REGIONAL SDC 25

BY Developers FOR Developers

Regional SDC Denver
April 30, 2025

Lions and tigers and MoEs, oh my!

Credit:MGM

Craig W. Carlson
craig.carlson@amd.com

SNIA®

About the speaker

- Storage and Network Architect at AMD
- Member SNIA Technical Council
- NVM Express board member
- Many years of storage and networking standards experience (I'm old!)

AMD

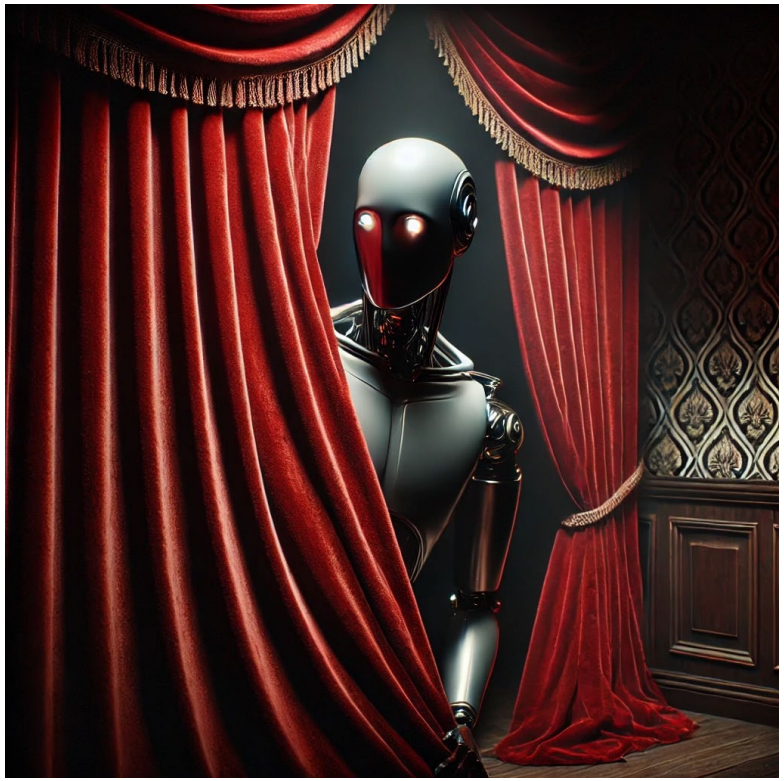
REGIONAL SDC

To the Emerald City (?)

- Are we marching towards the Emerald City?
  - And, Is the man still behind the curtain?
- The last year has been a wild ride in the AI arena
  - Proliferation of models and techniques
  - Increase in accuracy of model output (maybe?)
- Finally, how does all of this impact storage?

Models and AI providers galore (not a complete list!)

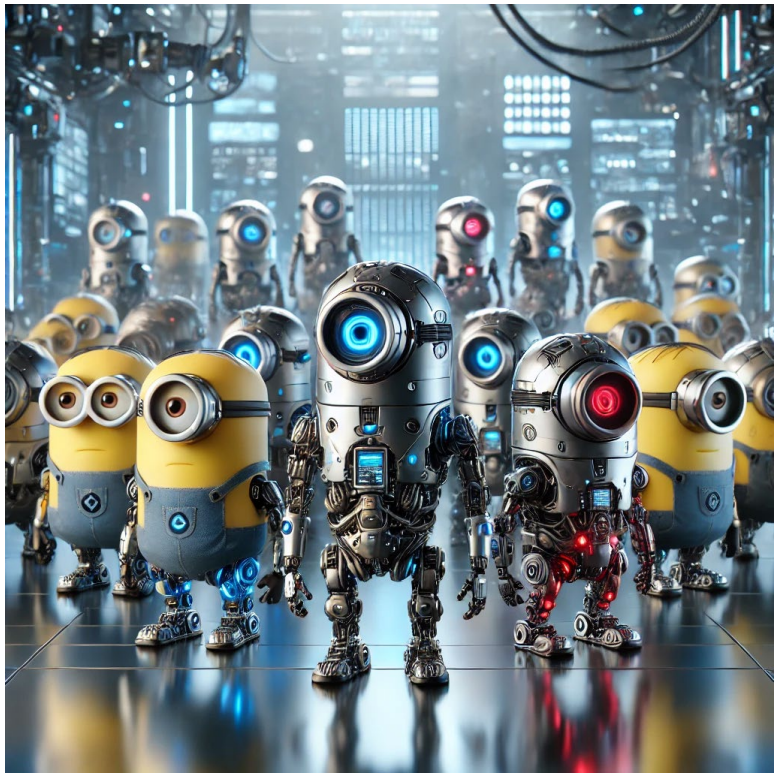
perplexity
Claude
IIElevenLabs
Gemini
Amazon Nova
LLaMA
Midjourney
sakana.ai
Grok
Leonardo.Ai
DeepAI
ChatGPT
BLACKBOX AI
THINKING MACHINES
Copilot
cohere
MISTRAL AI_
deepseek
World Labs



