SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

What can Storage do for AI?

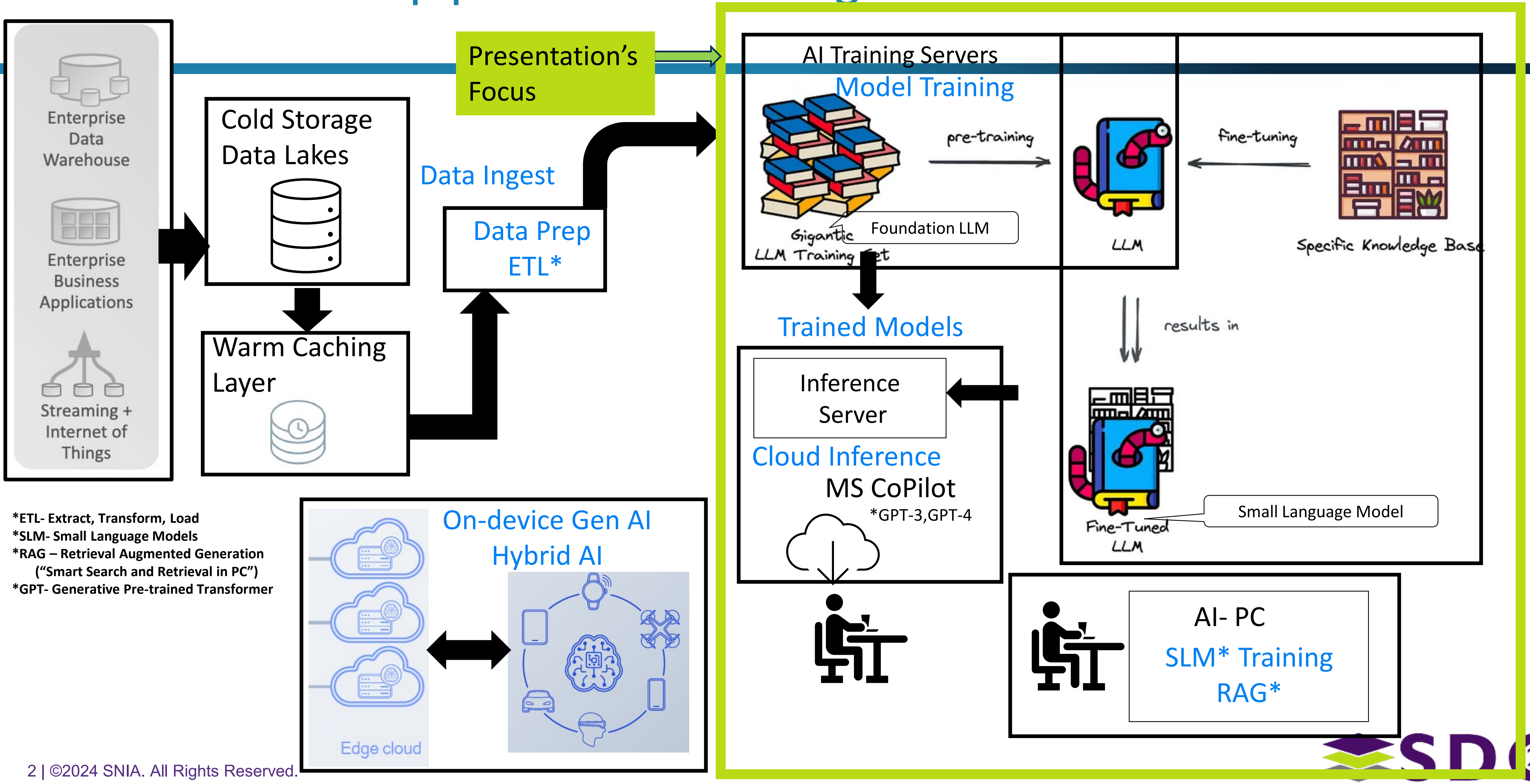
Suresh Rajgopal

Katya Giannios

Sujit Somandepalli

Micron

AI/ML pipeline and Storage Use cases



*ETL- Extract, Transform, Load
*SLM- Small Language Models
*RAG – Retrieval Augmented Generation
("Smart Search and Retrieval in PC")
*GPT- Generative Pre-trained Transformer

Outline

- Motivation
  - Why do we need (NVMe) Flash Storage to play a larger role in Training and Inference?
- Opportunities
  - Where can Flash storage contribute?
- Illustrated Example
  - What did we learn about flash storage in AI Training/Inference from our testing?

Cost, Power and Time impacts of Training [0]

Cost

- Each training run of GPT-3 cost 5M$[1]
- Cost of foundational model training is over 100M$[1]
- Largest models can cost **>1B$** to train by 2027[2]

