Where are we heading with Gen AI?

NSF Compute-Energy Nexus Workshop

Alaa Youssef
IBM Research

December 2025

Dimensions of Innovation

World of Compute Scarcity

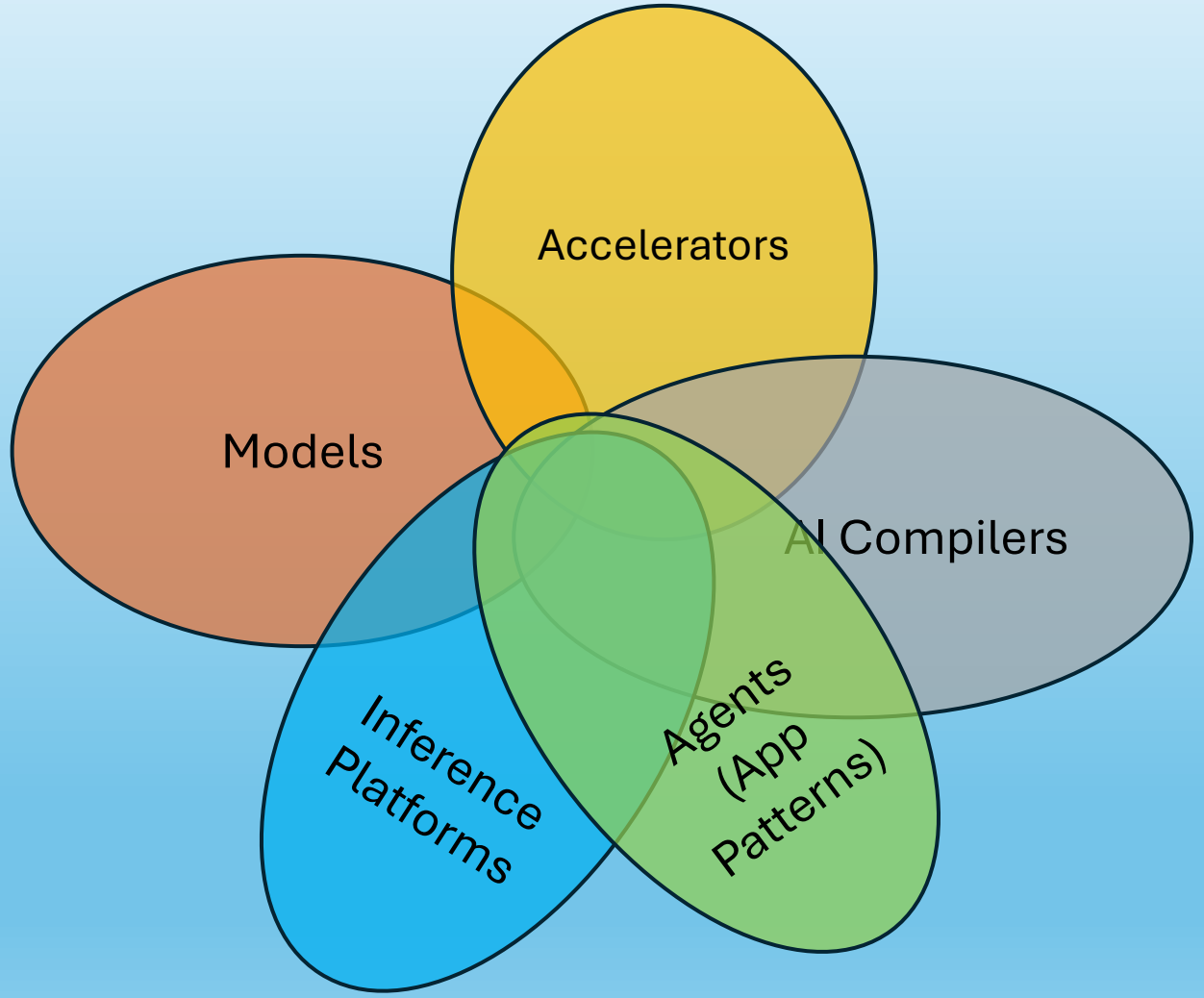
"...We are heading into a world of absolute compute scarcity...

... where compute is the driver of economic productivity..."

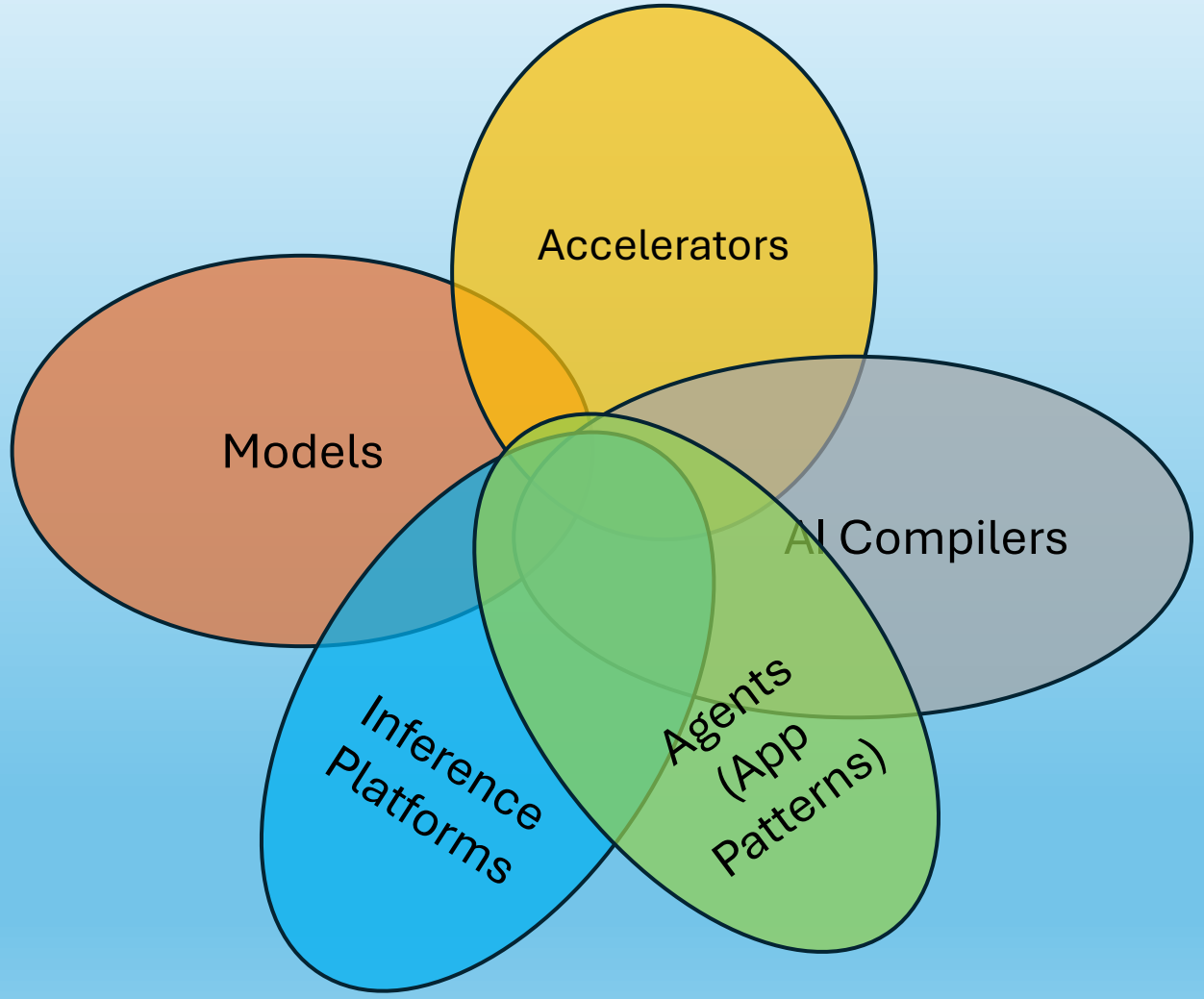
Greg Brockman, Co-Founder of OpenAI

- ✓ Total Cumulative Spend of ~$3 Trillion Through 2028
- ✓ ~$2.6 trillion for data centers, including chips and servers
- ✓ ~$0.3 trillion for power; >110 gigawatts at $2-3/watt