REGIONAL SDC

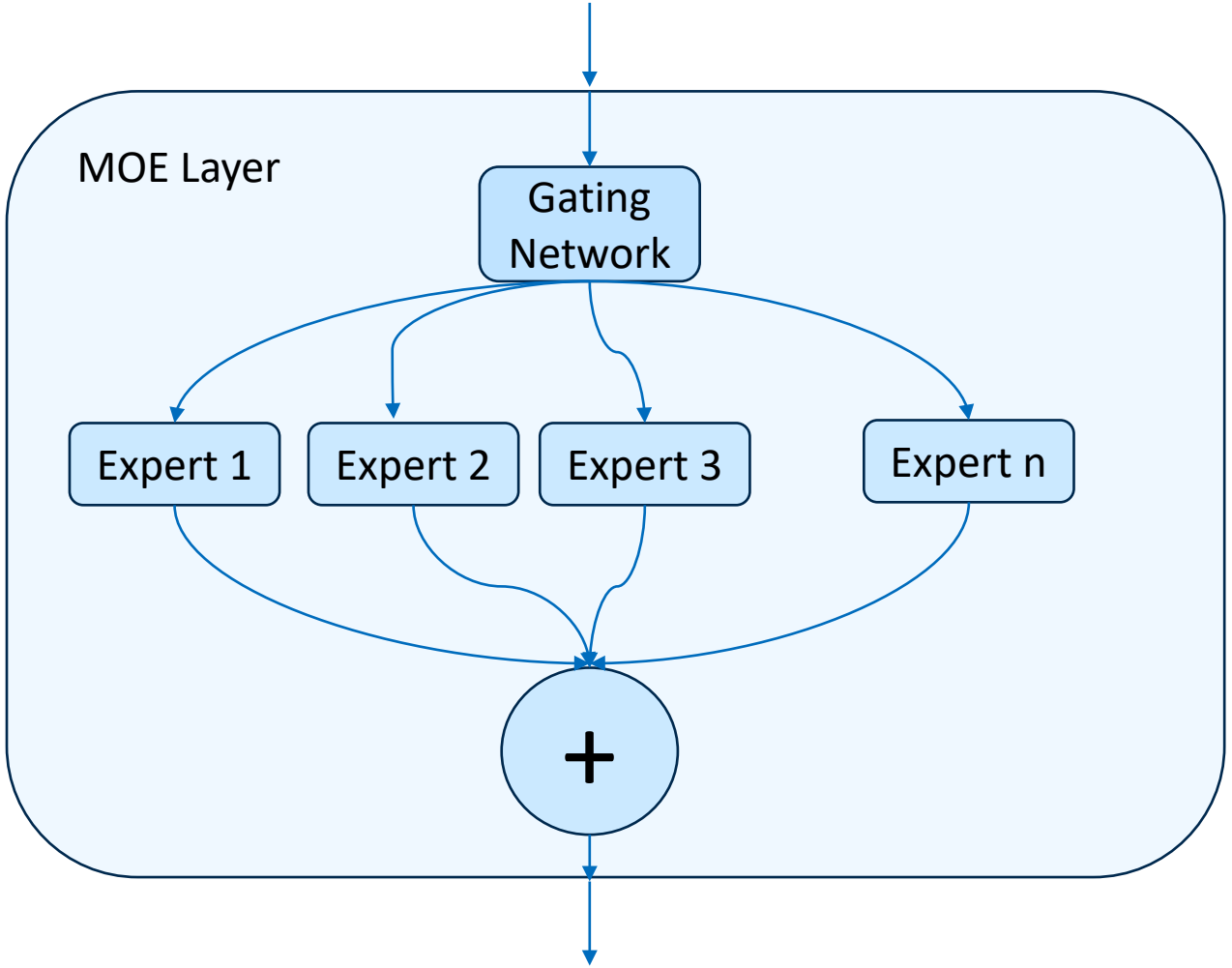
Techniques abound

- Some newer techniques
  - MoEs
  - RAG
  - AI Agents (Agentic)
  - SLMs
  - RL



MoEs (Mixture of Experts)