Time

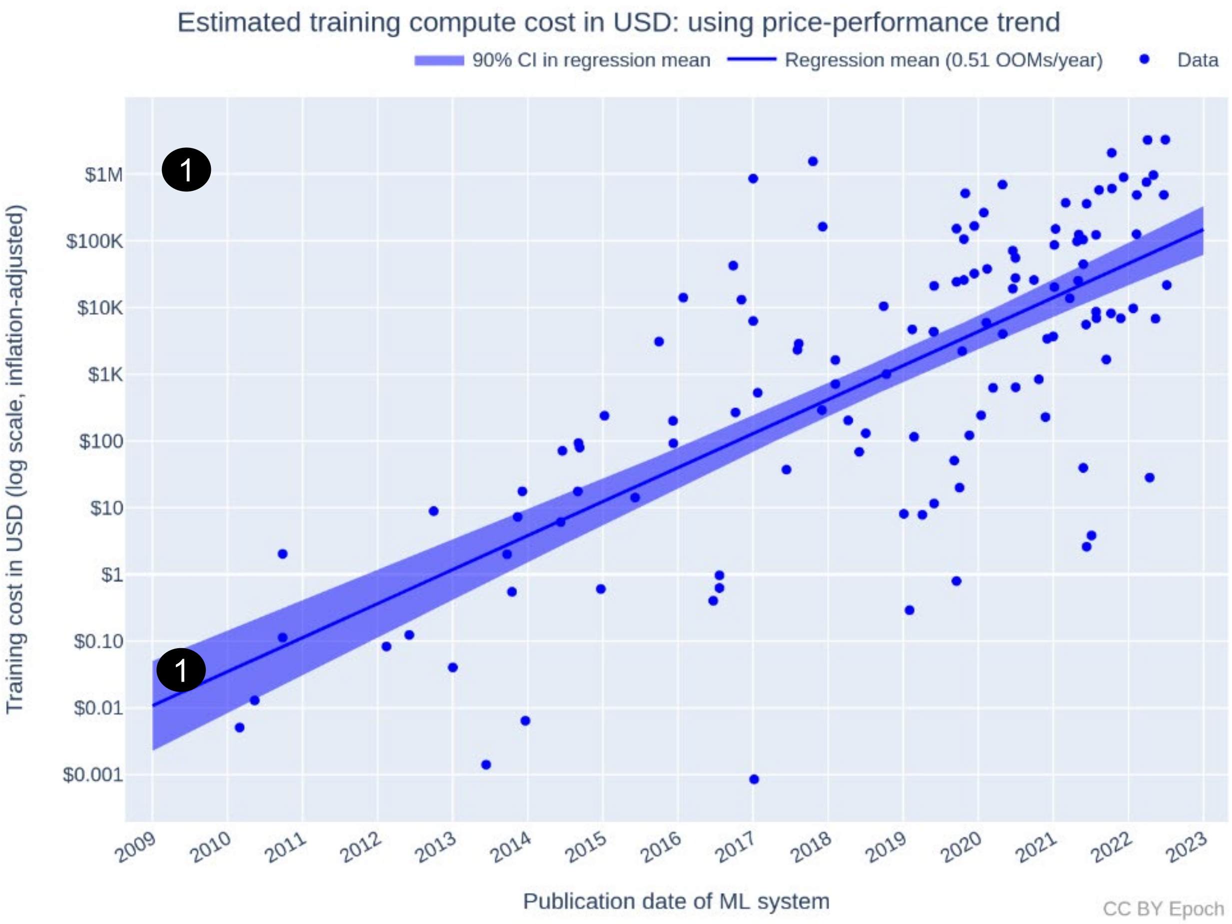
- Meta's Llama2 70B model took 1.7Mhrs[3]
- Palm-540B model took 8.8Mhrs[3]
- Training GPT-3 - **36yr**s with 8V100 GPUs/ or 7months with 512 GPUs[4]
- GPUs utilization is best-case **50%** usually much lower [0]

Power

- GPT-2 model training consumed 28MWhrs[5]
- GPT-3 consumed **10X** more 284MWhrs. [> 500 refrigerators running annually!!]
- Google just reported a 48% greenhouse gas increase due to AI in datacenters[6]

Foundational Model Training will be accessible to only a very few

The Need to Democratize Training



[Training Cost (EpochAI.org)]

- 0.5 order of magnitude cost increase ($10^{0.5}$) every year ~ 3X
- Cost = Hardware Cost + Energy Cost
  - – Upfront HW Cost and %age time spent on training?
  - – Energy Cost = Power x training time x Energy Rate
- 124 ML systems (not just LLMs)

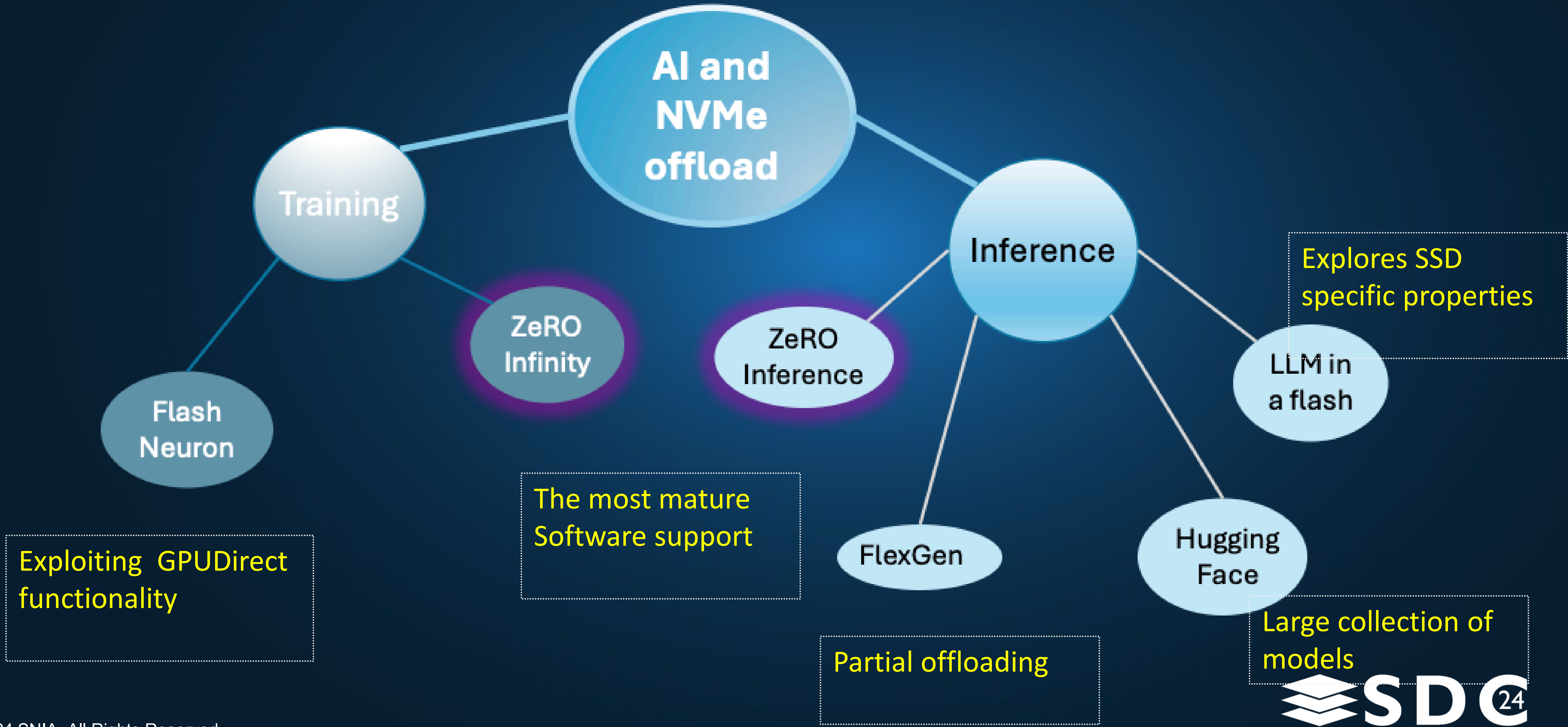
Making SLM Training accessible to more data scientists is a growing challenge

