Welcome to Lyon, France and the 2026 Common Europe Congress

Bienvenue à Lyon, en France, et au Congrès de Common Europe 2026

Power11 Announcements

Put AI into production

Lenez Jean-Manuel
IBM Switzerland

IBM

Jean-Manuel Lenez
jlen@ch.ibm.com
+41 79 278 92 44

Pre-Sales
IBM Power Systems
& Hybrid-Cloud

IBM Suisse
Esp. de Pont-Rouge 9A,
1212 Lancy, Suisse

16th June 2026

Topics covered

- Power11 overview
- Power11 Hardware announcement
- Whats new across Power11 Portfolio
- Put AI into production with Power11

Lenez Jean-Manuel
IBM Switzerland

IBM

Jean-Manuel Lenez
jlen@ch.ibm.com
+41 79 278 92 44

Pre-Sales
IBM Power Systems
& Hybrid-Cloud

IBM Suisse
Esp. de Pont-Rouge 9A,
1212 Lancy, Suisse

Power11 overview

Power11 Timeline

July 2025

E1180

- *Higher levels of availability and reliability*
- *New Quantum safe security enhancements*
- *Increased core / thread performance*

E1150

- *Higher levels of availability and reliability*
- *Significant Throughput Increase*
- *New Quantum safe security enhancements*
- *Increased memory bandwidth*

Spyre Tech Preview

S1124 and S1122

- *Higher levels of availability and reliability*
- *Significant Throughput Increase*
- *New Quantum safe security enhancements*
- *Increased memory bandwidth*

4Q 2025

Spyre

- *Enterprise Grade AI chip*
- *Competitive perf/watt*
- *Enables Generative AI capabilities*

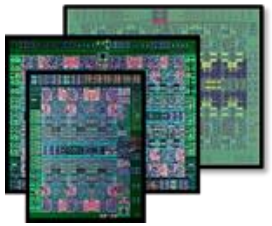
2026

S1112

- *1-Socket Optimized Server*
- *Retail Edge Server*
- *Reduced form factor (same as S1012)*
- *DDR5 DDIMM Memory support*

Power - Continuous System Innovation

POWER8 Family

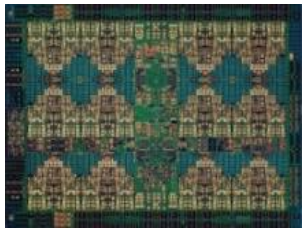


2014
22nm

Big Data Optimized

- Up to 12 Cores
- SMT8
- CAPI Acceleration
- High Bandwidth GPU Attach

POWER9 Family

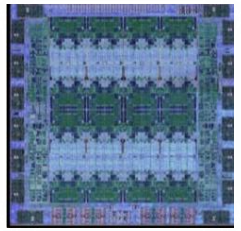


2017
14nm

Built for the Cognitive Era

- 12 / 24 core chips
- Slim or Fused core
- 4 / 8-way SMT
- 10MB L3 cache per core
- DDR4 RAM
- PCIe Gen4 on-chip
- CAPI 2.0, NVlink 2.0
- 25 GB/s interconnect

Power10 Family

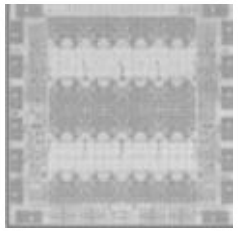


2021
7nm

Built for Hybrid Cloud

- Up to 15/24 cores per socket
- 4 / 8-way SMT
- AI-optimized ISA
- Energy efficiency focus
- HW enforced security
- PowerAXEON 2.0
- PCIe Gen5 on-chip
- Memory clustering

Power11 Family

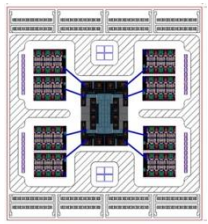


2025
7nm

Built for the AI Era

- Zero planned downtime
- Scalable and sustainable AI
- Energy Optimization
- Frictionless Hybrid Cloud
- Quantum Safe
- Best in class HANA TCO with SMT8

Power Next



202x
5nm

Built for the future

- Advanced Multi-chip Packaging
- Larger L3 cache
- Reduced system latency
- Improvement in single thread performance
- Improvement in core performance

Comparing Power11, Power10, and POWER9 CPUs

Characteristics	Power11	Power10	POWER9
Technology	7 nm	7 nm	14 nm
Die size	654 mm^2	602 mm^2	693 mm^2
Processor module size	68.5 mm x 77.5 mm	68.5 mm x 77.5 mm	68.5 mm x 77.5 mm
Number of transistors	30 billion	18 billion	8 billion
Maximum cores	16	15	12
Maximum hardware threads per core	8	8	8
Maximum static frequency / high performance frequency range	3.8 - 4.4 GHz	3.75 - 4.15 GHz	3.9 - 4.0 GHz
L2 Cache per core	2048 KB	2048 KB	2048 KB
L3 Cache	8 MB of L3 cache per core with each core having access to the full 128 MB of L3 cache, on-chip high-efficiency SRAM	8 MB of L3 cache per core with each core having access to the full 120 MB of L3 cache, on-chip high-efficiency SRAM	10 MB of L3 cache per core with each core having access to the full 120 MB of L3 cache, on-chip eDRAM
Memory technology	DDR5 and DDR4*	DDR4 and DDR5**	DDR4 and DDR3
I/O bus	PCIe Gen 5	PCIe Gen 5	PCIe Gen 4

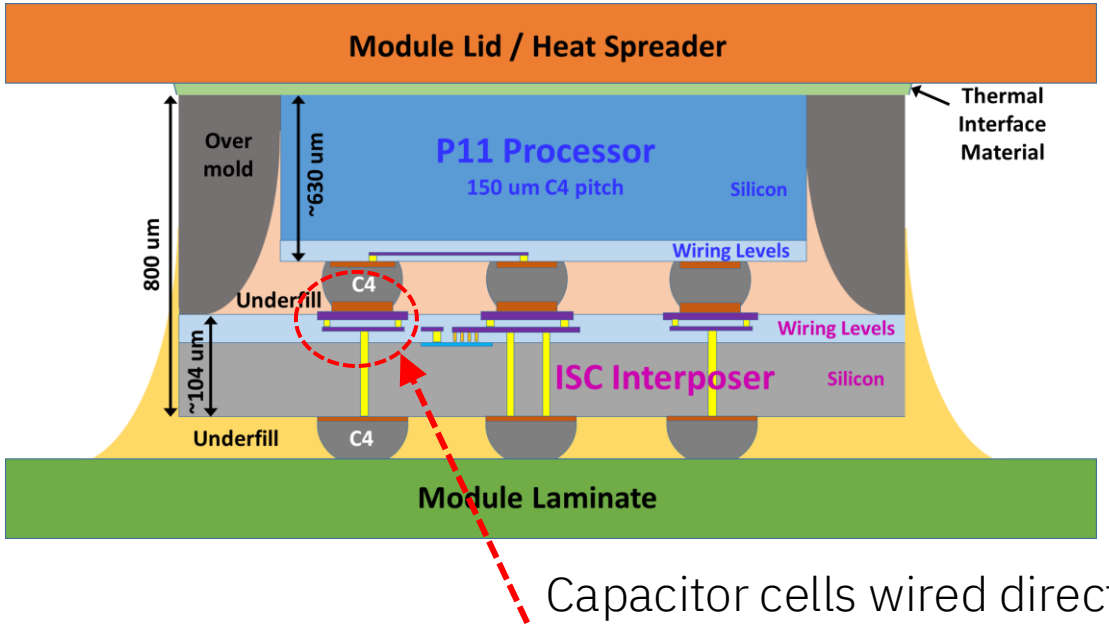
* DDR4 only as migration from Power10 during same serial number upgrade
** DDR5 support added after GA

7

Power11 - Processor

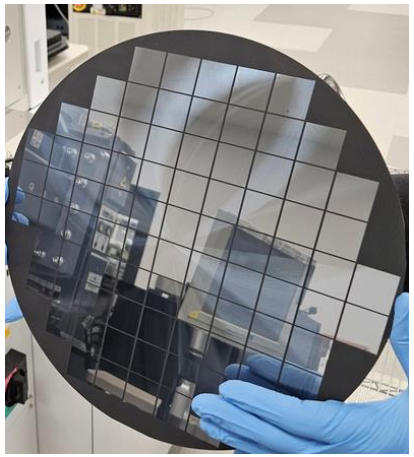
2.5D ISC (Integrated Stacked Capacitor) Structure

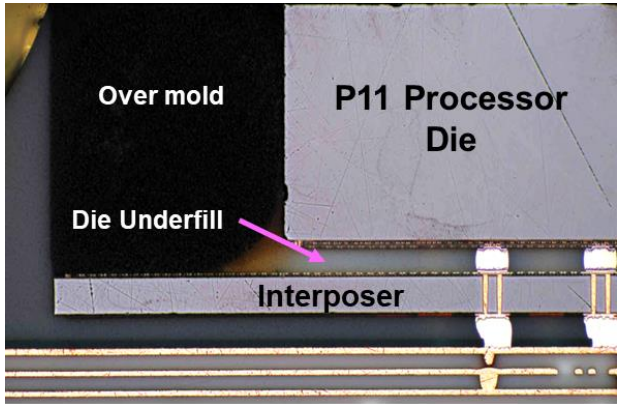
- P11 Processor stacked on ISC (Integrated Stacked Capacitor) Interposer by Samsung
- Integrating **capacitor cells directly into the interposer improves power stability** and reduces electrical noise, and enables a 50–100 MHz increase in processor frequency.



Capacitor cells wired directly into the power grid within the Interposer

P11 + ISC "Chip on Wafer" Samsung



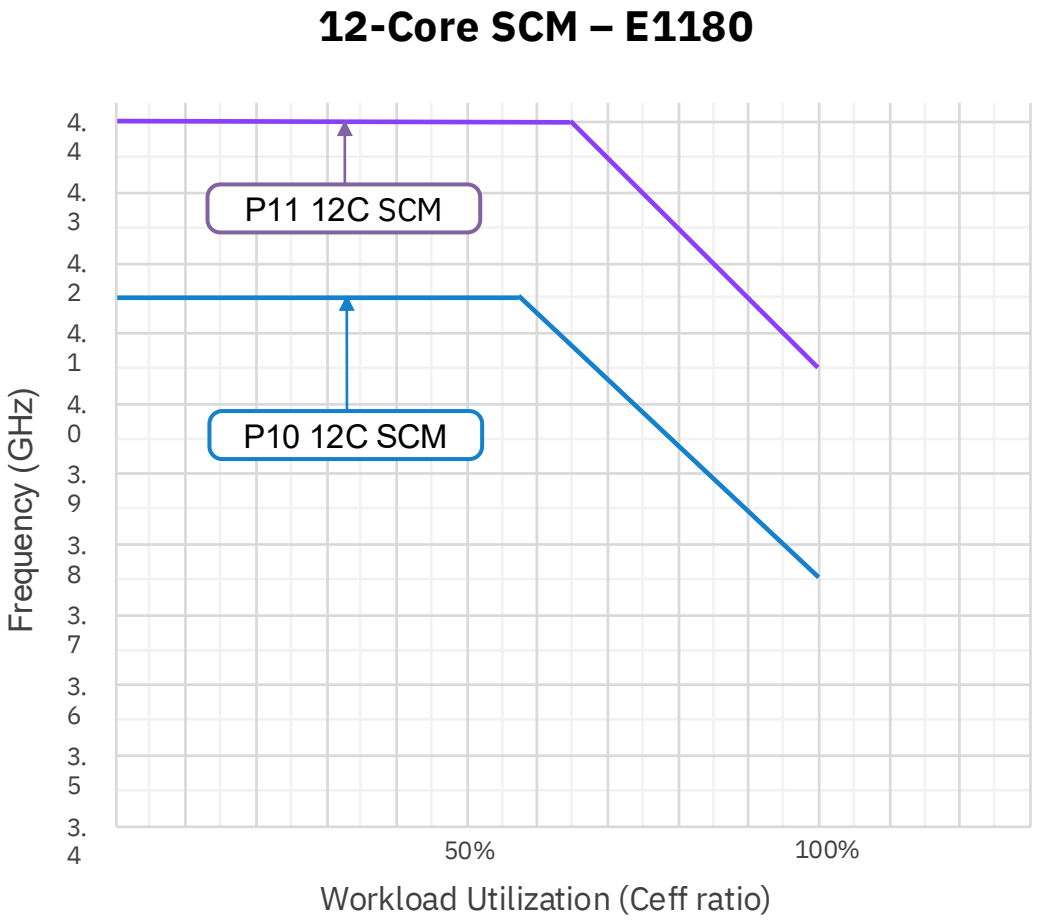


Hardware X-Section Image - EFK F/A team, T.W.

Based on Samsung iCube Si Interposer technology + DRAM / Capacitor experience

Power11 – Frequency improvements example

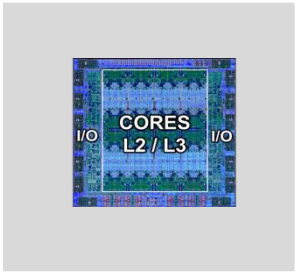
Power11 doesn't just deliver a higher peak frequency, **it sustains higher frequencies across a broader range of workloads**, providing more consistent and predictable performance.



- Depending on workload utilization, **frequency increase** is between 6-8%
- **Less performance variability** with reduced range of frequency operation

Power11 - Processor Options

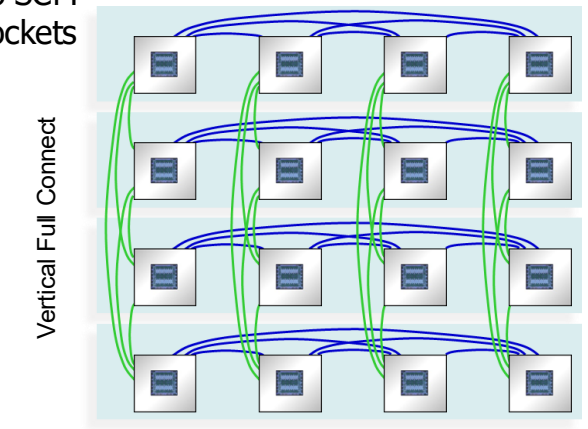
Max performance



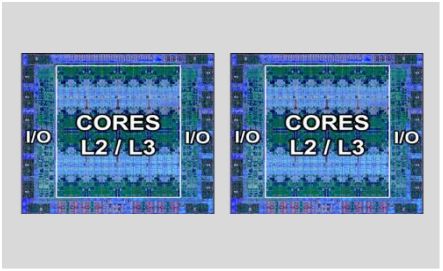
SCM - Single-Chip Module for E1180

- Enables highest frequencies and best core performance
- Provides up to 16 cores per socket (10/12/16 cores)
- 16 OMI Memory channels
- 32 PCIe lanes

Up to 16 SCM Sockets



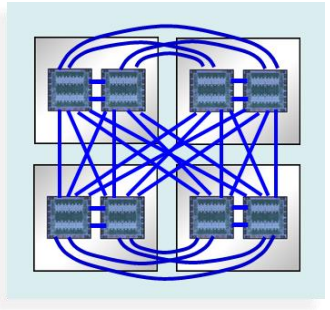
Max core density



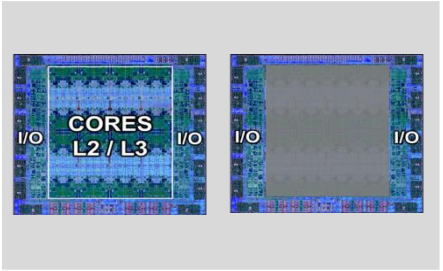
DCM - Dual-Chip Module for S1122/S1124/E1150

- Enables full range of offerings
- Provides up to 30 cores per socket (16/24/30 cores)
- 16 OMI Memory channels (8 per chip)
- 64 PCIe lanes

Up to 4 DCM Sockets



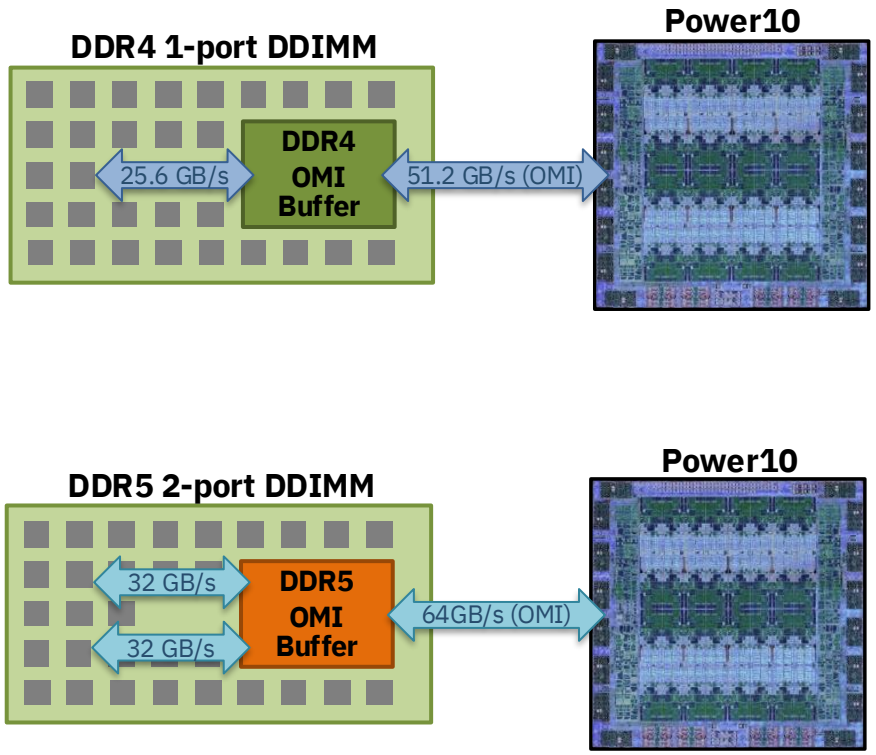
Lower entry price

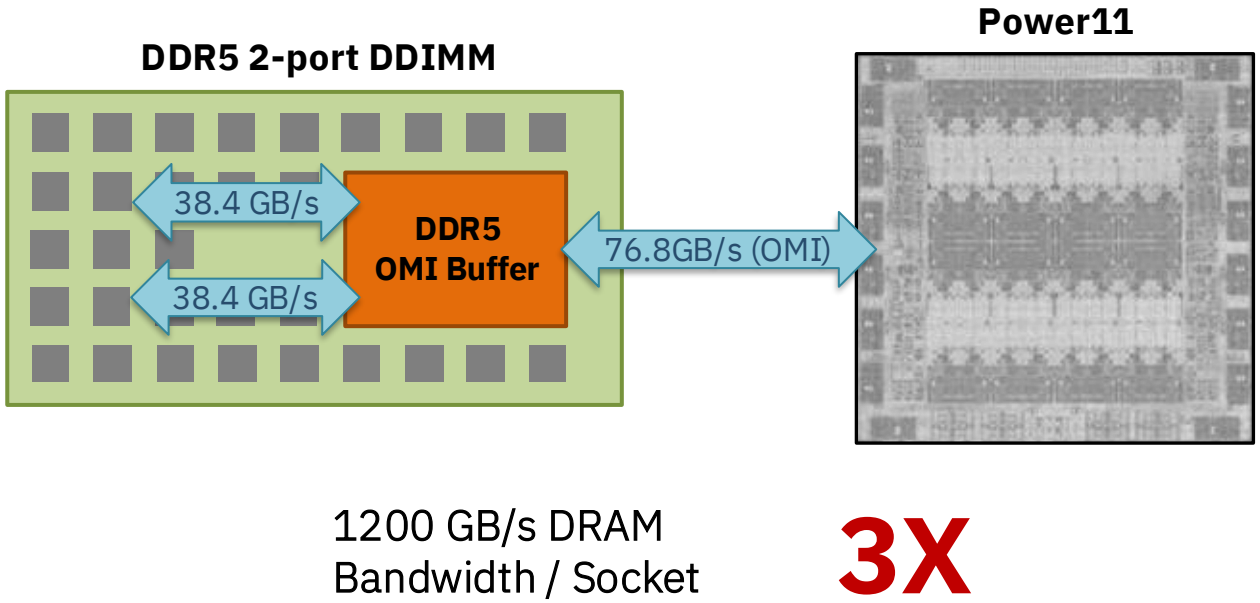


eSCM – entry Single Chip Module for S1122

- One die with compute, memory and I/O, 2nd die adds I/O only
- Enables low cost processor offerings
- Up to 10 cores per socket (4/10 cores)
- 8 OMI Memory channels
- 64 PCIe lanes
- **Available on S1122 + S1112**

Enhanced DIMM Technology with Power11





1200 GB/s DRAM Bandwidth / Socket

3X

Enhanced Memory Speeds for Power11

E1080

DDR4 @ 3200 / 2933 MHz
DDR5 @ 4000MHz

Up to 50% Increase for DDR5 Memory
Up to 80% Increase from DDR4 Memory

E1050

DDR4 @ 3200 / 2933 MHz
DDR5 @ 3200 MHz

→ **E1150: DDR5 @ 4000 MHz**

S1024

DDR4 @ 3200 / 2933 MHz
DDR5 @ 3200 MHz

→ **S1124: DDR5 @ 4800 / 4000 MHz**

S1022

DDR4 @ 3200 / 2666 MHz
DDR5 @ 3200 MHz

→ **S1122: DDR5 @ 4800 MHz**

S1012

DDR4 @ 3200 MHz

→ **S1112: DDR5 @ 4000 MHz**

Power11 Hardware announcement

IBM Power S1112 - Balcones

Planned Announcement: 7/7/26 , Planned GA: 7/24/26

- It’s a follow-on for **Bonnell** aka S1012, which originated from foreseen significant cost pressures in the lower-end portfolio line-up.