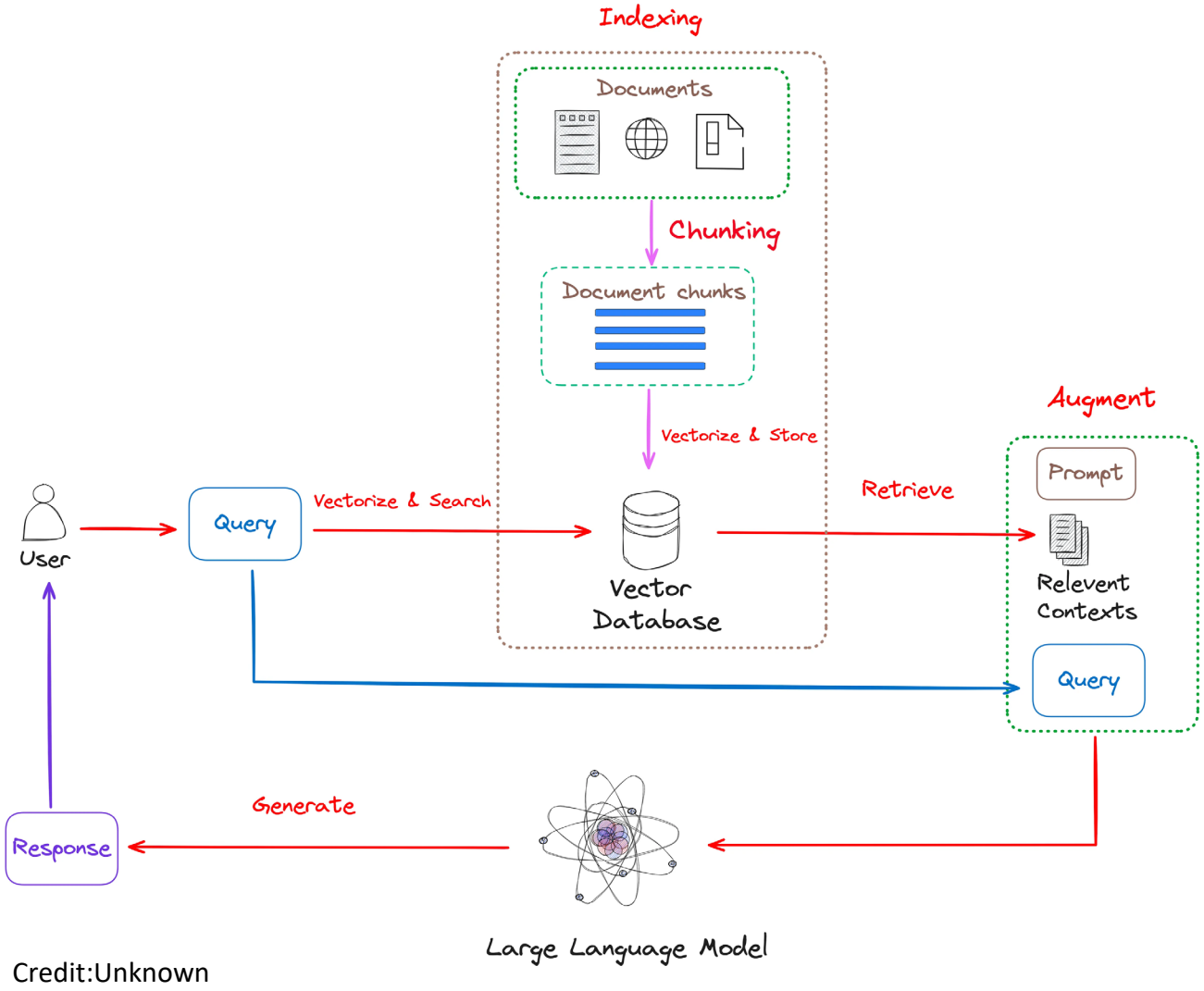
- A type of model that is divided into sub-models
  - The sub-models are trained to be an expert on a specific topic
  - Reduces computation cost during both training and inference
  - Mixture of experts offer a means to address the tradeoff between the greater capacity of larger models and the greater efficiency of smaller models



REGIONAL SDC

RAG (Retrieval-Augmented Generation)

- Augments LLM with external (more up to date) data
- Leads to more accurate, and timely, responses
- Prevents the need to constantly retrain models