Offload in AI Training – (NVMe Storage , CPU Memory)

- AI Training relies on keeping all training related data close to the GPU
  - Type of Data
    - Model Parameters (Weights and Biases)
    - Optimizer States (between training batches) and Gradients (parameter adjustments)
    - Checkpointing data (intermediate states)
    - Working Memory (during forward/backward passes)
  - For a 1T model, GPU requires ~30TB of operational training data – "Memory Wall"
  - Grows with model size and context size
- Today
  - Model scaling relies on aggregating GPU Memory (across several 100 GPUs)
  - 3D Parallelism – Data, Tensor or Pipeline parallelism
- Offload
  - Leverage heterogeneity in AI Servers – distribute training data in CPU/CXL/NVMe Flash

Effective Offload can provide a significant cost and power benefit

Offload Opportunities



DeepSpeed MSFT ZeRO Infinity - Training

Key Offloading Tenets

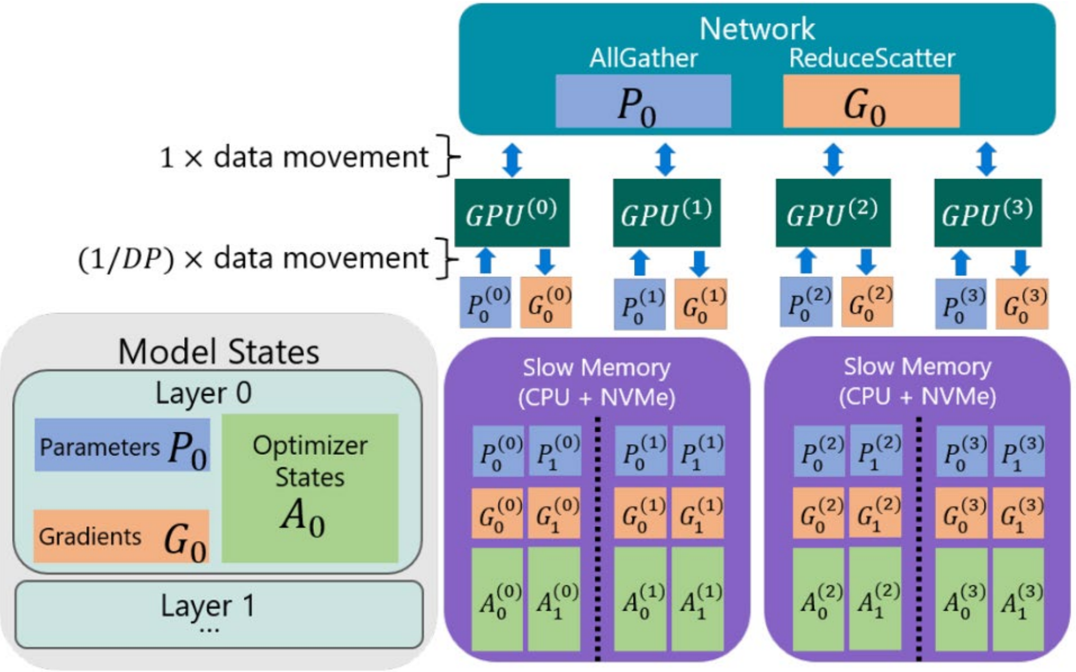
- Offloading partitioned model states (P_i, G_i, A_i) to CPU-DRAM or NVMe storage
- Enables **parallel memory access-** virtually unlimited heterogeneous memory bandwidth. (NVMe bw is ~100 GB/s per DGX-H100)
- Dynamic prefetching: overlapping read/write from NVMe to CPU with compute.

P_i – model parameters, G_i – gradients, A_i – optimizer states

DeepSpeed MSFT ZeRO Infinity - Training

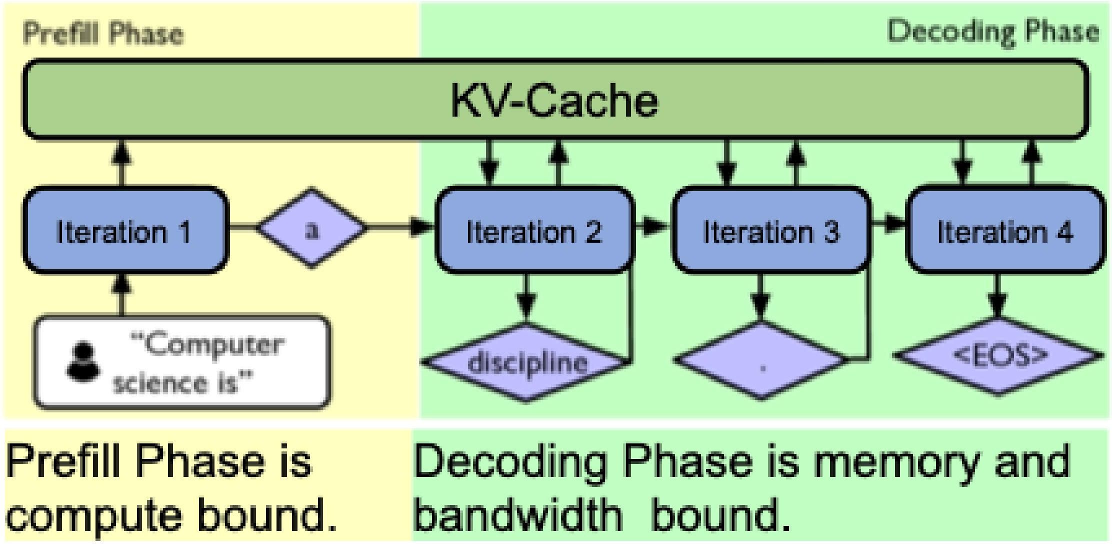
Key Offloading Tenets

- Offloading partitioned model states (P_i, G_i, A_i) to CPU-DRAM or NVMe storage
- Enables **parallel memory access-** virtually unlimited heterogeneous memory bandwidth.(NVMe BW upto ~100 GB/s per DGX-H100)
- Dynamic prefetching: overlapping read/write from NVMe to CPU with compute.



GPU memory is NOT the memory bottleneck. One can leverage CPU DRAM Memory and **NVMe Storage** for fine-tuning of Trillion parameter models!