- The base system design comprehends a **1-socket processor slot with 2U** height and half-rack width.
- Support for two processor core counts: 4 and 10 cores in an **eSCM** module architecture
- It will be offered in two form factors:
  - **Rack-mounted** using a bare-metal sliding drawer with no point of failure
  - **Deskside-mounte**d using a new tower design
  - Each form factor will have its own MTM

Power11 One-Socket System: S1112[1]

Modernize Today, Secure What Matters, Enable What's Next

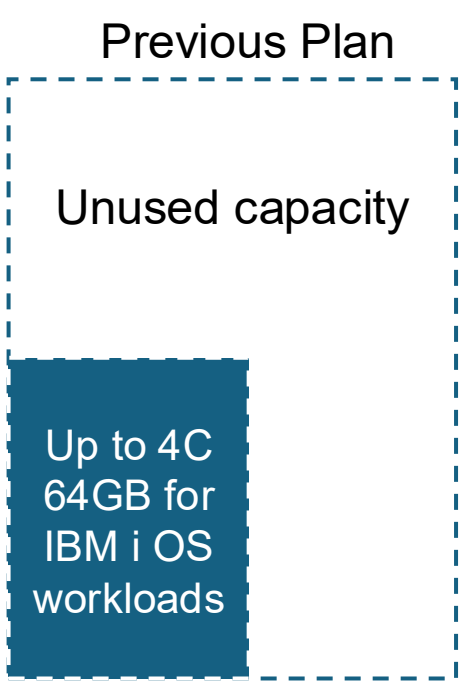
System Overview

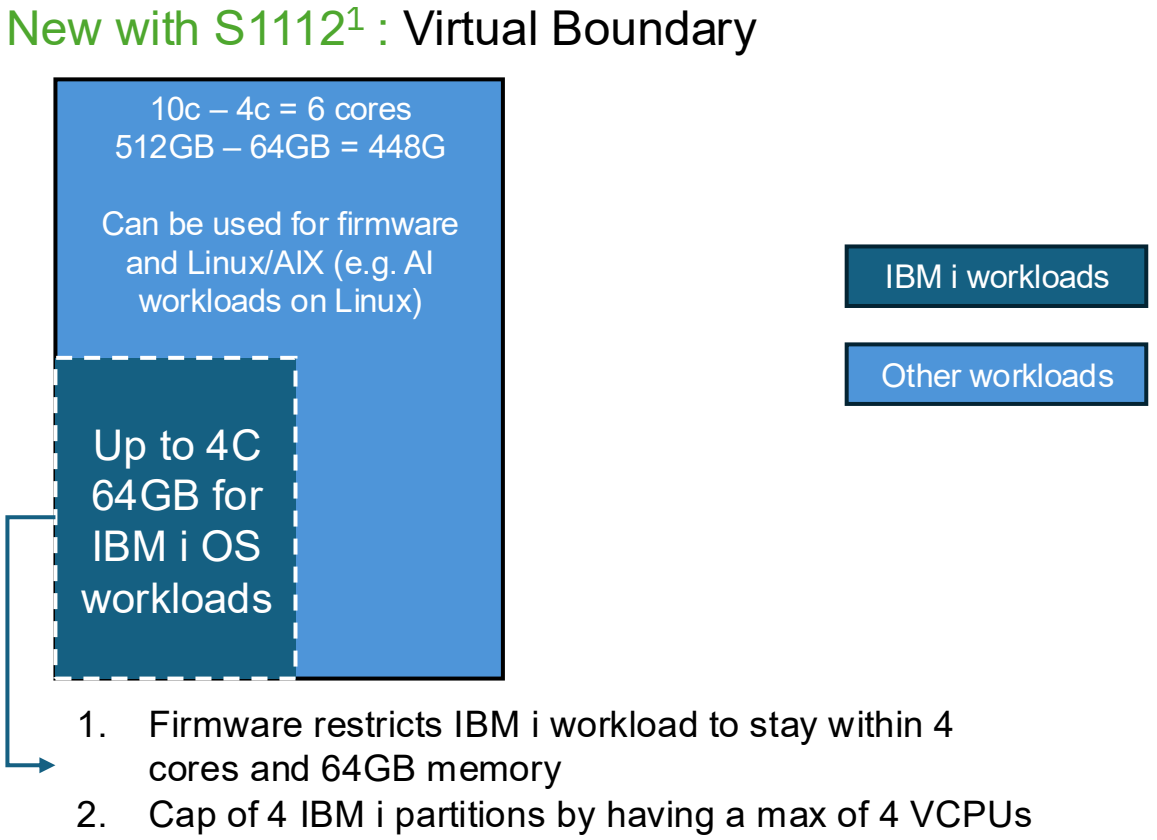
- **4 or 10 core** eSCM modules
- Up to **512GB**
- 4 DDR5 DDIMMs memory
- Up to **4 PCIe slots**
- Up to **4 NVMe drives**
- Up to **12.8 TB** total internal storage
- Support a **half I/O drawer**
- 2 form factors:
  - Rack-mounted (1/2 wide 2U)
  - Deskside-mounted

[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

S1112[1] for IBM i P05

1. Server can be purchased 4c and 10c
2. Up to 512 GB with no system memory restriction
3. Order 4c and deactivate up to 3c if fewer cores are needed





1. Firmware restricts IBM i workload to stay within 4 cores and 64GB memory
2. Cap of 4 IBM i partitions by having a max of 4 VCPUs

[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

Performance Benchmarks for S1112[1]

The S1112 delivers performance leadership at the system and core level in this space (1 Socket, 4 core).

CPW			S1112
			4c
			117,000
S1012	4c	111,300	1.05
S1014	4c	106,300	1.10
S914	4c	52,500	2.23
S814	4c	37,440	3.13

1.05X 4-core S1112 vs 4-core S1012

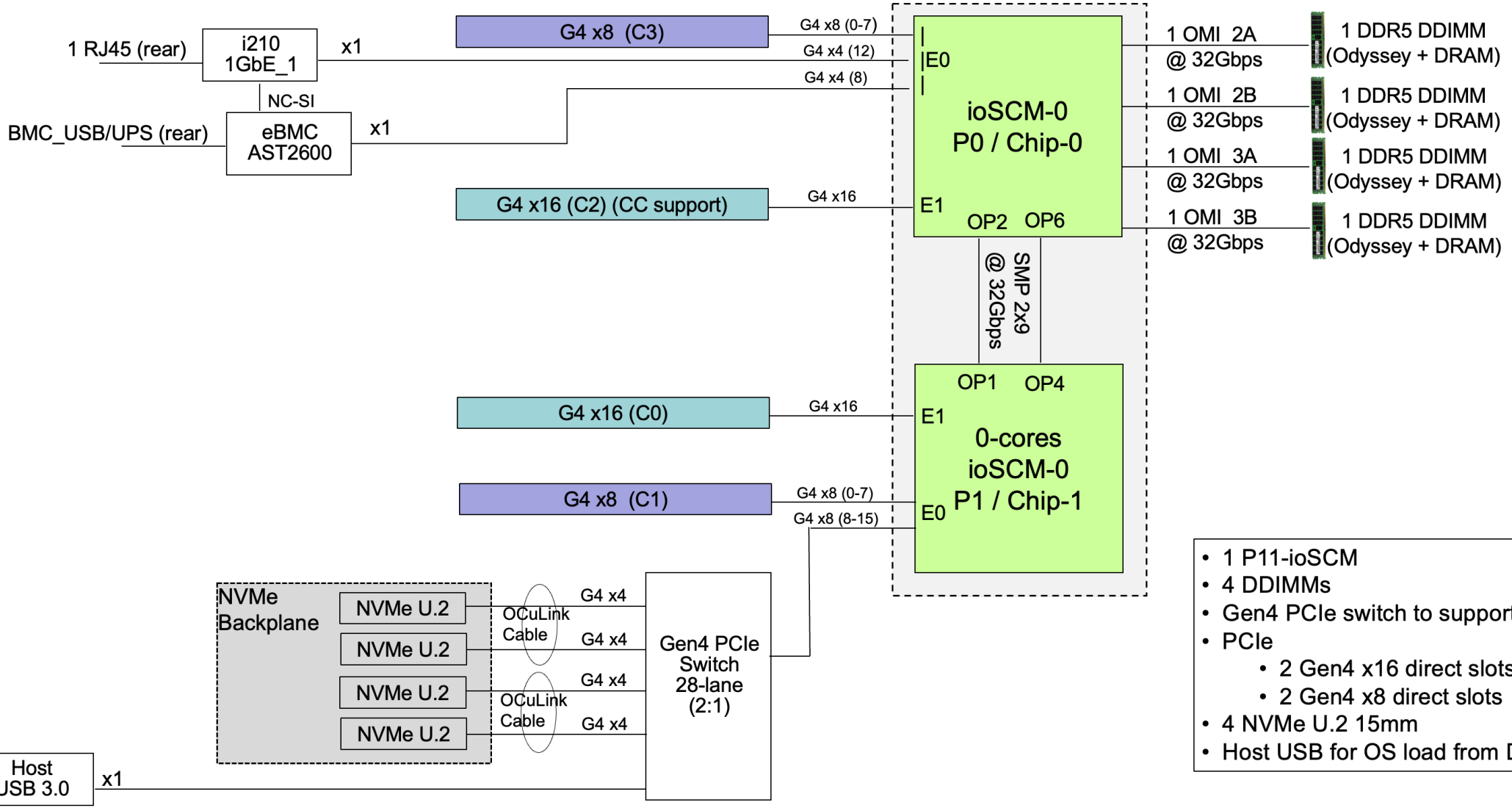
1.1X 4-core S1112 vs 4-core S1014

2.2X 4-core S1112 vs 4-core S914

3.1X 4-core S1112 vs 4-core S814

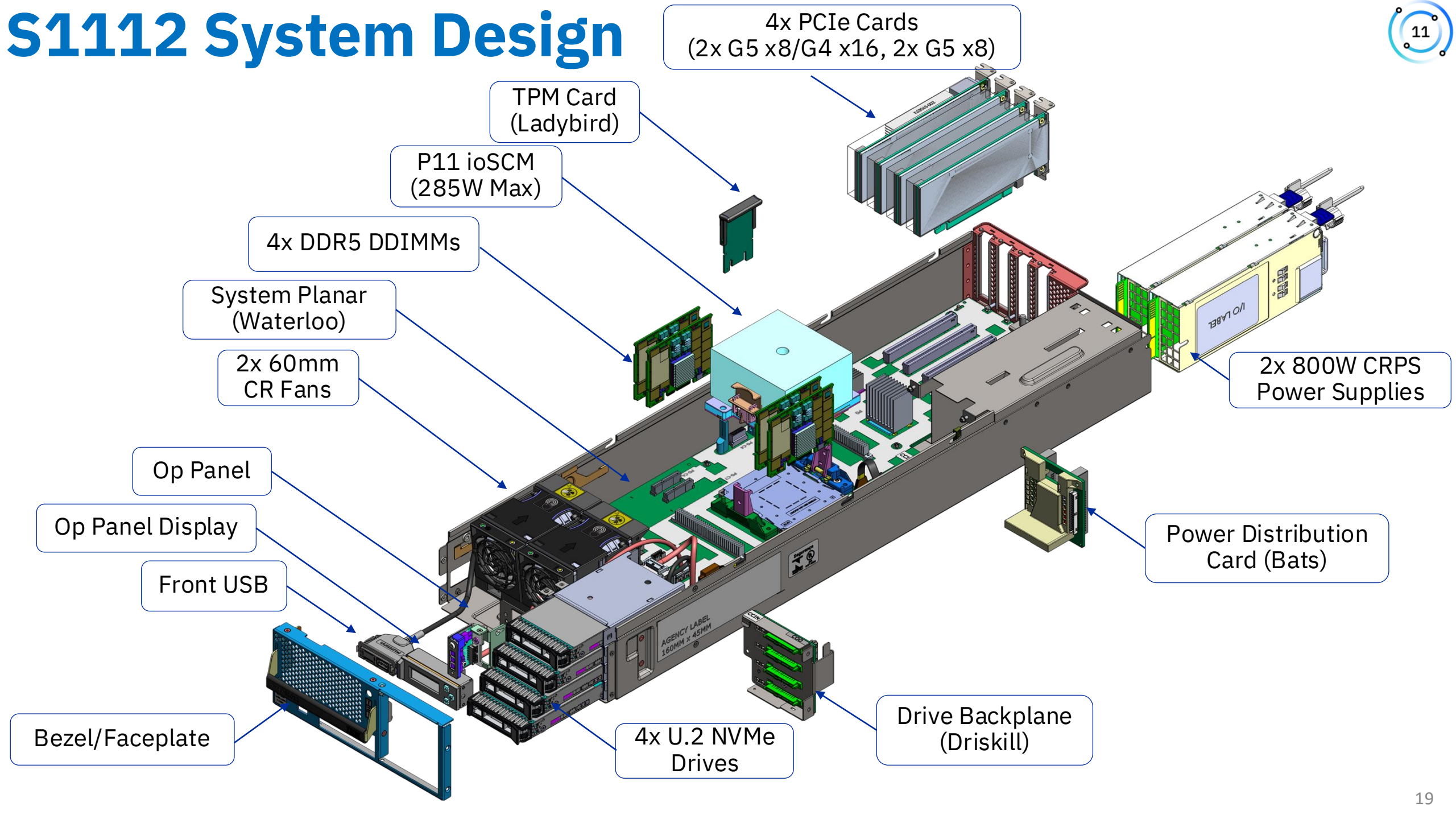
[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and should not be relied on in making a purchasing decision. 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 IBM products remain at the sole discretion of IBM.

Power11 S1112 1S2U Block Diagram



- 1 P11-ioSCM
- 4 DDIMMs
- Gen4 PCIe switch to support storage
- PCIe
  - 2 Gen4 x16 direct slots
  - 2 Gen4 x8 direct slots
- 4 NVMe U.2 15mm
- Host USB for OS load from DVD

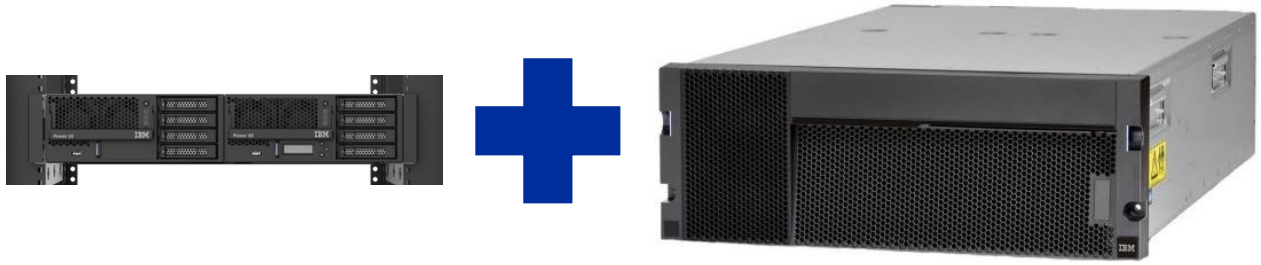
S1112 System Design



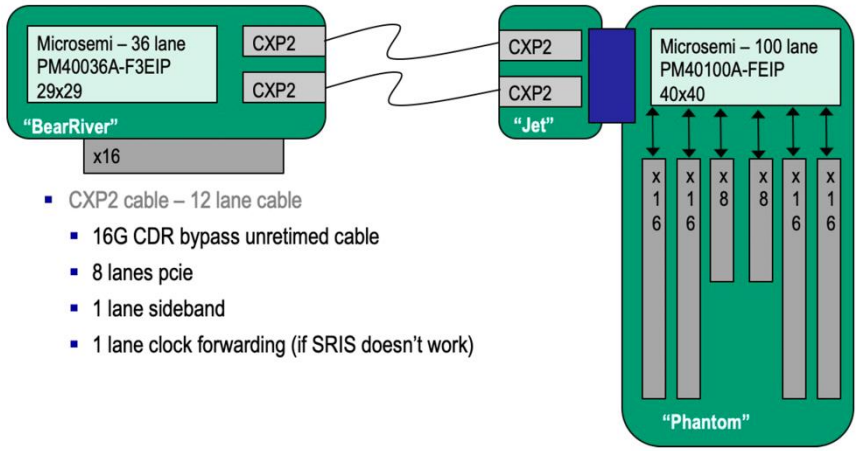
S1112 - PCIe I/O Expansion Design

There is support for one PCIe Gen4 I/O Expansion Drawer with one PCIe4 6-Slot Fanout Module providing **six PCIe GEN4 slots**.

The PCIe Gen4 I/O Expansion Drawer **is not supported on the Tower** (9242-21T) model.



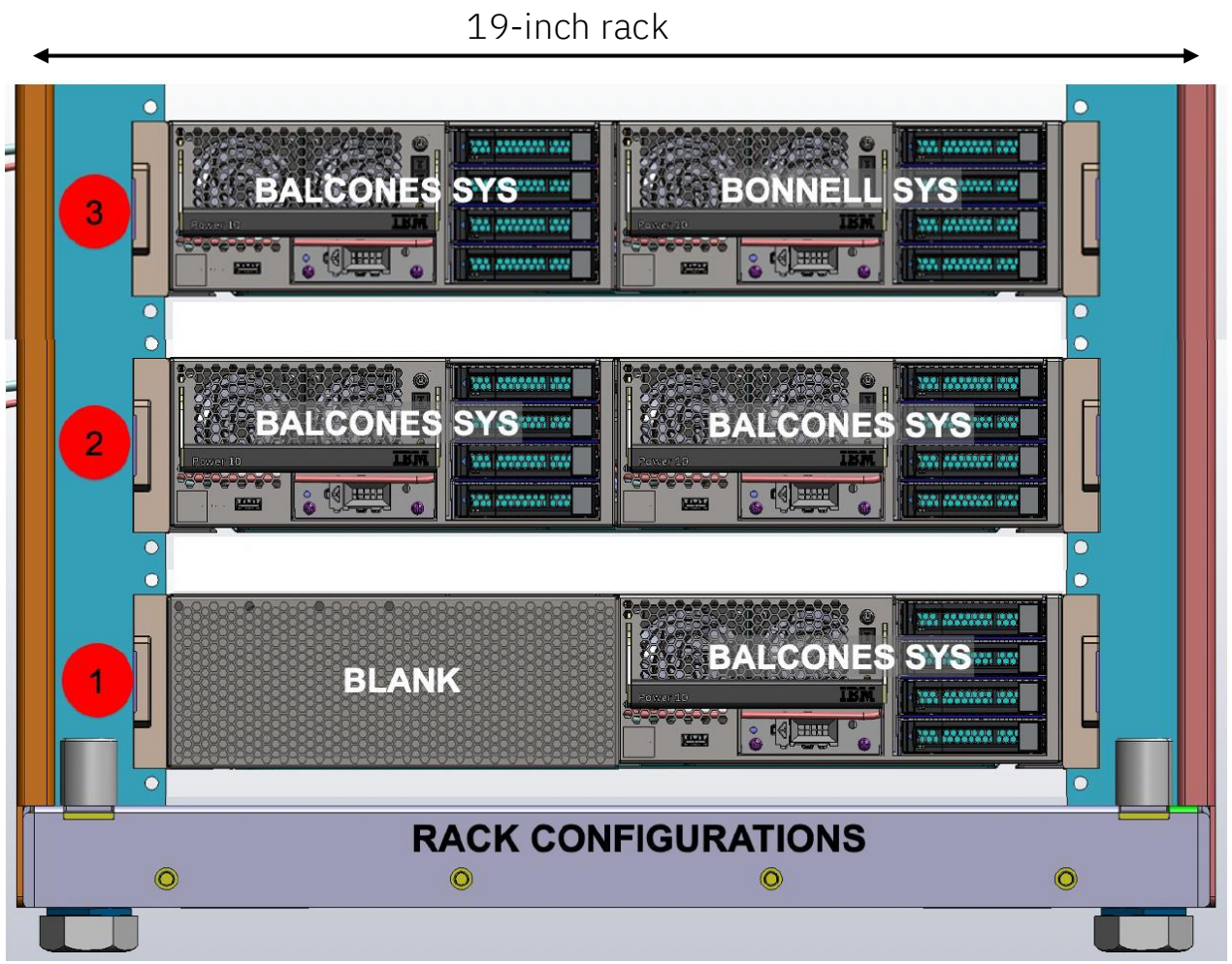
- Connectivity (Gen4)
  - Bear River = x16 to the host – x16 to the Fan out module



