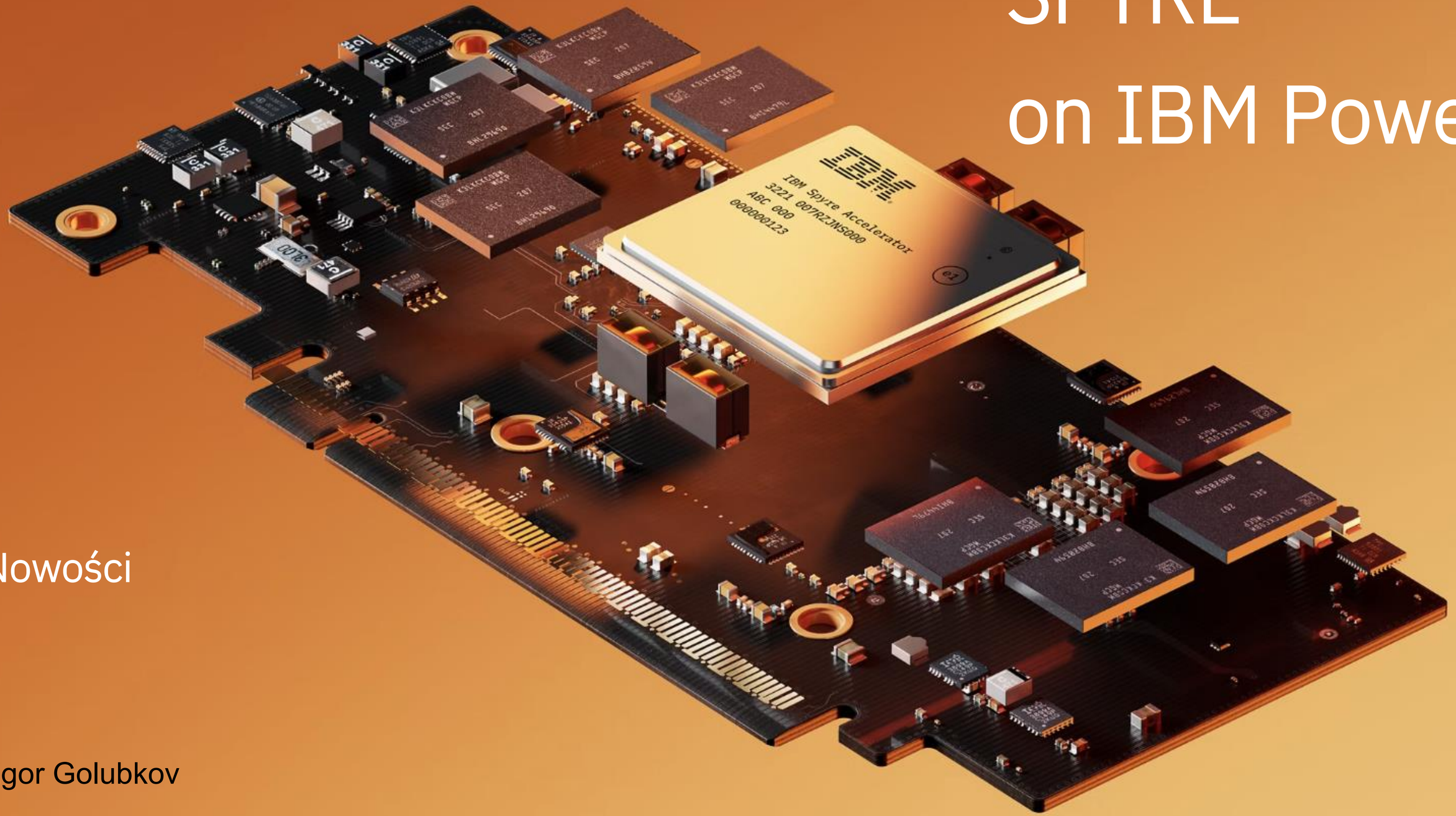
SPYRE
on IBM Power



Nowości

Egor Golubkov







Source: https://lakilakiidaman.wordpress.com/2016/04/06/19/

Training



Dane wysokiej jakości

Inferencing (Wnioskowanie)