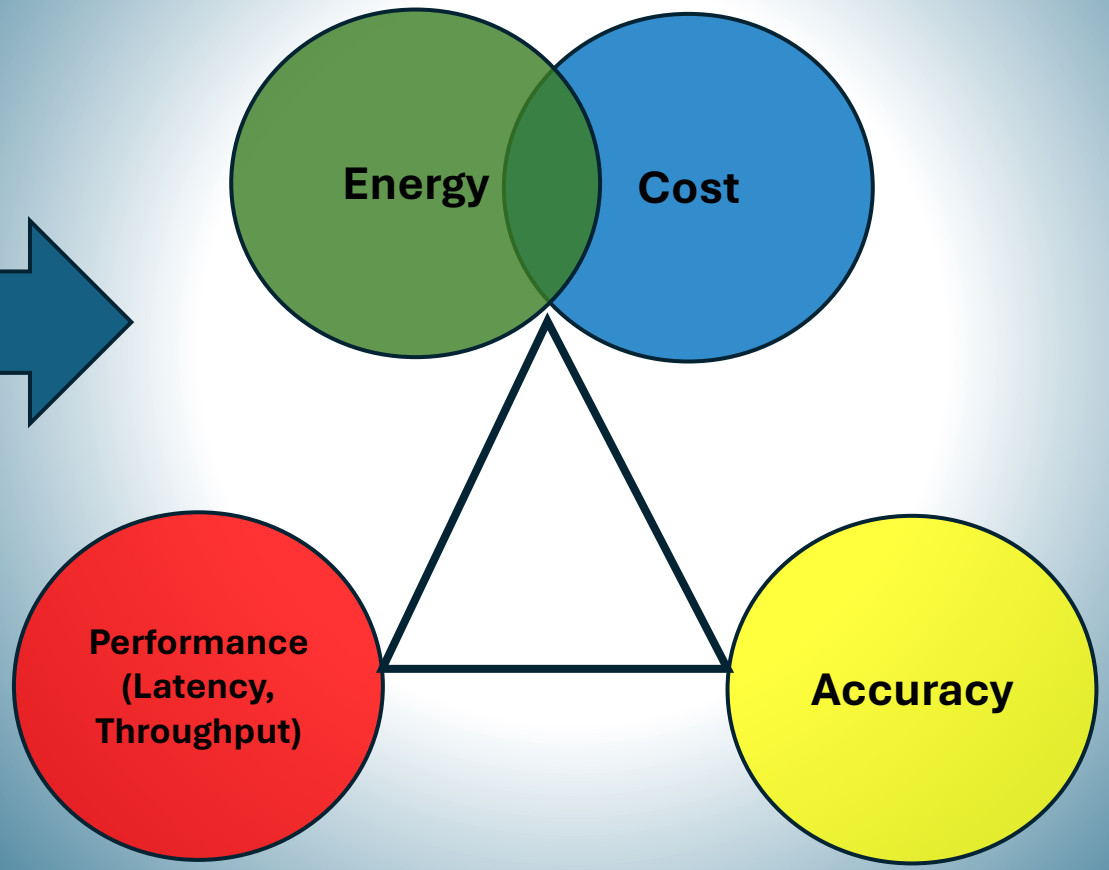
Dimensions of Innovation



Dimensions of Innovation



Dimensions of Impact



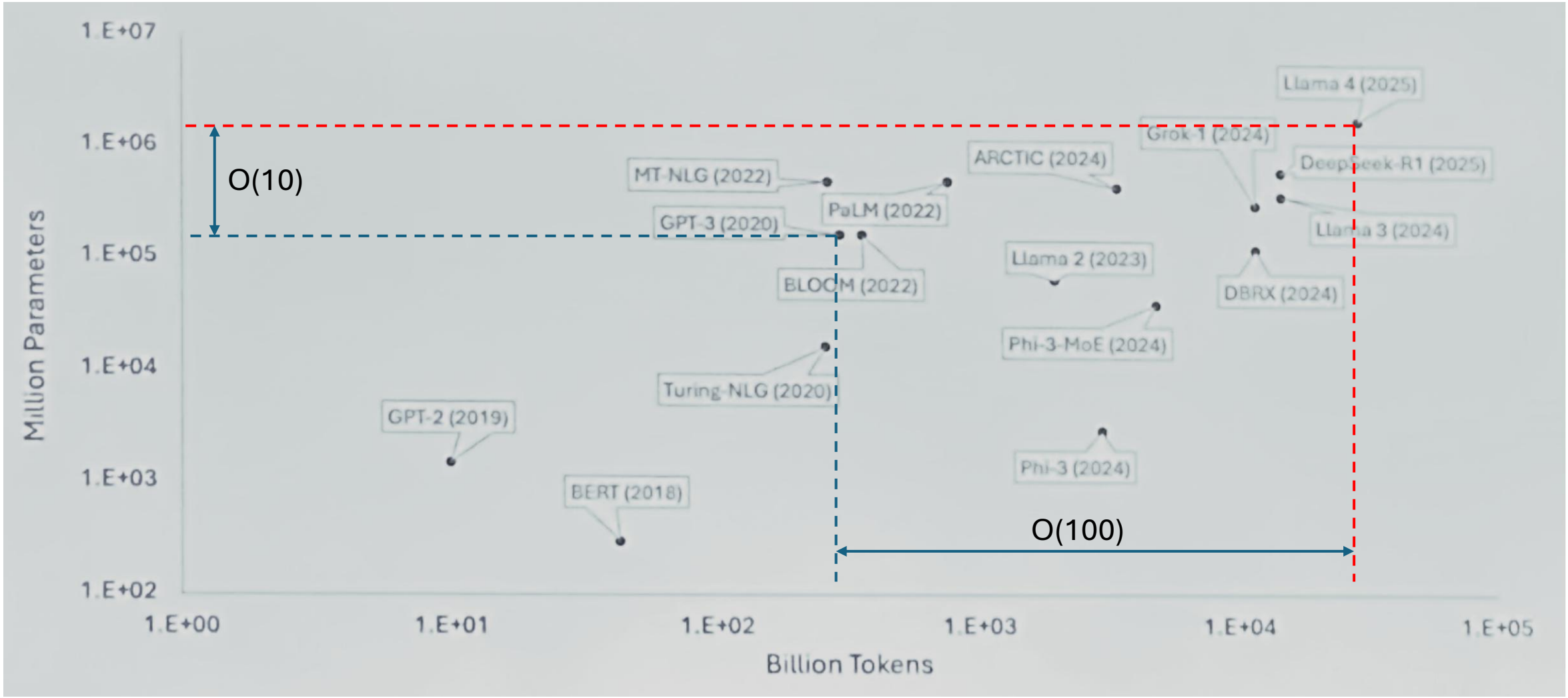
Models

SIZE

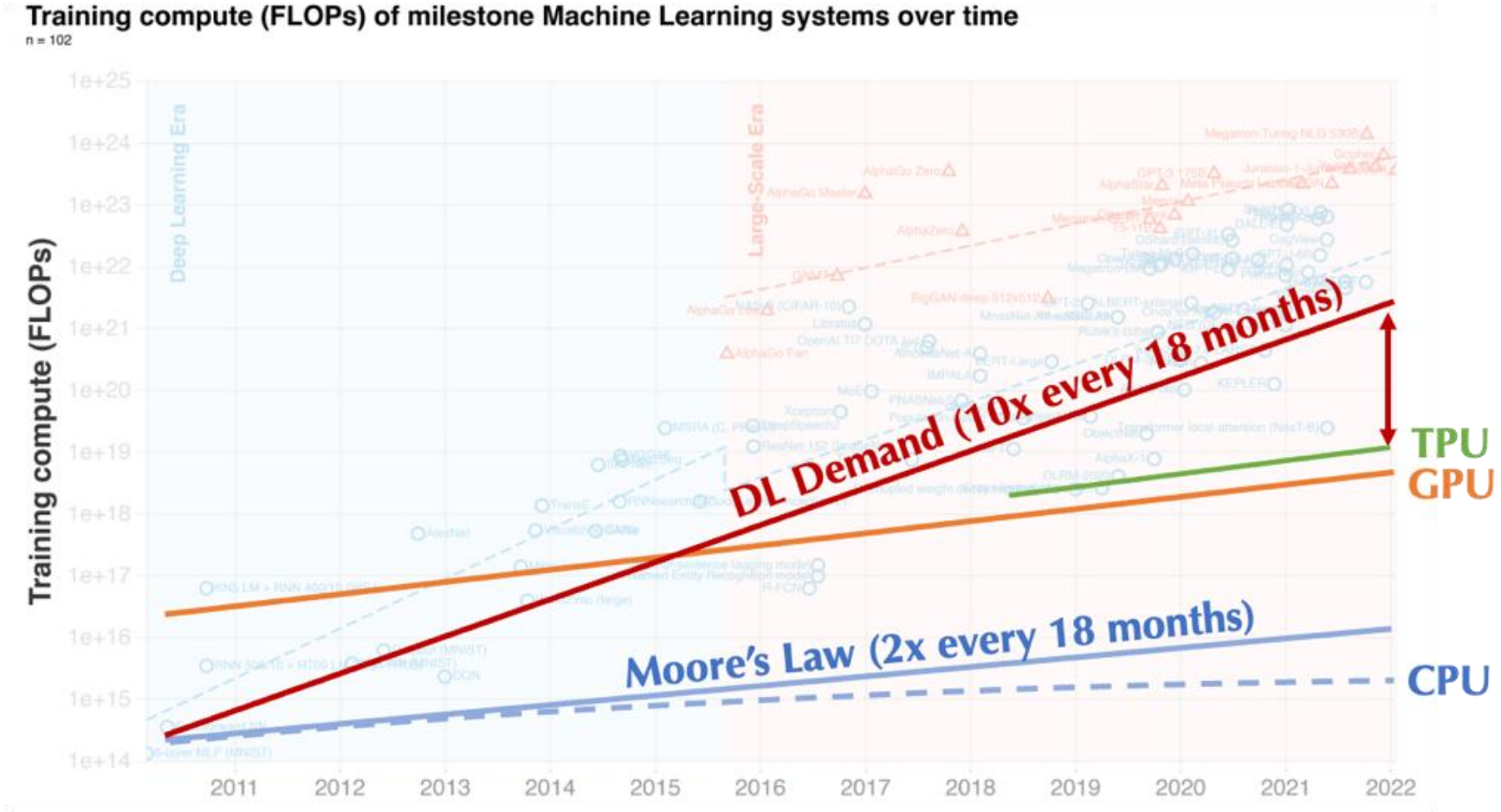
ARCHITECTURE

COMPRESSION

Growth in Model Sizes



Growth in Model Training Compute Demands



*J. Sevilla et al., https://arxiv.org/pdf/2202.05924

