PERSISTENT MEMORY + COMPUTATIONAL STORAGE SUMMIT 2022

VIRTUAL EVENT • MAY 24-25, 2022

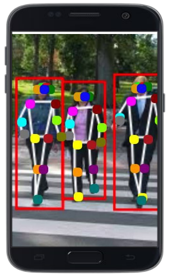
AI Memory at Meta: Challenges and Potential Solutions

Presented by Chris Petersen

AI at Meta

- Across many applications/services and at scale → driving a portion of our overall infrastructure (both HW and SW)
  - Ranking and recommendation - Video, Ranking, Search…
  - Content understanding - Computer vision, Speech, Translation, NLP, Video…
- From data centers to the edge

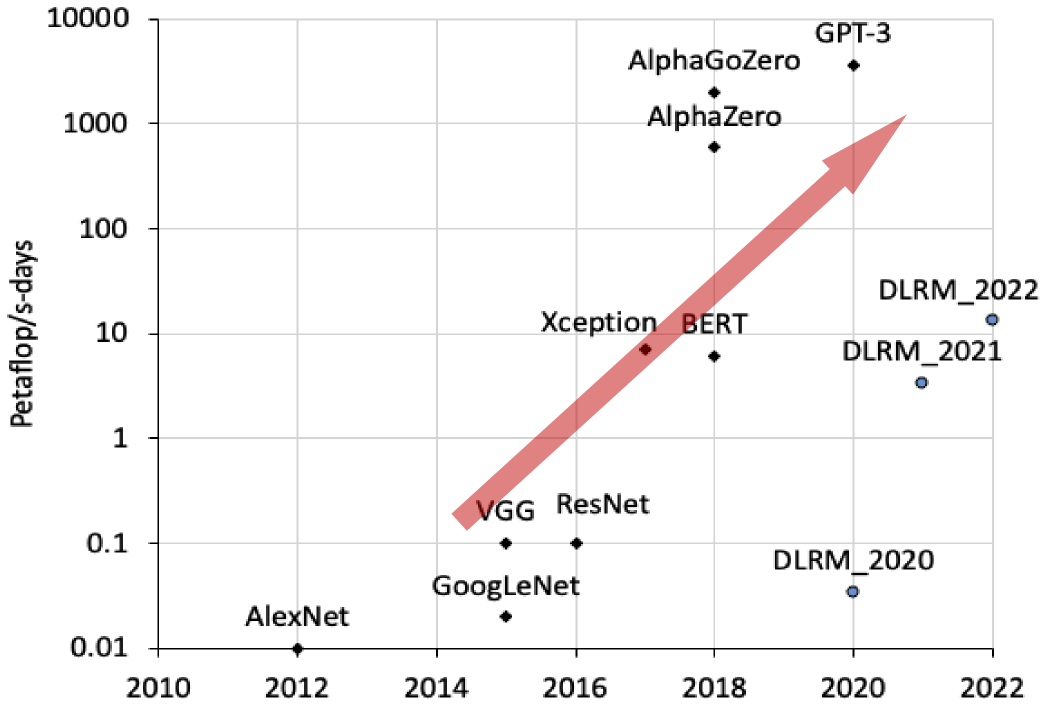
Keypoint Segmentation

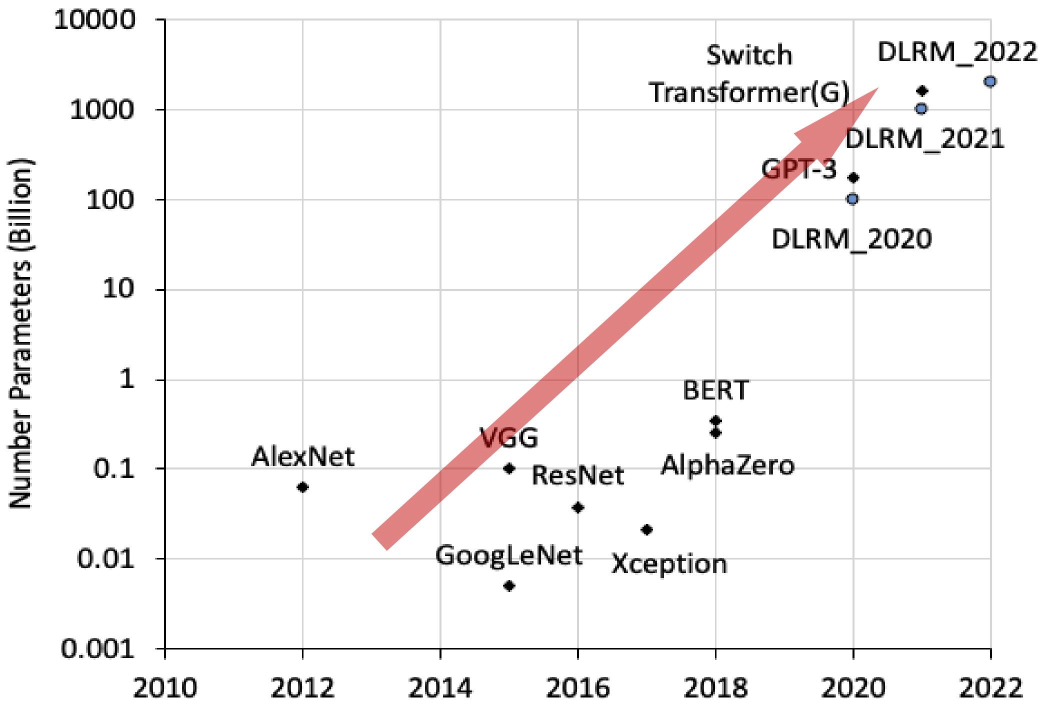


Augmented Reality with Smart Camera

Problem Statement: AI workloads scale rapidly

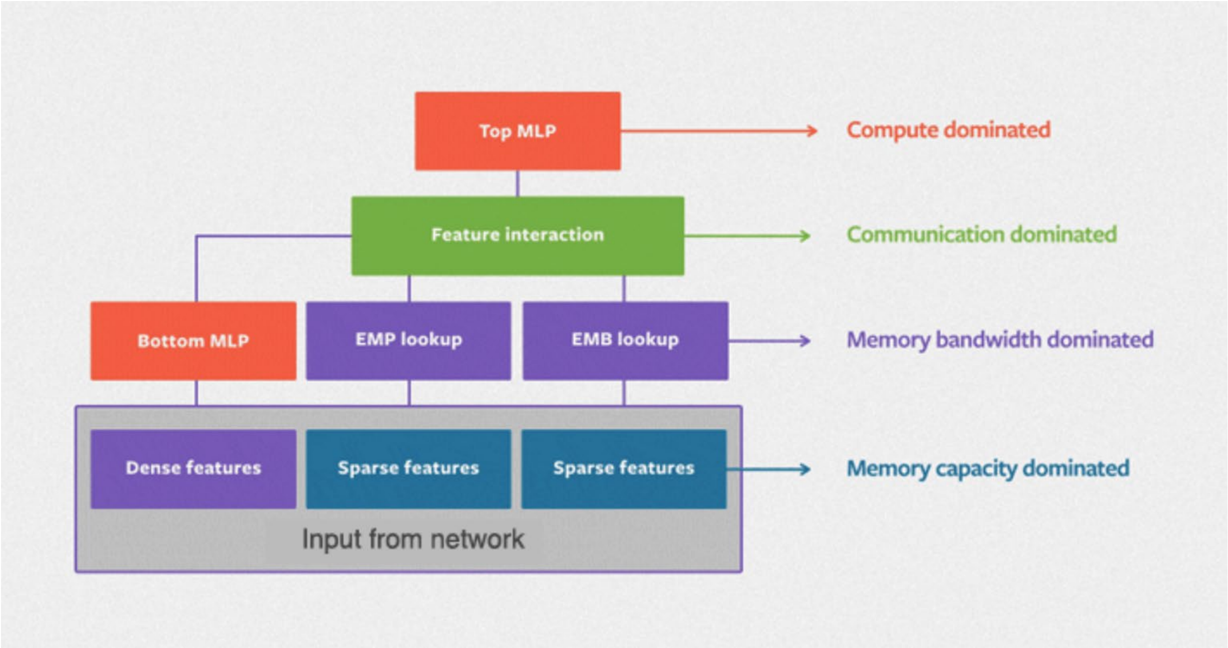
- Compute, Memory BW, Memory Capacity, all scale for frontier models
  - Scaling typically is faster than scaling of technology
- The rapid scaling requires more vertical integration from SW requirements to HW design