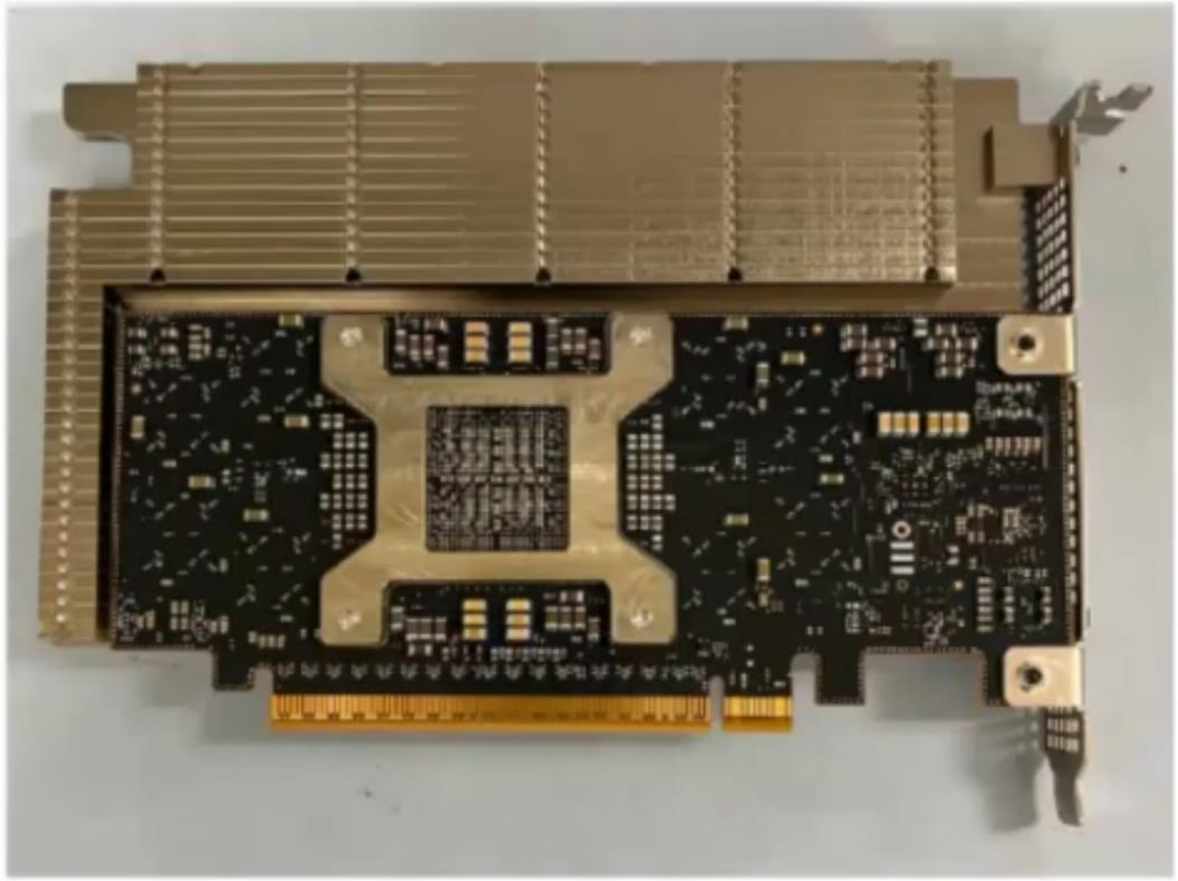


Dane różnej jakości

IBM Spyre™ Accelerator



Front

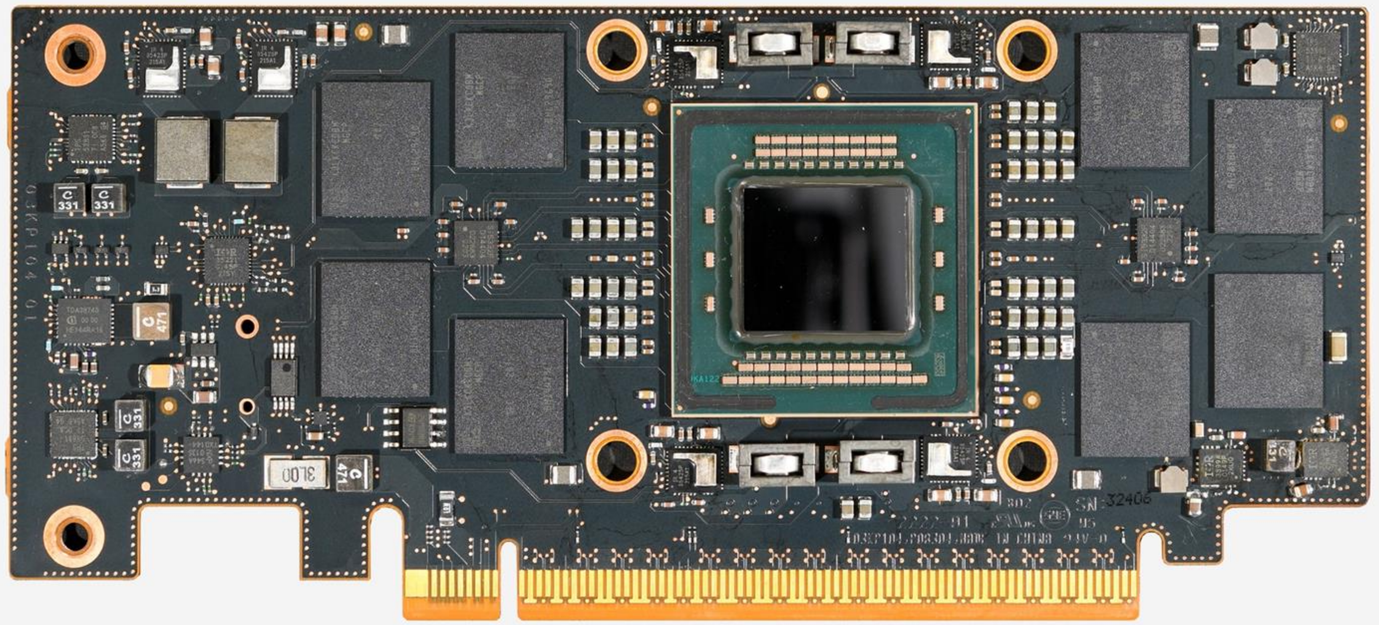


Back

IBM Spyre dedykowany pod GenAI

IBM Spyre™ Accelerator Adapter PCIe

- Wydajność: 300+ TOPS
- Pobór mocy: 75W
- Pamięć: 128GB LPDDR5



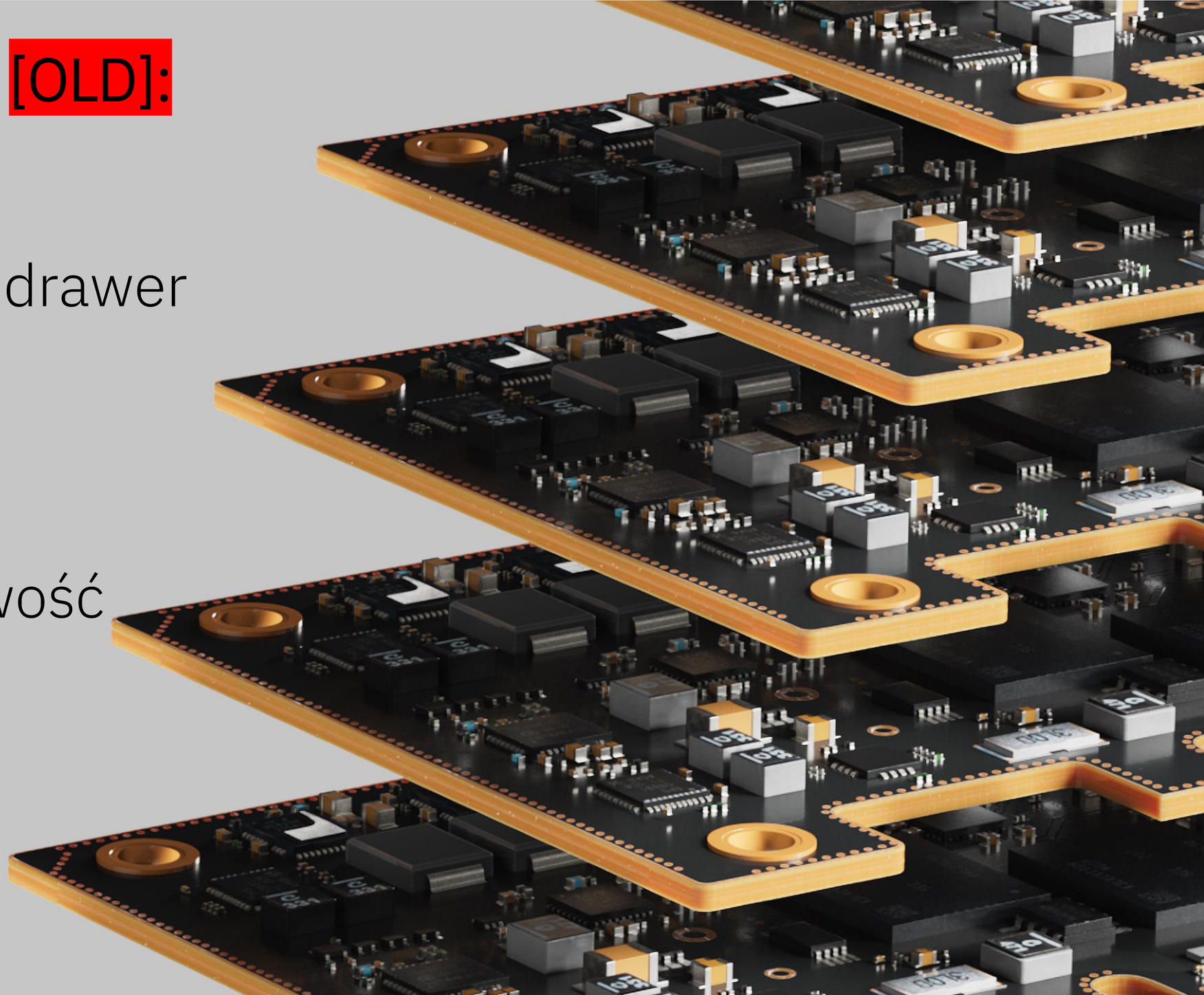
All projections are based on preview hardware.

Skalowalność Spyre [OLD]:

8 kart Spyre per I/O drawer

1 TB pamięci

1.6 TB/s przepustowość

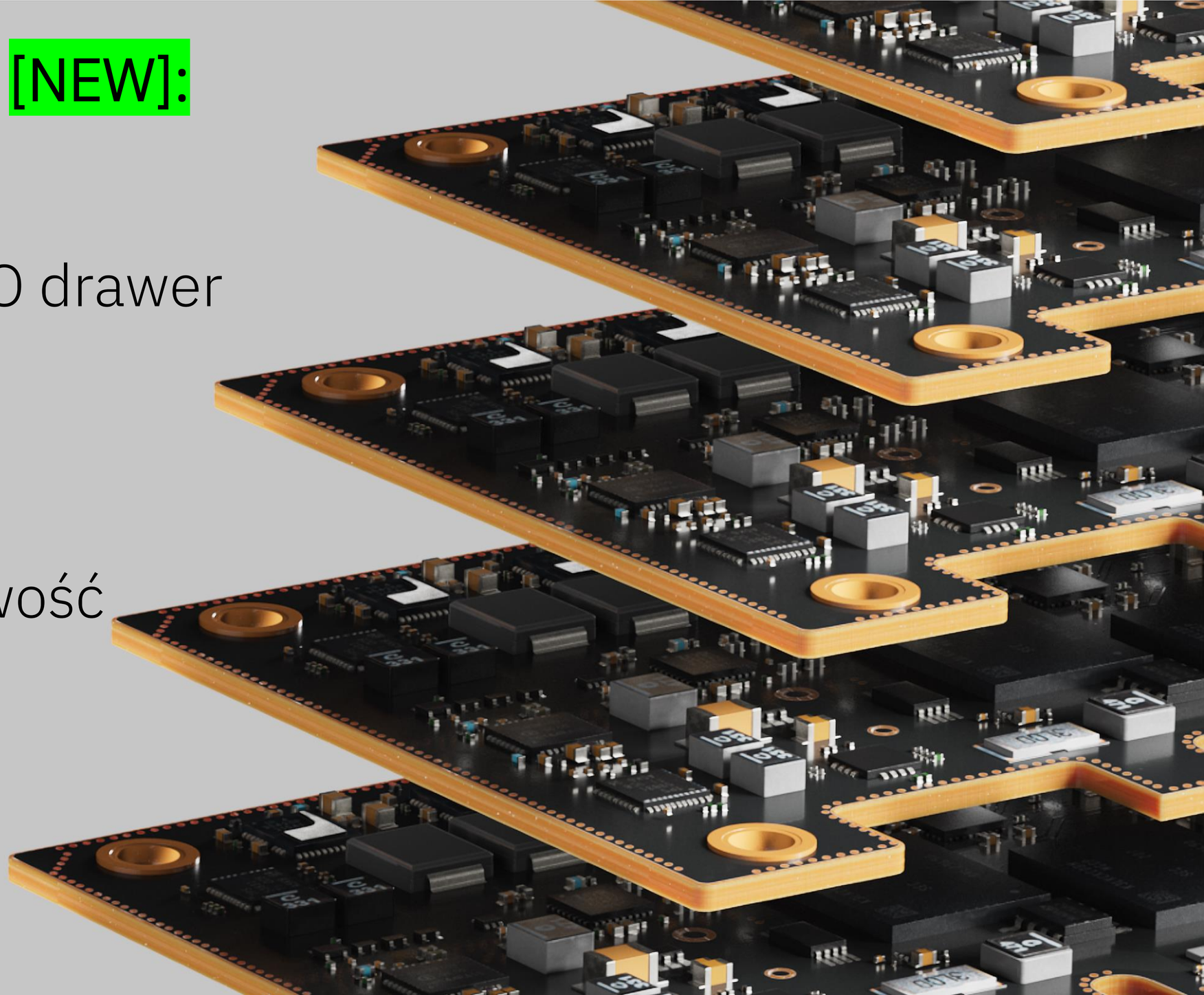


Skalowalność Spyre [NEW]:

12 kart Spyre per I/O drawer

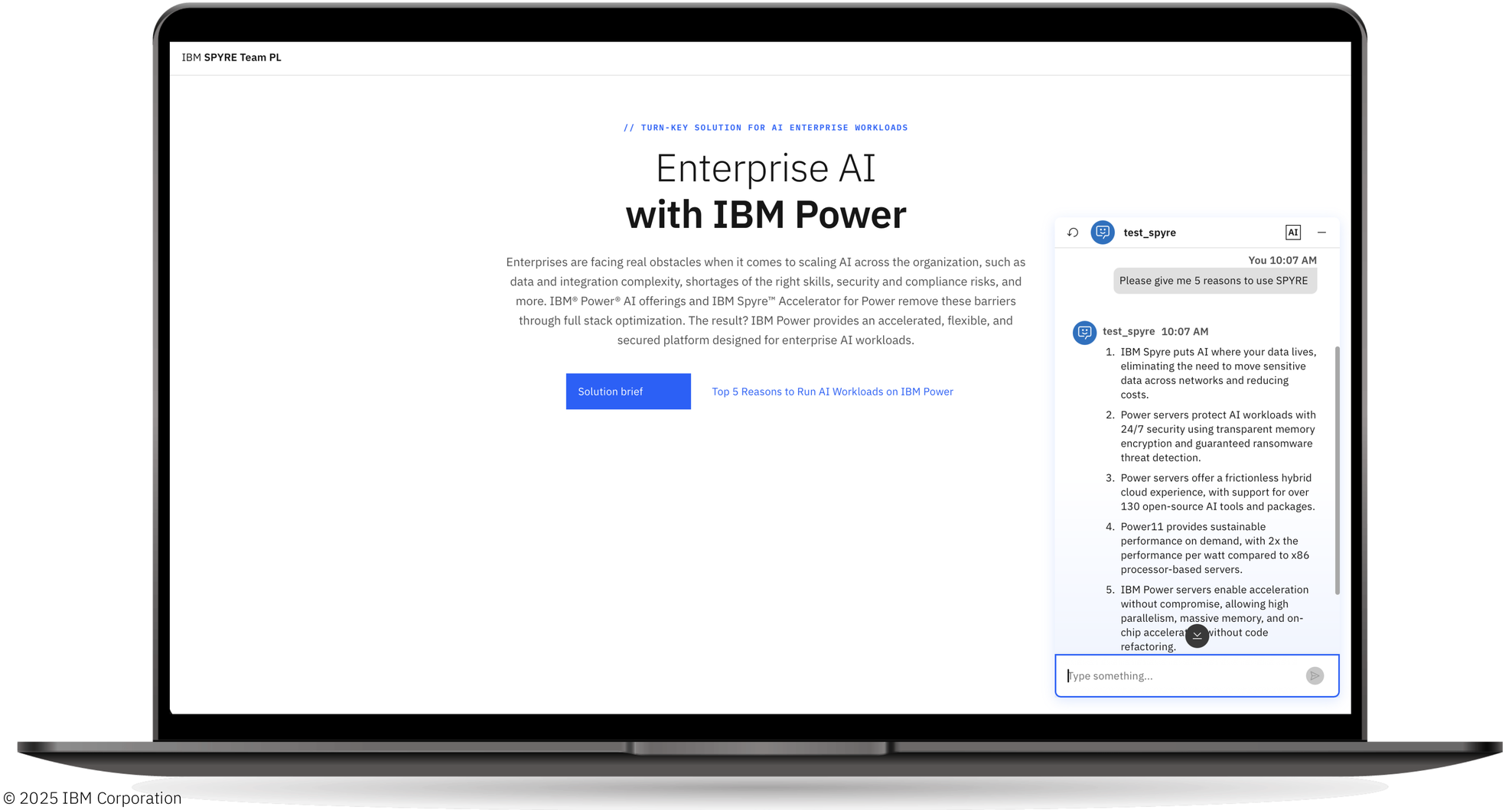
1.5 TB pamięci

2.4 TB/s przepustowość



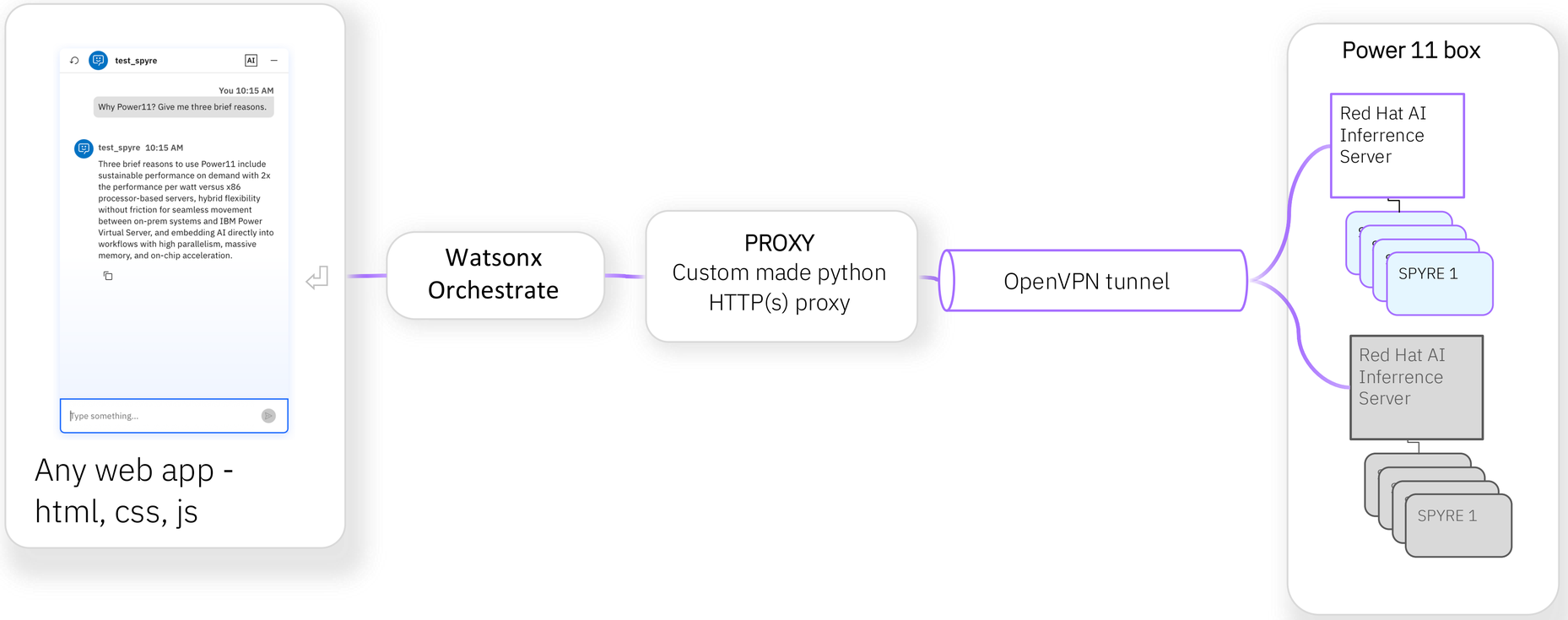
Yes, we can connect watsonx Orchestrate with an LLM model on SPYRE.

Q: Please give me 5 reasons to use SPYRE



http://localhost:8080

Jak to działa?



Architektura on-prem AI:



8 or 12 Spyres in a ENZ0 Expansion Drawer



ENZ0 Drawer



E1180



E1150



S1124

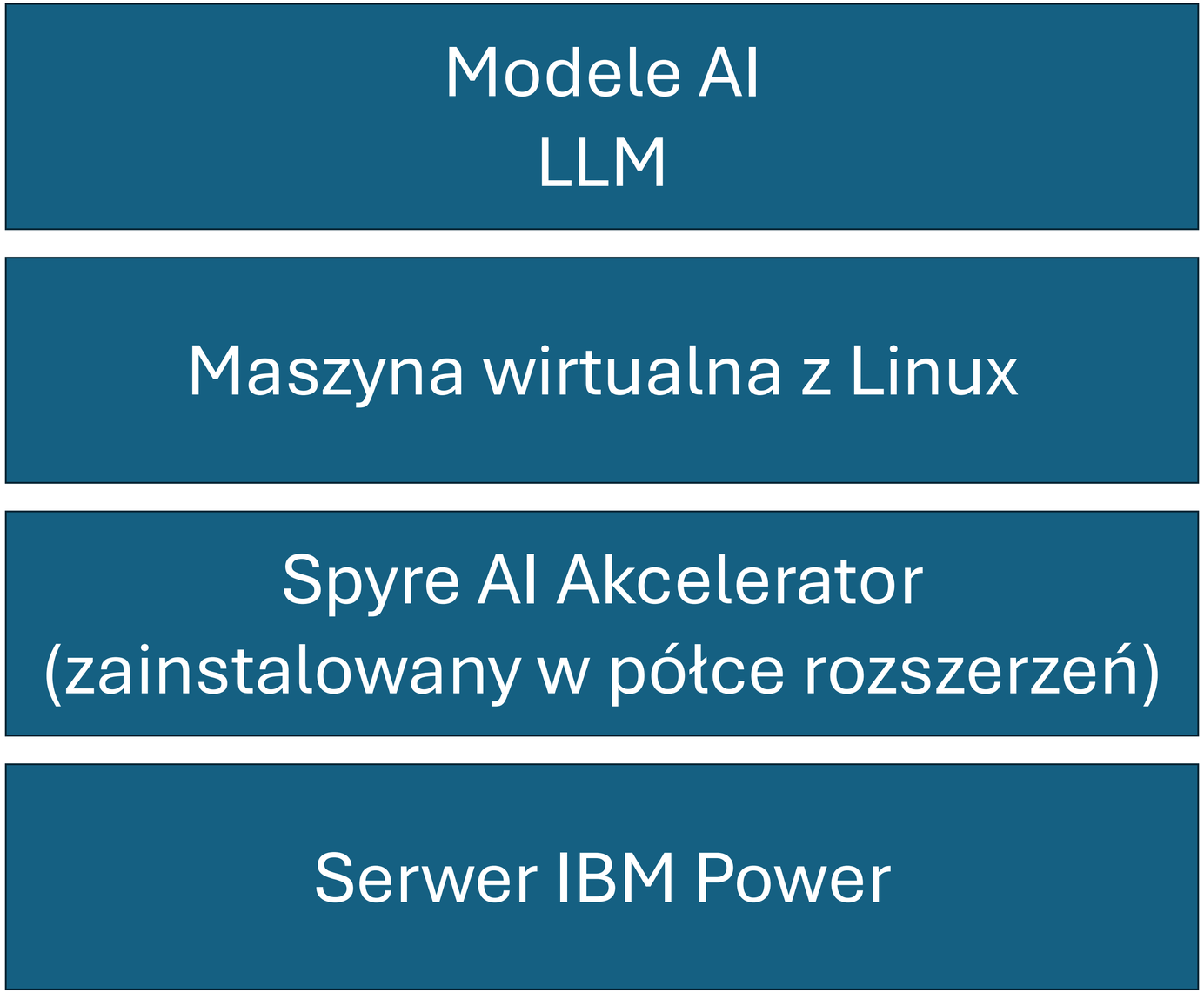


S1122

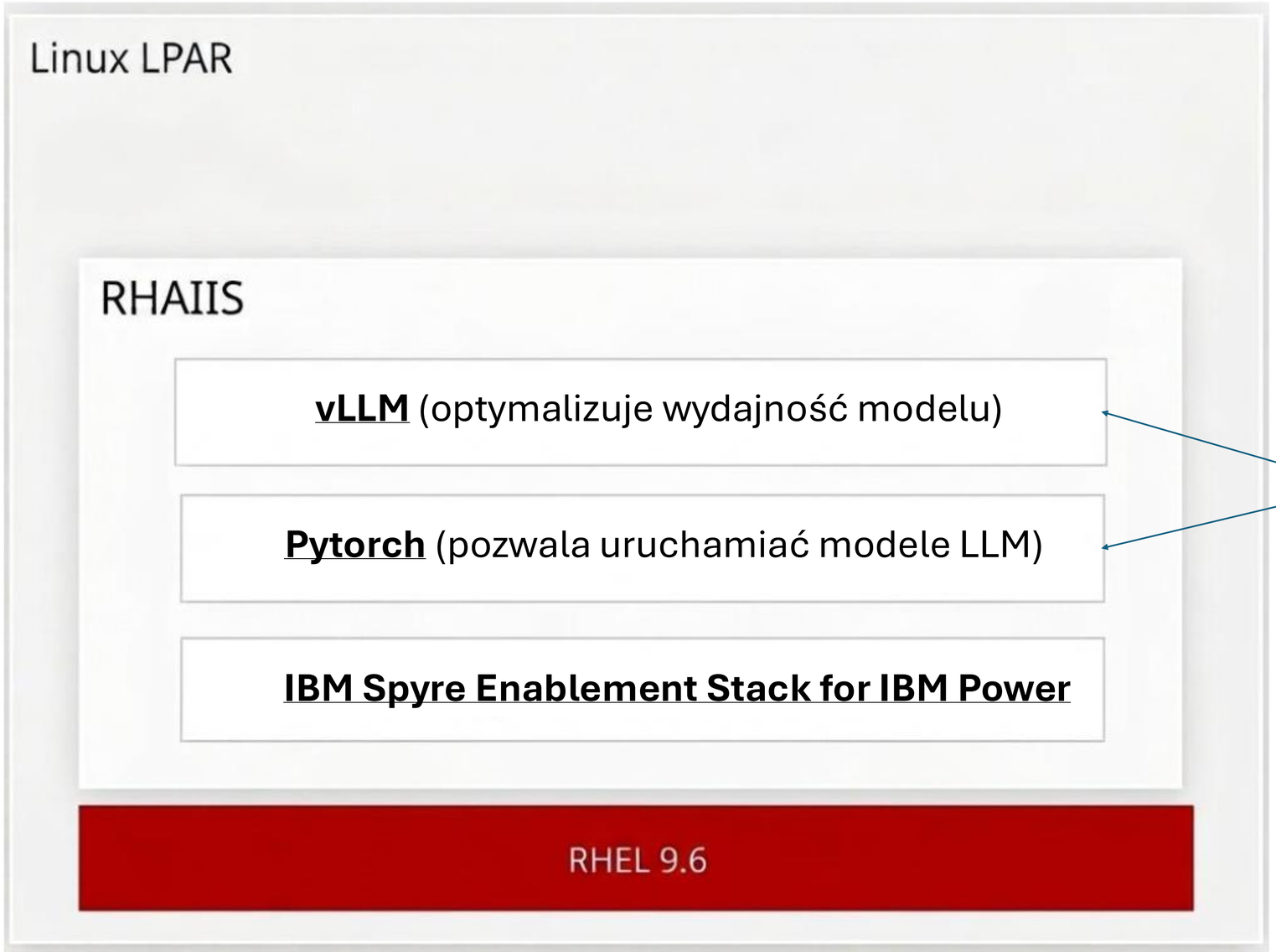


S1112

Architektura AI na IBM Power:



Architektura Red Hat AI Inference Server (RHAIIS)



Spyre on Power Use Cases

Use case	Model name	Batch size	Maximum input context size	Maximum output context size	Number of cards per container
Entity extraction	Granite3.3-8b-instruct	16	3K	3K	1
RAG embedding	Granite-Embedding-125m-English Granite-Embedding 278m-multilingual	Up to 256	51		
RAG embedding	Granite-Embedding-30m-English Granite-embedding-107m-multilingual	Up to 256	51		
RAG inferencing	Granite3.3-8b-instruct	32	32K (batch * context equals less than 128K)	32K (batch * context equals less than 128K)	4
Reranker	bge-reranker-v2-m3	Up to 4	8K	8K	

```
(APIServer pid=1) INFO 04-14 14:11:58 [loggers.py:257] Engine 000: Avg prompt
throughput: 0.0 tokens/s, Avg generation throughput: 16.6 tokens/s, Running: 1
reqs, Waiting: 0 reqs, GPU KV cache usage: 0.0%, Prefix cache hit rate: 0.0%
(APIServer pid=1) INFO:    172.25.25.81:37634 - "POST /v1/chat/completions
HTTP/1.1" 200 OK
(APIServer pid=1) INFO:    172.25.25.81:37634 - "POST /v1/chat/completions
HTTP/1.1" 200 OK
(APIServer pid=1) INFO 04-14 14:12:08 [loggers.py:257] Engine 000: Avg prompt
throughput: 4.3 tokens/s, Avg generation throughput: 15.2 tokens/s, Running: 1
reqs, Waiting: 0 reqs, GPU KV cache usage: 0.0%, Prefix cache hit rate: 0.0%
(APIServer pid=1) INFO 04-14 14:12:18 [loggers.py:257] Engine 000: Avg prompt
throughput: 0.0 tokens/s, Avg generation throughput: 15.3 tokens/s, Running: 1
reqs, Waiting: 0 reqs, GPU KV cache usage: 0.1%, Prefix cache hit rate: 0.0%
```

```
Model:              /src/granite-3.3-8b-instruct
Concurrent:         50
Total Requests:     100
Temperature:        0.7
Max Tokens:         2048
Authentication:     Enabled

Running non-streaming test...
Testing...
Load Test Results - Non-Streaming
Total Requests:              100
Successful:                  100 (100.0%)
Failed:                      0 (0.0%)
Total Duration:              263.58s
Avg Latency:                 108.44s
Tokens/Second:               114.6
Requests/Second:             0.38
Total Tokens:                34 723
  - Prompt Tokens:           4 510
  - Completion Tokens:       30 213

Running streaming test...
Testing...
Load Test Results
Total Requests:              100
Successful:                  100 (100.0%)
Failed:                      0 (0.0%)
Total Duration:              252.57s
Avg Latency:                 106.28s
Avg Time to First Token:     37.80s
Tokens/Second:               125.2
Requests/Second:             0.40
Total Tokens:                32 716
  - Prompt Tokens:           1 090
  - Completion Tokens:       31 626

✓ Results saved to: results.json
```