Inference

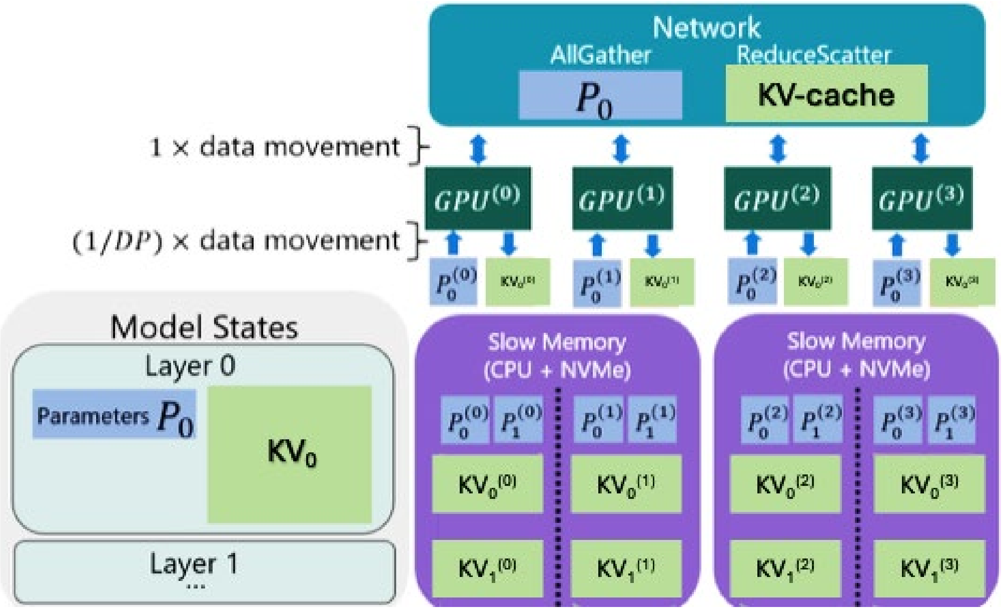


Offloading Opportunities

- Partitioned model parameters
- KV-Cache: offloading on NVMe possible, more significant performance degradation

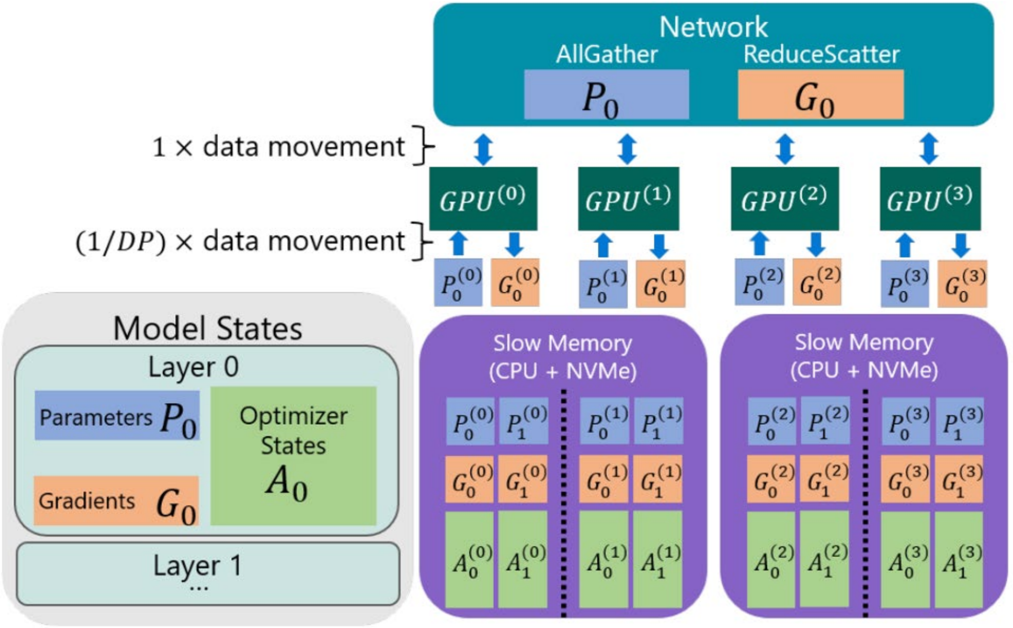
DeepSpeed MSFT ZeRO Inference

Inference



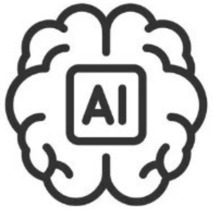
vs

Training



For Inference model parameters offload to NVMe storage, KV-cache offload to CPU DRAM memory!

Inference Applications



Online Inference

Important metrics:

- Time to First Token
- Total Latency

Applications:

✓ Realtime applications
✓ Chatbots

Offline Inference

Important metrics:

- Throughput

✓ **NVMe Storage**

Applications:

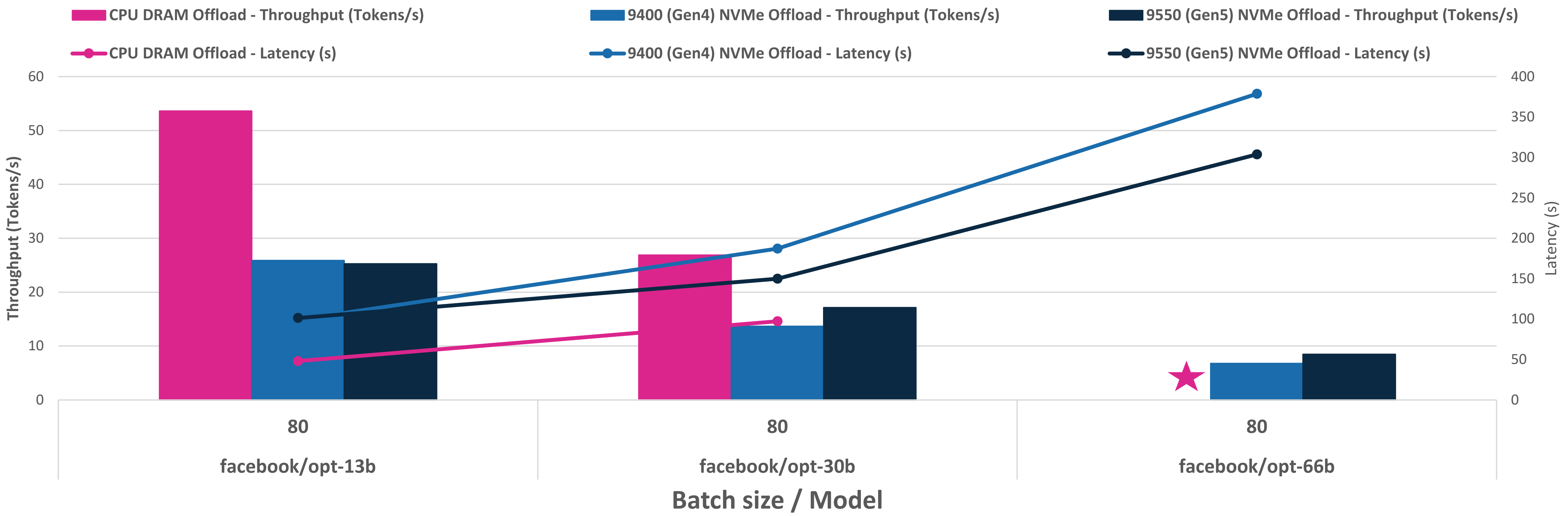
✓ Zoom AI companion summary
✓ Amazon reviewer's summary
✓ Daily updated recommendations for products/news/movies

