 hours ago	Ready	undefined	:	

5

Total Files

5

Ready

0

Processing

Turnkey Connectors (in planning)

Unified Data Management

Manage all your data sources in one place - SharePoint, local files, and more

Add Data +

Active Connections

✓ Corporate SharePoint

Connected

Last synced:

2 minutes ago

Next sync:

In 58 minutes

Files:

127

HEALTH METRICS

Uptime

99.9%

Avg Sync

45s

Error Rate

0.1%

Sync Now ↺

Test

Settings ⚙

All Sources (5) SharePoint (3) Local Files (2) By Tag Activity

☐	File name	Source	Size	Last updated	Version	Tag	
☐	 Product_Guide_2024.pdf	SharePoint	3.2 MB	Mar 10, 2026	v2	Products	⋮
☐	 Local_Test_Document.docx	Local	1.1 MB	Mar 15, 2026	v1	Testing	⋮
☐	 FAQ_2024.pdf	SharePoint	890 KB	Mar 12, 2026	v3	Support	⋮
☐	 Training_Manual.pdf	Local	2.5 MB	Mar 14, 2026	v1	Training	⋮
☐	 Policy_Document_2024.docx	SharePoint	1.8 MB	Mar 11, 2026	v2	Policies	⋮

Unified Management

All your data sources are managed in one place. Whether files come from SharePoint or local uploads, they're treated equally in your knowledge base.

Source Flexibility

Mix and match data sources based on your needs. Use SharePoint for documents that change frequently, and local uploads for static reference materials.

Smart Organization

Organize files by tags regardless of source. This makes it easy to find related content across all your data sources.

Turnkey Connectors (in planning)

Data Management Dashboard

Manage all data sources, files, and processing in one place

Connect Source



Upload Files



Active Connections

✓ Corporate SharePoint

Connected

Last synced: 2 minutes ago
Next sync: In 58 minutes
Files: 127

HEALTH METRICS

Uptime 99.9%
Avg Sync 45s
Error Rate 0.1%

Sync Now Test Settings

Processing Queue

Product_Guide_2024.pdf

Processing

Uploaded
0% complete

Uploaded
Extracting text
Chunking
Generating embeddings
Indexing
Ready

FAQ_Document.docx

Processing

Uploaded
0% complete

Uploaded
Extracting text
Chunking
Generating embeddings
Indexing
Ready

All Files (5) SharePoint (4) Local Files (1) By Tag Activity Usage & Queries

☐	File name	Source	Size	Last updated	Status	Version	
☐	Budget_report_2025_Q2.pdf	SharePoint	2.5 MB	March 12, 2026	Ready	vv3	⋮
☐	Budget_report_2025_Q1.pdf	SharePoint	2.5 MB	March 12, 2026	Ready	vv2	⋮
☐	Budget_report_2024_Q4.pdf	Local	2.5 MB	March 12, 2026	Ready	vv1	⋮

Go for days, not months.



“AI is a core enabler of our financial software solutions, accelerating innovation, enabling smarter decisions, and driving sustainable growth for our customers.”
– Jessica Kreuzmann, CEO, Portolan Commerce Solutions



LOB ISV Enterprise Architect



“The IBM Power team excels at rapidly showcasing innovation—from NLP-to-SQL to forecasting and agentic workflows—helping us identify the next generation of capabilities for our financial software solutions.”
– Alexander Nürk, CTO, Portolan Commerce Solutions



LOB ISV Enterprise Architect



“We installed the software stack within a *single morning*, and by the afternoon we were already playing with digital assistants. As IT administrator, I found the *deployment via the AI services catalog incredibly straightforward*.”
– Andre Lutz, Technical Leader Linux on Power, SVA



“I like using AI to build and execute plans for specific projects. *Bob's planning [...] provided a better understanding* of the steps through the project from beginning to end.”
– Bob Richardson, ERP and AI Analyst, Wynne Systems



“With supported, standardized APIs, we can deploy workloads across *cloud and on-premises environments*, aligned with our customers’ preferences.”
– Alexander Nürk, CTO, Portolan Commerce Solutions



ISV Enterprise Architect ITOps Manager



“In our Spyre environment, we’ve integrated a *digital assistant* directly into our ERP interface and use *intelligent information extraction* to convert customer inquiries into actual orders. We have also integrated our sales software with Spyre, which proactively supports *B2B sales with data-driven recommendations*.”
– Michael Klinger, CEO, KAPS GmbH



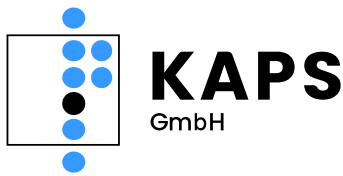
“This [...] allows us to expand our KOBE offering to include local LLM deployment, *bypassing the need for cloud-based services*.”
– Giancarlo Lui, CTO, Horsa Power S.r.l.