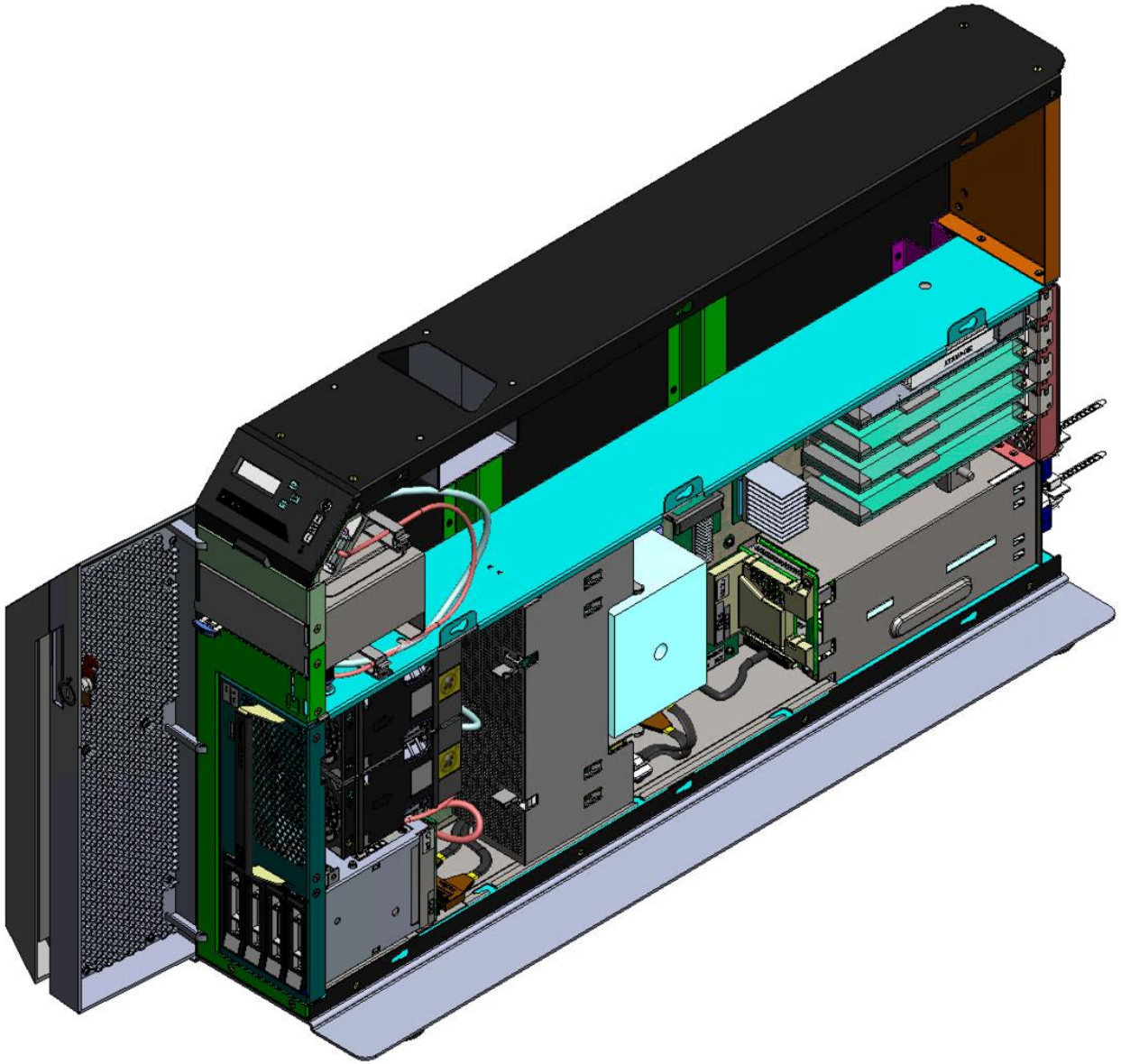
- CXP2 cable – 12 lane cable
  - 16G CDR bypass unretimed cable
  - 8 lanes pcie
  - 1 lane sideband
  - 1 lane clock forwarding (if SRIS doesn't work)

Bear River attach to Gen4 IO fanout module in Nimitz I/O drawer will use 1 internal PCIe in the S1112.

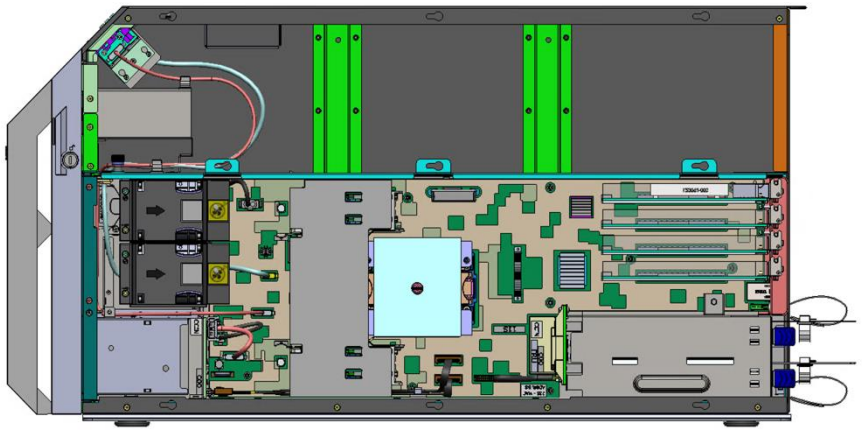
S1112 Rack configuration



S1112 Deskside configuration



Width: 11.2 cm (4.4")
Height: 41.1 cm (16.2")
Overall Width (including feet): 21.1 cm (8.3")
Maximum Depth (Front to Rear): 78.5 cm (30.9")



Power S1112 DDIMM Memory Options

Feature	Category	Description	Orderable/Supported
EM54	Memory	64GB (2x32GB) DDIMM (2U), 4000MHZ or 4800MHZ 16GBIT DDR5	Orderable
EM5B	Memory	128GB (2x64GB) DDIMM (2U), 4000MHZ or 4800MHZ 16GBIT DDR5	Orderable
EM5V	Memory	256GB (2x128GB) DDIMM (2U), 4000MHZ or 4800MHZ 16/32GBIT DDR5	Orderable

- Minimum config is 2x DDIMMs per socket
- All DDIMMs must be the same type

Power S1112 NVMe Options

FC (A/L), (i)	Category	Description	AIX/Linux IBM i	Orderable / Supported
ES5A	NVMe	800GB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ES5B	NVMe	800GB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O
ES5C	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ES5D	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O
ES4B	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ES4C	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O
ES4D	NVMe	3.2TB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ES4E	NVMe	3.2TB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O
EC7W	NVMe	800GB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
EC7Y	NVMe	800GB 4K NVME U.2 15MM SSD PCIe4 (IBMi)	IBMi	O
ECT0	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ECT1	NVMe	3.2TB 4K NVMe U.2 15mm SSD PCIe4 (AIX/Linux)	A/L	O
ECT4	NVMe	1.6TB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O
ECT5	NVMe	3.2TB 4K NVMe U.2 15mm SSD PCIe4 (IBMi)	IBMi	O

Power S1112 I/O Options

FC	Category	Marketing Name	S1112	ENZ0	Orderable / Supported	Replacement
EC71	ROCE	2-PORT 25Gb EN CONNECTX-6 Lx SFP28 NO CRYPTO PCIe4 x8 LP ADAPTER	X		O	
EC72	ROCE	2-PORT 25Gb EN CONNECTX-6 Lx SFP28 NO CRYPTO PCIe4 x8 LP CAPABLE ADAPTER		X	O	
EAPD	ROCE	PCIe4 4-Port 25Gb/10Gb RoCE SFP28 LP Capable Adapter		X	O	
EC73	ROCE	2-PORT 25Gb EN CONNECTX-6 Lx SFP28 CRYPTO PCIe4 x8 LP ADAPTER	X		O	
EC74	ROCE	2-PORT 25Gb EN CONNECTX-6 Lx SFP28 CRYPTO PCIe4 x8 LP CAPABLE ADAPTER		X	O	
EN2Y	LAN	PCIe LP 4-port 1GbE Adapter	X		O	
EN2Z	LAN	PCIe 4-port 1GbE Adapter		X	O	
EN2W	LAN	PCIe2 4-port 10GbE Adapter		X	O	
EN2X	LAN	PCIe2 LP 4-port 10GbE Adapter	X		O	
EN2J	Stg_ctrl	PCIe4 32Gb 2-port Fibre Channel Adapter		X	O	
EN2K	Stg_ctrl	PCIe4 LP 32Gb 2-port Fibre Channel Adapter	X		O	
EN2A	Stg_ctrl	2-PORT FIBER CHANNEL(16Gb/s), PCIe3 x8/SHORT/LP CAPABLE		X	S	EN1A
EN2B	Stg_ctrl	2-PORT FIBER CHANNEL(16Gb/s), PCIe3 x8/SHORT/LP	X		S	EN1B
EN1A	Stg_ctrl	PCIe3 32Gb 2-port Fibre Channel Adapter		X	O	
EN1B	Stg_ctrl	PCIe3 32Gb 2-port Fibre Channel Adapter	X		O	
EN1L	Stg_ctrl	PCIe4 32Gb 4-port Fibre Channel Adapter		X	O	
EN1N	Stg_ctrl	PCIe4 64Gb 2-port Fibre Channel Adapter		X	O	
EN1P	Stg_ctrl	PCIe4 64Gb 2-port Fibre Channel Adapter	X		O	
EJ2B	Stg_ctrl	SAS TAPE HBA W/4X HD MINISAS, LOW PROFILE CAPABLE PCIe3 12Gb x8		X	O	
EJ2C	Stg_ctrl	SAS TAPE HBA W/4X HD MINISAS, LOW PROFILE PCIe3 12Gb x8	X		O	
EC90	Graphics	PCIe4 x4 Graphics Adapter	X		O	
EPG4	Encryption	4770 CRYPTO COPROC PCI3 x4 LP	X		O	
EPG6	Encryption	4770 CRYPTO COPROC PCI3 x4 LP CAPABLE IN GEN3' CASSETTE FOR I/O DRAWER		X	O	
EJ24	Bus Expansion	PCIE X16 TO CXP CONVERTER CARD, SUPPORTS OPTICAL OR COPPER CABLES, SINGLE WIDE W/ LOW HS	X		O	
ENZ0	Drawer	19-INCH PCIe Gen4 4U I/O EXP DRWR	X		O	
ENZF	Drawer	PCIe Gen4 FANOUT MODULE	X		O	
	Drawer	Media Drawer 7226-1U3	X	X	O	
EUA5	DVD	STANDALONE USB DVD DRIVE WITH CABLE	X			

Energy Savings with Power S1112[1]

S1112 (10c/512GB) will deliver improved CPW per Watt

- **52% more CPW per Watt** compared to S914 (8c/256GB)
- **48% more CPW per Watt[1]** compared to S1014 (8c/256GB)

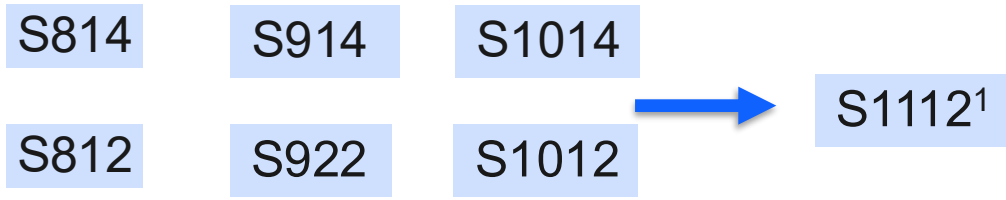
1. Power S1014 is 205,300 CPW @ 627 Watts (327 CPW/Watt), Power S1112 is 262,000E CPW @ 540E Watts (485 CPW /Watt); 485 / 327 = 1.48 more CPW/Watt
2. Power S914 is 122,500 CPW @ 383 Watts (319 CPW/Watt), Power S1112 is 262,000E CPW @ 540E Watts (485 CPW /Watt); 485 / 327 = 1.52 more CPW/Watt

[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

Offering Simplification for P05 & P10 Software Tiers

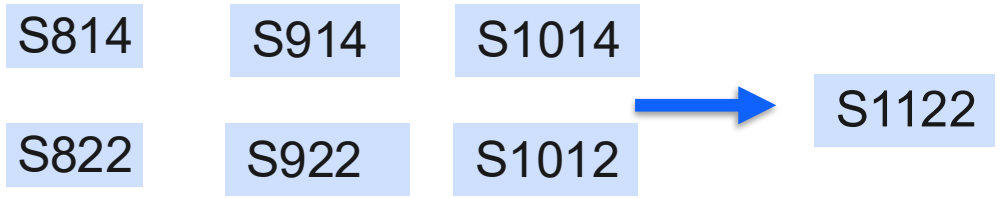
P05 Software Tier:

- S1112 is the only P05 tier system for Power11
- **Standardize on 1S2U to reduce cost, rack space and boost adoption**



P10 Software Tier:

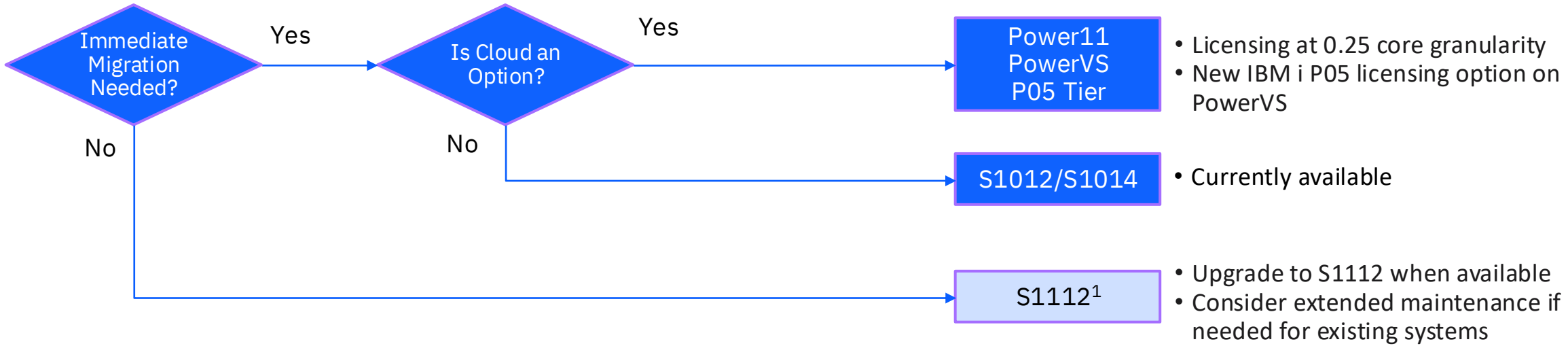
- S1122 is a better P10 option than previous generation one-socket offerings



[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

Upgrade options for IBM i P05 Power8/Power9

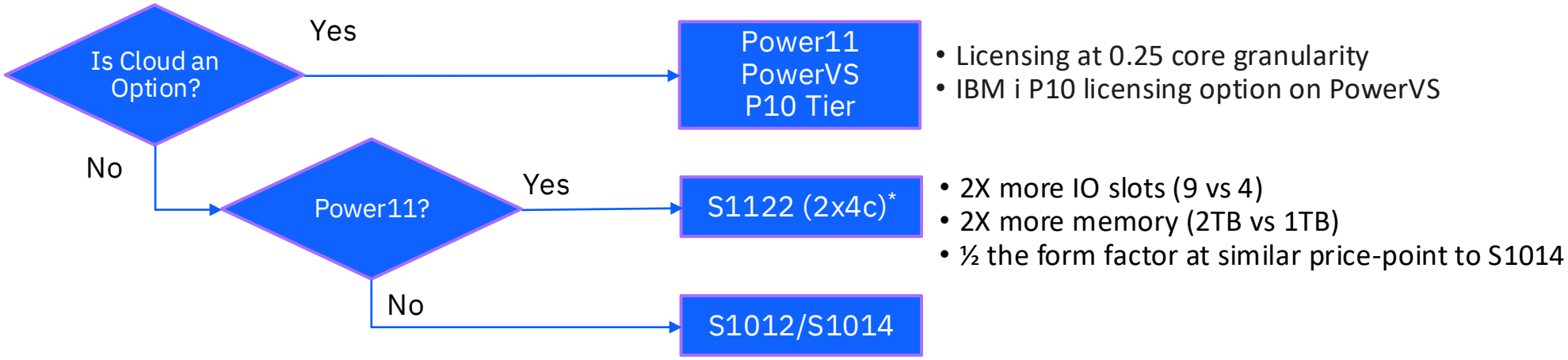
- IBM introduced the S1012 in 2Q2024 as a more space-efficient, cost-effective alternative to S1014



[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

Upgrade options for IBM i P10 Power8/Power9

- For IBM i P10 clients, **S1122** offers better capabilities as the Power11 server option, compared to **previous generation 1S4U and 1S2U systems**



* A very small number of clients upgrading to S1122 could be impacted as:
  - S1122 has less storage space but we offer an external drawer or larger NVMe drives
  - S1122 does not support high profile adapters – in most cases low profile options are available or an external IO drawer

The Power10 and Power11 range

S1012
9028-21B

- 1-socket, 2U rack and tower
- 1.4, 8 cores / SMT8 socket
- Max 8 cores
- 4 DDIMM slots
- 256 GB of memory (64 GB 1 and 4 Core)
- 4 PCIe FHHL slots (4 Gen5 capacity)
- 4 NVMe U.2 bays
- Optional RDX Media Bay
- ½ width

1C P05 | 4C P05 | 8C P10

S1014
9105-41B

- 1-socket, 4U rack and tower
- 4, 8 cores / SMT8 socket
- Max 8 cores
- 8 DDIMM slots
- 1 TB of memory (64 GB 4 Core)
- 5 PCIe FHHL slots (4 Gen5 capacity)
- 16 NVMe U.2 bays
- Optional RDX Media Bay

4C P05 | 8C P10

S1022S
9105-22B

- 1,2-socket, 2U rack
- Up to 16 cores per system
- 4, 8 cores / SMT8 socket
- Max 16 cores
- 16 DDIMM slots
- 2 TB of memory
- 10 PCIe HHHL slots (8 Gen5 capacity)
- 8 NVMe U.2 bays
- Max 51.2TB internal storage

P10

S1022 & L1022
9105-22A / 9786-22H

- 1,2-socket, 2U rack
- 12, 16, 20 cores SMT8 / socket
- Max 40 cores
- 32 DDIMM slots
- 4 TB of memory
- 10 PCIe HHHL slots (8 Gen5 capacity)
- 8 NVMe U.2 bays
- L1022: max 25% of cores with other operating systems

P10

S1024 & L1024
9105-42A / 9786-42H

- 1,2-socket, 4U rack
- 12, 16, 24 cores SMT8 / socket
- Max 48 cores
- 32 DDIMM slots
- 8 TB of memory
- 10 PCIe FHHL slots (8 Gen5)
- 16 NVMe U.2 bays
- Optional RDX Media Bay
- L1024: max 25% of cores with other operating systems

12C P20 | 2*12C P20 | 2*16C P30 | 2*24C P30

E1050
9043-MRX

- 2, 3, 4-socket, 4U
- 12, 18, 24 cores SMT8 / socket
- Max 94 cores
- 64 DDIMM slots
- 16 TB of memory
- 10 PCIe FHHL slots (8 Gen5)
- 10 NVMe U.2 bays

~~IBM i~~

E1080
9080-HEX

- 1 to 4 CEC 5U + 2U SCU
- 10,12,15 cores SMT8 / socket
- Max 240 cores
- 64TB memory, 16TB /drawer
- 8 PCIe FHHL slots (8 Gen5)
- 4 NVMe U.2 bays

P30

S1112
9149-21B, 9149-21T

- 1-socket, 2U rack and tower
- 1.4, 10 cores / SMT8 socket
- Max cores 10 cores
- 4 DDR5 Dimm slots
- 512 GB of memory (64 GB 1 and 4 Core)
- 4 PCIe FHHL slots (4 Gen5 capacity)
- 4 NVMe U.2 bays
- Support for a 6-slot ½ drawer
- ½ width

1C P05 | 4C P05 | 10C P10

S1122 & L1122
9824-22A / 9856-22H

- 1,2-socket, 2U rack
- 16, 24, 30 cores SMT8 / socket
- Max 60 cores
- 32 DDIMM slots
- 4 TB of memory
- 10 PCIe HHHL slots (8 Gen5 capacity)
- 8 NVMe U.2 bays
- L1022: max 25% of cores with other operating systems

P10

S1124 & L1124
9824-42A / 9856-42H

- 1,2-socket, 4U rack
- 16, 24, 30 cores SMT8 / socket
- Max 60 cores
- 32 DDIMM slots
- 8 TB of memory
- 10 PCIe FHHL slots (8 Gen5)
- 16 NVMe U.2 bays
- Optional RDX Media Bay
- L1024: max 25% of cores with other operating systems

16C P20 | 2*16C P20 | 2*24C P30 | 2*30C P30

E1150
9043-MRU

- 2, 3, 4-socket, 4U
- 16, 24, 30 cores SMT8 / socket
- Max 120 cores
- 64 DDIMM slots
- 16 TB of memory
- 10 PCIe FHHL slots (8 Gen5)
- 10 NVMe U.2 bays

~~IBM i~~

E1180
9080-HEU

- 1 to 4 CEC 5U + 2U SCU
- 10,12,16 cores SMT8 / socket
- Max 256 cores
- 64TB memory, 16TB /drawer
- 8 PCIe FHHL slots (8 Gen5)
- 4 NVMe U.2 bays

P30

AIX IBM i Linux Red Hat OpenShift

All servers based PowerVM

IBM Power S1112[1] for Secure Web-Tier Server

Harden and isolate the most exposed layer of the IT stack

Designed to protect the enterprise "front door" (web tier)

Why Now?