Results

- ZeRO-Inference
  - Supermicro SYS-521GE-TNRT
    - 2x Intel Xeon Platinum 8568Y+
    - 256GB DDR5 DRAM
    - 1x Nvidia L40s
    - 1x Micron 9550 Pro NVMe SSD 7.68TB
- Models Tested
  - OPT 13b, OPT 30b & OPT 66b
- Batch Sizes Tested
  - 80, 128 and 140
- Prompt /Output Length: 512 /32
- Offload – CPU DRAM Offload and NVMe Offload
  - PCIe Gen4 – Micron 9400 SSD 7.68TB
  - PCIe Gen5 – Micron 9550 SSD 7.68TB

Testing ZeRO – Inference on a workstation class system

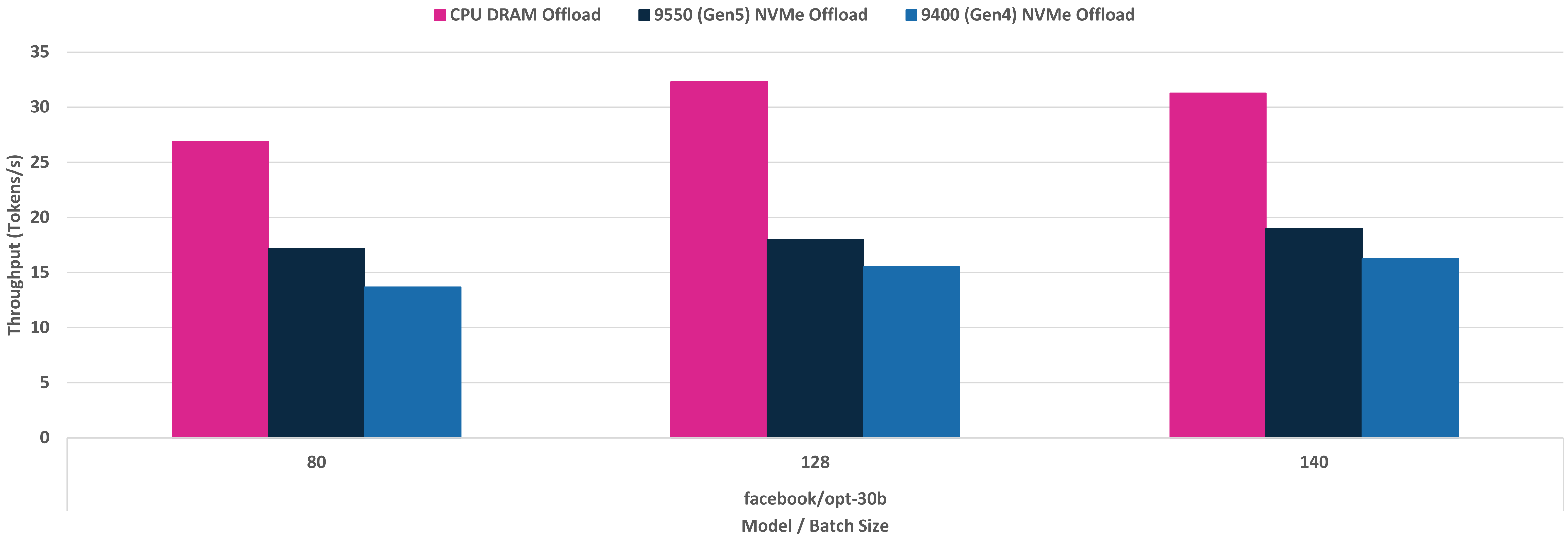
Inference Performance with Model Size Scaling



CPU Memory Offload provides best performance with lowest latency.
NVMe offload allows you to run larger models at the same batch size

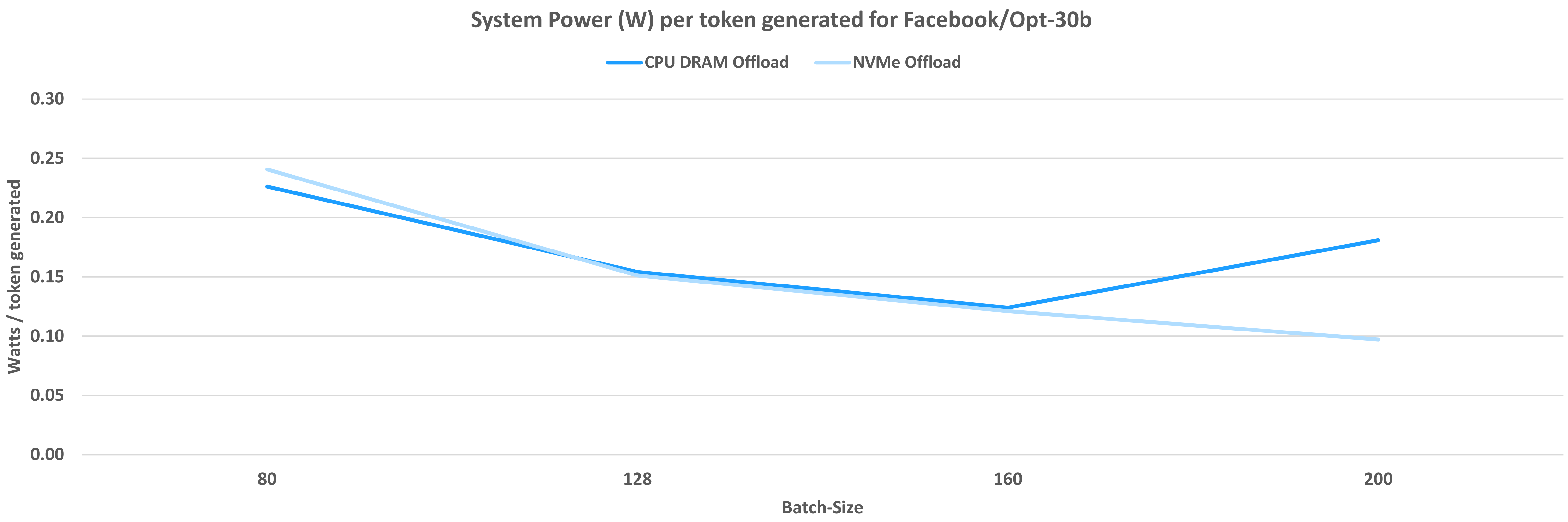
★ CPU DRAM Offload ran out of memory for facebook/opt-66b