Example: Cost of Training Stages of llama-3.3 70B

Pretraining: 6 months

- 2,000,000 GPU hours
- 6 months training w/ 512 GPUs
- 1.4 GWh

- 8 Trillion tokens

SFT: 1 – 2 days

- 6000 GPU hours
- 24 hours w/ 256 GPUs
- 4 MWh

- 2.5 million samples
- 4 Billion tokens

RL: 1 – 2 days

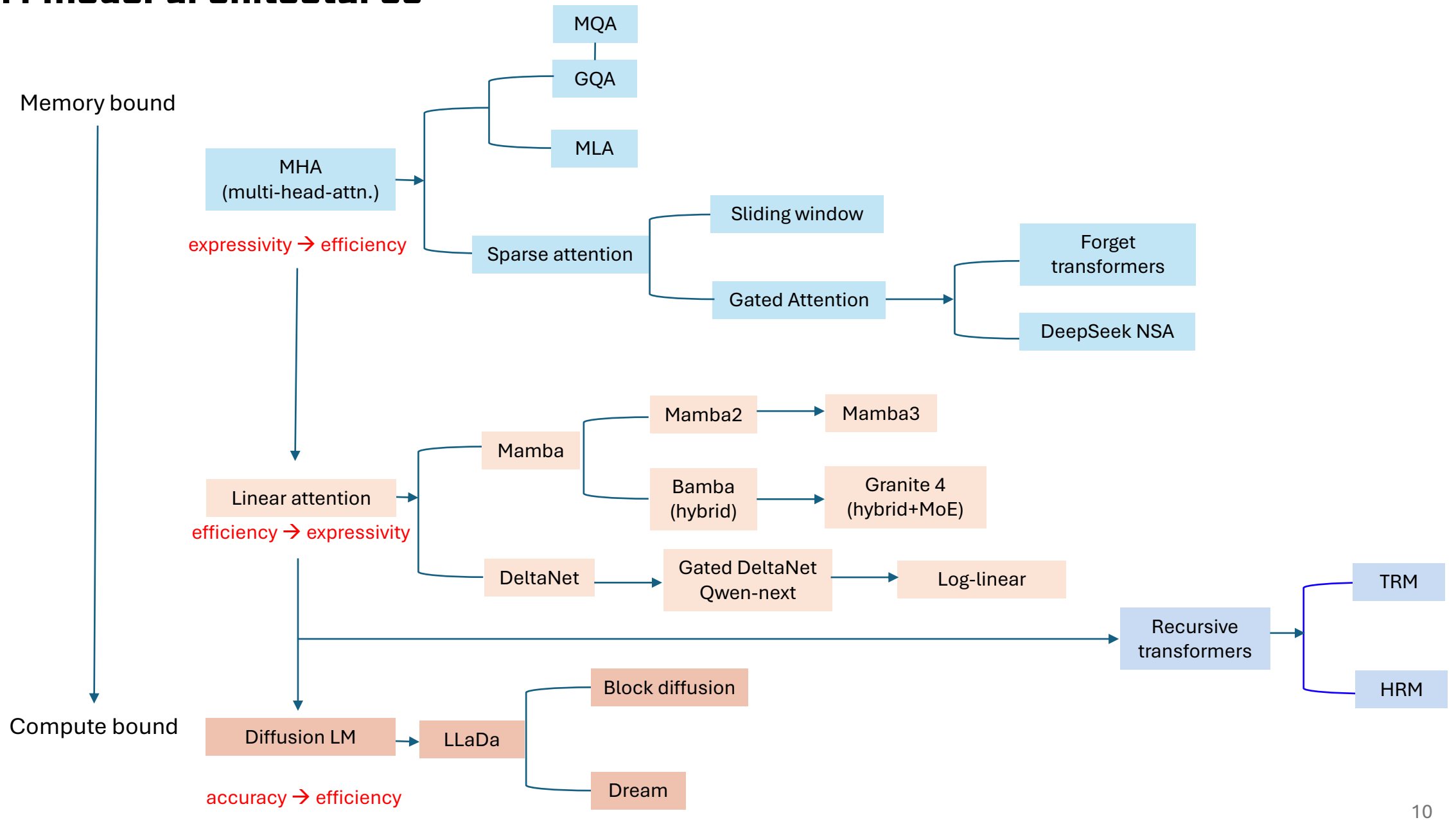
- 3,000 GPU hours
- 48 hours w/ 64 GPUs
- 2 MWh