- **Web servers are the most common entry point for cyberattacks**
- ~50% of breaches start at the web tier
- Average cost of a data breach: $4.4M (2025)
- **Common hypervisors are frequent attack targets**

Flexible Software Stack support on S1112

- Red Hat Enterprise Linux
- Apache
- PowerSC
- HTTP Server for i ((powered by Apache)

Why S1112 vs x86 at the Web Tier

- Strong isolation, simple design
  - **Power11 Quantum-safe Capabilities**
  - Hardware-backed LPAR isolation
  - **No shared hypervisor exposure across workloads**
- Built-in secure virtualization
  - **PowerVM has 98% fewer vulnerabilities (CVEs) than Vmware (NIST)[2]**
  - Fewer patches → lower outages → reduced breach risk
- Predictable, right-sized economics
  - Ideal for always-on web workloads
  - **No recurring hypervisor licensing costs**

[1] Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and 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 IBM products remain at the sole discretion of IBM.

[2] https://nvd.nist.gov/vuln/search#/nvd/home?resultType=records / **National Institute of Standard and Technology**

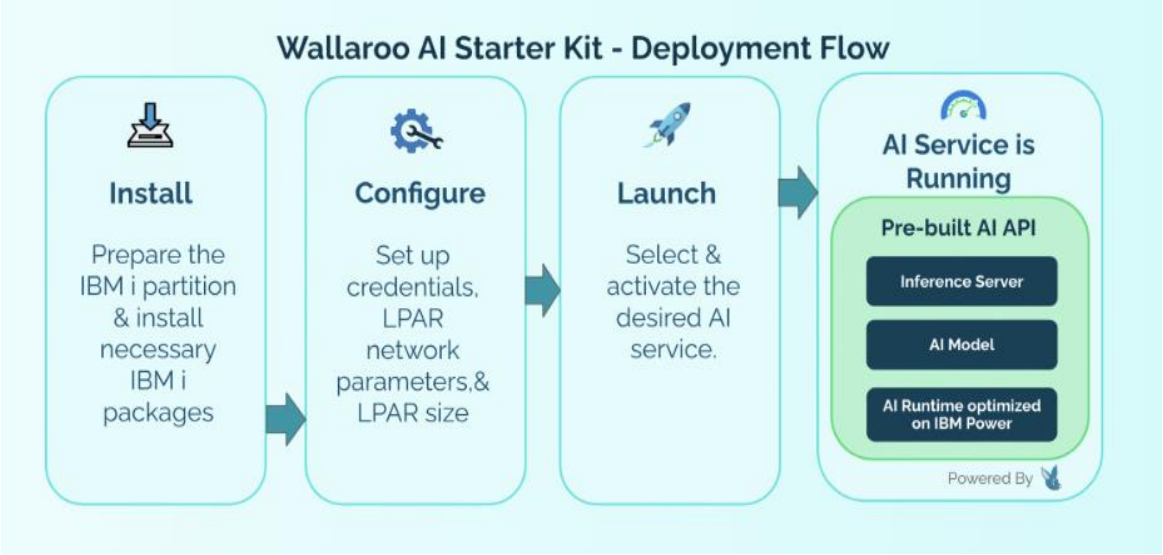
IBM Power S1112 for ai

Example : Powered by IBM Power with wallaroo AI Starter Kit(WASK)

The **Wallaroo AI Starter Kit (WASK)** is a platform that makes it easy to deploy and use AI models on **IBM Power**.
It helps **IBM i, AIX, and Linux** users to quickly run **AI assistants, RAG applications, and AI inference services** with high performance.

Perfect for Entry-level AI Deployment

- Low-cost entry point
- **Easy deployment and management**
- **Minimal risk – AI next to your data**
- **No requirement to modify IBM i application**
- Can scale up cores / memory later
- First step into AI use cases:
  - ✓ Digital Assistant for HR, sales, IT, operations
  - ✓ Internal document Q&A for employees
  - ✓ Content summarization
  - ✓ Information extraction
  - ✓ Code assistant



https://wallaroo.ai/wallaroo-ai-starter-kit-for-ibm/

IBM i Partition Preserved

- 4 cores / 64 GB retained on S1112 (Per IBM i P05 tier)
- AI runs in a separate Linux LPAR for safety and isolation

IBM Power S1112[1] OS requirement

AIX

Operating system LPAR with any I/O configuration

- AIX Version 7.3 with the 7300-04 Technology Level and Service Pack 1
- AIX Version 7.3 with the 7300-03 Technology Level and Service Pack 3
- AIX Version 7.2 with the 7200-05 Technology Level and Service Pack 12

If installing the AIX operating system with Virtual I/O only LPAR (one of these)

- AIX version 7.3 with the 7300-04 Technology Level
- AIX version 7.3 with the 7300-03 Technology Level
- AIX version 7.2 with the 7200-05 Technology Level and Service Pack 8

VIOS

VIOS 4.1.2.10
VIOS 4.1.1.30

IBM i

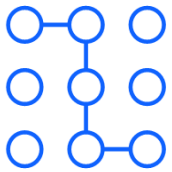
IBM i 7.6 TR2 – Both native and as a client of VIOS
IBM i 7.5 TR8 – Both native and as a client of VIOS
IBM i 7.4 TR12 – Both native and as a client of VIOS

Linux

RHEL 9.6, RHEL 9.8
RHEL 10.2
SLES15 SP7
SLES16
KVM on PowerVM supported with SLES16
OpenShift (v4.17) support will be based on RHEL 9.6

S1112

Modernize Today. Secure What Matters. Enable What's Next.



IBM i System Upgrade

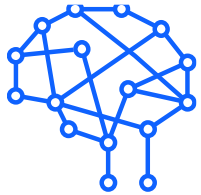
Upgrade IBM i environments to a modern Power platform that delivers higher performance, smaller footprint, and stronger resilience

- Improve application uptime and **resilience** at remote/branch locations
- **Reduce up to 75% IT footprint** vs. previous 1S4U models, less power, cooling and space
- **Enterprise-grade security** plus cost-effective HA/DR and backup options protect mission-critical workloads.

Small Apps and DBs

Enterprise-grade security, availability, and performance for smaller database workloads

- Secure Oracle, DB2, and OSDB environments with enterprise-grade protection
- Ensure consistent availability and reliability for "small but critical" workloads
- **Optimize performance per-core** while avoiding additional licensing cost
- Maintain compliance with data protection standards



Edge AI Inferencing

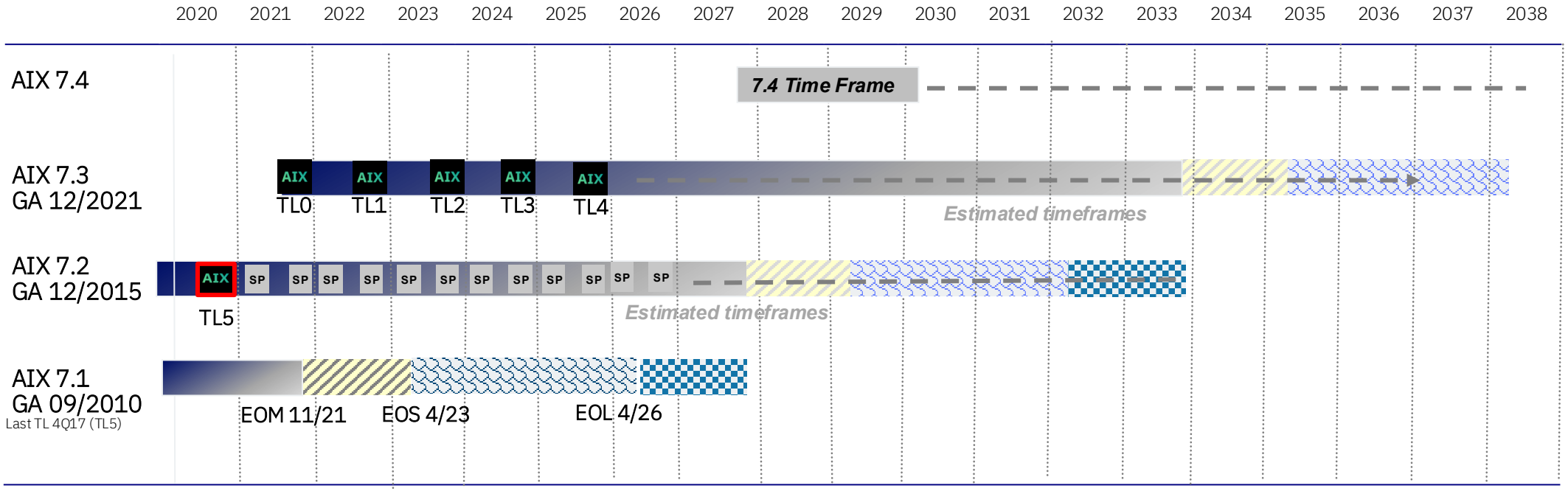
Run AI inferencing securely at the edge—closer to data, users, and operations without building new infrastructure stacks.

- **Run AI workloads locally** at ROBO sites for faster insights
- Deliver competitive performance SLAs with **reduced latency**
- **Protect sensitive data** and results flowing through AI models
- Scale inferencing capacity efficiently with minimal infrastructure

Whats new across Power11 Portfolio

AIX, IBM Concert, PEP2.0, Upgrade path..

AIX Release Streams



- All AIX releases run on Power10
- AIX 7.2 and 7.3 will support Power11
- Standard support for AIX Technology Levels is three years with the possibility of a special bid support extension
- See the IBM Support document for Best Practices with AIX Maintenance

Marketed and serviced

Serviced only

Fee-based service extension

AIX Technology Level (TL)

AIX Last TL for a release

SP Service Pack

EOM = End of Market Availability

EOS = End of Standard Service

EOL = End of Life

Strategic Drivers for AIX

Industry leading Innovations, Performance, Reliability, and Security – battle tested by world's leading enterprise class workloads and applications. We strive to achieve…

Zero Planned Downtime
(Reduced Downtime)

Maximize Business Continuity & Cost Efficiency

- Minimal/No Operational Disruption
- Reduce IT Downtime
- Lower TCO

Zero Tune Performance
(Workload Optimization)

37

Workload Optimization with Out of the Box Performance

- Effortless Optimization
- Business Value unleashed Instantly
- Efficiency from Day One

Zero Trust Security
(Optimized Security)

Guarding future of Enterprises with total protection

- Integrated & Advanced Security Capabilities
- Layered Security for Trust, Compliance, and Continuity
- Future-Ready Protection

Zero Touch Automation
(Administrative Efficiency)

Extreme automation for Administrative Efficiency

- AIX Ansible Collection

- Simplified Operations
- Lower Human Error, Smarter Resource Use
- Boost Productivity

AIX Live Update Innovation with Live Library Update (LLU)

- **Live Library Update (LLU)**

  Provides for live activation of library updates without requiring application restarts

  Controlled via the raso command or the LDR_CNTRL env var

  Initial support covers libc and libpthreads

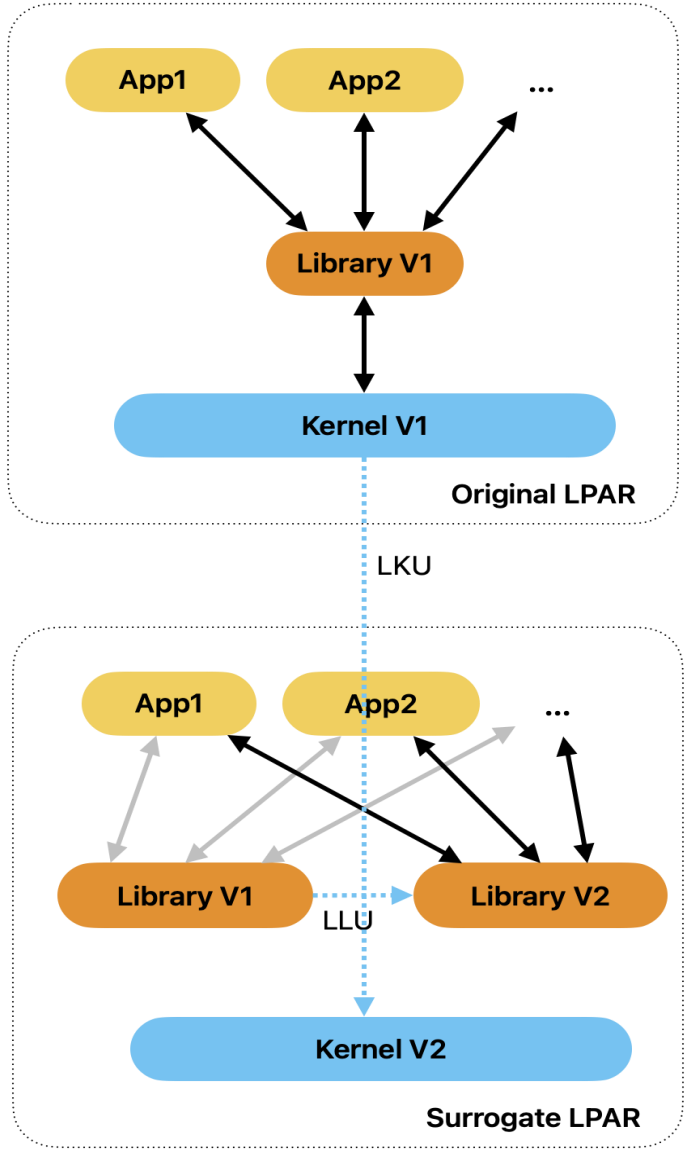
- **Live Kernel Update (LKU) and LLU**

  LKU and LLU can be performed together to live update both kernel and user space AIX components

  LLU can also be used independently

- LLU extends LKU to allow for live update of AIX without AIX reboots or application restarts

  Works with i-fixes, SPs, and new AIX TLs



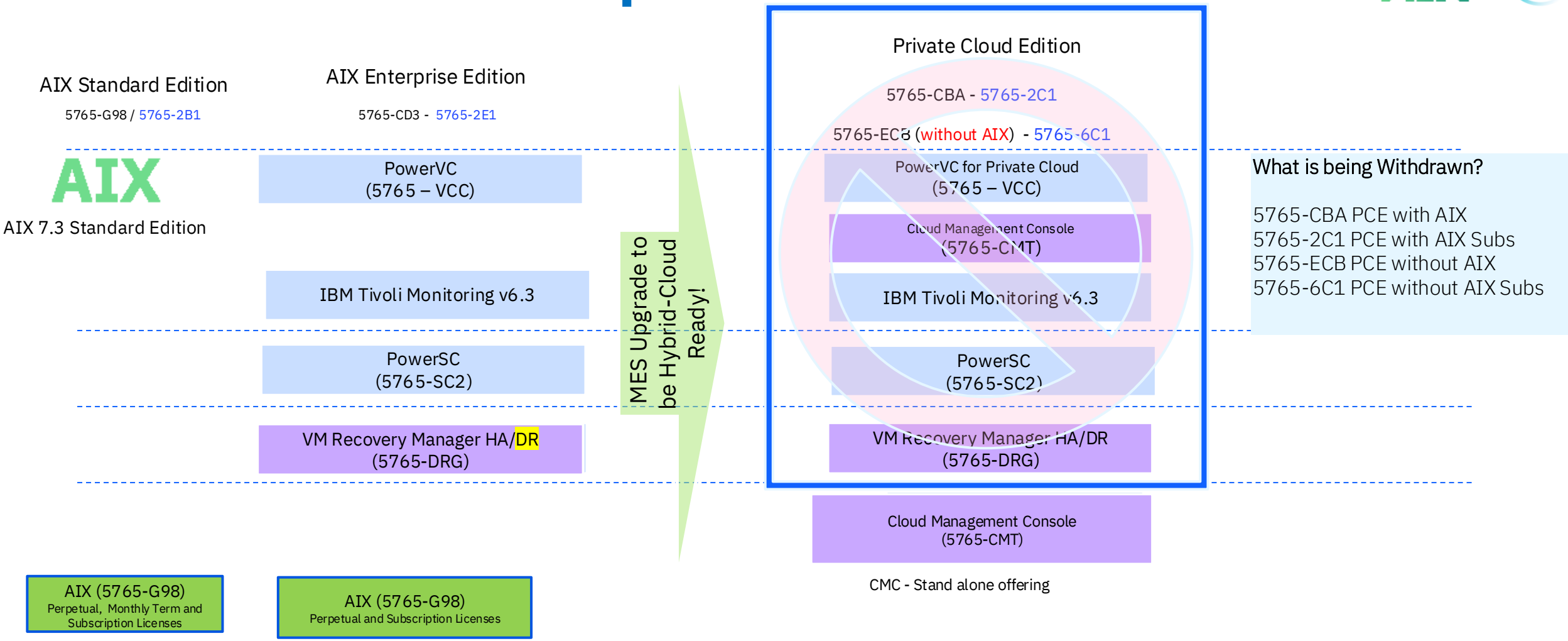
Read the LLU blog

Planned AIX Changes for Software Update Access Keys

- What is it
  - AIX on Power10 and later server includes an AIX Update Access Key
  - The AIX Update Access Key (UAK) **contains the SWMA expiration date**
  - AIX verifies if an active Software Maintenance Agreement (SWMA) is present
  - Starting with AIX 7.3 TL4, **updates and migrations will fail** if the update image is beyond the SWMA expiration

- Outcome for Users
  - Strict enforcement, updates & migrations are blocked if UAK is expired.
  - **i-fixes are not affected**
  - Action required, renew SWMA to maintain a valid UAK.

AIX PCE EOL and EE Update



IBM Concert for Power

IBM Concert for Power uses AI **to identify risks, recommend fixes, and automate remediation**, keeping Power environments secure and resilient.

Concert for Power helps …

Resilient Power Systems - Concert **keeps systems patched**, current, and secure. Closing the gap that leaves 29% of Power systems below IBM's recommended security patch level.

Identify - Concert provides **a unified view** of multiple aspects of the Power systems into a single view with one prioritized picture.

Understand and recommend – Concert understands the exposure and explains it to the user for quicker comprehension and time to remediation. **It will recommend best-in-class remediation actions for the user**.

Remediates - Concert **can automatically remediate identified exposures**, adapting to any brownfields environment and meeting the client's change management needs using their own tools.

What platforms are supported for Power?

Delivers automated vulnerability management across the entire Power stack by discovering systems, assessing risks, recommending fixes, and on Power11, automating remediation with Zero Planned Downtime.

Supported platforms

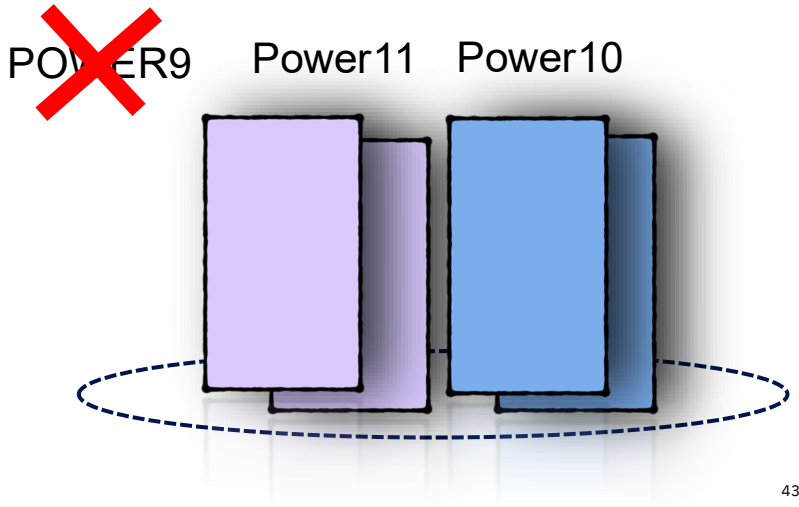
- **IBM Power 11 systems**:
  - Fully supported for discovery, advisory and assessment.
  - Automated remediation through Concert Workflows and ZPD.
- **IBM Power 10 systems**:
  - Supported for discovery, advisory and assessment.
  - Automated remediation through Workflows only.

Supported Power components

- **Systems**: Automated discovery of Power servers and their firmware levels, hardware-level CVE correlation, impact assessment, remediation recommendations, and approval-driven remediation via system firmware and IO adapter updates.
- **HMC**: Automated discovery of managed consoles, CVE and advisory correlation, impact assessment, remediation recommendations, and approval-driven remediation via HMC upgrades.
- **VIOS**: Automated discovery based on installed VIOS versions and fixes, CVE correlation, impact assessment, remediation recommendations, and approval-driven remediation via VIOS updates.
- **AIX**: Automated discovery based on installed AIX levels and fixes, CVE correlation, impact assessment, remediation recommendations, and approval-driven remediation via iFix updates.
- **Power Linux (RHEL/SLES)**: Automated discovery based on installed versions and errata, CVE correlation, impact assessment, remediation recommendations, and approval-driven patching.
- **IBM i**: Automated discovery based on installed versions and fixes, impact assessment, remediation recommendations, and remediation via PTF updates.