Inference Performance with Batch Size Scaling



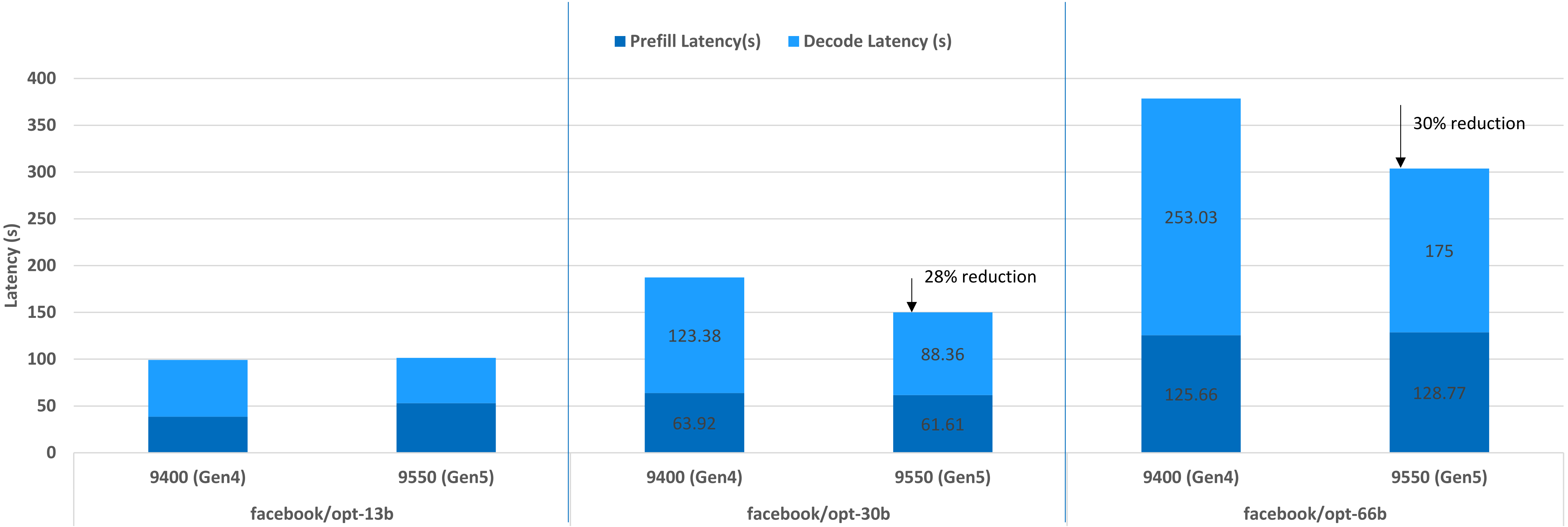
As batch size increases, NVMe offload allows you to increase performance while CPU DRAM offload starts to plateau.

Power Efficiency benefits of NVMe Offload



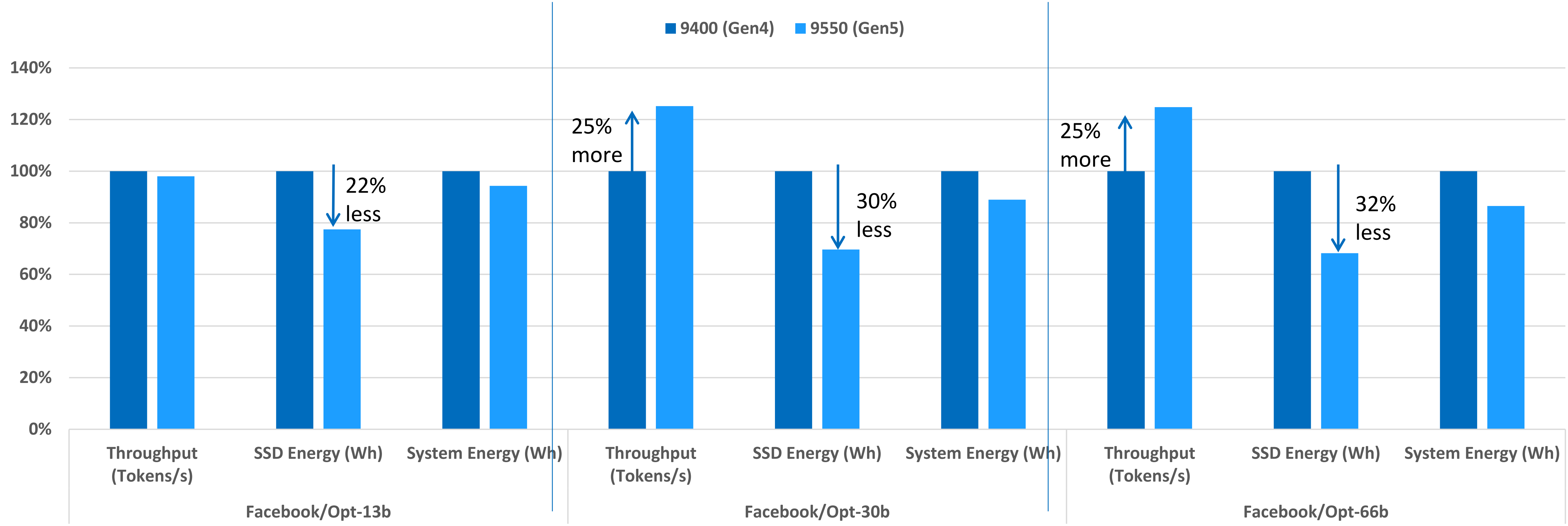
Power per generated token improves with batch size for NVMe Offload
But, where are the NVMe efficiency improvements coming from?