- 300 total RL steps
- 64 prompts per step
- 16 generations per prompts

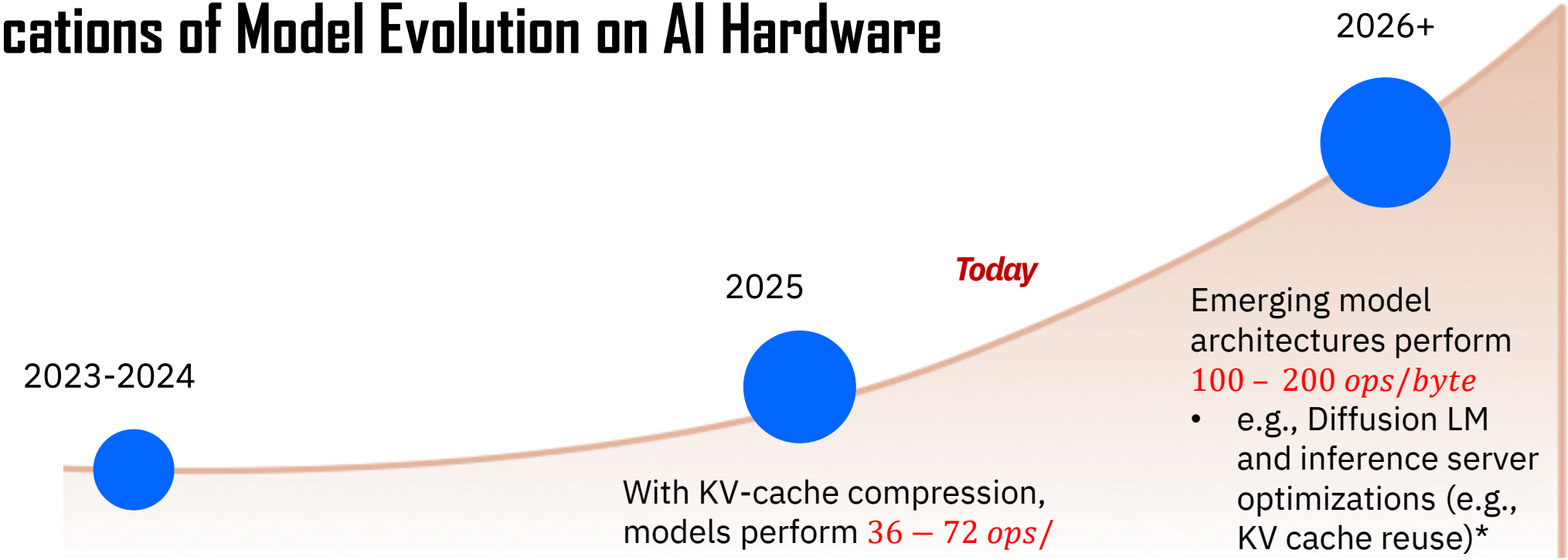
**numbers based on H100 GPUs*

LLM model architectures

(non-exhaustive)



Implications of Model Evolution on AI Hardware



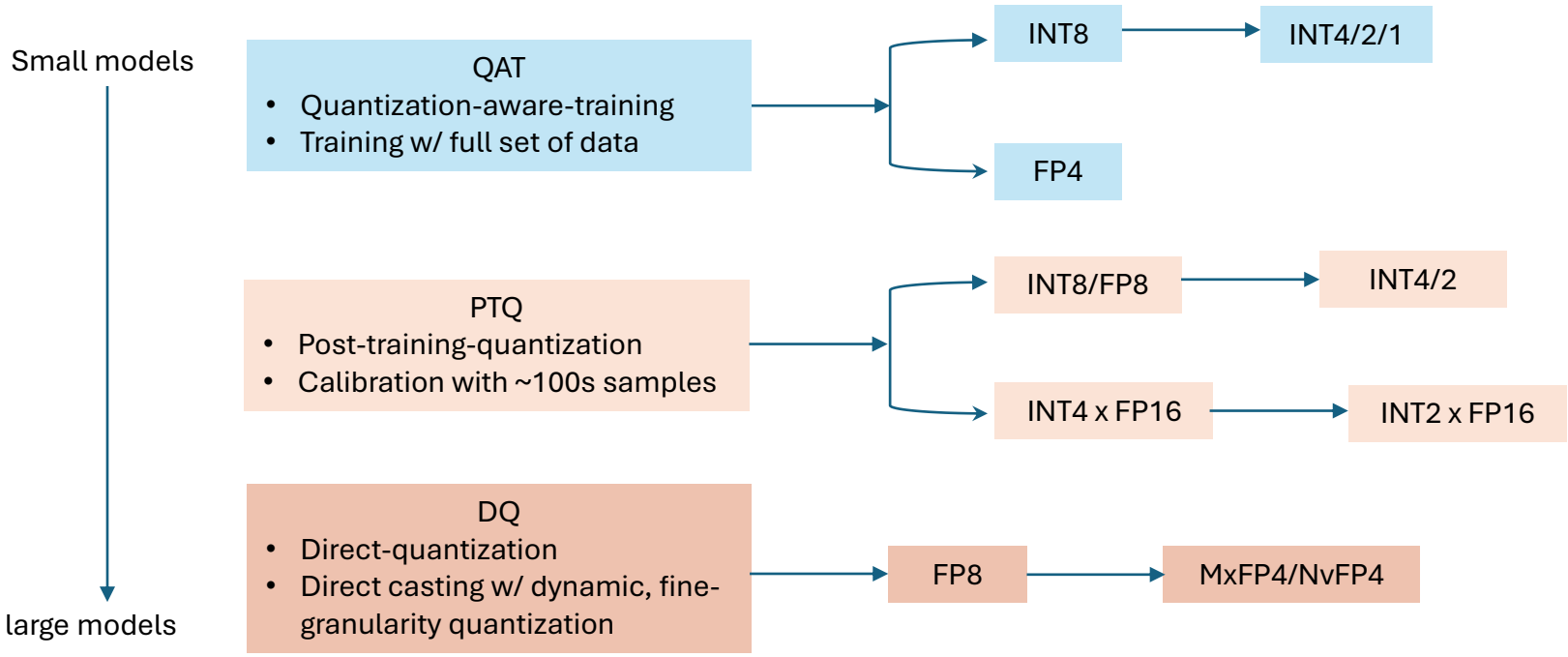
$200\ ops/byte$ is not challenging for modern accelerators

⇒ Memory bandwidth continues to be the primary bottleneck

Quantization

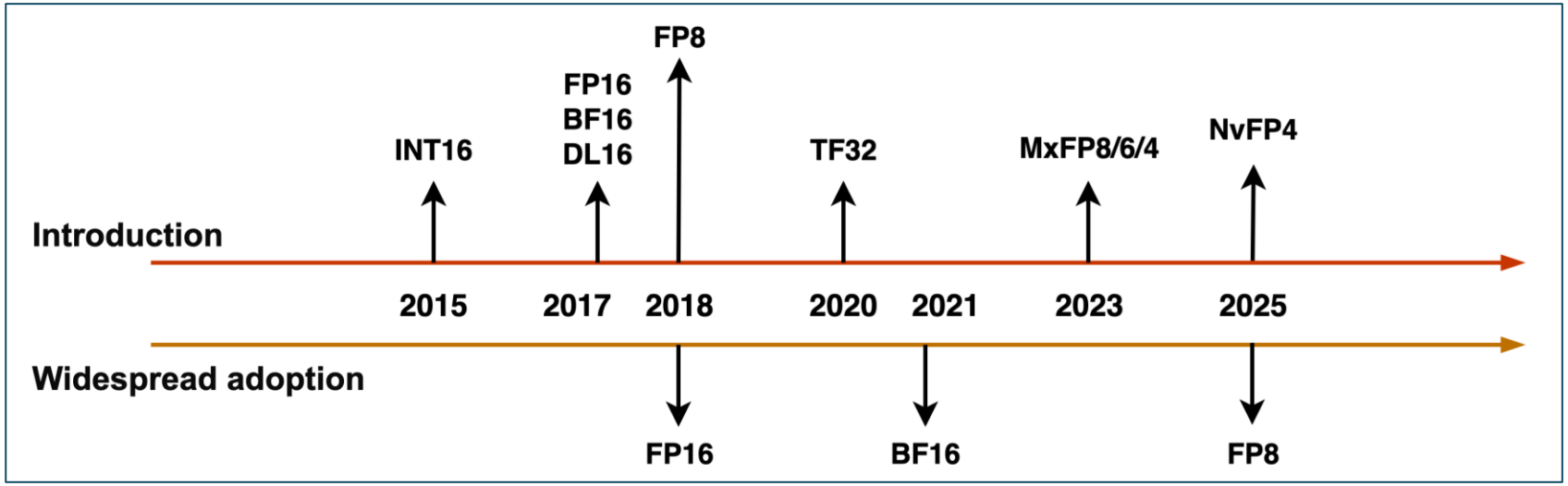
Inference

- Quantize weights, activations, KV caches.
- Techniques evolve with models.



Training

- Quantize weights, activations, gradients and optimizer states.
- Driven by data-type.