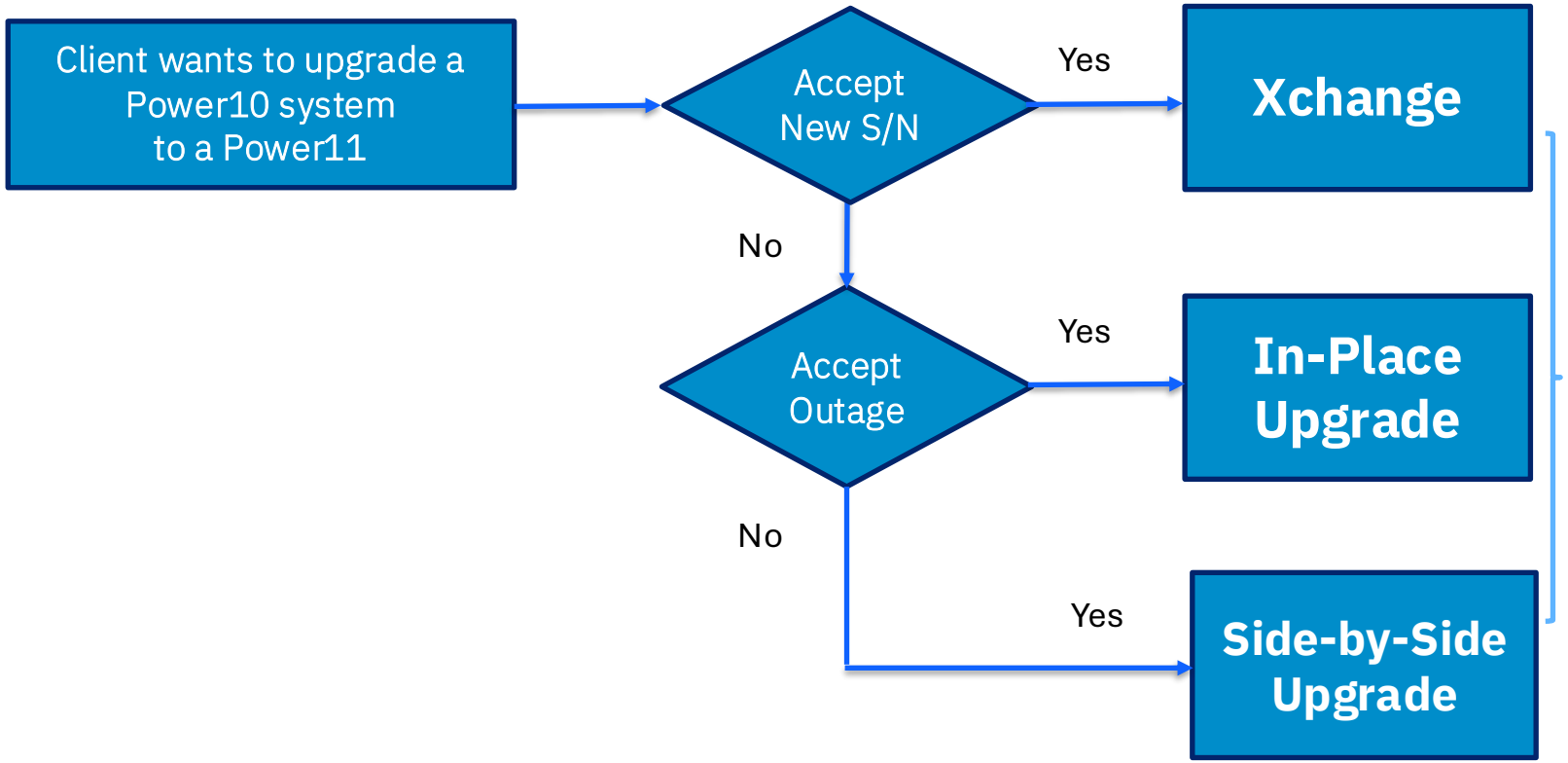
Requirements for Power Enterprise Pools 2.0

- A valid Power Enterprise Pools Subscription (5765-P2E) is required on each Power11 system being deployed in a pool.
- ONE single machine class is allowed per pool
  - For example: 5 x E1150 in one pool, 3 x E1180 in another pool.
- Power11 and Power10 systems may interoperate within a single pool.
- All machines within a pool must be in the same enterprise and geopolitical country
- A minimum of 1 Processor Activation is required on each system within a pool.
- A minimum of 256GB Memory Activation is required on each system within a pool.



43

Power 11 upgrades options*



Availability

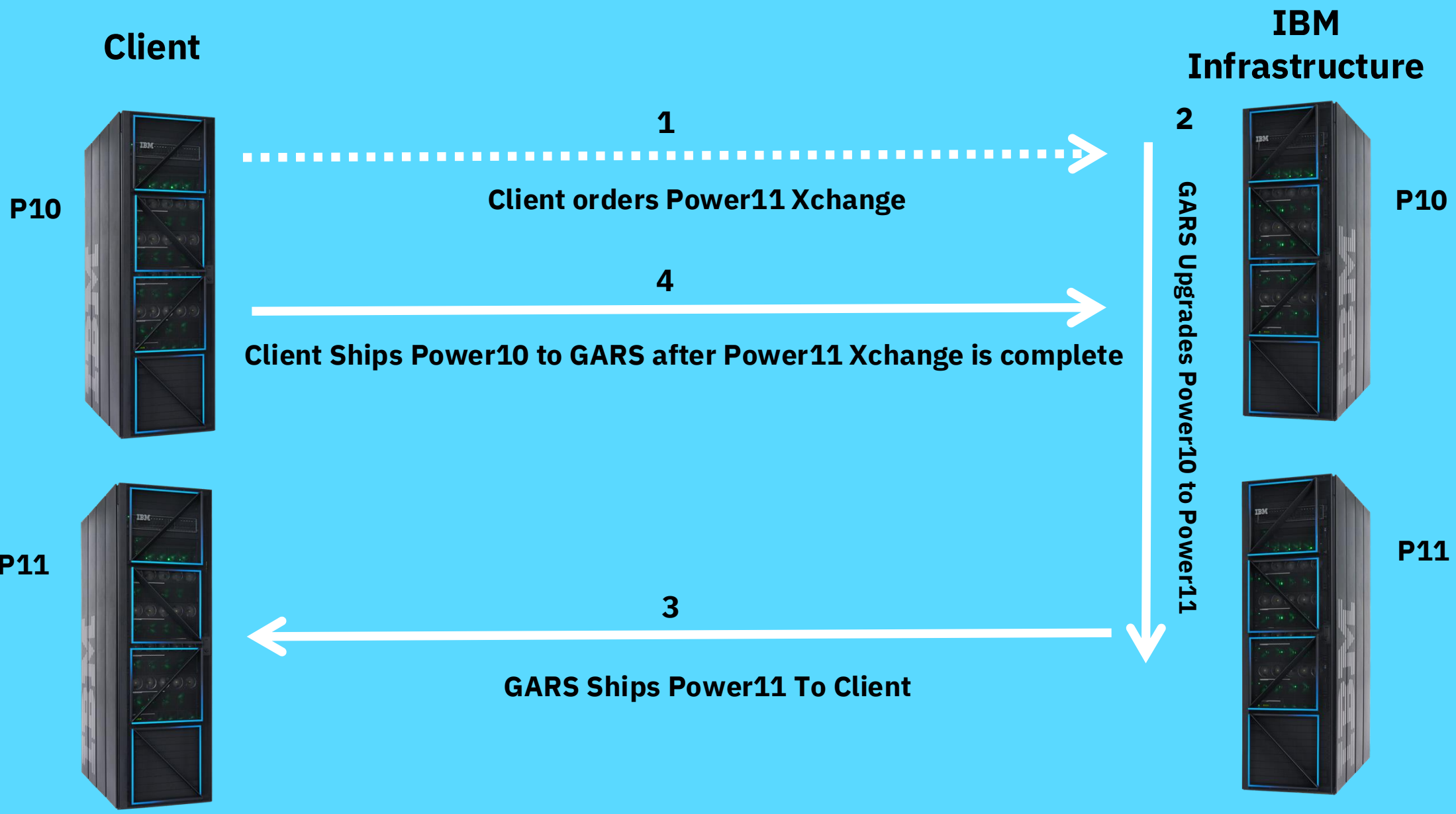
E1180
Announced: 7/8/2025

E1150
Announce: 10/28/2025

Xchange:
Announce: 11/11/2025

*Mid-Range or High-end only

IBM Power Xchange Offering



IBM Power Xchange

Unlock the advantages of a side-by-side exchange for approximately the same total cost of ownership as an MES upgrade.

Migrate to a fully tested IBM Power11 Machine
Includes a 90-day migration period for the following machines:

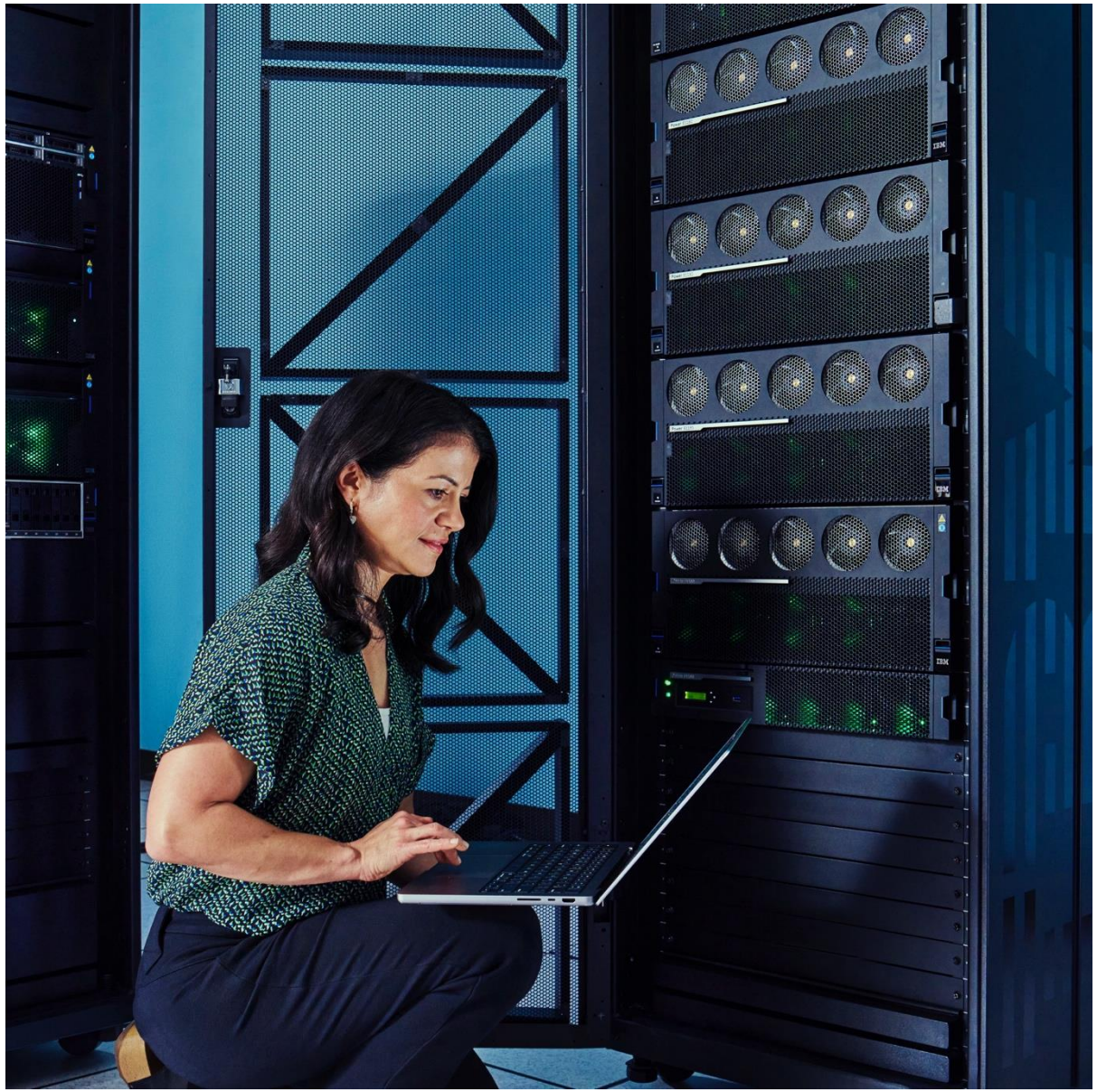
- 9080 HEX (1080) → 9080 HEU (1180) and
- 9043 MRX (1050) → 9043 MRU (1150)

Minimize Application Downtime
Zero downtime with end-to-end maintenance.
Latest technology and NEW capabilities with no disruption to operations.

Backed by IBM Technology Lifecycle Services
Includes IBM warranty and installation services

Maintenance & Financing Eligible
Qualifies for IBM maintenance and financing options, making it a flexible and secure investment.

Aligned with Sustainability Goals
Supports IBM's commitment to sustainability and the circular economy.

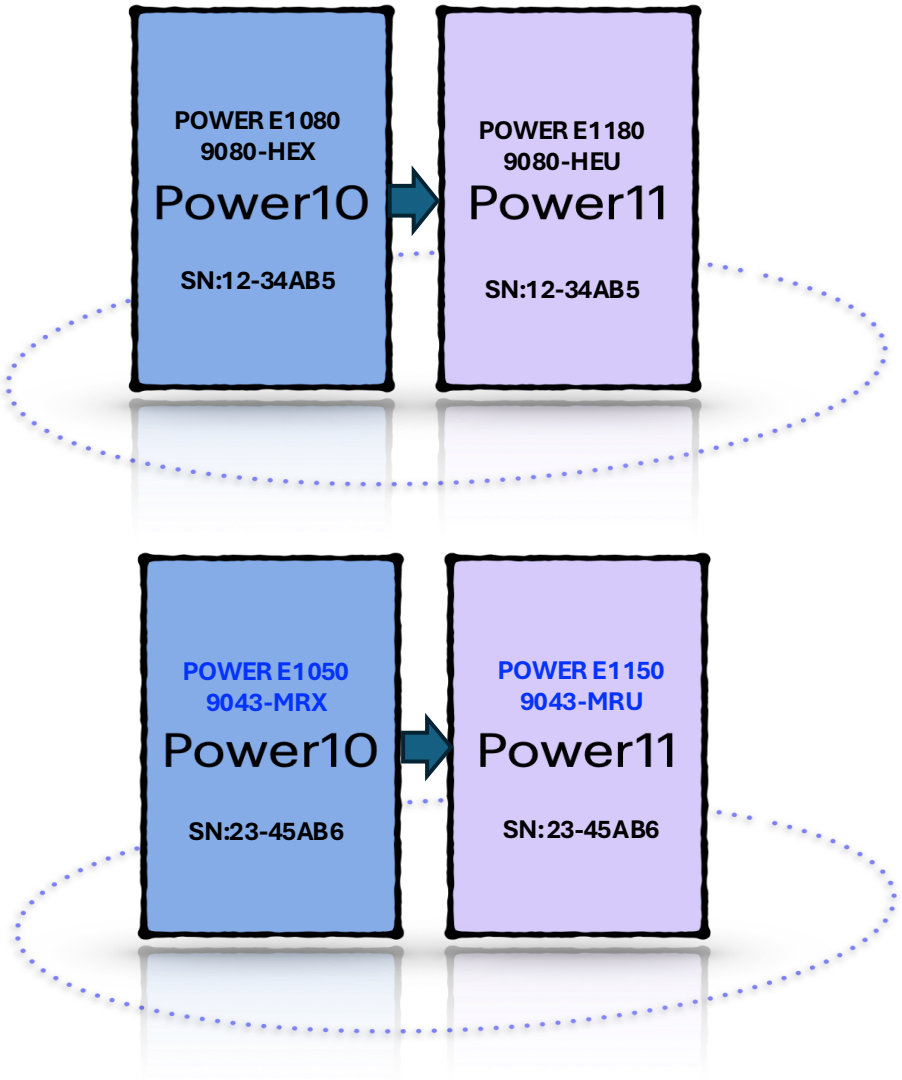


Return your Power10 for a Power11 NOW!

Serial Number Preserving model upgrades

- Serial Number Preserving model upgrades provides
  - Flexibility and cost savings by **retention of select memory and IO features**
  - Retaining some existing software and services entitlements
  - **Conversion credits for activations**
  - Enabling to preserve warranty/maintenance contract while leveraging the latest technology.

.



Serial Number Preserving comparison

Comparison

Feature	In-Place Upgrade	Side-by-Side Upgrade
Serial number preserved	✅	✅
Software activation credits preserved	✅	✅
Existing software/service entitlements preserved	✅	✅
Reuse of existing memory	✅ Yes	❌ Generally no
Reuse of existing I/O adapters	✅ Yes	❌ Generally no during migration
Additional new hardware required	Low	High
Planned downtime	Yes	Minimal or none
Cost	Lower	Higher

- **In-Place Upgrade → Best for minimizing cost** by reusing existing memory and I/O hardware, but requires planned downtime.

- **Side-by-Side Upgrade → Best for minimizing downtime**, since a new system runs alongside the existing one, but it requires significant additional hardware investment.

Announcement Summary

- IBM announced the sunset (end-of-market) for all the remaining **Power10 servers.**

- **July 31th 2026**, is the effective end-of-market date. IBM won't accept orders for new products afterward.

- **Support services will continue accordingly**. The End Of Service (EOS) dates will be defined and announced further.

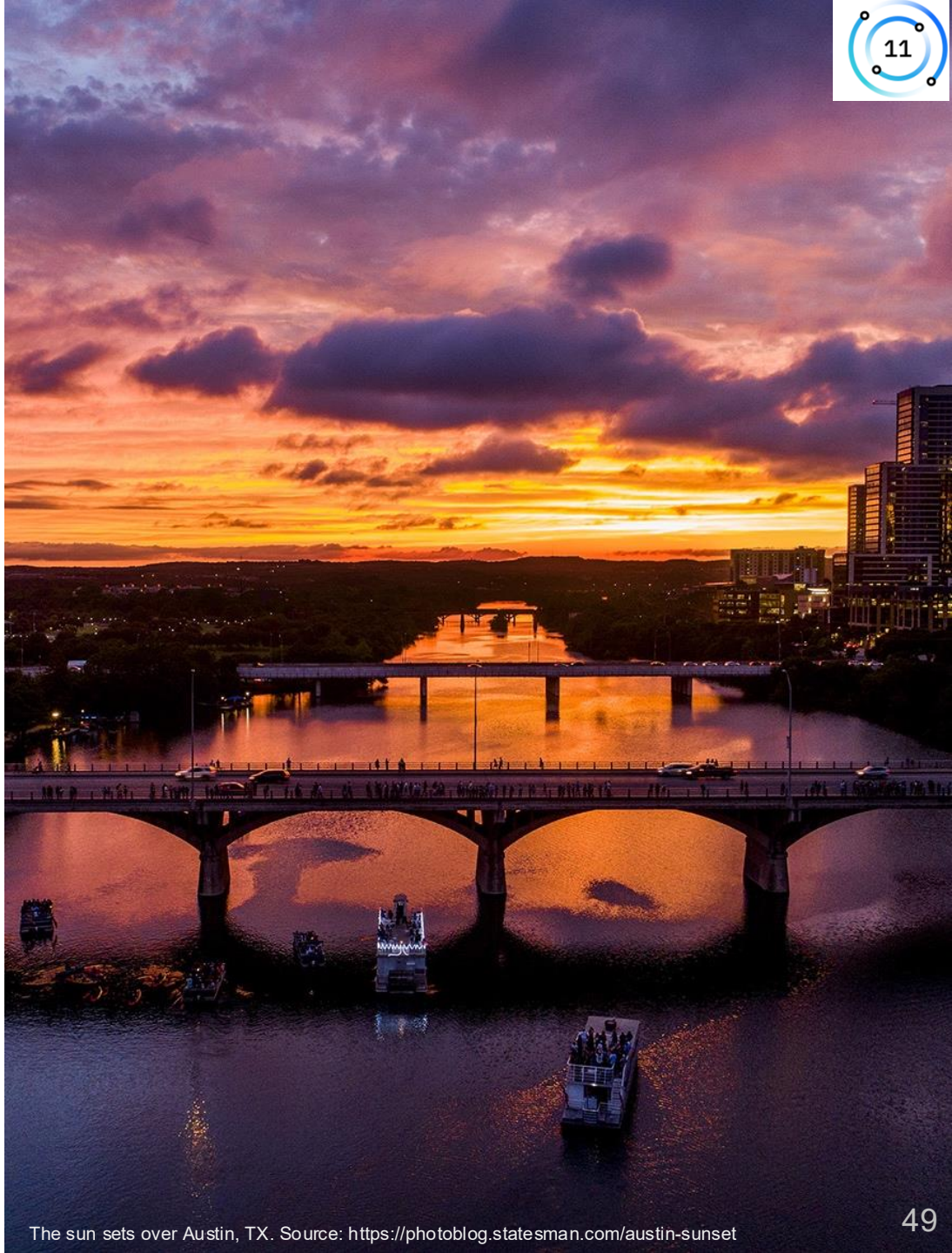
The offerings to be withdrawn from the market are:

- IBM Power System E1080 (9080-HEX)
- IBM Power System E1050 (9043-MRX)
- IBM Power System S1024 (9105-42A)
- IBM Power System S1022 (9105-22A)
- IBM Power System S1022s (9105-22B)
- IBM Power System S1014 (9105-41B)
- IBM Power System L1022 (9786-22H)
- IBM Power System L1024 (9786-42H)

Why the withdrawal of Power10 systems?