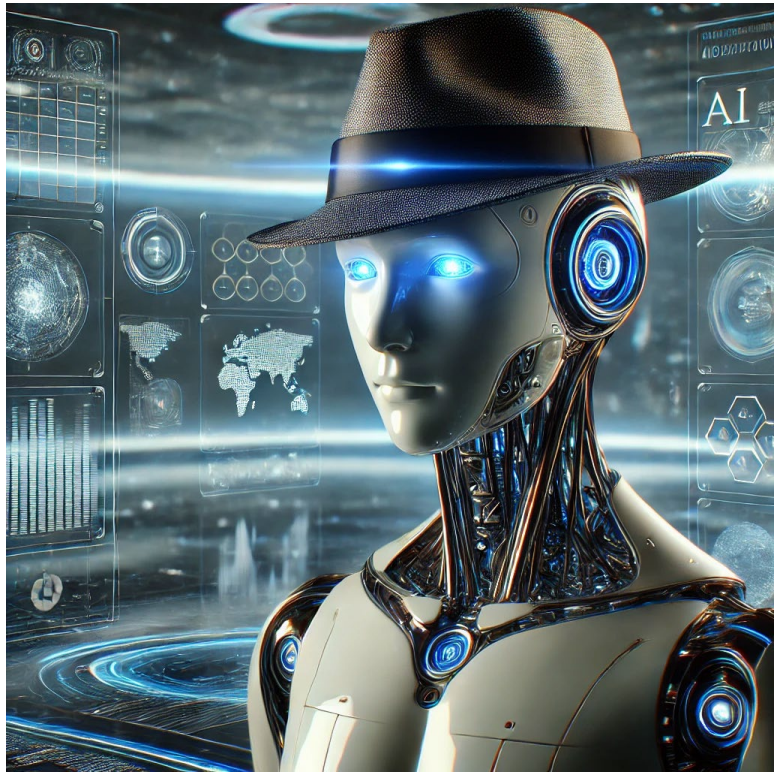


Credit:Unknown

REGIONAL SDC

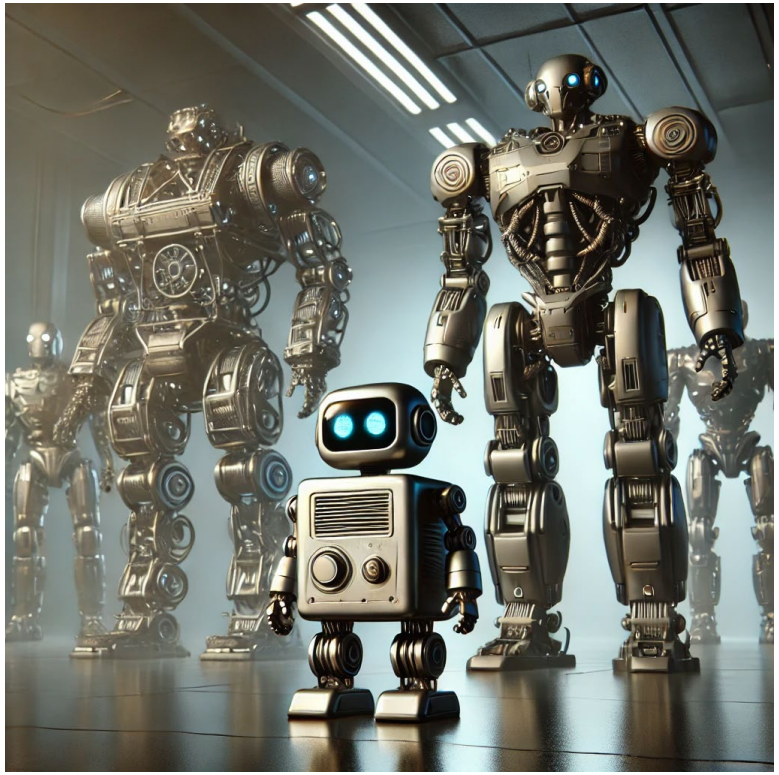
AI Agents (or Agentic models)

- A model that is capable of autonomously performing a task for a user
- Examples of uses include:
  - Code generation
  - IT automation
  - Conversational assistants
  - Customer interaction



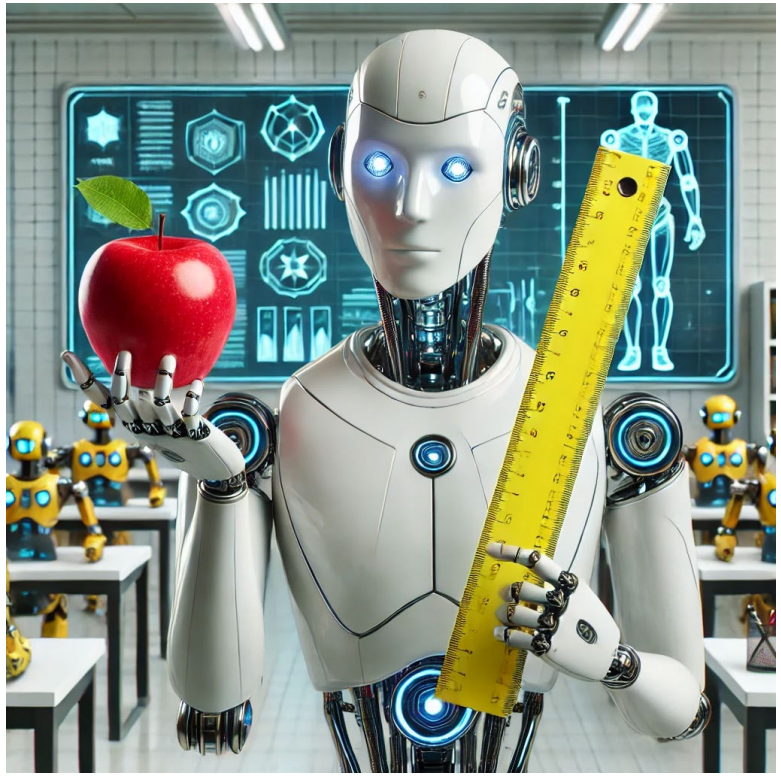
SLM (Small Language Models)

- Built with simplified versions of neural networks used for LLMs
  - 100 million vs trillions of parameters
- Trained on more specialized topics
  - Save power and time in training
- Used for specific tasks (agents)
  - Easier on-device inference



RL (Reinforcement Learning)