Recommendation models (e.g. DLRM)

- One of the main drivers of AI HW platforms
- Most of the memory capacity is contributed by sparse features (embedding tables)
  - Dense
    - Requires high BW at low capacity
  - Sparse
    - Requires high capacity at high BW



https://ai.facebook.com/blog/dlrm-an-advanced-open-source-deep-learning-recommendation-model/

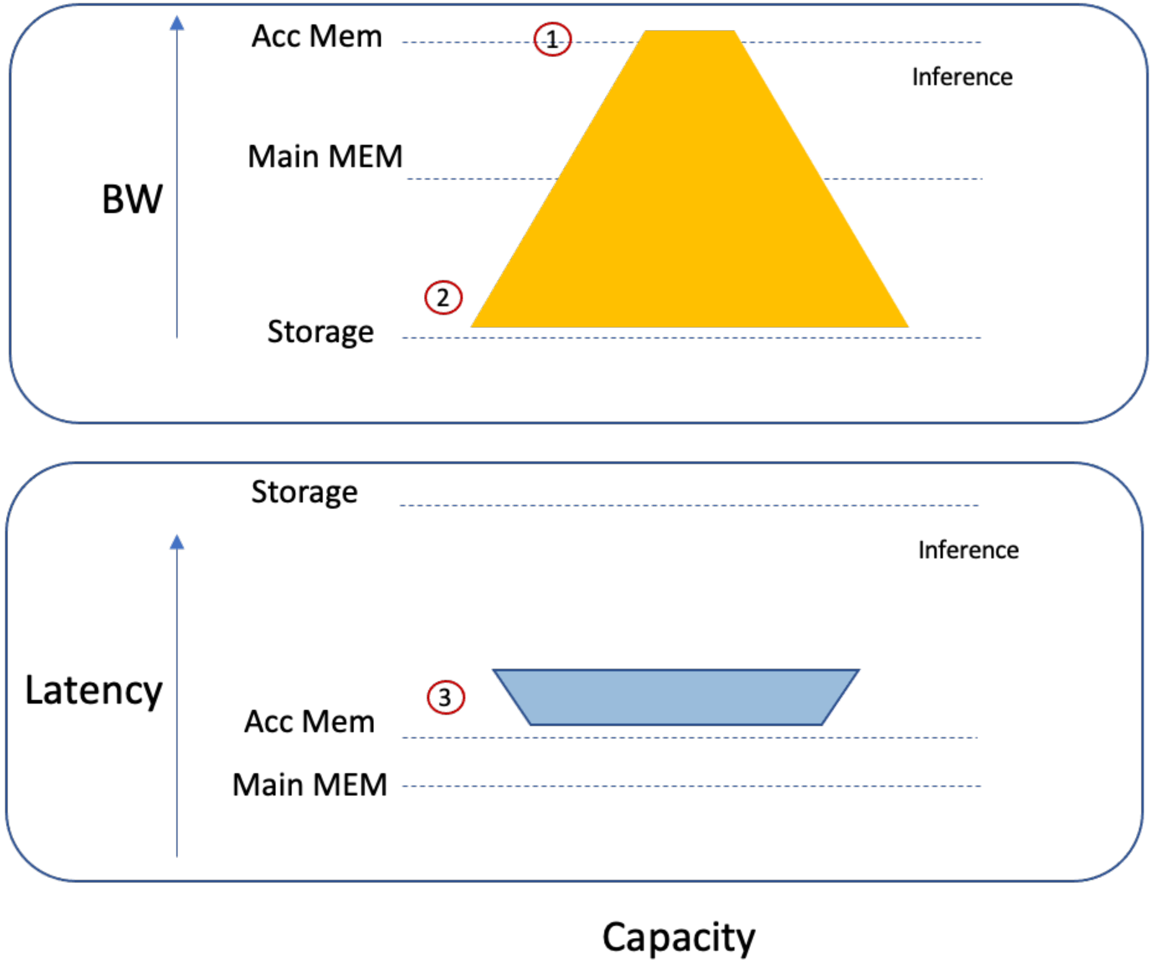
DLRM Requirements

Bandwidth

1. Considerable portion of capacity needs high BW Accelerator memory.
2. Inference has a bigger portion of the capacity at low Bandwidth. More so than training.