- A new set of systems featuring the Power11 technology was announced in July 2025, giving it 1y of portfolio transition from Power10.
- Power11 is a new, superior product and is fully available in the market

- **S1012 servers remains in the market**

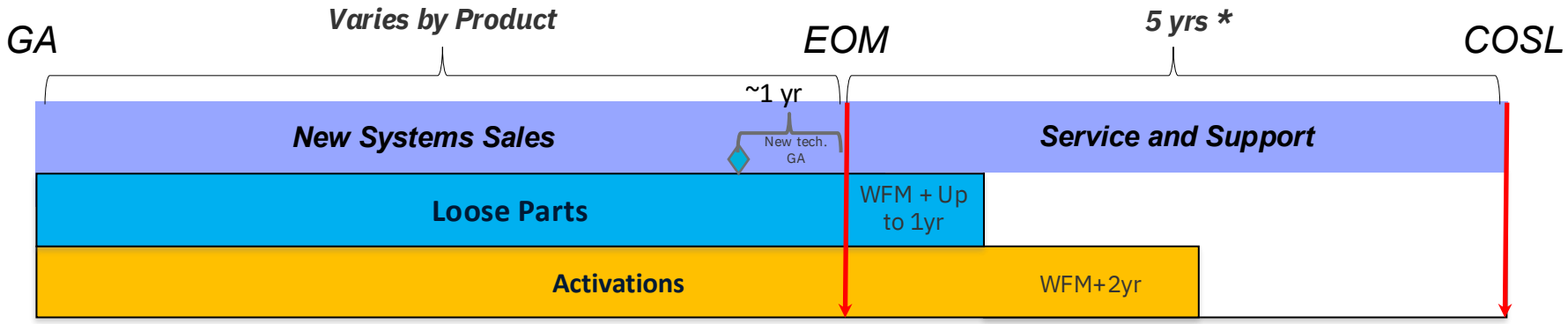


The sun sets over Austin, TX. Source: https://photoblog.statesman.com/austin-sunset

Power Systems Lifecycle

HE and MR Lifecycle

Power
E1080
E1050

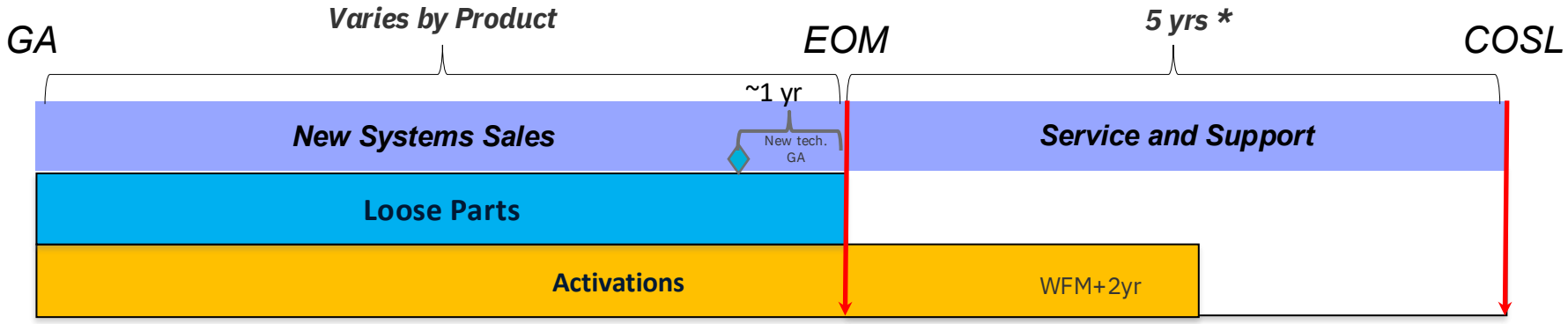


GA to WFM generally 5 years, but can vary based on supply and market factors

5 years standard from WFM to EOS

SCO Lifecycle

Power
S1014
S1022S
S/L1022
S/L1024



GA to WFM generally 5 years, but can vary based on supply and market factors

5 years standard from WFM to EOS

Post EOS Support - Service Extensions will be offered in lieu of Hardware Maintenance Agreements or Expert Care
- Separate pricing and more limited coverage
- Coverage is based on several factors, including availability of parts, skills, etc.
- Contracts available in one-year increments

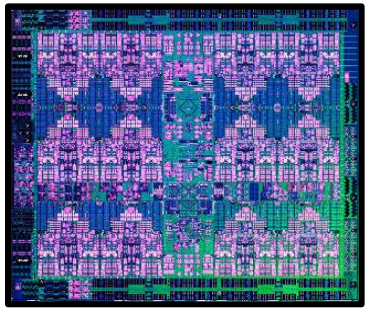
Legend:
SCO = Scale-Out
EOM = End of Market
COSL = Change of support level
Ann = Announcement
GA = General Availability
* Estimated dates based on historical data

Put AI into production with Power11

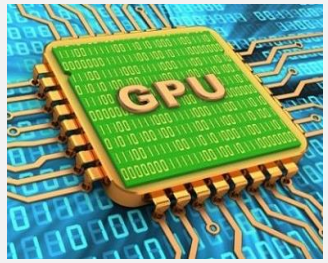
AI POWER – Power11

Inference → Large Inference → Fine Tuning → Training

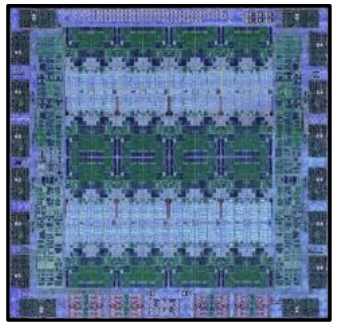
POWER9
(release: 2018)



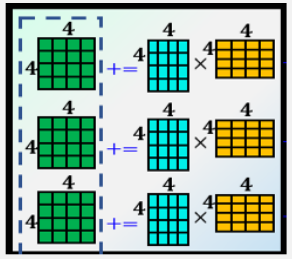
External



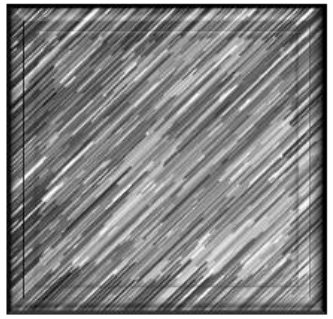
Power10
(release: 2021)



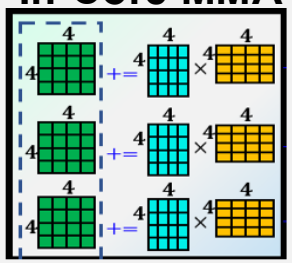
In-Core MMA



Power11
(release: 2025)



In-Core MMA



IBM Spyre

Power Full stack optimization

Users get **out-of-the-box acceleration** without code changes, easing deployments of AI models such as large language models

* Partnering with Rocket Software

ISV AI solutions

Use case & industry-specific AI solutions

Data & AI Platforms, Off-the-Shelf AI solutions

IBM stacks

Red Hat stacks

Open-source stacks

Data science frameworks & runtimes

PyTorch

Tensor-Flow

ONNX Runtime

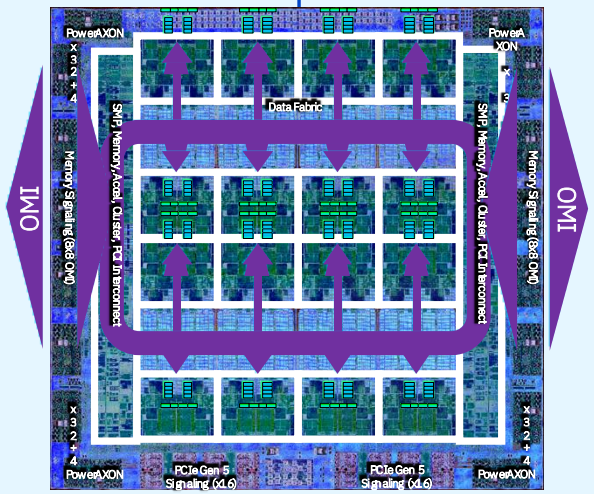
Math libraries optimized for IBM Power10 / Power11*

Open-BLAS

Eigen

MLAS

AI acceleration in every IBM Power10 / Power11 core



4 Matrix Math Accelerator (MMA) engines per core

8 SIMD engines per core

High bandwidth data-path

Commercial support
3,000+ open-source tools & packages.

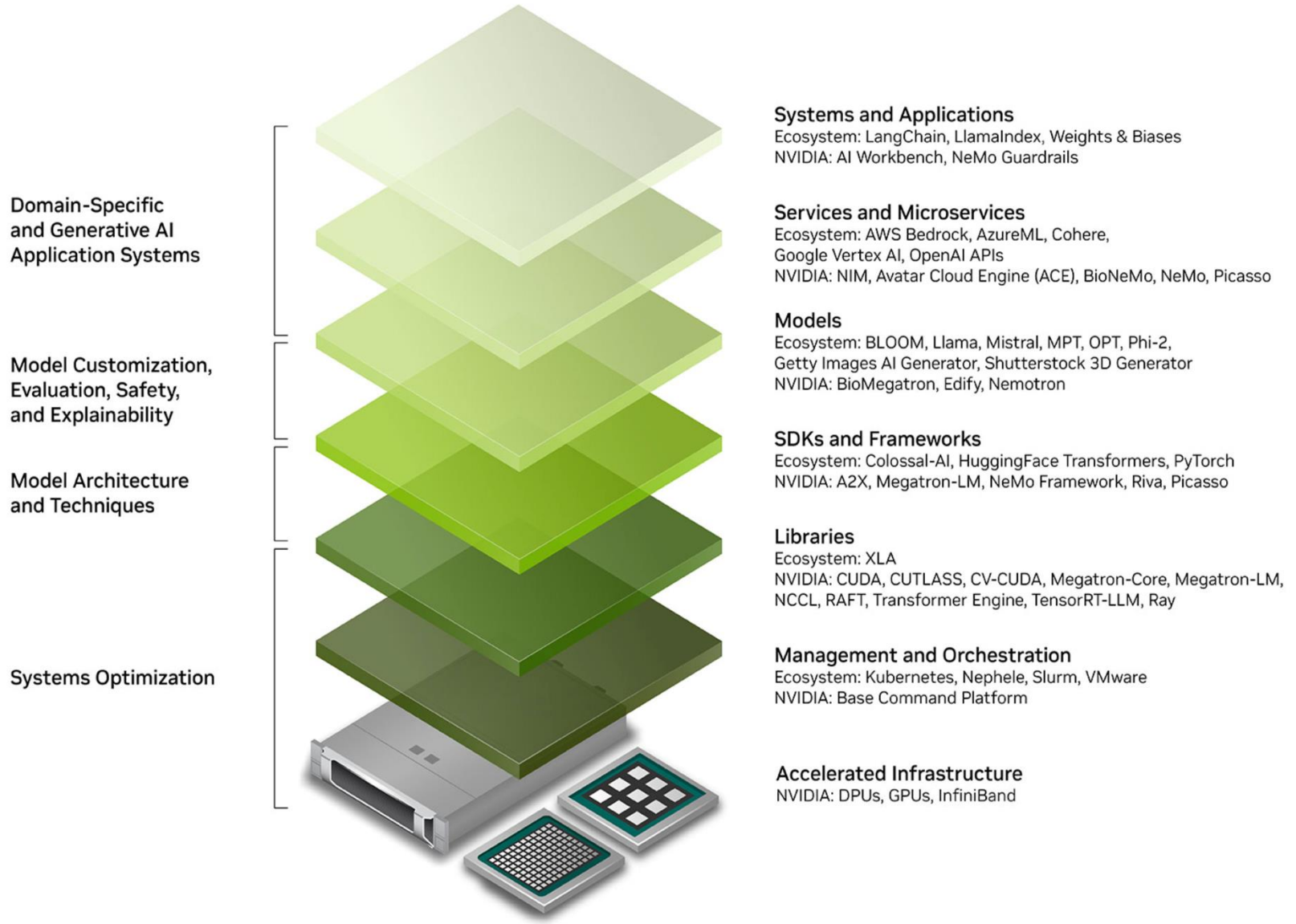
Kubeflow, KServe, NVIDIA Triton Inference Server, jupyter, kafka, Spark, SELDON, ALIBI EXPLAIN, trino, Hugging Face, EVIDENTLY AI, Horovod, Katib, Label Studio, Prometheus, Grafana, PyTorch, TensorFlow, ONNX, milvus, OpenSearch, weaviate, PostgreSQL, Chroma, LLaMA C++, vLLM

...and more.

Generative AI stack and options

This is the vision of NVIDIA

- *Other options exist, commercial or open source, cloud or on-premise*
- *On each floor of the stack, many possibilities, all constantly evolving*
- *So many choices to make!*
- *Strong skills required*

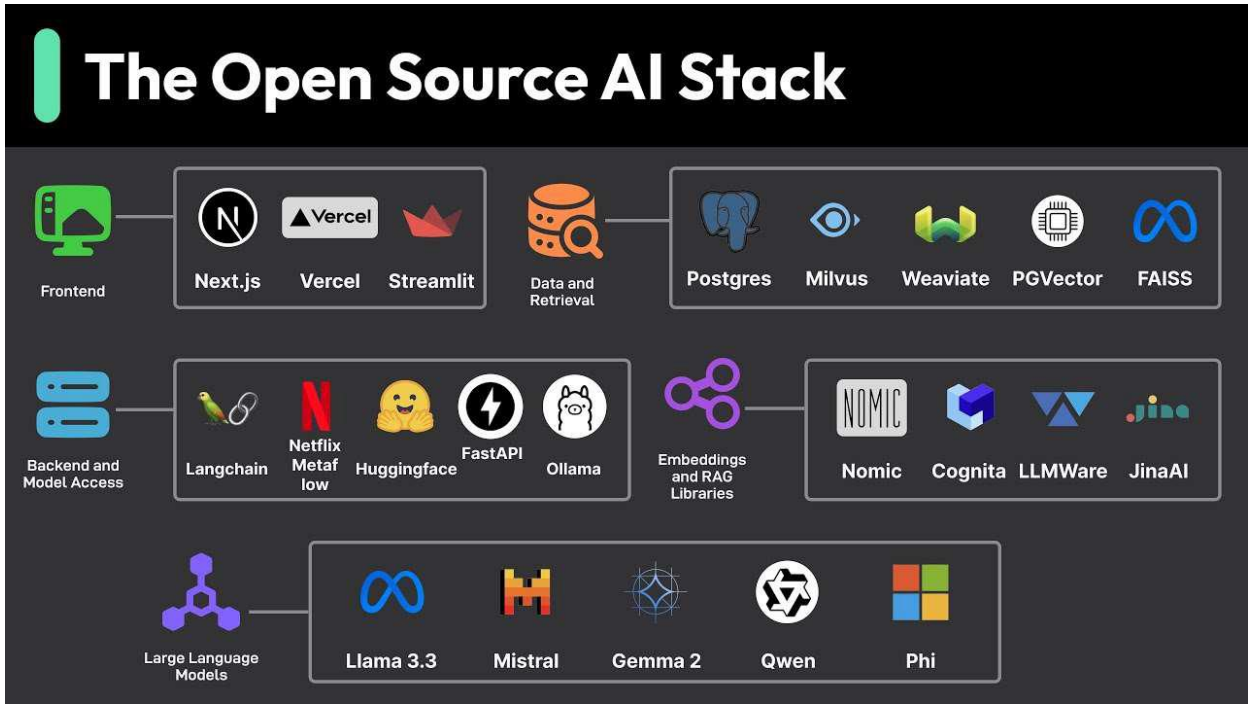


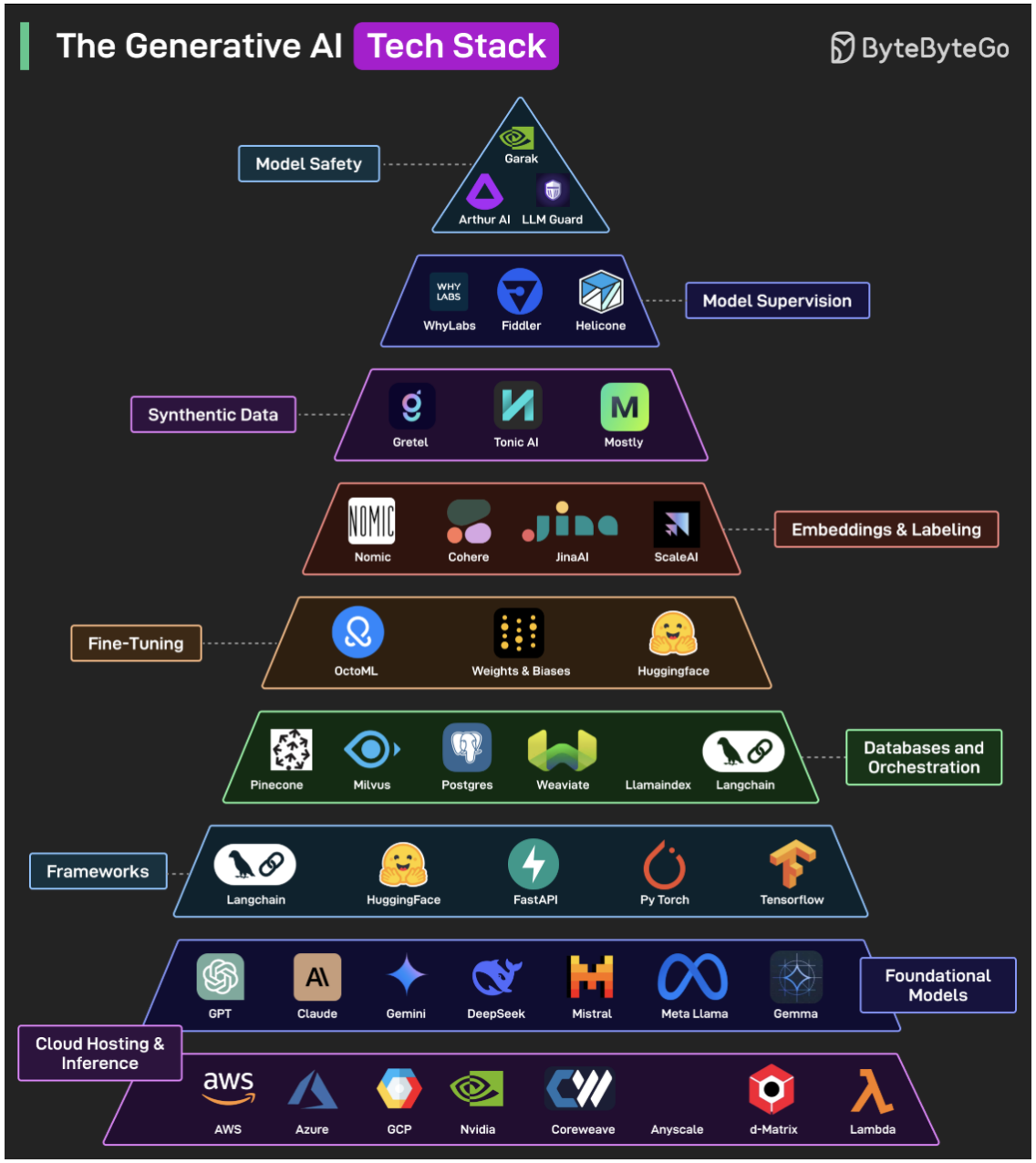
https://developer.nvidia.com/topics/ai/generative-ai?

Open source Generative AI Stack

There is a vast ecosystem of open-source products available for building a generative AI stack.

The challenge lies **in selecting the right software components**, deploying them effectively, and then managing the ongoing **complexity of updates**, maintenance, and evolution.

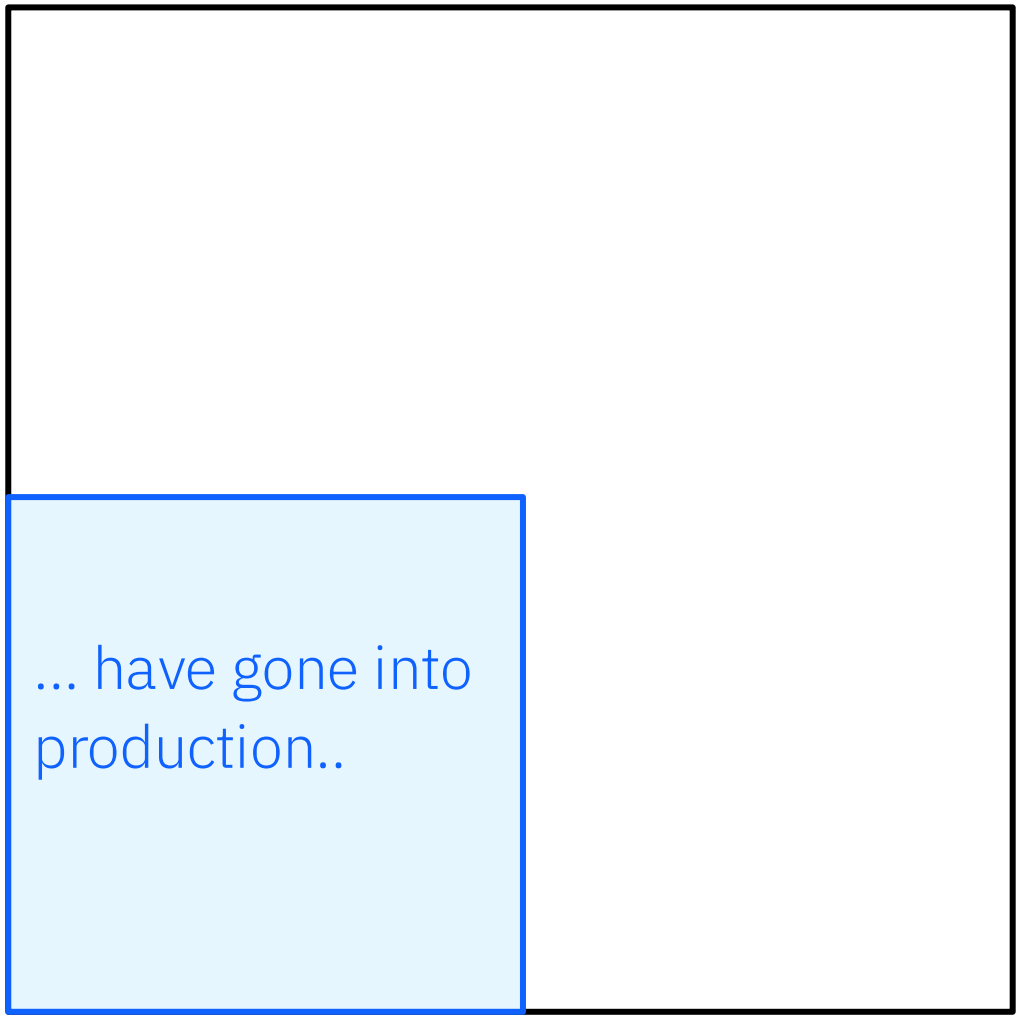




98% of enterprises have experimented with gen AI

But only ...