Florian Dahlitz Apr 15th at 2:18 PM

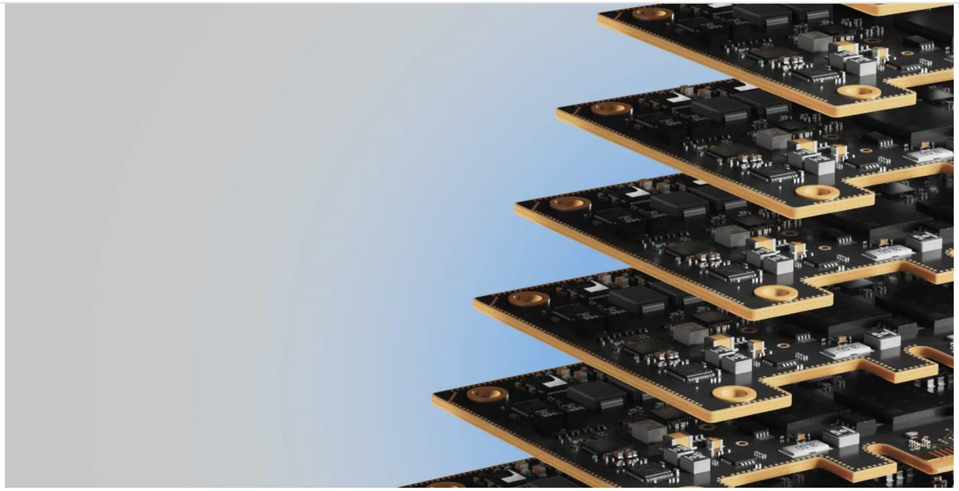
If you refer to the token generation phase, I would say yes, this is what you can expect to have. Spyre shines in the prompt intake phase. @raji maybe you have more insights into the performance numbers in the token generation phase?



Rajalakshmi Srinivasaraghavan Apr 15th at 2:22 PM

Yes 15 tok/sec is in the expected range for single request for granite8B model . When we move to fp8 model in next release we can expect some improvements .

IBM Research



Authors

Viji Srinivasan · Raghu Ganti · Mudhakar Srivatsa · Elpida Tzortzatos · Tabari Alexander · Vaidyanathan Srinivasan

Topics

AI Hardware

We recently published our 1H 2026 roadmap for enabling the IBM Spyre Accelerator in the PyTorch ecosystem. In a companion technical note, we described the hardware — its 32 active AI cores, mixed-precision SIMD-systolic arrays, and programmable dataflow design.

Here, we wanted to walk through the key themes of the roadmap and how we are building first-class PyTorch support for a dataflow accelerator in an ecosystem largely shaped by GPUs. Our philosophy is ecosystem-first — we're leveraging upstream mechanisms, minimizing custom code, and contributing back the pieces that make it easier for the next accelerator to follow the same path.

https://research.ibm.com/blog/pytorch-support-ibm-spyre

PyTorch Dev Discussions

Sign Up · Log In

IBM Spyre Accelerator: PyTorch Enabling Status and Feature Plan - 1H 2026

hardware-backends

raghukiran1224 — Mar 2

IBM Spyre Accelerator: PyTorch Enabling Status and Feature Plan - 1H 2026

Introduction

This roadmap outlines IBM's plan for integrating the Spyre accelerator with the PyTorch ecosystem during the first half of 2026. Our goal is to provide seamless PyTorch support for Spyre by building on existing PyTorch ecosystem components, ensuring minimal runtime overhead while maximizing performance and developer productivity. Equally important, we are committed to contributing back to the community — generalizing dataflow accelerator enablement in torch.inductor and lower layers, contributing OpenReg testing infrastructure into PyTorch core, and establishing out-of-tree CI/CD infrastructure — so that our work benefits not just Spyre but the broader ecosystem of dataflow accelerators.

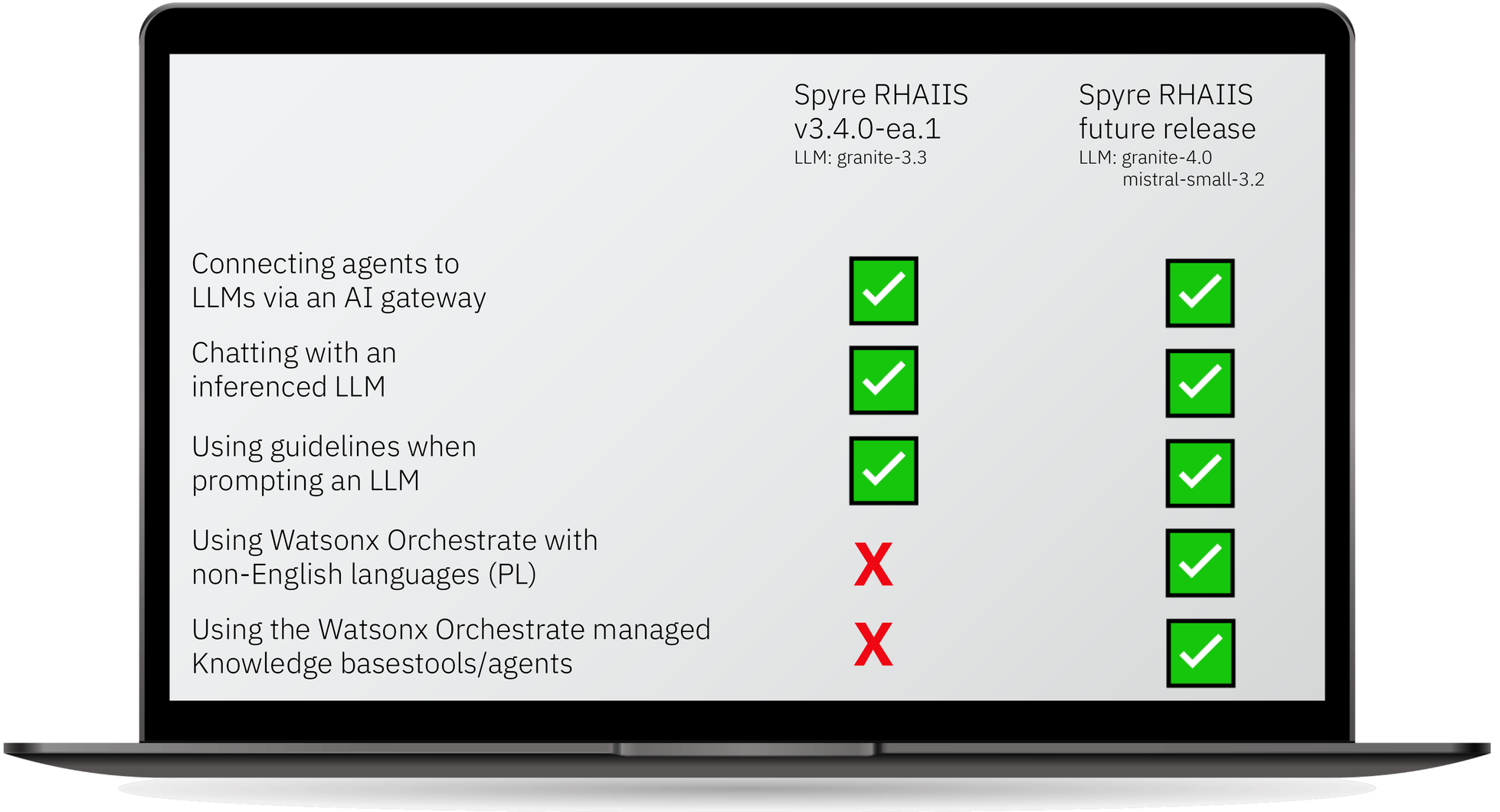
Scope for 1H 2026: All work in this roadmap is scoped to **inference only** at **FP16 and FP8** precisions, starting with the following **priority models**:

1. GPT-OSS (20B)
2. Granite 4 Hybrid (30B)
3. Mistral-small
4. Qwen 2.5 VL 7B
5. Llama 3.1 8B-instruct
6. Ministral 8B
7. Ministral 14B

https://dev-discuss.pytorch.org/t/ibm-spyre-accelerator-pytorch-enabling-status-and-feature-plan-1h-2026/3319

What Watsonx Orchestrate capabilities are available with SPYRE inferenced models?