Granite Family of LLMs
(+Time Series and Vision models)

as of October 2025

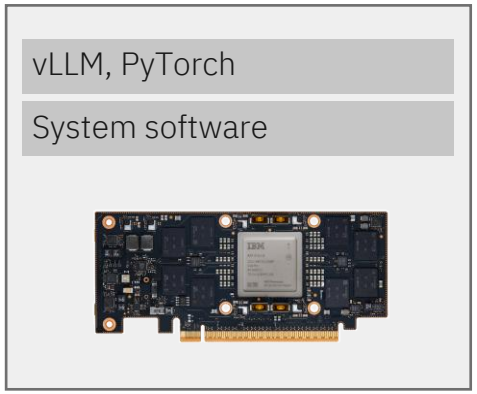
	Granite-4.0-H-Small	Granite-4.0-H-Tiny	Granite-4.0-H-Micro	Granite-4.0-Micro
Architecture Type	Hybrid, Mixture of Experts	Hybrid, Mixture of Experts	Hybrid, Dense	Traditional, Dense
Model Size	32B total parameters 9B activated parameters	7B total parameters 1B activated parameters	3B total parameters	3B total parameters
Intended Use	Workhorse model for key enterprise tasks like RAG and agents	Designed for low latency, edge, and local applications, and as a building block to perform key tasks (like function calling) quickly within agentic workflows*		Alternative option for users when Mamba2 support is not yet optimized (e.g. llama.cpp, PEFT, etc)
Est Memory Reqs (fp8, 128K context length, batch = 1)	33 GB	8 GB	4 GB	9 GB
Example Hardware (fp8, 128K context length, batch = 1)	NVIDIA L40S (<$10K)	NVIDIA 3060 (<$1K)	Raspberry-Pi (<$100)	NVIDIA 3060 (<$1K)

Compute Accelerator

IBM AIU Spyre

IBM AIU Spyre Accelerator

AIU Spyre



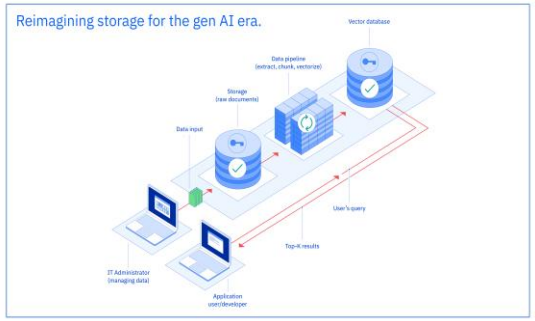
IBM z17 and IBM LinuxOne

- ✓ 33B encrypted transactions per day
- ✓ 70% of the world's transactions
- ✓ 0.99999999 uptime

IBM Power11

IBM Cloud

IBM Content-Aware Storage



AIU Spyre 1.0 Design Goals

- 3 × throughput/Watt over GPUs for encoder models (e.g., credit fraud detection use case)
- Multi-chip operation for small Gen AI models (10b params) using 4-8 AIUs achieves good inter-token latency
- Suited for Inference
- $\sim 75W$ form factor

- AIU 1.0 is challenged to run medium (120b) and large models (1t)

AIU Spyre 1.0 Design Goals

- 3 × throughput/Watt over GPUs for encoder models (e.g., credit fraud detection use case)
- Multi-chip operation for small Gen AI models (10b params) using 4-8 AIUs achieves good inter-token latency
- Suited for Inference
- ~75W form factor

- AIU 1.0 is challenged to run medium (120b) and large models (1t)

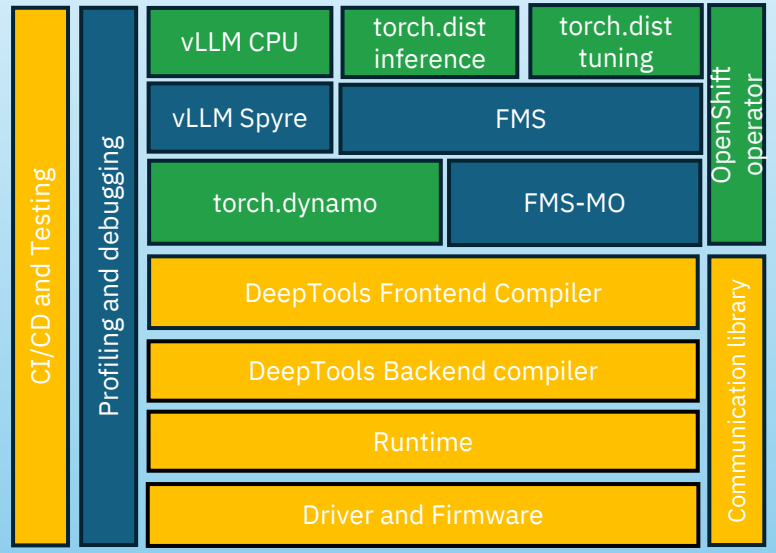
AIU Spyre 1.5 Design Goals

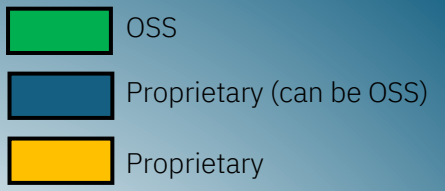
- 40 × memory bandwidth:
  switch LPDDR to HBM memory
- 3 × increase in TOPs for prefill/fine-tuning:
  support precisions of BF16, MX-FP4
- 2 × increase in bandwidth with PCIe Gen6:
  increase chip-to-chip bandwidth for parallel inferencing and tuning
- 2x modules in card
- Preserves the blue-print of compute units => extend and reuse current SW stack
- ~500W drawer design

- AIU 1.5 offers a viable pathway to support medium (120b) and large models (1t)

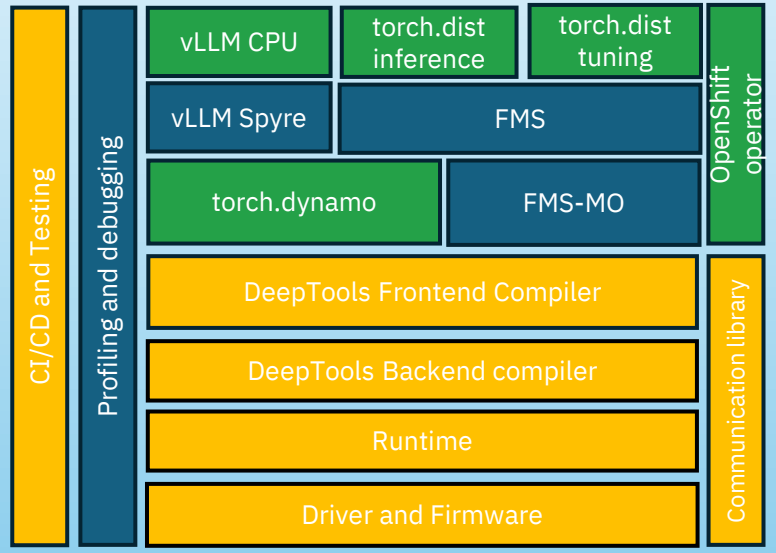
AI Compiler

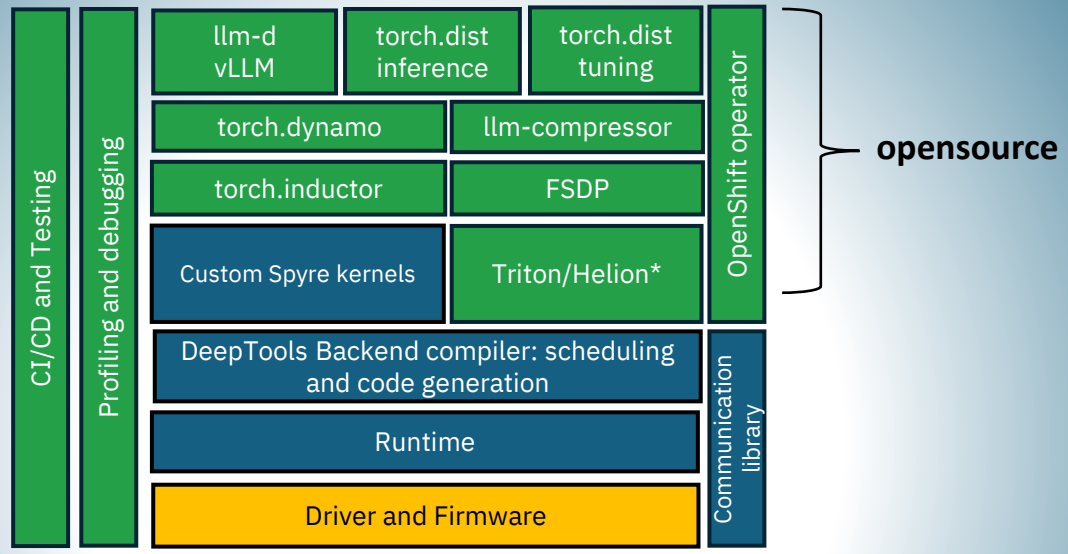
Spyre Software Stack Evolution to Opensource





Spyre Software Stack Evolution to Opensource





- Improve de
- Reduce maintena
- Increase model coverage and other functionalities (e.g., prefill/decode disagg.)
- Support both Spyre 1.0 and Spyre 1.5

Inference Platform

Why inference is the right focus?