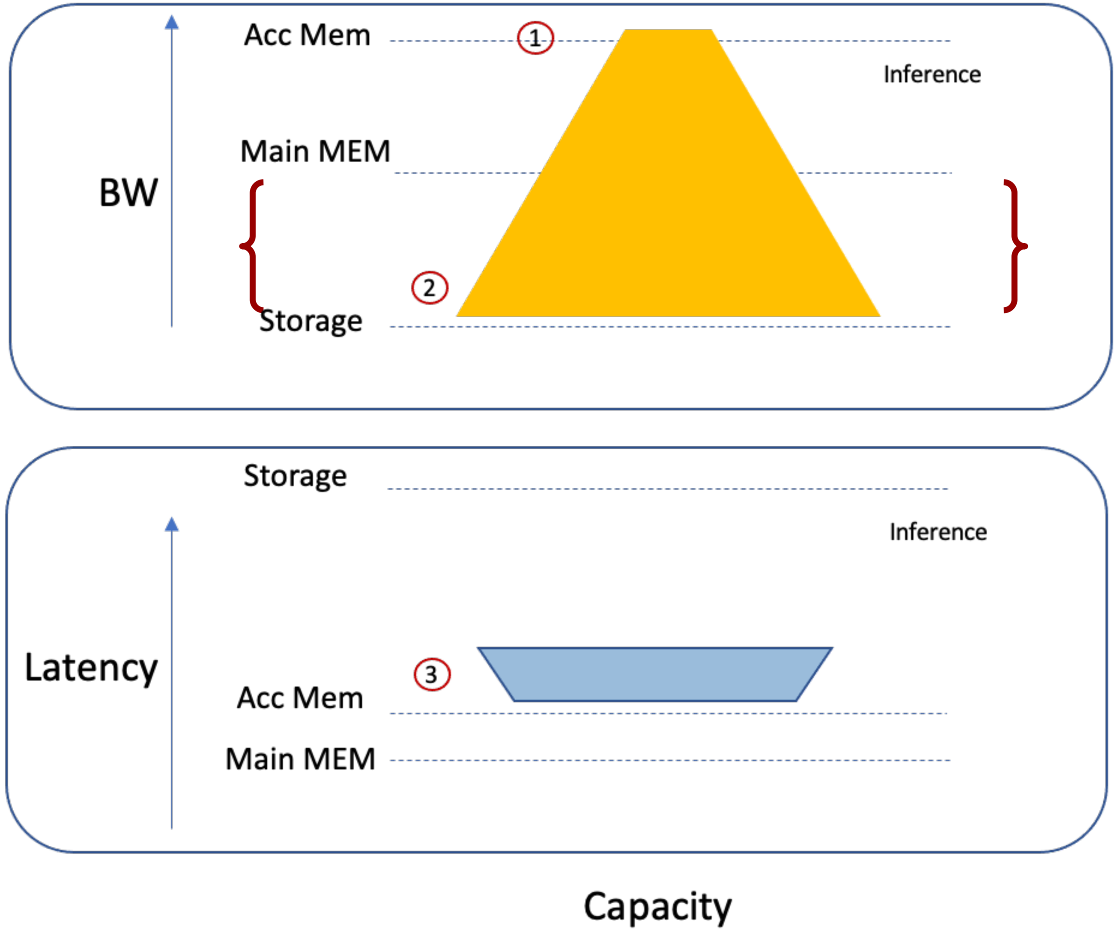
Latency

3. Inference has a tight latency requirement, even on the low BW end



System Implications of DLRM Requirements

- A tier of memory **beyond HBM and DRAM** can be leveraged, particularly for inference
  - Higher latency than main memory. But still tight latency profile (e.g TLC Nand Flash does not work)
  - Trade off performance for capacity
  - This does not negate the Capacity and BW demand for HBM and DRAM



How does it fit in the whole e2e system?