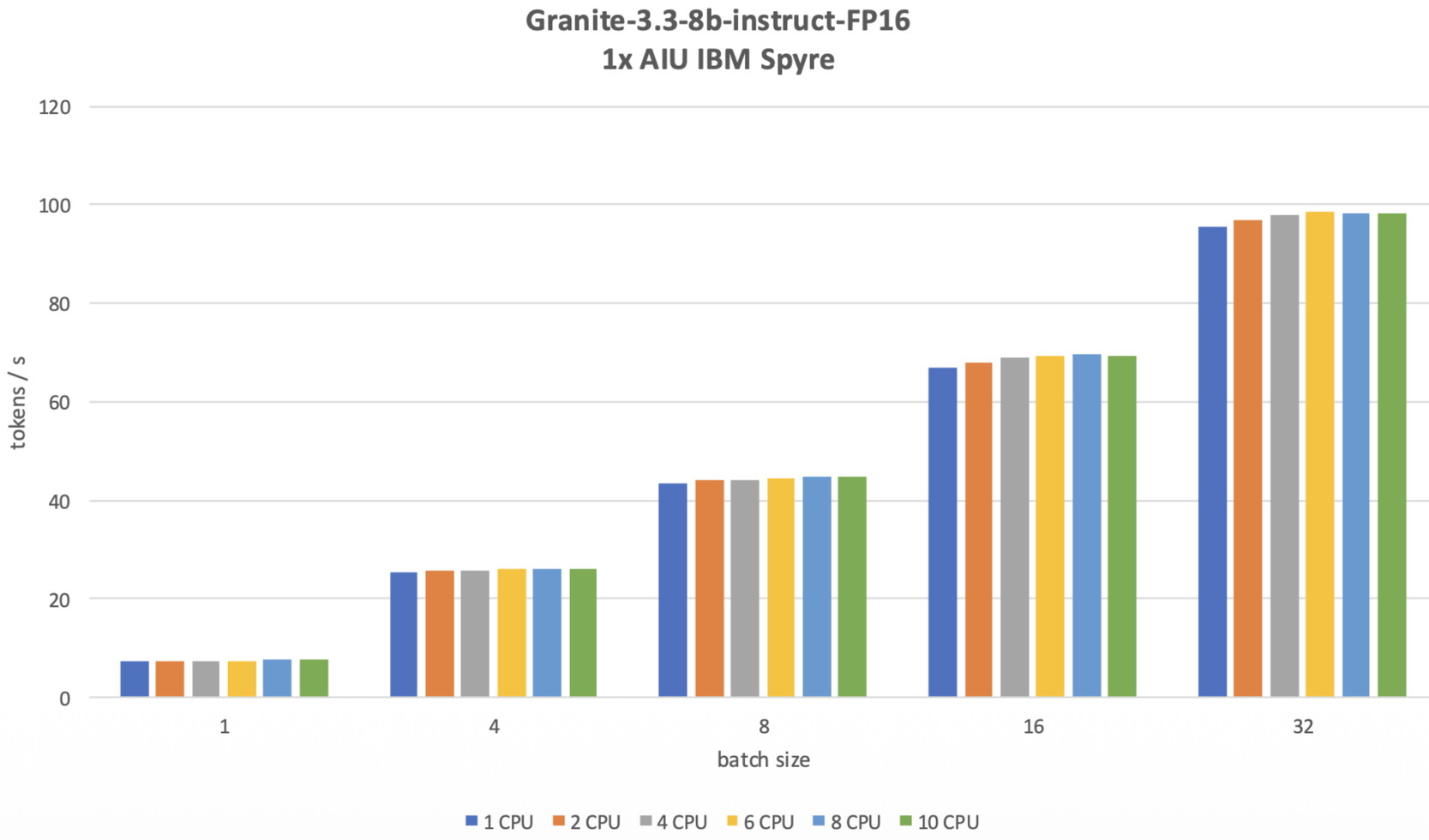
Q: What can I do now and in the future?



Wyniki kolejnych testów

vLLM Inference with 1x IBM Spyre adapter

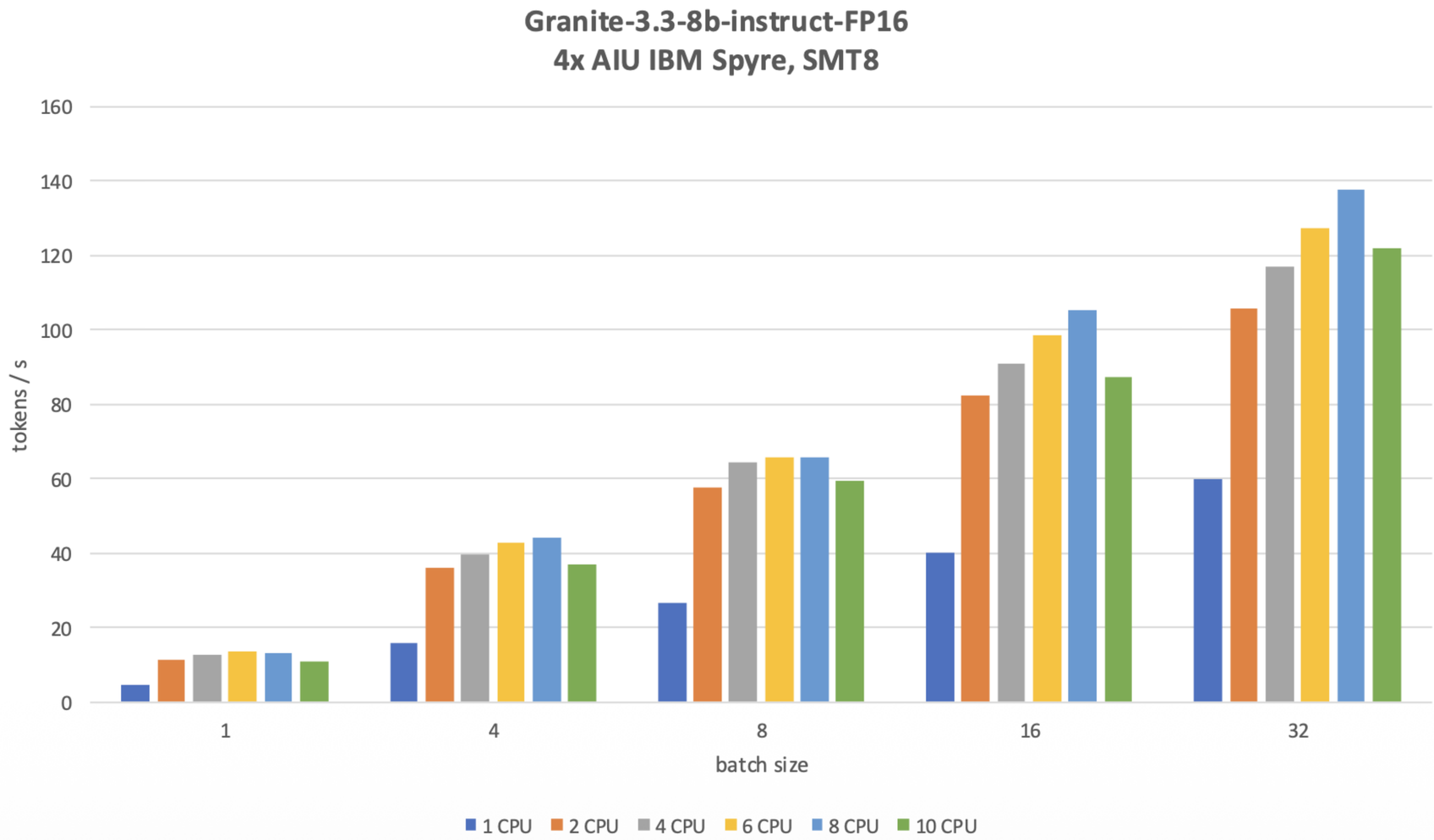
- Inference computation is offloaded to IBM Spyre adapter
- Throughput tokens/s does not depend on CPU count
- Use min. 1 CPU (SMT8) for inference with 1x IBM Spyre adapter
- <u>vLLM benchmark (CB):</u>
  - random input 1024 tokens
  - output 512 tokens
  - 1–10 dedicated Power11 CPU
  - 256 GB RAM (1 NUMA node)