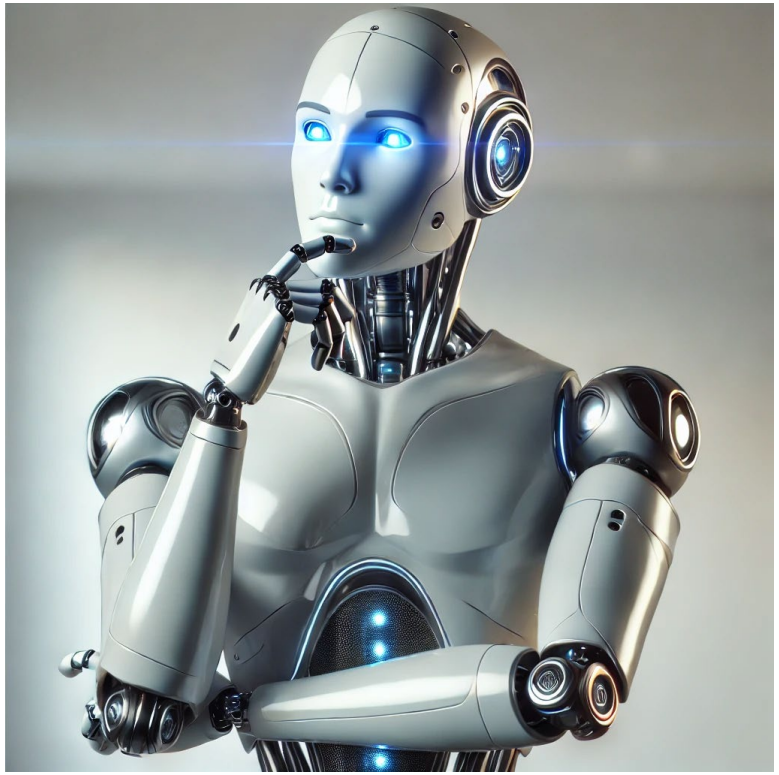
- Use a trial-and-error system, with rewards, to increase model accuracy
  - Idea is to mimic the punishment/reward learning process common in real world human learning
- Benefits of this process include
  - Excels in complex environments
  - Requires less supervision
  - Optimizes for long term goals
- Uses
  - Marketing personalization
  - System optimization
  - Financial predictions



REGIONAL SDC

The data problem

- This proliferation of models is driving data and storage challenges
  - Models are increasing in size/complexity and no longer fit in GPU memory (or host system memory)
  - GPU and data center resources are more in demand
  - Increasing data center sizes are pushing power consumption through the roof
    - Thermal and power density going up
- While many of the newer techniques attempt to mitigate this, the increasing size and complexity does present storage challenges

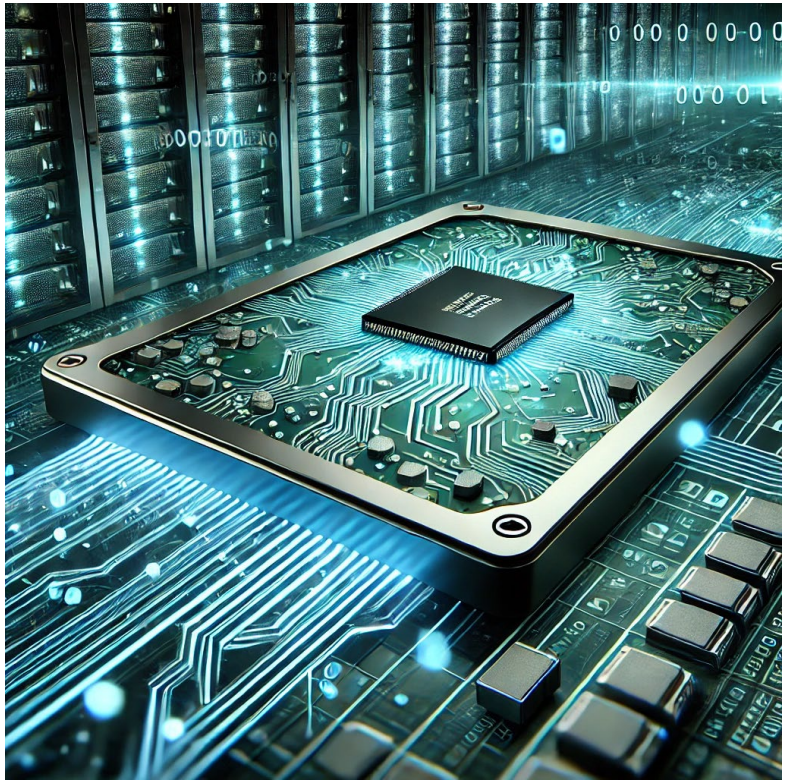


Storage Trends and Impacts

- Accelerator accessed storage
  - CPU or Accelerator directed
- Device initiated I/O
- Content Aware Storage (CAS)
- Taking a closer look at storage requirements for AI
  - Many times systems are waiting on storage – How can we fix that

Accelerator accessed storage

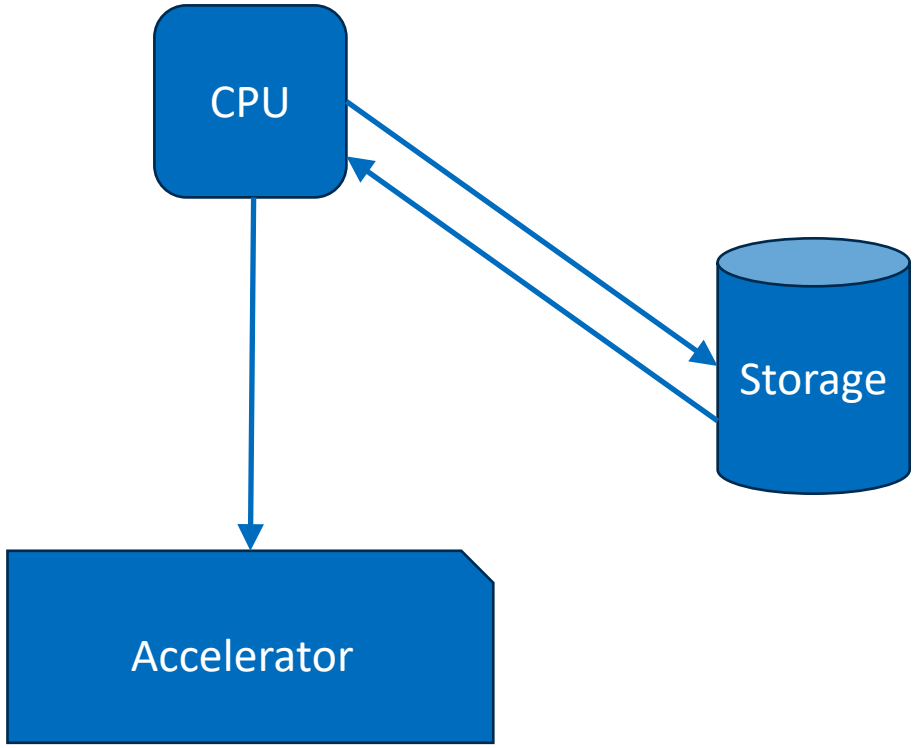
- Three flavors
  - CPU initiated
  - CPU initiated, data directed to Accelerator
  - Device initiated