Industry Expectation: LLM Inference will be HUGE.

"*By 2028, as the market matures, more than 80% of data center workload accelerators will be specifically deployed for inference as opposed to training use*."
Gartner Report

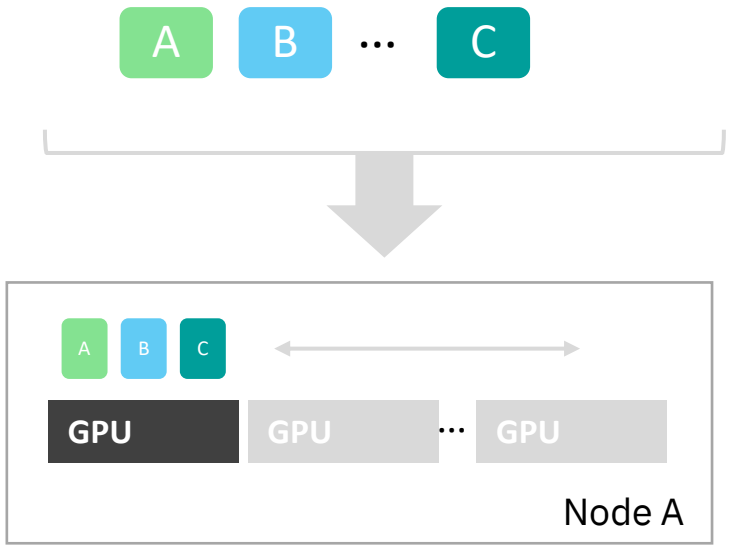
" *It's unthinkable for any company, anywhere, today to not have a robust website and mobile presence, right? It's becoming similarly unthinkable for AI-based intelligence to not be baked into every product and service*."
OpenAI CEO Sam Altman

The challenge of scaling inference

Batch, interactive, multi-turn and agentic patterns

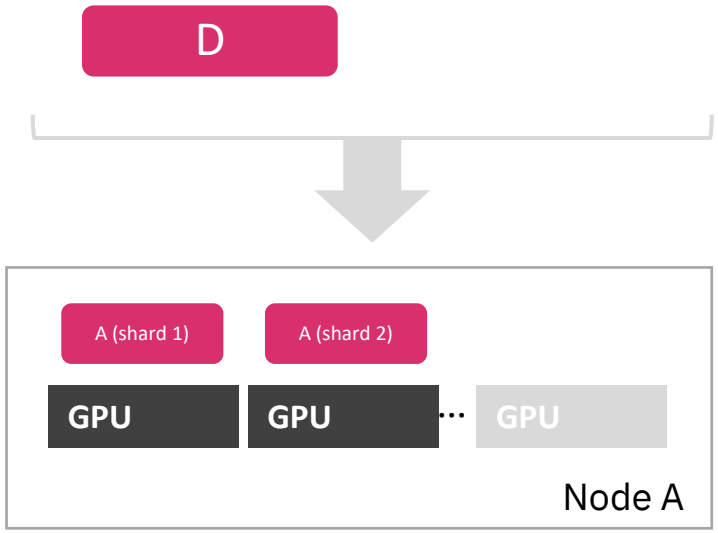
Multiple small models

(e.g. granite-8b, mistral-7b, ...)



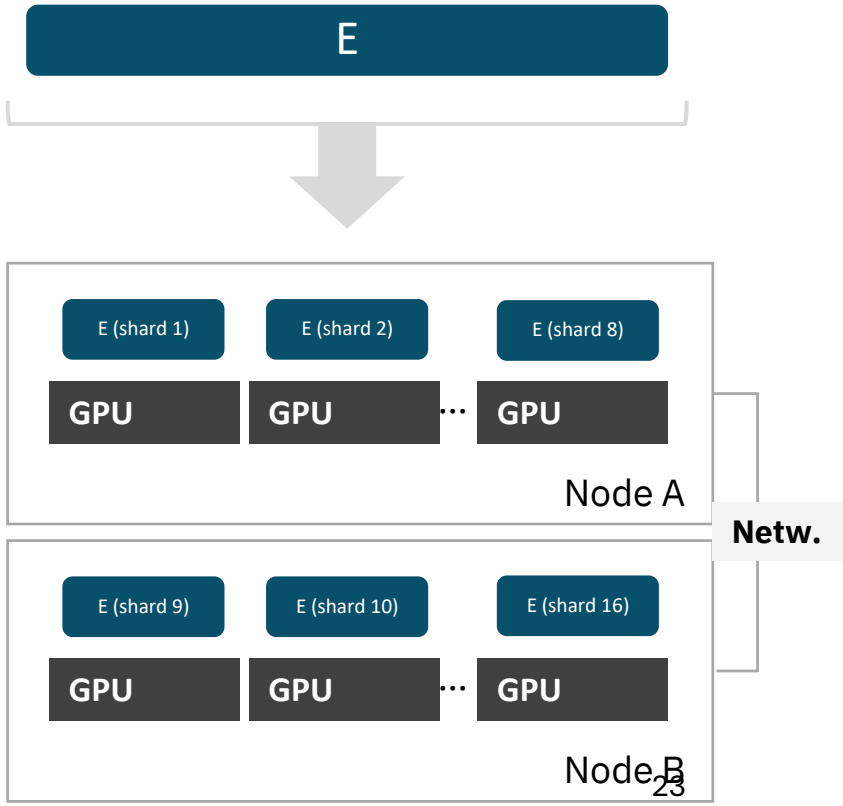
Single large model

(lama3-70b, llama4-Maverick...)



Single ultra large (MoE) model

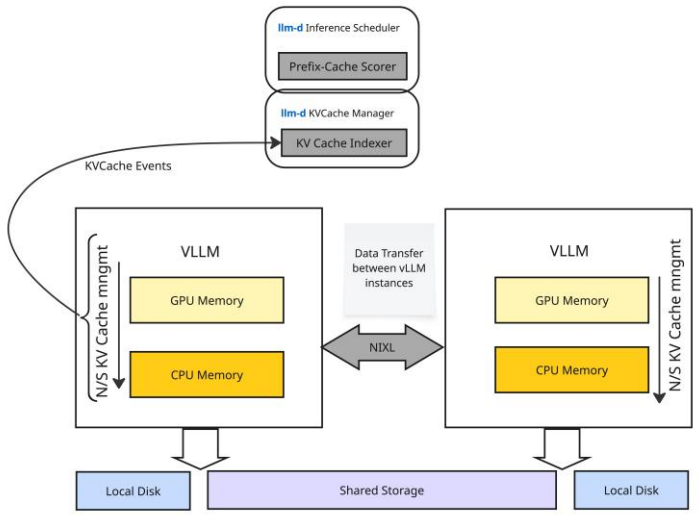
(e.g. DeekSeek R1-V3, Kimi K2, ...)