“MMA *reduced production planning optimization time by ~46%* for Becom, enabling them to securely integrate optimization workflows with their IBM i data and ERP system.”
– Clemens Zauchner, Senior Data Scientist, IT-PS



LOB ISV Enterprise Architect ITOps Manager



“Early customer previews have generated strong excitement, and we are now *preparing for general availability*.”
– Michael Klinger, CEO, KAPS GmbH



“Given these results, *we are officially incorporating Spyre hardware into our commercial proposal for KOBE*.”
– Giancarlo Lui, CTO, Horsa Power S.r.l.



“Forecasting and optimization are reshaping our entire production chain: Cutting volatility, optimizing batch sizes, and enhancing coordination between materials and manufacturing capacity. We have successfully *rolled out the solution across our ERP systems in Europe*.”
– Johann Katona, Head of Organizational Development & IT, BECOM Electronics GmbH



LOB ITOps Manager

Identify AI needs and objectives

- Clarify needs and pains
- Outline opportunities
- Align compliance context

Align and prioritize use cases

- Assess ROI potential
- Assess feasibility
- Specify success criteria

Plan and engineer solutions

- Scope and run POC/MVP
- Design and test solution
- Review continuously

Evaluate ROI and day 1 ops

- Assess success criteria and incorporate feedback
- Deploy and initialize
- Implement change management

Use and day 2 ops

- Employ innovation within everyday business tasks
- Validate in-production ROI
- Ensure day-2 operations

Disclaimers: (1) SVA has externally shared experiences with Spyre on LinkedIn (https://www.linkedin.com/posts/udo-sachs_ibmchampion-ai-ibmpower-ugcPost-7442243262792343552-xWHX/7utm_source=share&utm_medium=member_desktop&rcm=ACoAABafyM0B3VuT78eQAUTiufiZT0ADhZmimB4) and within a digital (German) live event, published at YouTube: (1) “KI-Strategie: Warum sie scheitert – und wie Erwartungsmanagement gepaart mit schlüsselfertiger KI sie retten kann” (https://youtu.be/vqksbP0_uCQ) and (2) “Schlüsselfertige KI mit IBM Spyre im technischen Fokus: Architektur und KI-Workloads.” (<https://youtu.be/creYkigango>). The quotes on this slide come from these venues, were freely translated into English, and then approved by SVA. (2) Client Becom and their partner IT-PS have internally measured the average total execution times when running their production planning optimization (based on linear programming) over a given input data set (static data sample with all assembly units that involve several hundred individual optimizations). Their test setup resembles the production environment with a Linux LPAR on IBM Power with 3 physical IBM Power cores (SMT8, 24 Linux virtual cores, nproc=24), 23.8GB memory, and an external storage (IBM Storwize Flash Storage). On IBM Power9-based servers (IBM S922), they have measured 9.95 minutes and on IBM Power10-based servers (IBM S1022), they have measured 5.3 minutes, leading to a reduction of approximately 46%. More detailed client stories are available at <https://www.ibm.com/downloads/documents/us-en/15db45f02d520e7a> and <https://www.ibm.com/it-infrastructure/us-en/resources/power/ibm-i-customer-stories/> (BECOM Electronics GmbH).

Documentation

- IBM Spyre Accelerator for Power Documentation
 - <https://www.ibm.com/docs/en/ibm-spyre-for-power?topic=spyre-accelerator-power>
- Red Hat AI Inference Serving with IBM Spyre Accelerators
 - https://docs.redhat.com/en/documentation/red_hat_ai_inference_server/3.2/html/getting_started/inference-rhaiis-with-podman-spyre-power_getting-started
- IBM Power Systems Service And Productivity Tools
 - <https://www.ibm.com/support/pages/service-and-productivity-tools>
- Spyre Accelerator Hardware Documentation
 - <https://www.ibm.com/docs/en/power11/9856-42H?topic=ad-pcie5-x16-ai-accelerator-adapter-fc-ecse-ccin-2e1f>
- AI Services
 - <https://www.ibm.com/docs/en/aiservices?topic=ai-services>
- Implementing AI on Power11: Introducing the IBM Spyre Adapter Redbook
 - <https://www.redbooks.ibm.com/redpieces/pdfs/sg248592.pdf>

Thank you

© 2026 International Business Machines Corporation

All rights reserved. The information contained in these materials is provided for informational purposes only and is provided AS IS without warranty of any kind, express or implied. Any statement of direction represents IBM's current intent, is subject to change or withdrawal, and represent only goals and objectives. IBM, the IBM logo, and Power are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