NVMe Bandwidth Improvement = Decode Latency Improvement



Decode latency improves with PCIe Generations – enabling power efficiency

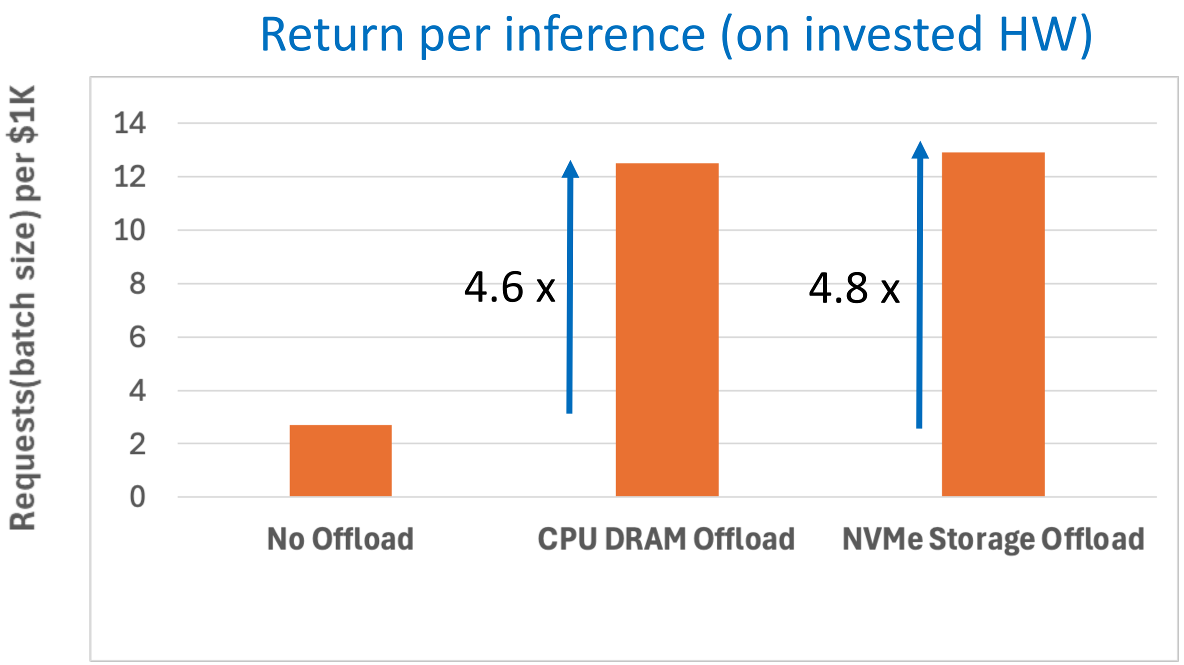
Inference Power Efficiency Gen4→Gen5



Micron 9550 Gen5 SSD is 20-30% more energy efficient and improves inference throughput by 25% compared to the previous generation

Cost Benefits of Offloading

- Inference on a 30b param model (OPT30b) with a batch size of 200
- Configuration:
  - No Offload 6xL40S GPUs, 256GB of DRAM
  - CPU offload:1xL40S GPU, 512GB of DRAM
  - NVMe Offload:1xL40S GPU, 256GB of DRAM



Offloading can yield 4.8x cost efficiency improvement

Conclusion and Call to Action

- Power and Cost considerations for AI-at Scale deployment are real
- NVMe offload can be a cost and power efficient alternative
  - Accommodates larger models –better quality responses
  - Support larger batch sizes – more inference requests per unit time, better GPU utilization
- Offload libraries like ZeRO Inference should be leveraged
- Enabling NVMe Offload requires
  - Careful model optimizations to hide storage latency behind compute
  - Large blocks sizes and use of multiple threads further accelerate SSD performance
- Storage for AI – Call to Action
  - Move to faster PCIe interfaces on SSDs – Gen4, Gen5,..
  - Focus on Read performance, optimize bandwidth
  - Understand the software stack above to build latency tolerant solutions

Please take a moment to rate this session.

Your feedback is important to us.

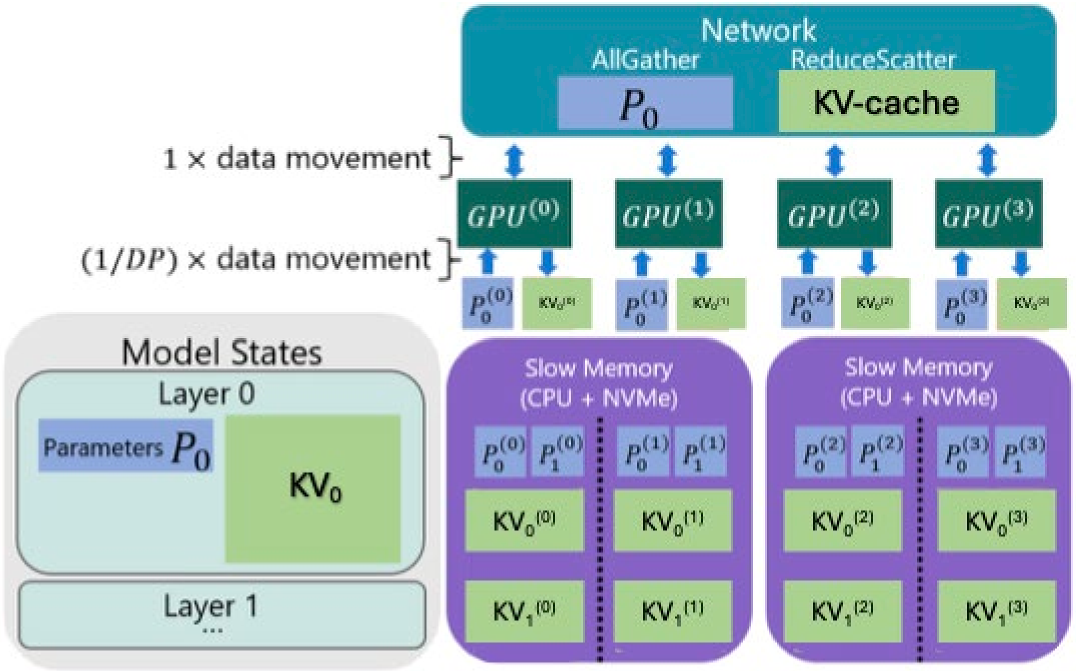
DeepSpeed ZeRO Infinity:

DeepNVMe

- **Asynchronously reading and writing** tensors to NVMe storage at near-peak NVMe bandwidth in PyTorch.
- **Data transfers** between persistent storage and DL application memory through optimizations built on NVMe SSDs and **NVIDIA GDS** (NVIDIA GPU direct storage).
- Both **intra**-request (I/O from one user thread) and **inter**-request parallelism (I/O requests from multiple user threads) are leveraged by the applications.