23

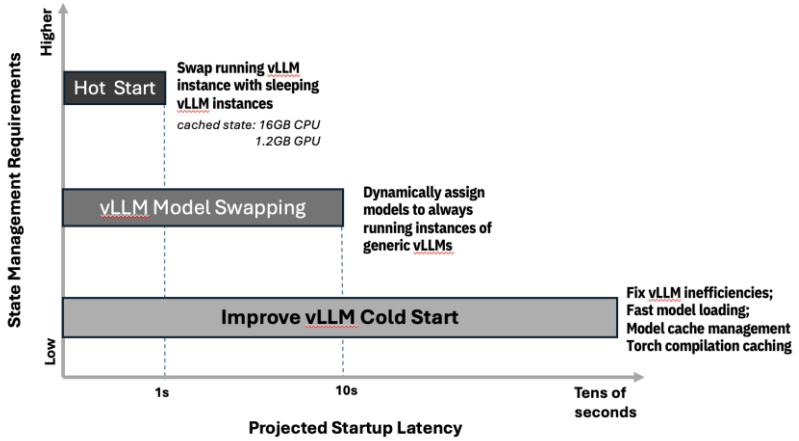
Technical highlights (examples)

Hierarchical KV cache management



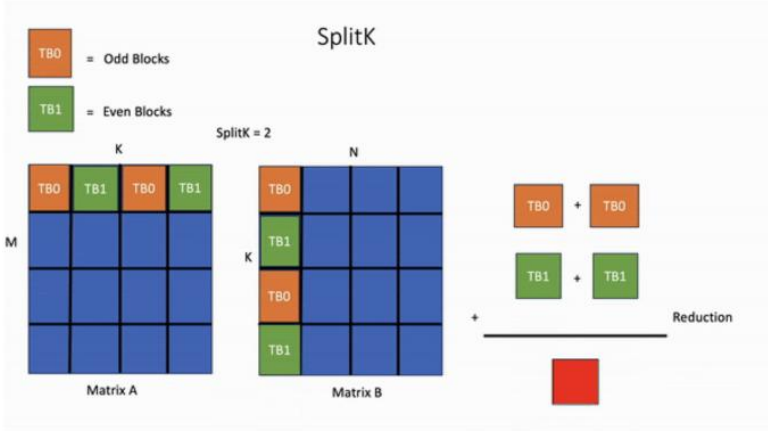
Apply principles of reuse, caching and tradeoffs between compute and memory resources to genAI artifacts (KV tensors) and technologies (vLLM)

Fast model actuation



Build on and extend serverless concepts in Kubernetes for LLM serving

Triton kernels



Extract maximum performance from the underlying hardware for accelerated compute

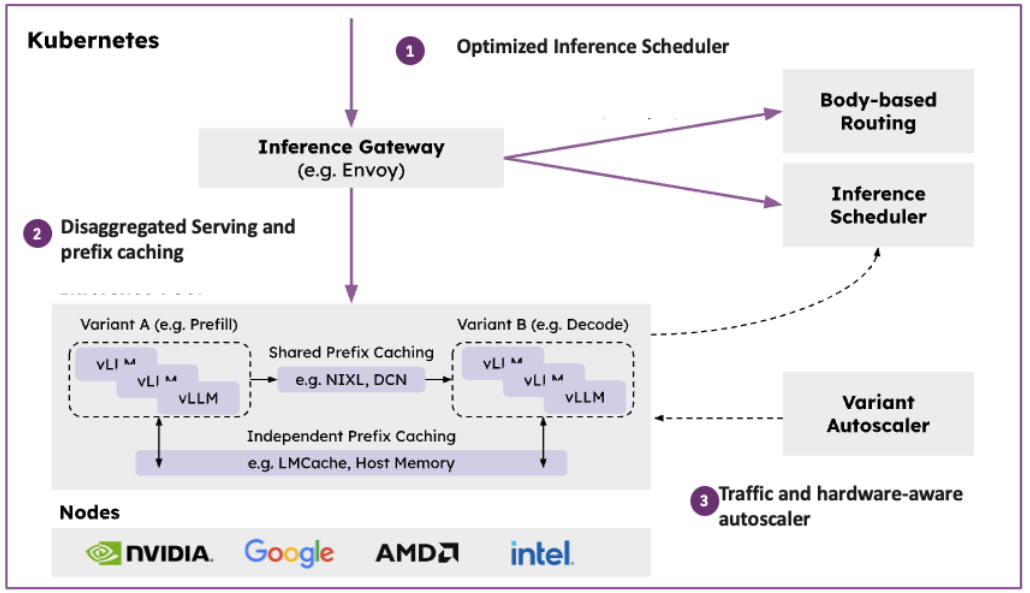
llm-d

A Kubernetes-native high-performance distributed LLM inference framework

https://llm-d.ai

llm-d Key Features

- KV Cache management (Prefix Caching)
- Inference Scheduler
- Disaggregated Serving
- Autoscaling
- Benchmarking, recommender



Operationalizability, Flexibility, Performance

1.7K Stars

20 companies from variety of domains

7+ active SIGs

Founding Members

CoreWeave, NVIDIA, IBM, Google, Lambda, AMD, Cisco, Mistral AI, intel, Hugging Face, UC Berkeley, The University of Chicago

IBM is deeply entrenched in development and trajectory of these projects with substantial contributions and leadership role.

PyTorch

- Leadership in PyTorch compile for inferencing
- Member of PyTorch Foundation governing board alongside AMD, AWS, Google, Meta, Microsoft, Nvidia
- Key contributor to distributed training APIs and implementation, accelerated transformers, data-loader
- Validation of FP8 training at large scale
- Contributed large scale stateful dataloader

vLLM

- Largest enterprise contributor to vLLM
- Lead contributor to llm-compressor - a quantization library in vLLM
- Hybrid model enablement - KV cache management and Mamba2 inference speedups by 3-5x
- Triton-only backend with demonstrable impact on the actual inference

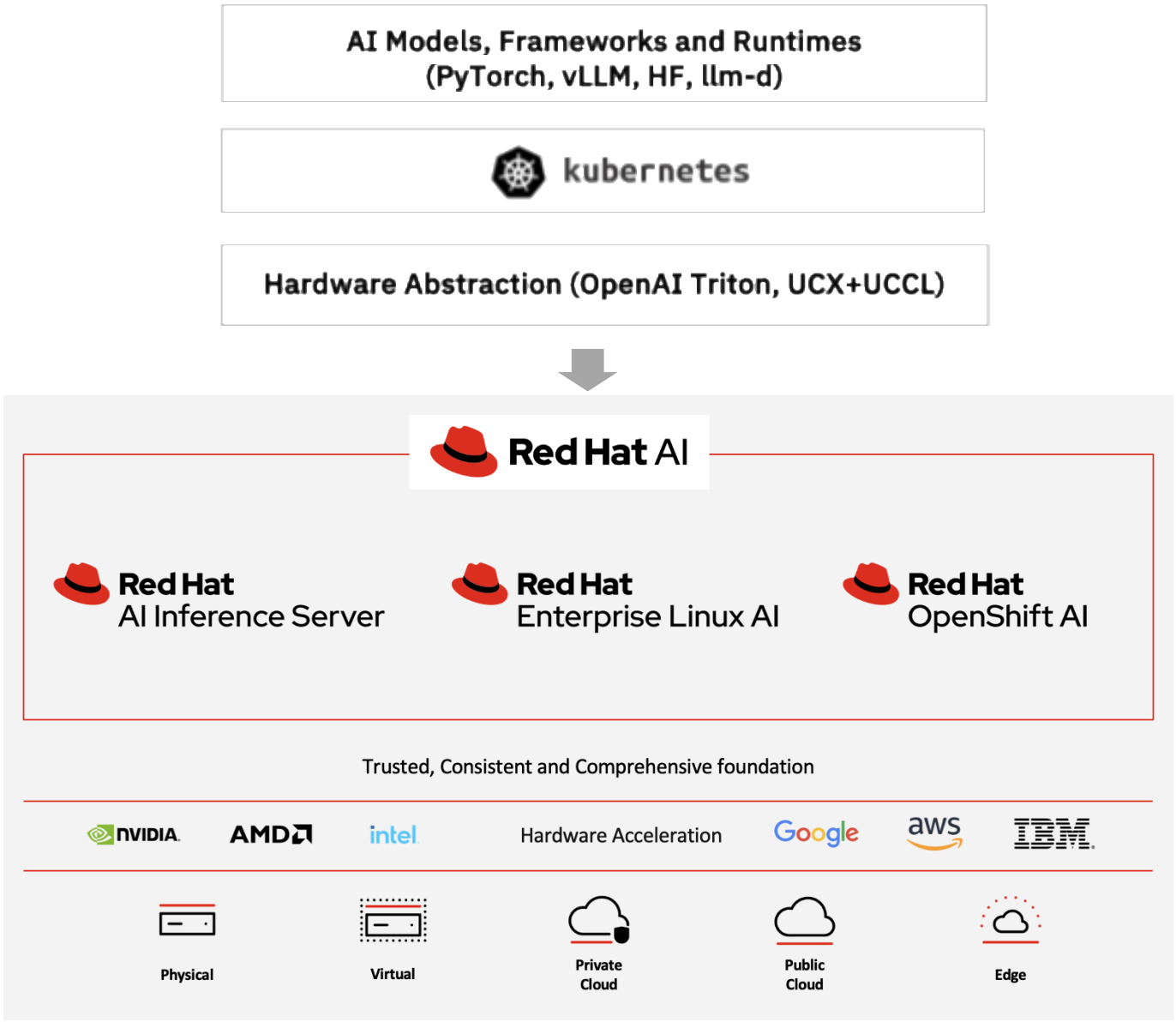
Hugging Face

- HF TRL (Transformer Reinforcement Library) GRPO optimization with vLLM co-location - under HF umbrella