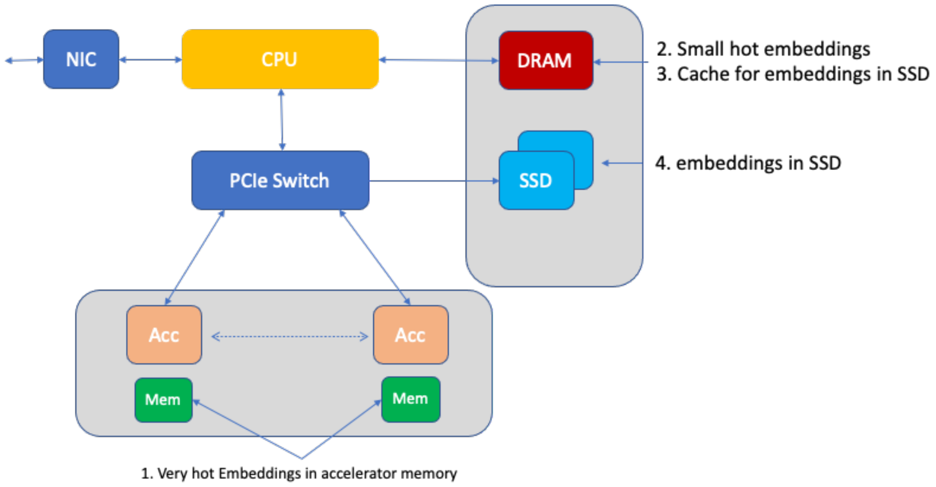
- Perf / Watt and then Perf/$ are important metrics
- Different scenarios in real use cases
  - Simpler HW
  - Avoiding scale out
  - Facilitate Multi-tenancy

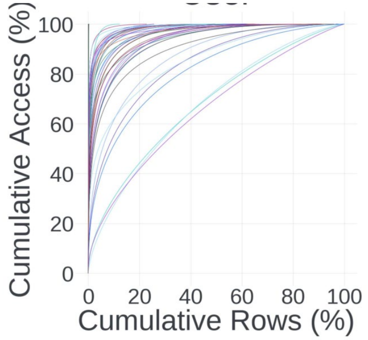
An Example Implementation

SUPPORTING MASSIVE DLRM INFERENCE THROUGH SOFTWARE DEFINED MEMORY

Ehsan K. Ardestani[1] Changkyu Kim[1] Seung Jae Lee[1] Luoshang Pan[1] Valmiki Rampersad[1] Jens Axboe[1] Banit Agrawal[1] Fuxun Yu[2] Ansha Yu[1] Trung Le[3] Hector Yuen[1] Dheevatsa Mudigere[1] Shishir Juluri[1] Akshat Nanda[1] Manoj Wodekar[1] Krishnakumar Nair[1] Maxim Naumov[1] Chris Peterson[1] Mikhail Smelianskiy[1] Vijay Rao[1]

- The published work below mainly focuses on the lowest memory tier (using NVMe SSDs)
- Software Defined Memory backed by SSDs
  - The BW demand required SCM SSDs
    - High IO rate at Smaller access granularity
  - Application level Caching in main memory
    - Row cache due to lack of spatial locality
  - Fast IO (io_uring)
  - Placement policies among DRAM and SSDs
    - To improve overall performance





 |

Impact Scenario #1: Save Power with Simpler HW

Deployment of a **143 GB** model with SDM enabled system, with simpler HW, can reach the same latency as deployment on a more complex model with more DRAM resulting in **20% power savings**.