Wyniki kolejnych testów

vLLM Inference with 4x IBM Spyre adapter

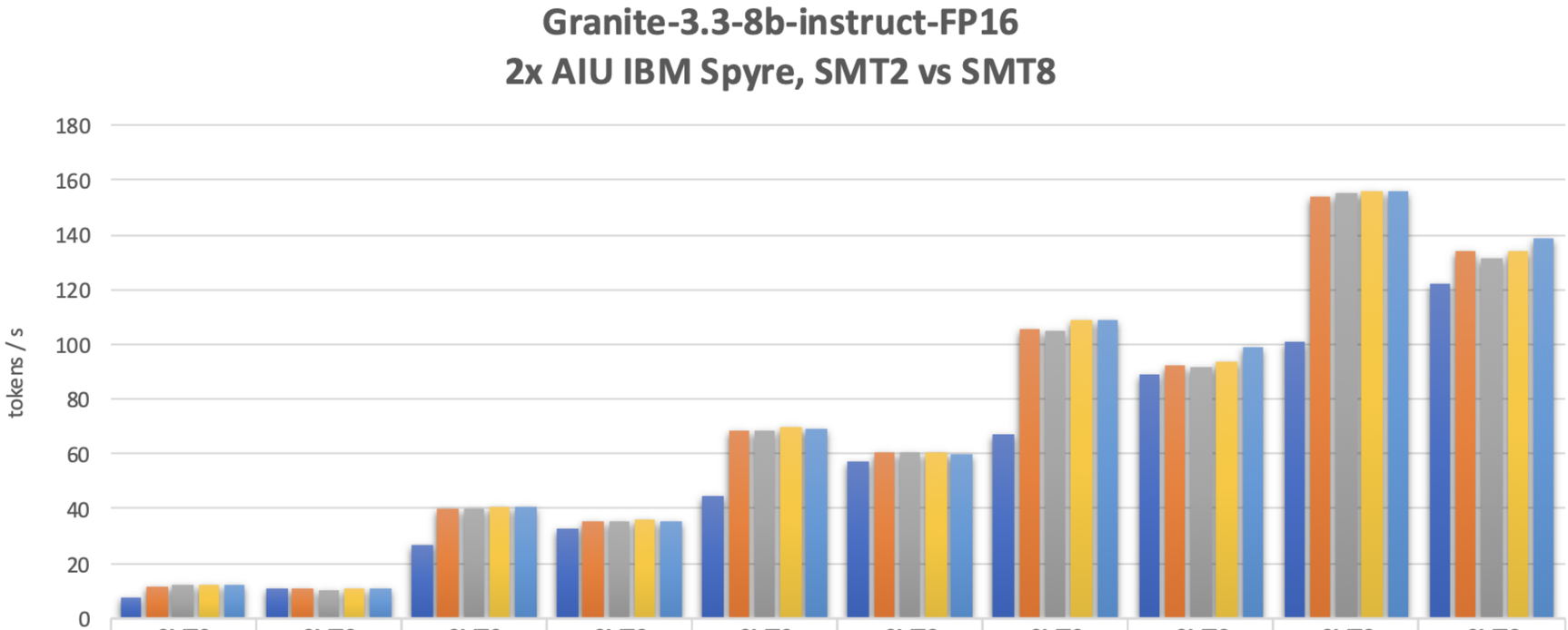
- Distributed inference computation with 4x IBM Spyre adapters
- Throughput tokens/s depends on CPU because of tensor parallelism in vLLM
- <u>vLLM benchmark (CB):</u>
  - random input 1024 tokens
  - output 512 tokens
  - 1–10 dedicated Power11 CPU
  - 256 GB RAM (1 NUMA node)



Wyniki kolejnych testów

SMT2 vs SMT8 with 2x IBM Spyre adapter

- Distributed inference computation with 2x IBM Spyre adapters
- SMT2 improves throughput tokens/s from 4 CPU
- Use min. **4 CPU (SMT2)** for inference with **2x IBM Spyre adapter**

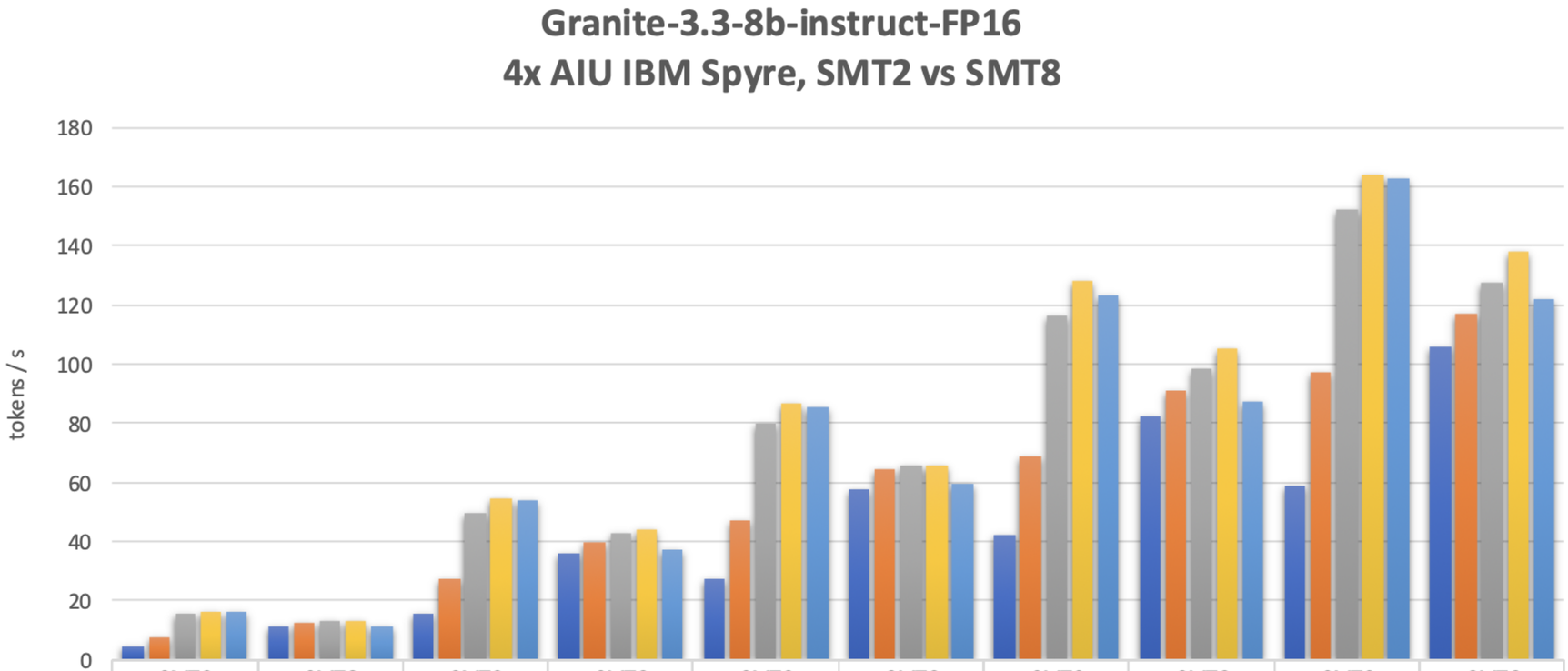


	batch size 1		batch size 4		batch size 8		batch size 16		batch size 32	
	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8
2 CPU	7.84	10.62	26.68	32.46	44.85	57.43	67.06	88.89	101.14	122.1
4 CPU	11.78	10.6	40.08	35.48	68.32	60.33	105.61	92.22	153.64	134.35
6 CPU	11.92	10.47	39.8	35.08	68.3	60.34	104.91	91.99	155.1	131.46
8 CPU	12.08	10.56	40.95	35.83	70.12	60.73	108.9	93.42	155.74	134.41
10 CPU	12.05	10.59	40.81	35.42	69.3	59.54	108.74	98.8	156.25	138.55

Wyniki kolejnych testów

SMT2 vs SMT8 with 4x IBM Spyre adapter

- Distributed inference computation with 4x IBM Spyre adapters
- SMT2 improves throughput tokens/s from 6 CPU
- Use **8 CPU (SMT2)** for inference with **4x IBM Spyre adapter**

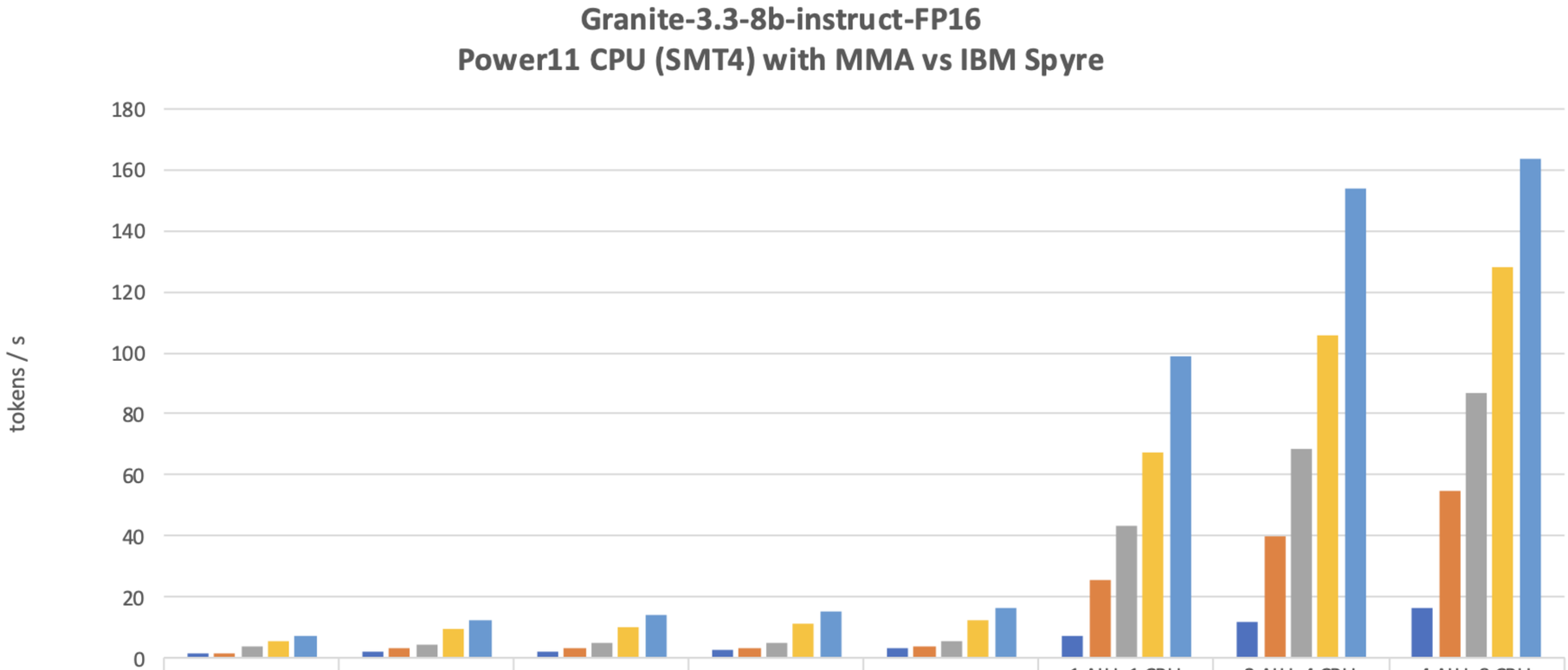


	batch size 1		batch size 4		batch size 8		batch size 16		batch size 32	
	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8	SMT2	SMT8
2 CPU	4.31	11.21	15.31	36.27	27.17	57.66	41.96	82.36	58.99	105.71
4 CPU	7.7	12.64	27.56	39.85	47.28	64.23	68.53	90.99	96.92	116.96
6 CPU	15.42	13.4	49.64	42.65	80.06	65.96	116.08	98.53	152.19	127.24
8 CPU	16.42	12.99	54.68	43.95	86.92	65.91	127.88	105.47	163.77	137.87
10 CPU	16.04	11.12	53.89	37.07	85.66	59.5	123.46	87.33	162.43	121.93