- Additional optimizations including low-overhead multi-threading and smart work scheduling, **avoiding data copying**, and memory pinning.

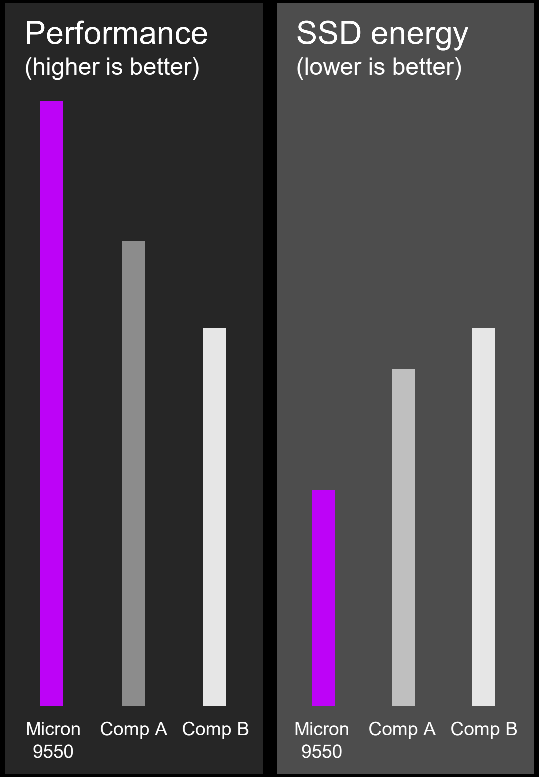
DeepSpeed MSFT ZeRO Inference



- Offloading partitioned model states (P_i, KV-cache) to CPU-DRAM or NVMe storage (only P_i)
- Enables parallel memory access
- Dynamic prefetching

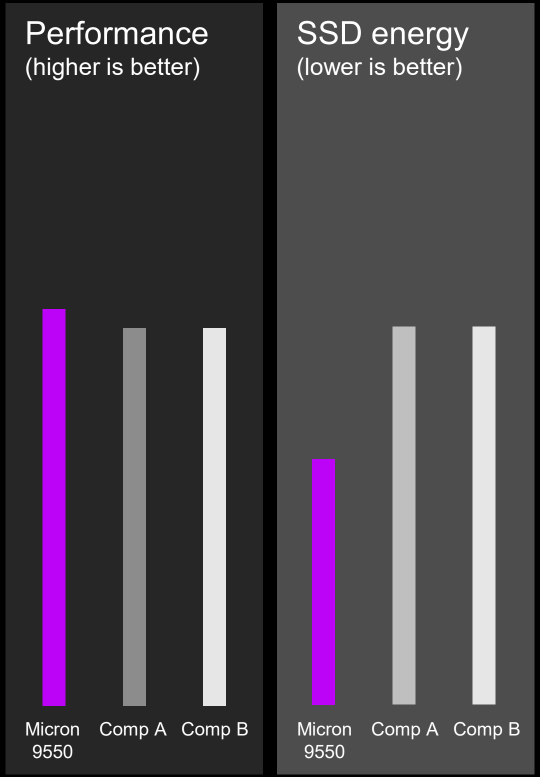
Micron 9550 – built for AI

Graph neural network training
(Big accelerator Memory)



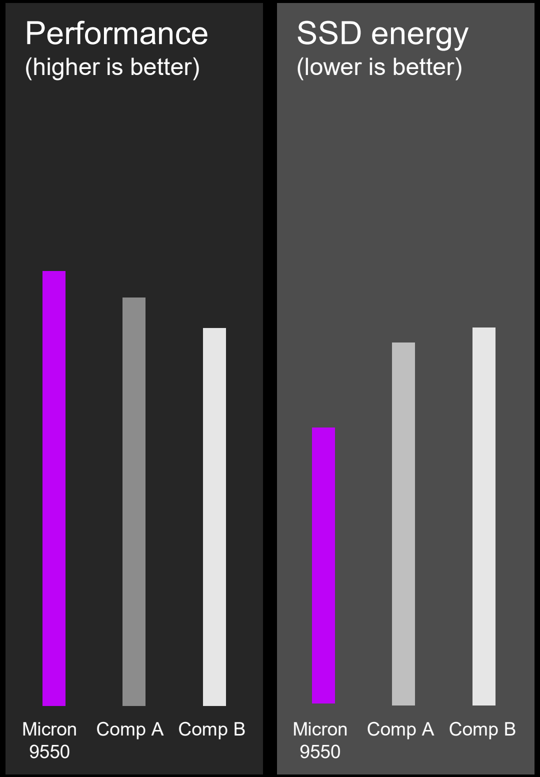
Up to
60% higher performance
43% less energy

Unet3D medical image training
(Deep learning IO)



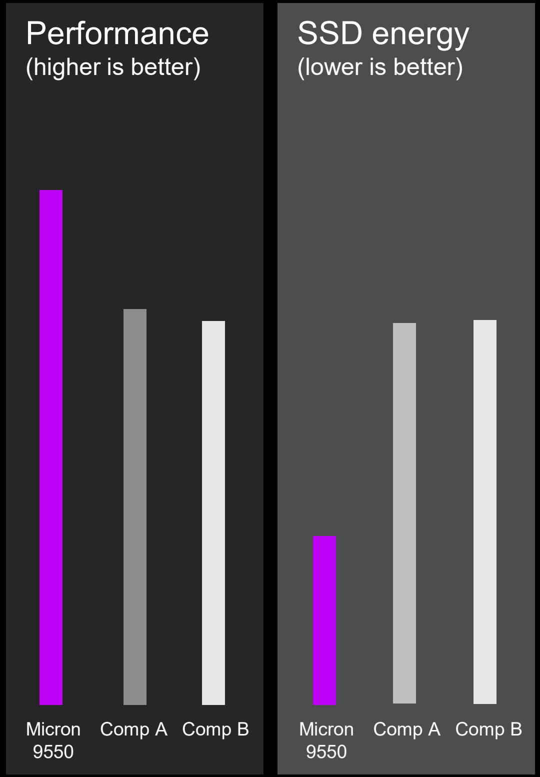
Up to
5% higher performance
35% less energy

Large language model inference
(DeepSpeed ZeRO-Inference LLM)



Up to
15% higher performance
27% less energy

NVIDIA GPUDirect® Storage



Up to
34% higher performance
56% less energy