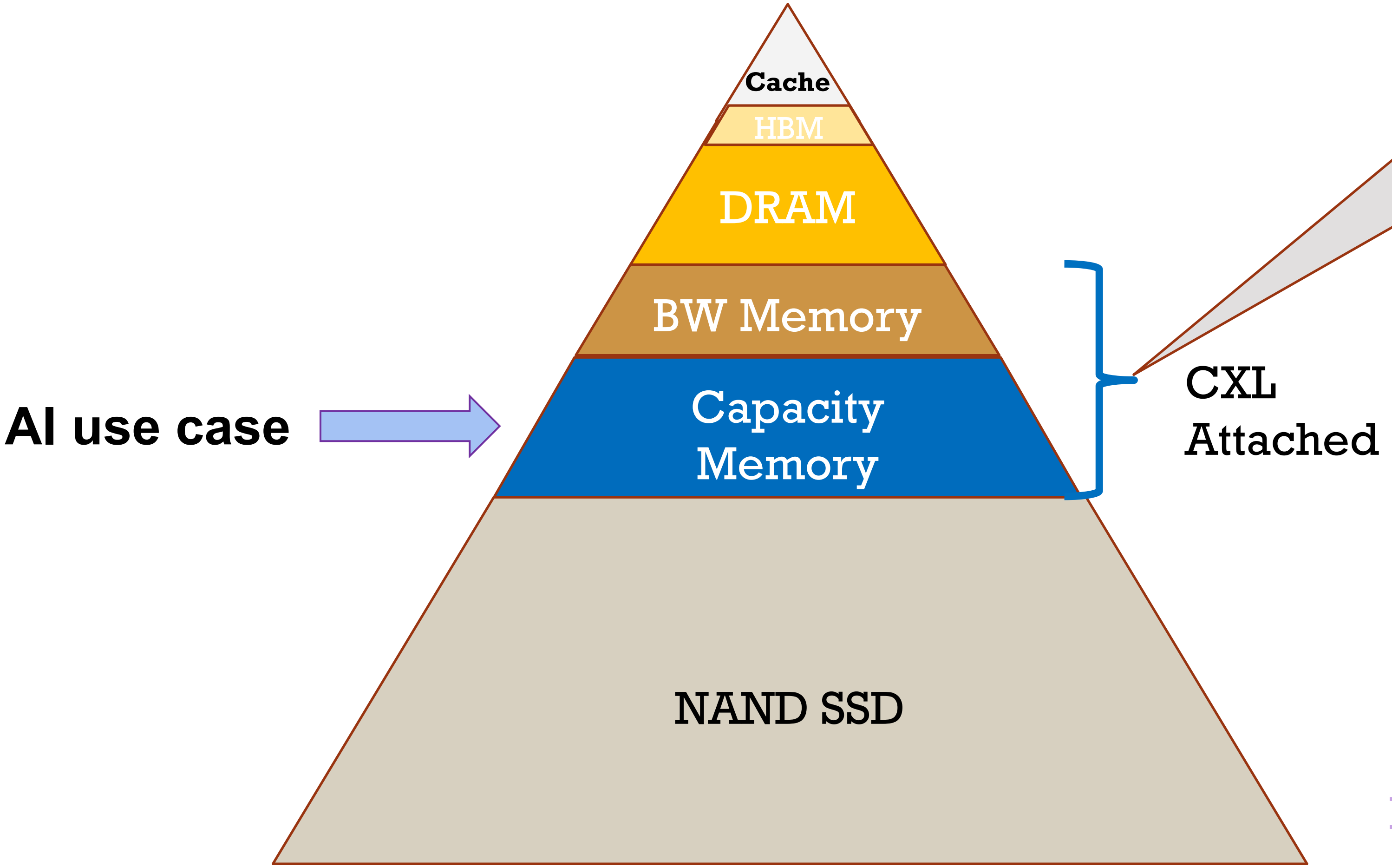
Scenario	QPS	Power	Total Hosts	Total Power
Baseline: 2-Socket, High Mem Capacity	240	1.0	1200	1200
SDM system: 1-socket, Low Mem Capacity + SDM	120	0.4	2400	960

Impact scenario #2: Save Power by Avoiding Scale out

- Systems with higher compute (e.g. using accelerators) require higher BW throughout the memory hierarchy
- Using SDM with a SCM SSD for a 150 GB model prevents scale out, **saves power by 5%**, and allows for a simpler serving paradigm

Scenario	QPS	Host Power	Total Hosts	E2E Power
Baseline: Scale Out	450	1.0+0.25	1500+300	100%
Option 1: Nand Flash	230	1.4	2978	189%
Option 2: SCM SSD	450	1.0	1500	95%

Memory Tiers at a High Level



- Load/store interface
- Cache line read/writes
- Scalable
- Heterogeneous
- Standard interfaces

Ref: OCP 2021 Presentation

Need to **store** more data per node while improving TCO

Summary

- AI Models scale faster than the underlying memory technology
- Additional tiers of memory beyond host DRAM can help (aka Capacity Memory)
- This memory tier can trade off some performance for capacity
- CXL provides a viable option to enable this new memory tier

Please take a moment to rate this session.

- Your feedback is important to us.