Wyniki kolejnych testów

vLLM Inference IBM Spyre vs CPU

- Inference on dedicated Power11 CPU with MMA
- vLLM inference on CPU includes 20–30% MMA instructions (batch size dependent)
- Throughput with 8 CPU (SMT2) and 4x IBM Spyre adapter



	1 CPU	2 CPU	4 CPU	6 CPU	8 CPU	1 AIU, 1 CPU, SMT8	2 AIU, 4 CPU, SMT2	4 AIU, 8 CPU, SMT2
batch_size 1	1.42	1.98	2.28	2.58	3.18	7.45	11.78	16.42
batch_size 4	1.7	3.03	3.22	3.41	3.68	25.4	40.08	54.68
batch_size 8	3.47	4.48	4.79	5.06	5.4	43.49	68.32	86.92
batch_size 16	5.24	9.24	10.32	11.12	12.14	67.09	105.61	127.88
batch size 32	6.93	12.23	13.78	14.99	16.46	98.85	153.64	163.77

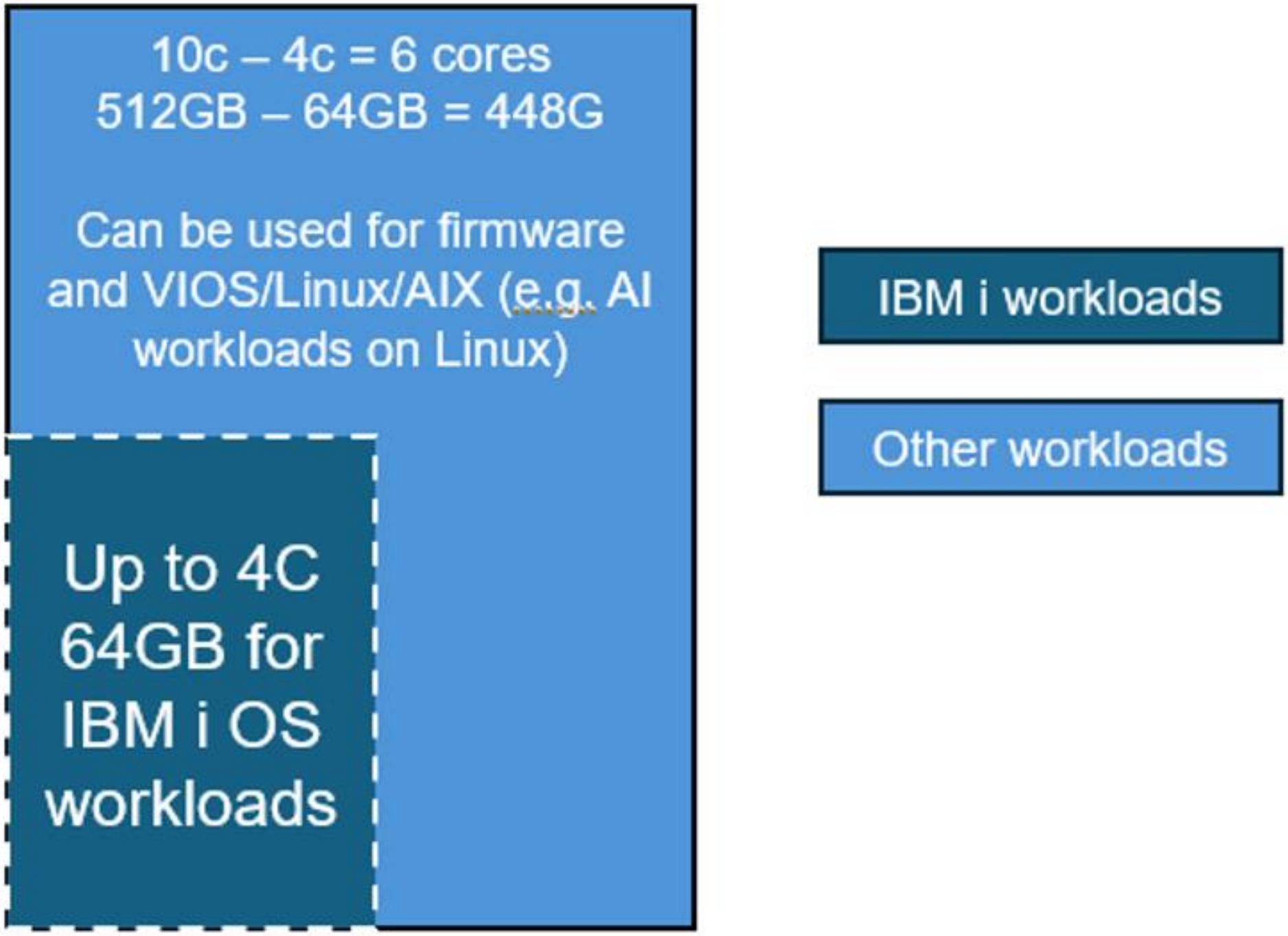
S1112



10 rdzeni
512 GB RAM
4 sloty PCIe
4 dyski NVMe
+ opcjonalna półka rozszerzeń
+ wsparcie dla Spyre
+ wsparcie dla Oracle Standard Edition 2 (SE2)






10c – 4c = 6 cores
512GB – 64GB = 448G
Can be used for firmware and VIOS/Linux/AIX (e.g. AI workloads on Linux)
Up to 4C 64GB for IBM i OS workloads
IBM i workloads
Other workloads

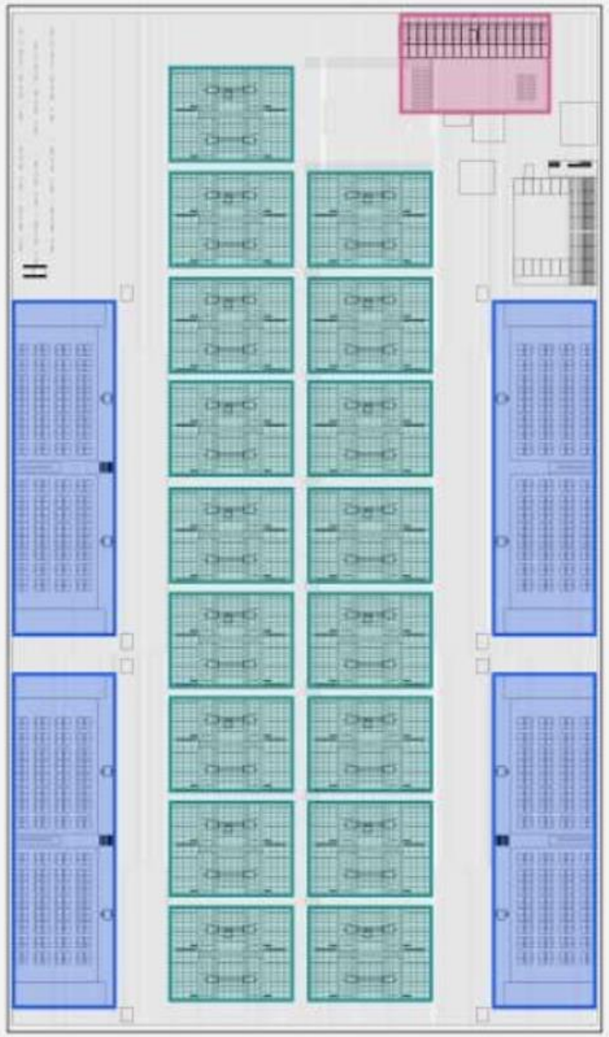
Spyre 2.0

Enterprise GenAI inference acceleration

Designed for larger enterprise inference workloads
Power envelope and performance vary by configuration

Tak jak w kartach GPU NVidia

HBM3e Memory
– 96GB HBM3e memory, up to ~4TB/s (20x)
– 20x vs current generation



AI cores
– 16 active AI cores + 1 redundant core
– AI core changes for performance and FP4/MXFP4, up to 4x TOPS

PCIe Gen6
– PCIe Gen6
– Low-latency peer-to-peer interface