26%



BCG “where’s the value in AI” Oct 2024, N=1000

Boston Consulting Group (BCG) octobre 2024, basée sur un échantillon de 1000 entreprises, qui analyse la création de valeur par l’IA et le degré de maturité des projets (POC vs production).

Top Barriers to scale AI projects into production

Skills

- **Lack of data science or AI engineering skills to integrate AI into current workflows**
- Technology
- Security
- No AI-powered assistance to help transform exiting processes

Data Complexity

- Data resides **in different locations, different formats with no standardized connectors**
- Integrating and transforming enterprise data for AI use cases is complex

AI Integration Complexity

- Concerns of **re-designing existing workflows and disrupting established processes**
- Limited AI software availability

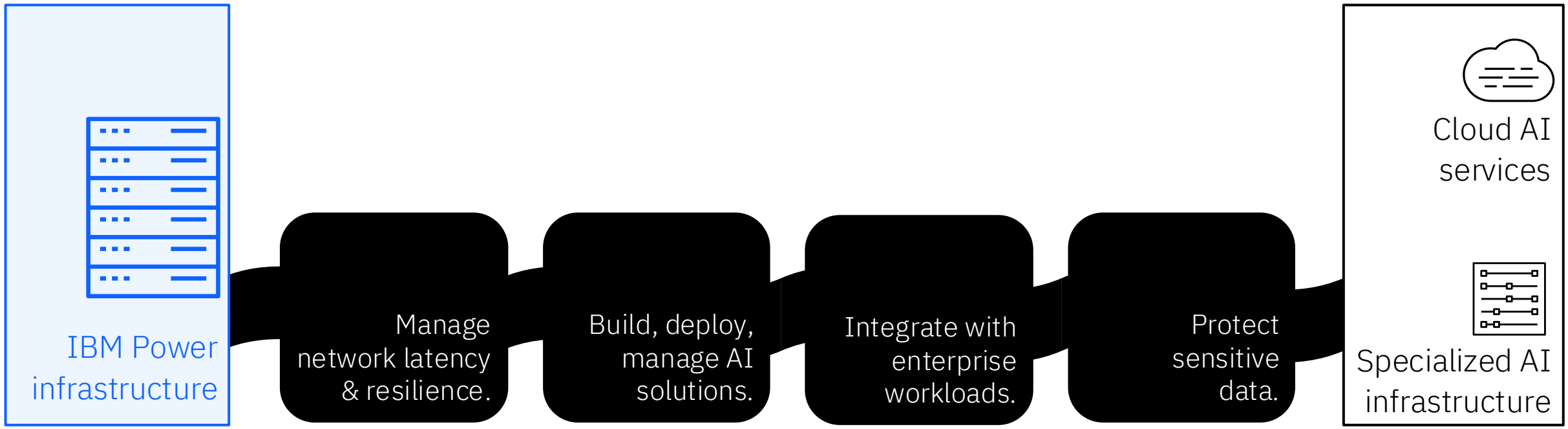
Cost

- Concerns of **expensive infrastructure needed to meet performance requirements**
- Concerns of cost associated with operationalizing and developing newer skills for AI services

Privacy & Security

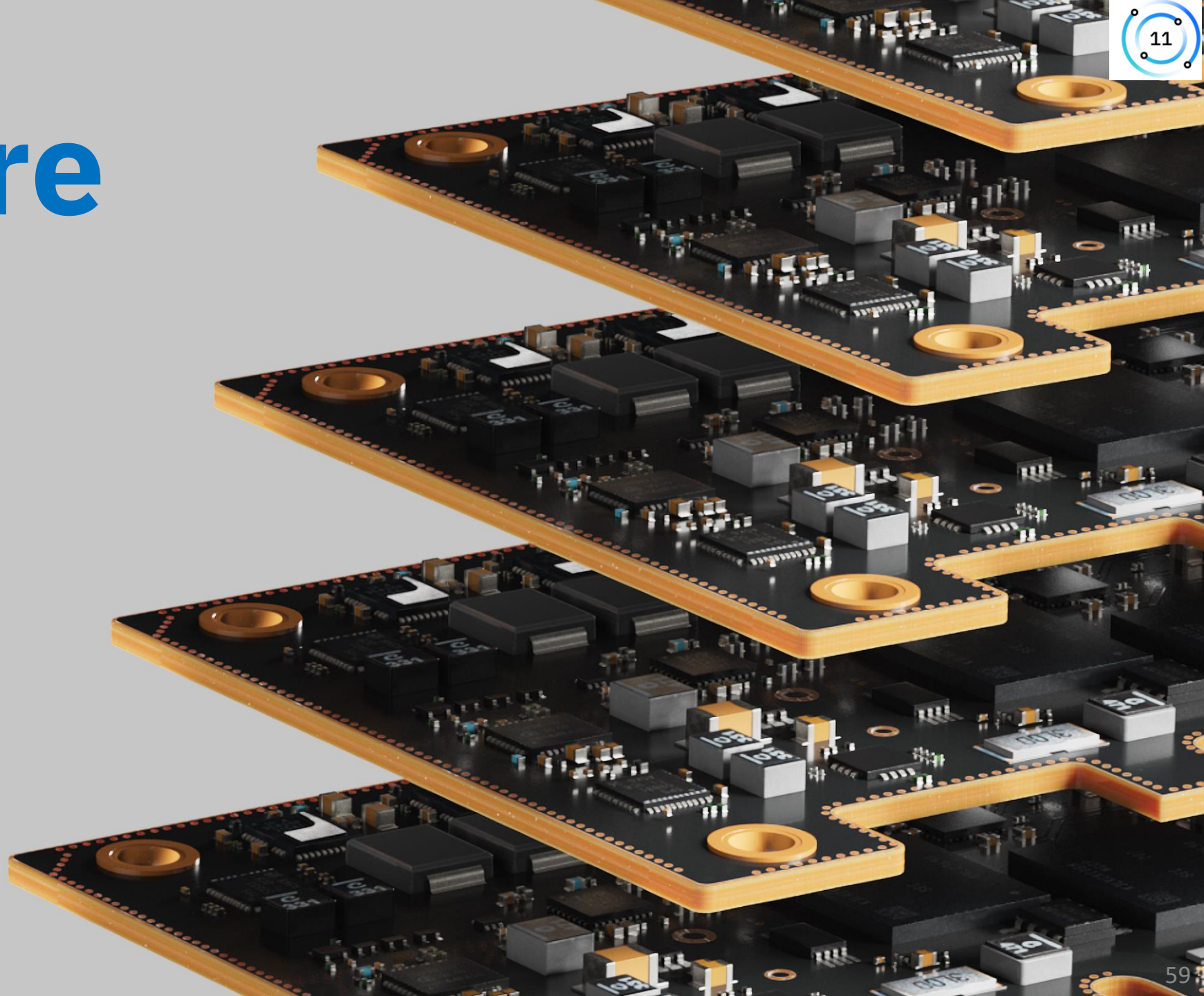
- Data used for training or fine tuning **can include sensitive information and moving data off the trusted platform can lead to vulnerabilities**
- AI models need to be protected to prevent adversarial attacks

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



IBM Spyre

AI Turnkey for companies



IBM Spyre™ Accelerator - PCIe attached card

Not a GPU !!

Dedicated to inference, it is not a GPU

32 AI Units core - 5nm

300+ TOPS FP16

128GB of LPDDR5 memory

75W PCIe gen5 x16 adapter << of major GPUs

Excellent perf/watts ratio

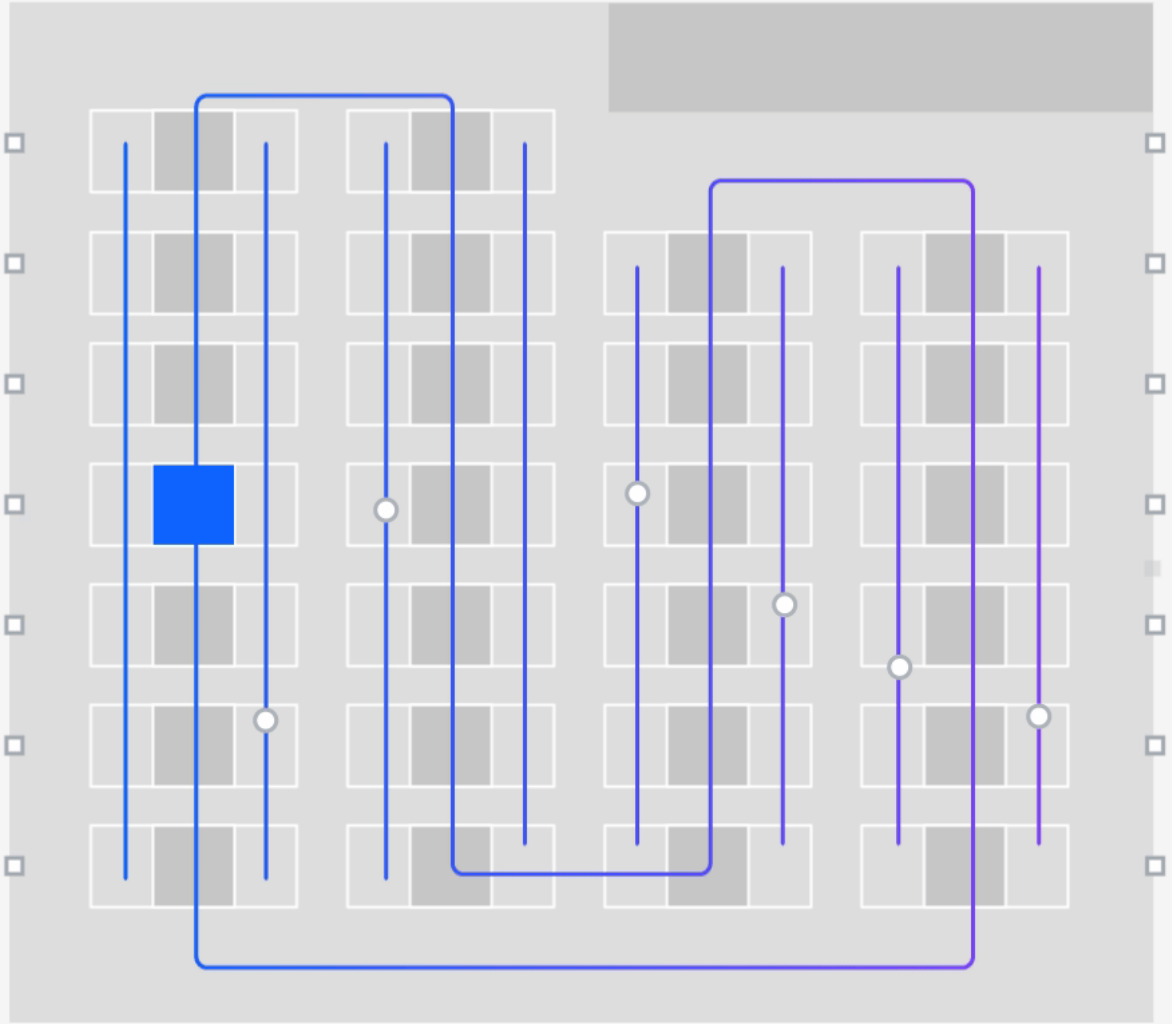
Multi-precision inference FP16/8

GPU Model	CUDA Cores	Tensor Cores	Power Usage (W)
NVIDIA V100	5120	640	250
NVIDIA T4	2560	320	70
NVIDIA A100	6912	432	400
NVIDIA RTX A6000	10752	336	300
NVIDIA A10	9216	288	150
NVIDIA H100	14592	456	**700**
NVIDIA L40S	18176	568	350
NVIDIA L4	7,424	240	72
NVIDIA B200	18944	592	1000

Bidirectional Ring Interconnect Architecture

Each Spyre card integrates **32 AI Units** (AIUs), each with its own **2 MB local scratchpad** memory.

These AIUs are interconnected through a high-speed bidirectional ring, enabling **ultra-fast data exchange between compute units and local memory, eliminating data bottlenecks**.



200GB/s
LPDDR5 connects to scratchpad

128B
bidirectional ring connects scratchpads

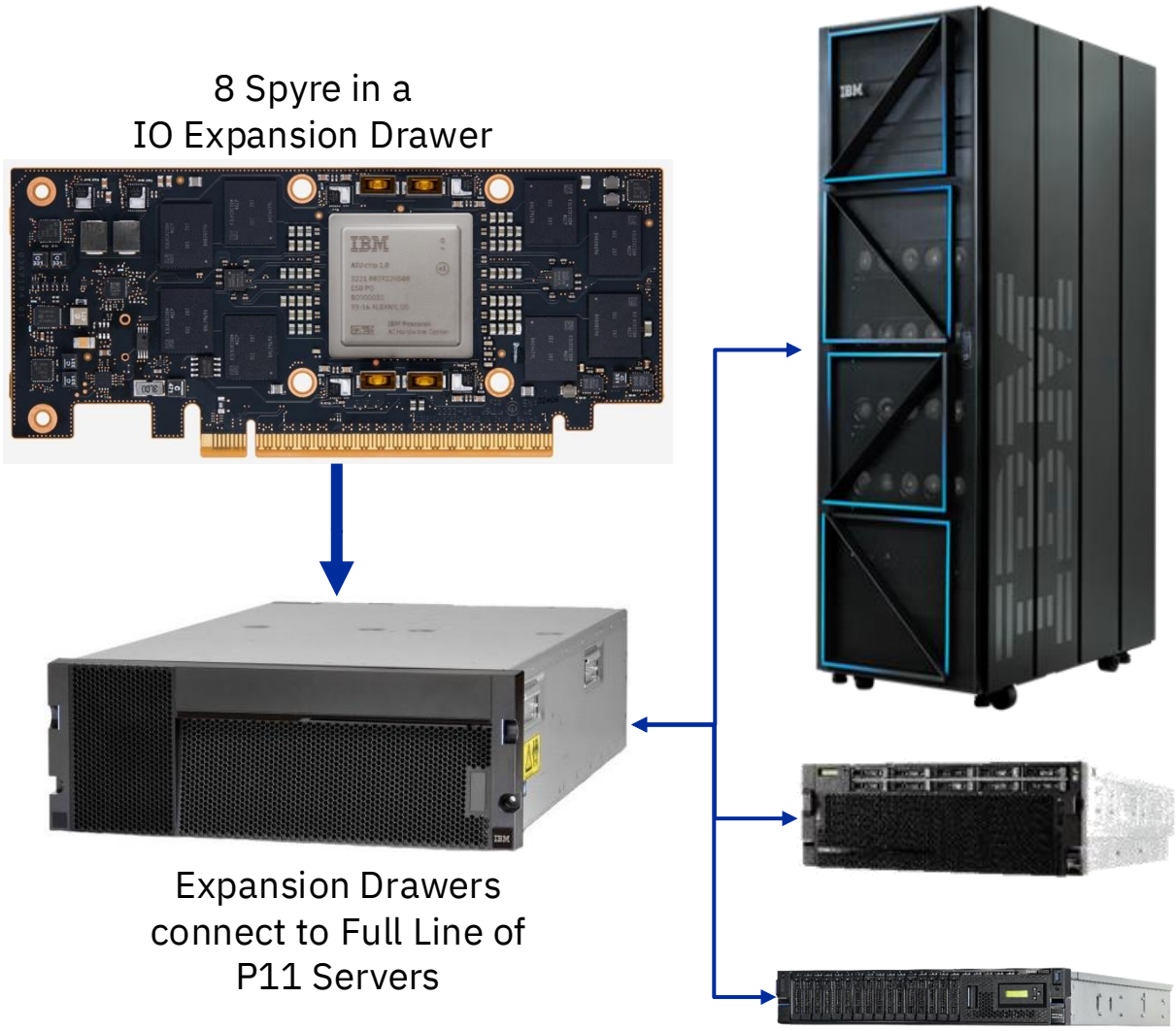
32B
bidirectional ring connects the cores

GPU vs IBM Spyre Card

Feature	GPU	IBM Spyre Card
Primary purpose	Originally designed for graphics; now a general-purpose accelerator	Designed specifically for AI inference
Flexibility	Highly versatile	Optimized for LLM inference
AI workload	Training and inference	Primarily inference
Compute architecture	Thousands of parallel cores	Specialized AI Compute Units
Memory	VRAM (HBM/GDDR)	Dedicated VRAM + local Scratchpad Memory
Power efficiency	High performance but often high power	Optimized performance per watt
Examples	NVIDIA H100, AMD MI300	IBM Spyre

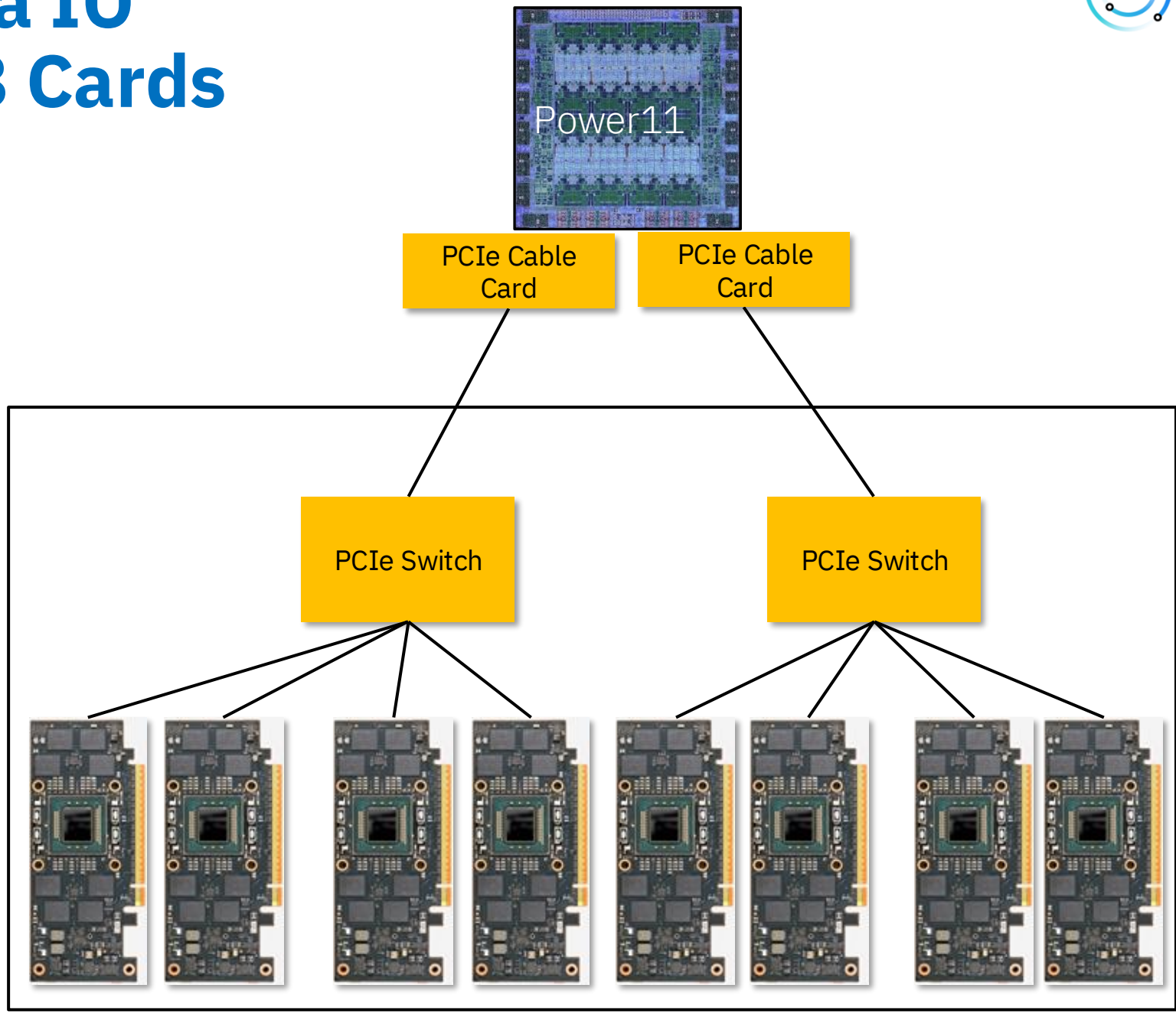
IBM Spyre on Power

- **8 cards in I/O Drawer form a logical cluster**
- **1TB of Memory**
- **1.6TB per second aggregate memory bandwidth**
- **2400+ TOPS FP16**
- **Will be Integrated into AI SW Stacks, drivers, runtimes, optimized frameworks…**