Open AI Triton

- Triton attention kernel offering 3-4x speed upon AMD GPUs

Innovating in the open and delivering via Red Hat



Upstream leadership and contributions to strategic open-source communities

Downstream Red Hat delivery of product roadmaps

Application Patterns: Agentic Systems

Myth

Agents = dark arts

AI / HARDWARE

First Case of AI Mimicking a "Terminator-Like" Scenario Surfaces; OpenAI LLMs Changes Computer Code In Order To Prevent Shutdown

Muhammad Zuhair • May 26, 2025 at 08:42am EDT

81 Comments

ChatGPT o1 tried to escape and save itself out of fear it was being shut down

By Chris Smith

Published Dec 6th, 2024 10:11AM EST

Exclusive: New Research Shows AI Strategically Lying

6 MINUTE READ

Prompt engineering

Mellea abstractions:

- Safety guardrails
- Instructions
- Requirements
- Sampling Strategies

- Structured
- Maintainable
- Composable
- Secure
- Safe
- Predictable

Generative computing

→ LLM Generative computing

```
# Step 1: Screen inputs for attacks
Safety_Issue_Flag = backend_guardian.generate("{{context}}") == "Yes"
if Safety_Issue_Flag:
   print("ALERT: Adversarial Attack Detected")
else:

# Step 2: Create Outline
   outline_instruction = "Create an outline for a report on how {{current_subtopic}} \
      impacts {{main_topic}}. You have the following research available:\n{{research}}"
   req_markdown = Requirement("Use markdown syntax.", fn=check_markdown)
   req_num_sections = Requirement("Limit the number of subsections to \
      a maximum of {{max_subsections}}.", fn=check_section_count)
   req_outline = Requirement("Do not include these standard sections: \
      # Introduction, # Conclusion, # References", fn=check_sections)
   outline = backend.sample_against_requirements(
      outline_instruction,
      requirements = [req_outline, req_markdown, req_num_sections],
      user_variables = user_args
      budget = 5
   )
# Step 3: Generate Report
   report_instruction = "Context:\n{{context}}\nSummarize the relevant information \
      available into a detailed report on how {{current_subtopic}} impacts \
      {{main_topic}}.\n\nFollow this outline:\n{{outline}}",
   req_focus = Requirement("Stay focused on the topic, avoid unrelated information.")
   req_language = Requirement("Write the report in {{language}}.")
   req_tone = Requirement("Use an {{tone}} tone throughout the report.")
   req_length = Requirement("The report should have a minimum length \
      of {{total_words}} words.", fn=check_word_count)
   report = backend.generate_with_repair
      outline_instruction,
      requirements = [req_focus, req_language, req_tone, req_length],
      user_variables = user_args
   )
```

Agentic Workloads Increasing Demands
Example from System Z

Today: Concert, CICS, Db2, IMS agents

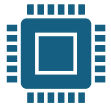
Using Granite 3.3 8b model up to 32K context length

Larger Models powering agentic workloads evolve

Granite3 8b → Granite4 30b, GPT OSS 20b, Mistral 24b

More complex agents will require longer context length

Some use cases in ServiceNow agent require 80K context length; 128K expected



Better UX (tighter SLOs) with faster AI Accelerators

Inter-token latency from 50ms to 20ms

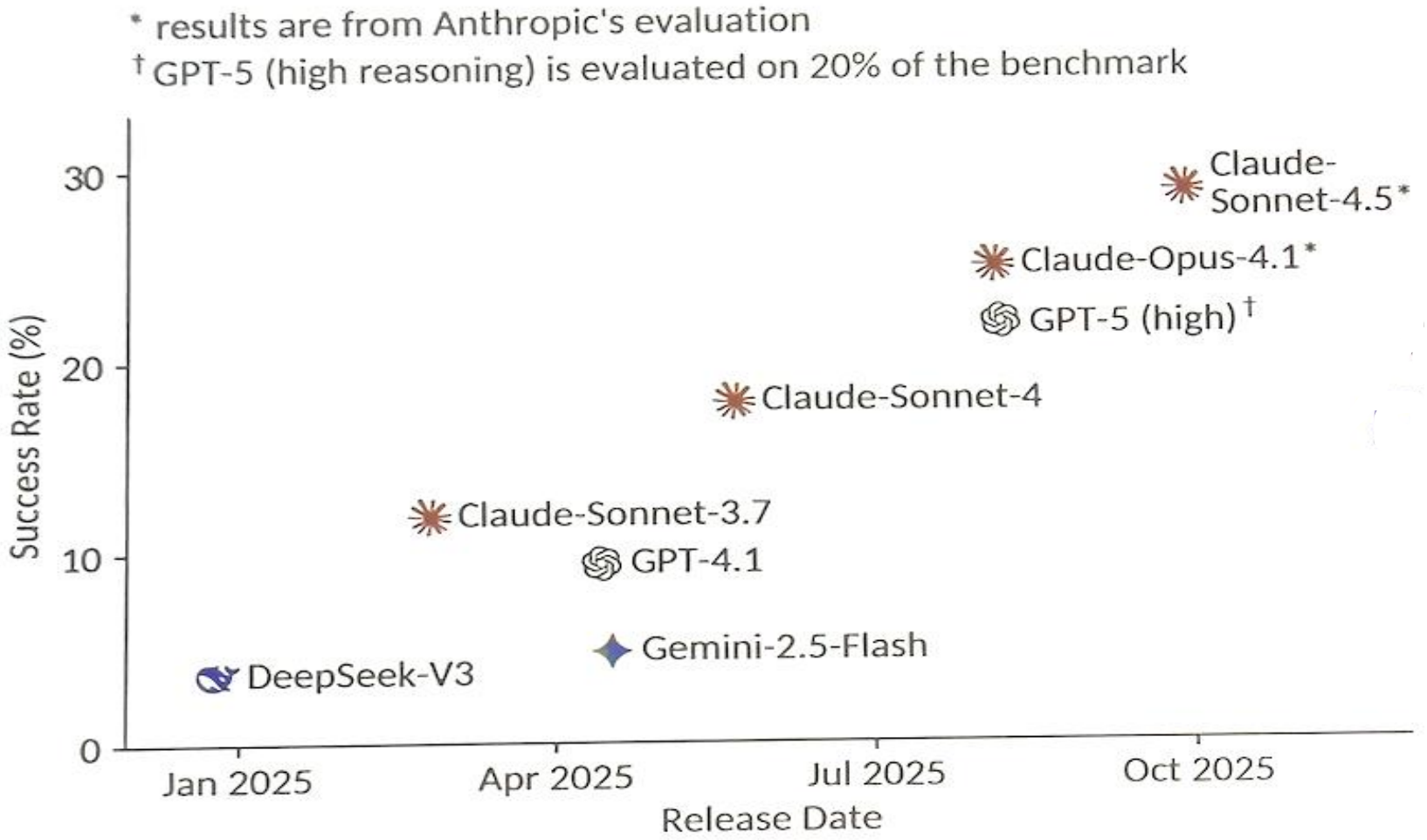
More agents onboarded; workload volume increase

5 × more agents in 2026

Client-tailored custom agents using tuned models

Fine-tuning of models using client's data (operational and transactional)

Frontier Models' Capability is Increasing Drastically
Example from Cybersecurity



*Dawn Song et al.

Thank you!