CPU Initiated

- Traditional model
- CPU initiates I/O with Storage
- Data response is sent to CPU memory
- Data is then copied to Accelerator memory

- Requires extra data copy

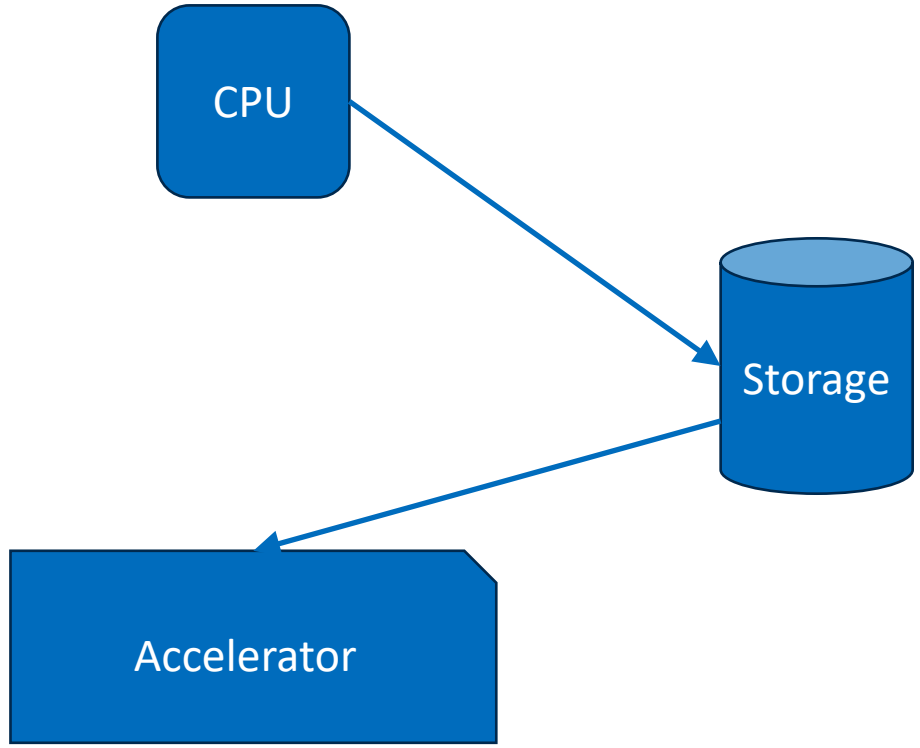


REGIONAL SDC

CPU initiated, data directed to Accelerator

- CPU initiates I/O with Storage
- Data response is sent to Accelerator memory

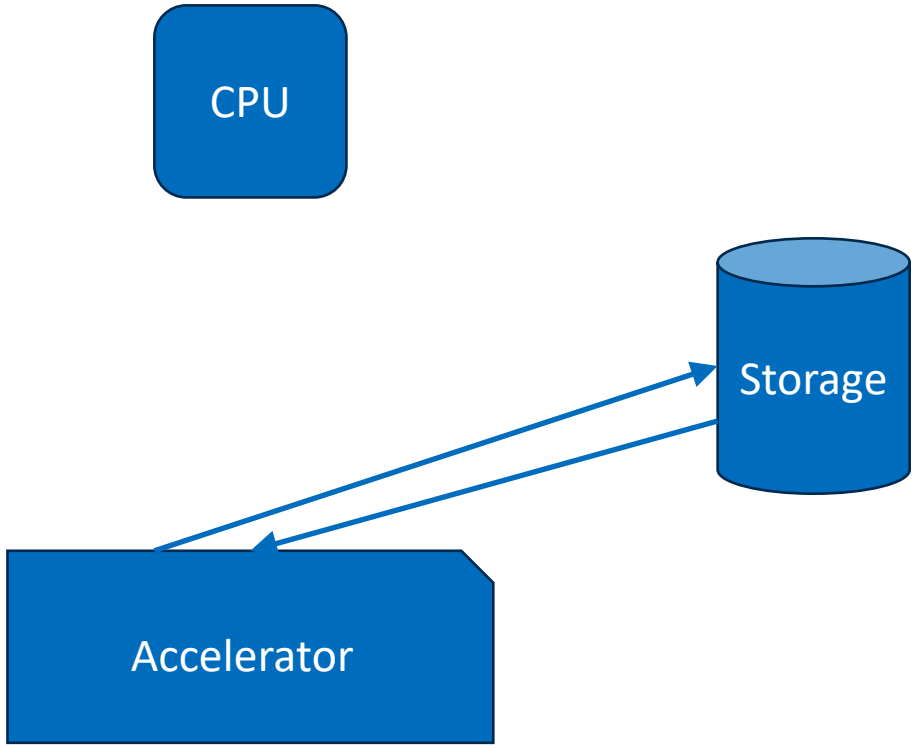
- Removes extra data copy
- Still requires CPU to handle I/O command processing
- This concept has been around for a while