System Integration via IO Drawer in a Batch of 8 Cards

- Two host connecting cable cards per drawer
- Groups of 4 cards are behind a PCIe Switch (Fanout Module)
- Software stack **can share the models to distribute work across multiple Spyre cards** to reduce inference latency
- Each Spyre card works on part of the problem

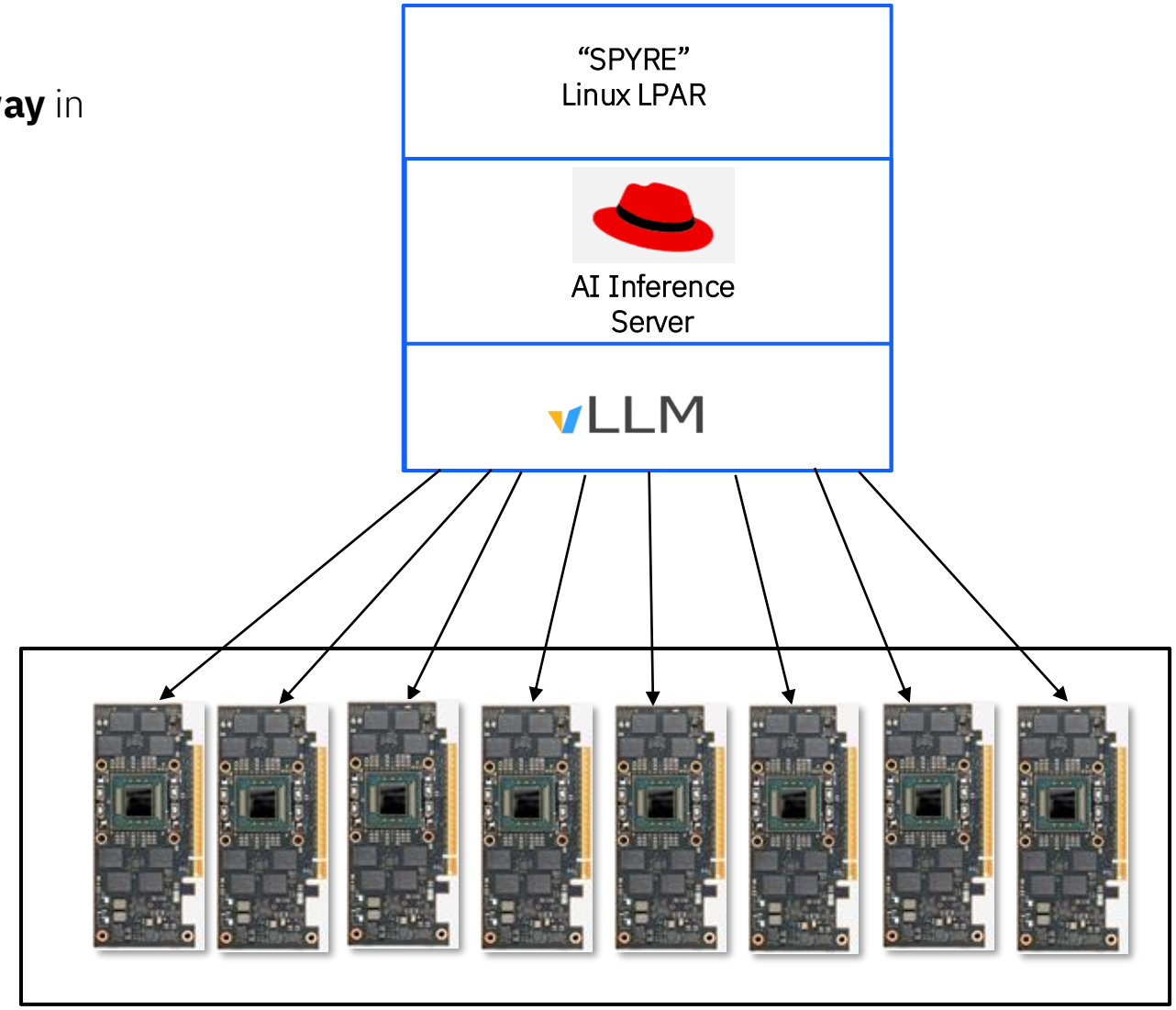


vLLM organizes, Spyre speeds up

vLLM is an open-source inference engine that allows to **execute AI models in a fast, scalable and efficient way** in memory (optimized KV cache)

vLLM dynamically groups requests, and maps IBM Spyre run batches in parallel.

Results:
- Better bandwidth
- Less latency
- Better use of resources



Spyre Stack

external

mandatory

optional

*removal allowed if purchased via Red Hat. Support for those components will then come via Red Hat.

Announcements
- Software & services (5639-SPY, 5639-SPA, 5639-RIS, 5639-SAI)
- Hardware (ECSE)

Licenses & policies
- 5639-SAI: IPLA (warranted)
- 5639-RIS: RH Policies (warranted)
- 5639-OAI/5639-OAA/5639-OCP/5639-OPP: RH Policies (warranted)
- 5639-SPR: IPLA (warranted)
- 5639-SPY: ILAN (non-warranted)
- 5639-SPA: IPLA (warranted)

Services
Support

Applications
Core applications, ISV solutions, ...

Application integration
5639-SAI
IBM Open-Source AI Foundation for Power
(catalog with pre-built AI services, adoption patterns, and integration packages & tools such as Python wheels)

either on RHEL ...or on OpenShift

Inference server
5639-RIS
Red Hat AI Inference Server*

Red Hat OpenShift AI*

5639-OAA
Red Hat AI Accelerator for IBM Spyre™*

Operating system
5639-1RE/3RE/5RE
RHEL 9.6+*

5639-SPR
IBM Spyre™ Operator for Power

5639-OCP, 5639-OPP
Red Hat OpenShift*

Middleware
5639-SPY
IBM Spyre™ Enablement Stack for Power
(driver, backend, runtime, firmware, models, ...)

5639-SPA
IBM Spyre™ Stack Updates for Power
(for driver, backend, runtime, firmware, models, ...)

HMC

PowerVM

Hardware
9080-HEU, 9043-MRU, 9856-42H, 9824-42A, 9856-22H, 9824-22A
IBM Power11 server

ENZ0 + 2xENZA + ECLR / ECLS / ECLX / ECLY / ECLZ + EJ24 / EJ2A
I/O drawer + cabling + adapter

8xECSE
8x IBM Spyre™ Accelerator for Power

It's a go shopping in a catalog of production-ready use cases.

And you could tailor them without expensive customization.

IBM Spyre is designed to accelerate **real enterprise AI use cases**, not just AI models.
The platform provides pre-built services and adoption patterns that enable organizations to quickly deploy many AI applications across multiple industries.

Clients are aligning AI use cases with core workloads on IBM Power.

Cross-industry Use cases

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
- Gate Global Logistics — Order processing assistant
- Large retailer (US) — Product sales assistant

Industry-specific use cases

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
- SPH — Medical transcription assistant
- Claims & EHR matching agent
- SPH — 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 — Real estate assistant
- ...

integrated Intelligent apps Adoptoion Patterns

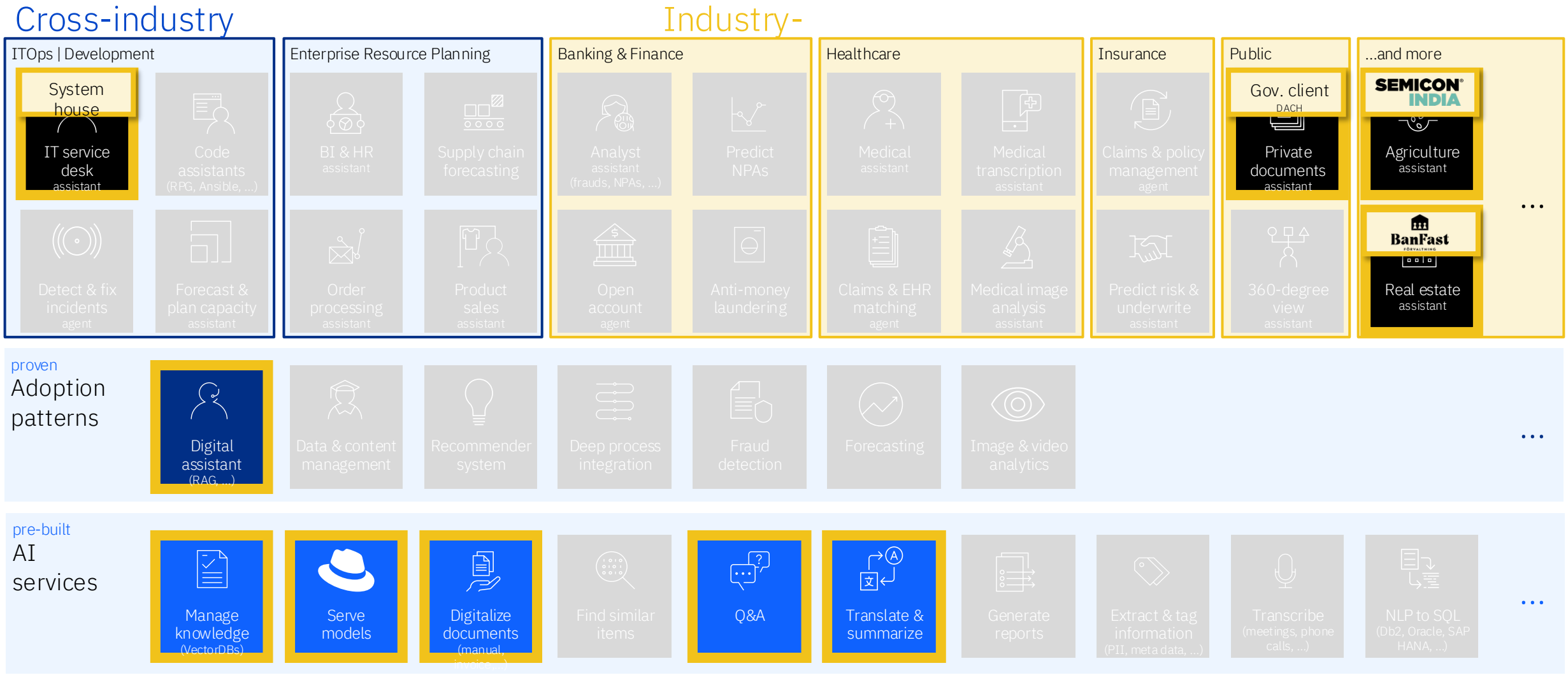
- Assisted workflows (deep integration of AI services into enterprise workflows)
- Digital advisors (retrieval augmented generation, chat interactions, ...)
- Agentic collaborators (tools & skills, autonomous orchestration, ...)

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, ...)
- ...

Digital assistants

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

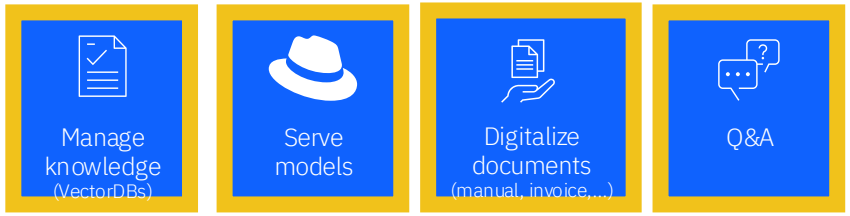


RAG out-of-the-box

being validating in IBM Spyre™ for Power.

proven
Adoption
patterns

pre-built
AI
services



Open-source
Technologies

milvus

Red Hat AI
Inference Server

Docling

carbon-components-react

Granite

Entity extraction is the process of automatically identifying and classifying relevant information (entities) from unstructured text

Deep process integrations

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

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]

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.

IBM Spyre™ Solutions catalog

Optimized

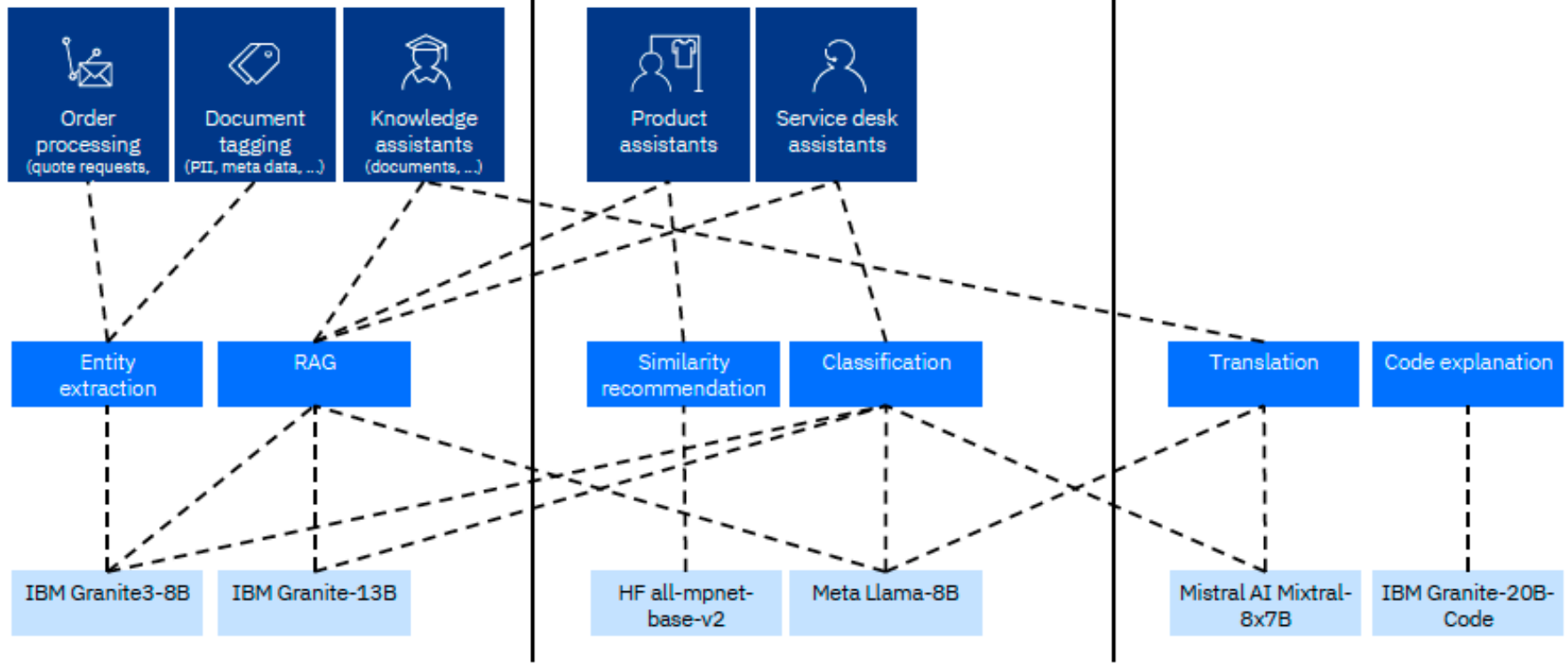
IBM has optimized accuracy & performance for enterprise use cases.

Tested

Functionally working and exploring for enterprise use cases.

Available

DIY exploration and experimentation of AI capabilities.



Final determination of use cases, capabilities, and models subject to change until GA.

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

FINALLY, SPYRE SOLUTION IS :

1. An integrated hardware AND software offering
2. Turnkey solution with a catalog of ready to deploy use cases, AI services and deployment patterns.
3. IBM Spyre is designed to accelerate real enterprise AI workloads, not just AI model execution
4. Deploy AI as close as possible to your data to leverage Data Gravity
5. End-to-end IBM support for both hardware and software

Redbook : https://www.redbooks.ibm.com/redpieces/abstracts/sg248592.html

Roadmap (subject to change)

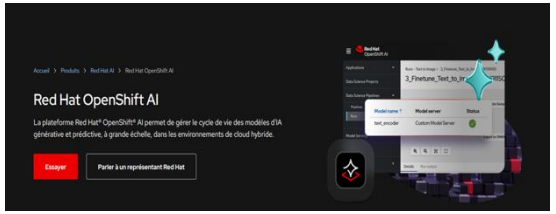
AI service updates
- Translate
- Extract & Tag Information
- NLP to SQL
- Transcribe

Model updates
- LLMs: mistral-small-3.2-24b (Q2), granite4.1-8B (Q2), meta-llama-3.1-8B, mistral-3-14B-instruct(Q3), granite-3.3-2b-vision, mistral-small-3.2-24b-instruct-2506
- Image-2-Text: granite-vision-3.3-2b(Q3)

General updates
- Performance improvements (FP8 support) (Q2)
- +50% concurrent users

General
- Bring-your-own AI service
- Experiment with AI services in IBM PowerVS
- Bring your own models (2027)

Hardware
- 12 cards in a Drawer instead of 8
- 2 X more I/O Drawer per server :
  E1180 : 1-4, Mid-Range. Scale-out : 1-2



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

Delivery Roadmap

	4Q 2025 – Mar. 31st GA (delivered)	2Q 2026 (in plan)	3Q 2026 (working plan)	4Q 2026 (early assessment)	1Q 2027+ (HW & SW advances)
proven adoption patterns	IBM **Open-Source AI Foundation for Power** • Simplify integration (Power-optimized Python wheel ecosystem) • Support for 1Q26 catalog (containers released for AI services)	• Simplify integration continued (Python wheel currency & additions) • Support for 2Q26 catalog	• Simplify integration continued (Python wheel currency & additions) • Support for 3Q26 catalog	• Simplify integration continued (Python wheel currency & additions) • Support for 4Q26 catalog	
catalog with pre-built AI services	Open-source innovation • AI services ▪ Digitize Documents (texts, tables, images) ▪ Knowledge Management ▪ Q&A ▪ Summarize ▪ Similarity Search • UIs ▪ Digital assistant prototype ▪ Document ingestion UI ▪ CLI-based installation ▪ UI-based installation (prototype) • General ▪ Multi-language support (English, German) ▪ Integration APIs (REST) ▪ Deployment options (RHEL, OpenShift) ▪ Air-gapped installation	• AI service updates ▪ Q&A: accuracy improvements ▪ Translate (English-German) • UI updates ▪ UI-based installation • General updates ▪ Try AI services in IBM PowerVS ▪ Try AI services on-chip (MMA, SIMD) ▪ Custom weights for supported models ▪ Support for additional languages (Italian, French) ▪ Backup and restore all data	• AI service updates ▪ Extract & tag information ▪ Forecast • UI updates ▪ ERP auto-form-filling prototype ▪ Sizing calculator • General updates ▪ Additional languages (Japanese) ▪ Integration with Power Mission Control ▪ Offer AI services through MCP • Connector updates ▪ Fetch & digitize documents (i/AIX/Linux filesystems, Box, OneDrive)	• AI service updates ▪ NLP to SQL ▪ Detect anomalies • UI updates ▪ Live monitoring & audit logs • General updates ▪ Additional languages (Spanish, Portuguese, Swedish) ▪ Bring-your-own AI service ▪ Agents create and manage architectures • Connector updates ▪ Fetch structured data from Db2 for i ▪ Call AI services via Db2 for i SQL	
inferencing platforms	RH **AI Inference** | RH **OpenShift AI** Models ▪ **LLMs**: granite-3.3-8b-instruct ▪ **Reranking**: bge-reranker-v2-m3 ▪ **Embedding**: granite-embedding-30m-English, granite-embedding-125m-English, granite-embedding-107m-multilingual, granite-embedding-278m-multilingual	RH **AI Inference** | RH **OpenShift AI** | External platforms (IBM **watsonx.ai**, OpenAI, ...) Model updates ▪ **LLMs**: granite-4.1-8b-fp8, llama-3.1-8b-instruct (incl. support for custom weights), mistral-small-3.2-24b-instruct-2506 ▪ **Embedding**: multilingual-e5-large, multilingual-e5-large-instruct	▪ **LLMs**: ministral-3-14b-instruct-2512-bf16 ▪ **Image-2-Text**: granite-3.3-2b-vision, mistral-small-3.2-24b-instruct-2506 ▪ **Embedding**: qwen3-embedding-0.6b (or qwen3-embedding-4b)	▪ Unlock significantly broader model ecosystem in 2027 (preparation for new model stack)	
accelerated infrastructure	IBM **Spyre™ Accelerator for Power** + IBM **Power** on-chip acceleration • Base performance (FP16 support) • Performance improvements (chunked prefill, prefix caching)	IBM **Spyre™ Accelerator for Power** | IBM **Power** on-chip acceleration | IBM **Power Virtual Servers** (with watsonx.ai aaS as external platform) • Performance improvements (FP8 support)	• +2x concurrent users (12 cards in a drawer; expanded drawers: 1-4 HE, 1-2 MR&1124)	• RAS improvements (Enhanced Error Handling)	

Thank you for your attention.

Any Question

Jean-Manuel Lenez
jlen@ch.ibm.com
+41 79 278 92 44

Pre-Sales
IBM Power Systems
& Hybrid-Cloud

IBM Suisse
Esplanade de Pont-Rouge 9A,
1212 Lancy, Suisse