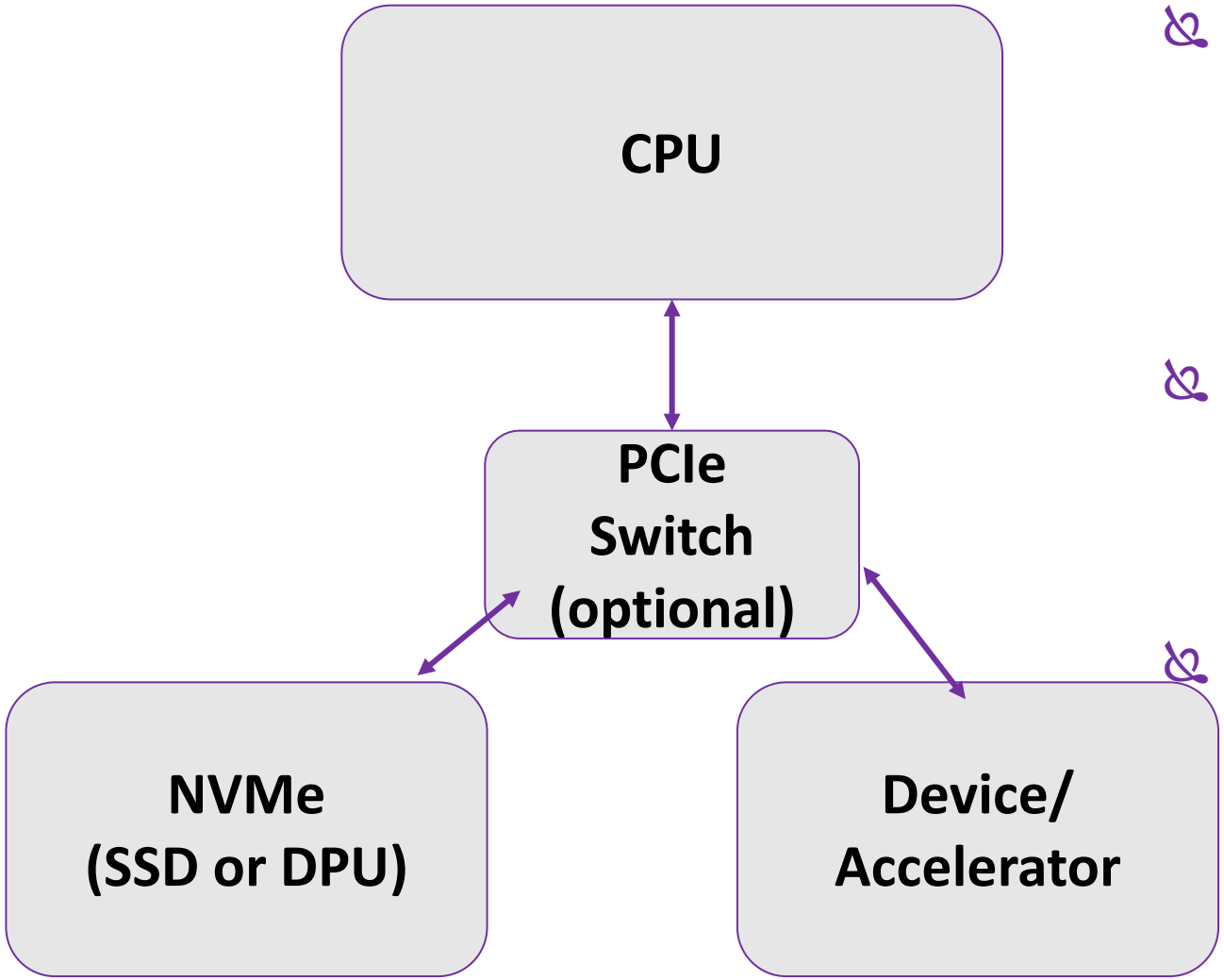
Accelerator initiated

- Accelerator initiates I/O with Storage
- Data response is sent to Accelerator memory

- Removes extra data copy
- Removes CPU from processing path
- Requires a basic OS on the accelerator side
  - This could be a DPU



Device-initiated I/O

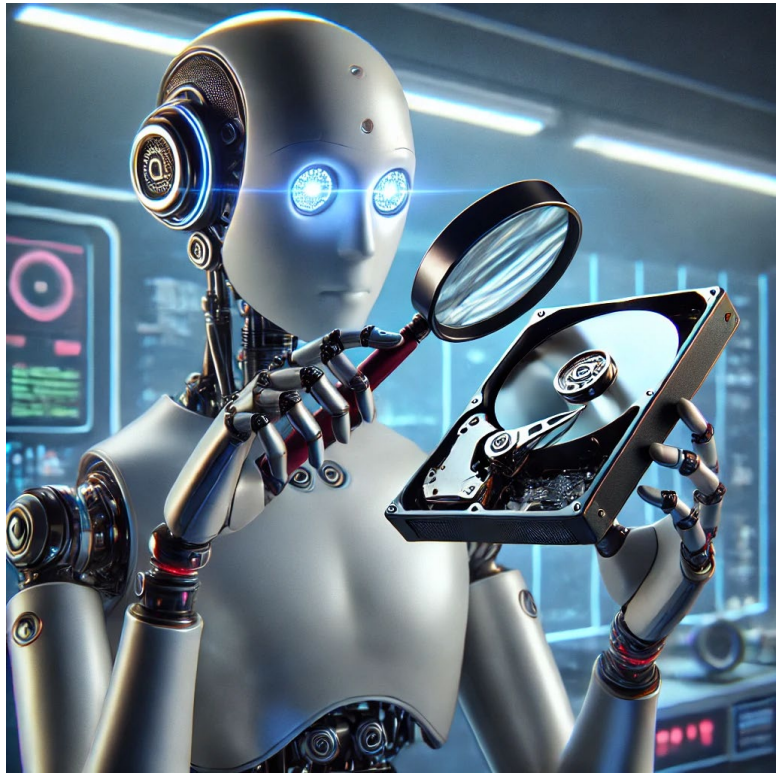


- Extend Point-to-point DMA to allow a device to formulate NVMe SQEs and put them on a NVMe SQ (in host or device memory) and ring the appropriate doorbell.
- A recent academic paper (NVIDIA+Illinios) coined “BAM” did a PoC of this using libnvme, nvcc and a GPU.
- Now the IO instructions do not have to run on the CPU!

https://arxiv.org/pdf/2203.04910

Content Aware Storage

- Similar to older Content Addressed Storage
  - Uses AI to identify content
  - Major use case is to build vector databases for RAG
- Tag based data retrieval



Storage designed for AI workloads

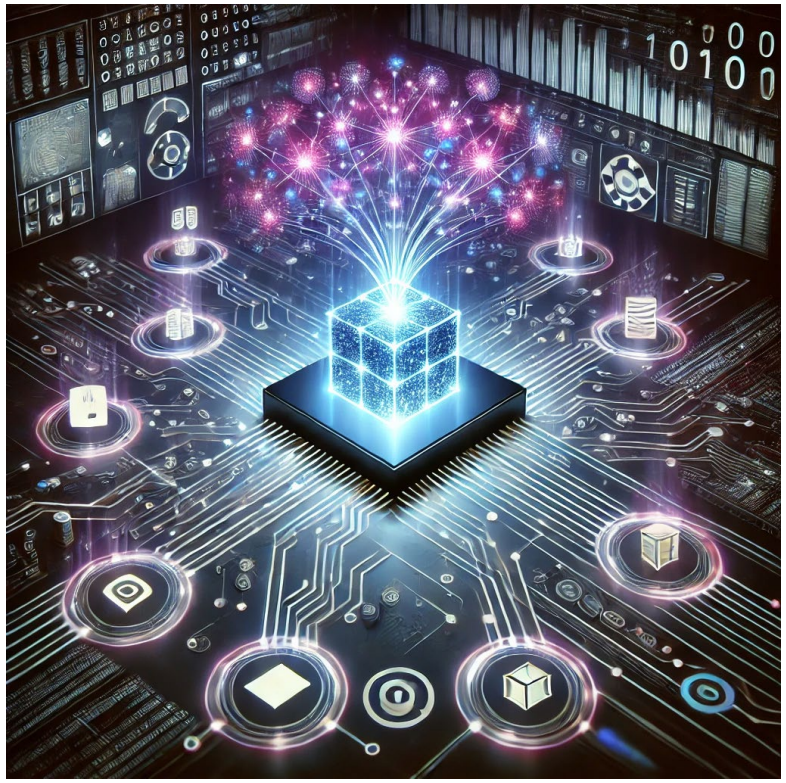
- What does storage look like going forward considering AI workloads?
  - NAND page sizes to better match AI workload writes (reduce write amplification)
  - Higher performance devices in both IOPs and bandwidth to reduce the impact of "waiting for storage"
  - Higher power efficiency – Large scale of AI datacenters means that every components power efficiency is important
  - Intelligent Storage Devices – Standardization of intelligent storage devices
- SNIA could be a good place to define these requirements
  - SNIA members are already talking about these requirements

REGIONAL SDC

SNIA Activities

- SNIA TC formed the AI Task force in 2024

- Goal is to look at where SNIA could apply its data and storage expertise for AI
  - Some of the items the task force is looking at:
    - AI Whitepaper for AI best practices
    - Areas in data retrieval and storage for AI
    - Data archiving for AI
    - Requirements for AI storage
    - Others? (Let us know what you think!)



Final thoughts

- Are we at the Emerald City yet?
  - Ask the man behind the curtain ☺
- Advances in AI continue to put pressure on storage systems
  - But, storage is many times the last thing that system developers think about
  - It is up to our community to have answers to the upcoming challenges ready before they are asked