OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

November 8, 2021 | San Jose, CA

Multi-Tier Memory in Windows and Azure

Scott Lee
(Ray)Jui-Hao Chiang

Agenda

- Directional sharing on multi-tier memory support on Windows and exposure to applications.
- Potential usage of multi-tier memory in Azure

Multi-Tier Memory in Windows

Multi-Tier Memory

- Memory devices on the horizon that have different performance characteristics than DDR DRAM DIMMs.
  - HBM (High Bandwidth Memory): expected in some servers
  - CXL memory device: may contain DRAM and/or Storage Class Memory (SCM)
- Performance characteristics
  - Latency
  - Bandwidth
- Two memory categories
  - General Purpose Memory
  - Dedicated Memory

General Purpose Memory

- Memory available on current systems
- Allocatable by anyone
- Contributes to system commit
- Specific performance characteristics for General Purpose Memory is TBD.
  - General guidance: any performance equal to or up to a NUMA node difference for DDR DRAM DIMM can be General Purpose Memory

Dedicated Memory

- Available only through dedicated APIs
- Does not contribute to system commit
- Have performance characteristics different from General Purpose Memory
  - Can be lower or higher performance
- Can have many different Dedicated Memory types on a system

OS Discovery of Dedicated Memory

- Leverage mechanisms in ACPI and UEFI to discover Dedicated Memory on a system
- Memory-only NUMA nodes in ACPI SRAT
- Report of performance characteristics through ACPI HMAT table
- Dedicated Memory marked as reserved or special purpose memory to OS

Software Discovery of Dedicated Memory

- System can have many different Dedicated Memory types
- Each Dedicated Memory type has a unique combination of attributes
  - Current attributes: read/write latency and bandwidth
  - Number of attributes likely to grow over time
- Performance characteristics reported will be the speed of the media outside of any intermediate caches and from the closest CPU node
- New APIs
  - Enumeration of Dedicated Memory types available on a system
  - Allocate or free memory from a specific Dedicated Memory type

OS Usage of Dedicated Memory

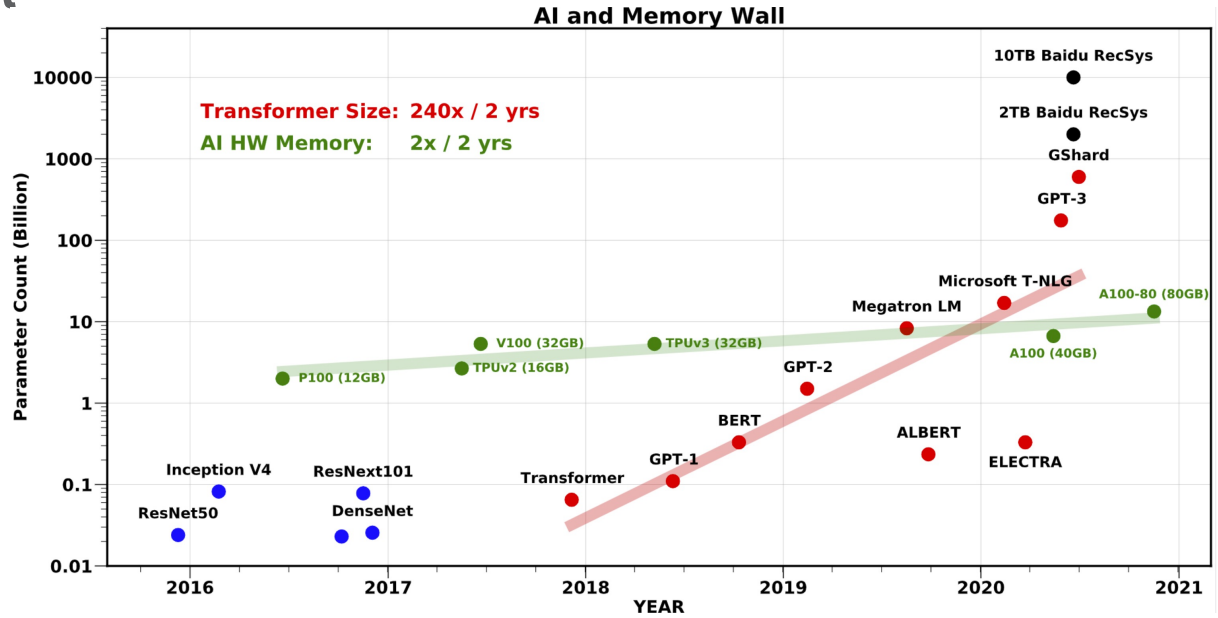
- See usages of Dedicated Memory by OS to augment the General Purpose Memory (e.g. DRAM) available on a system
- Potential usages within OS memory manager
  - Secondary store for standby pages
  - Faster pagefile
  - Secondary store for compressed memory pages

Multi-Tier Memory in Azure

Potential Scenarios of Multi-Tier Memory on Azure

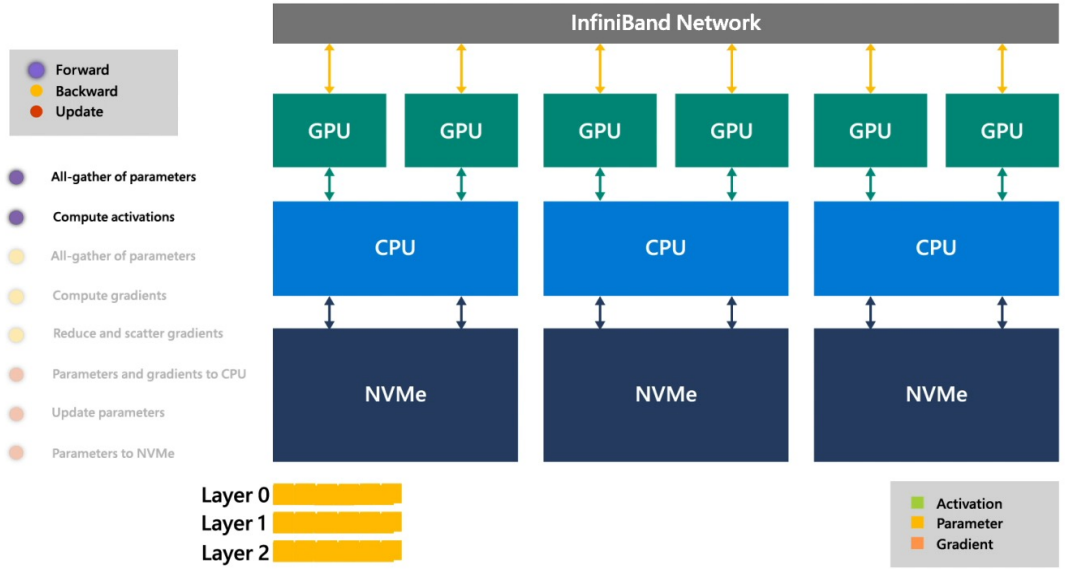
- Zero-Infinity
  - **Microsoft DeepSpeed** [paper] [github] [website]
  - Zero Infinity software stack (Linux support) manages different classes of media
  - No model code refactoring is required to run on Zero-Infinity
- Far Memory OS Pagefile
  - Studied with **Microsoft Research Asia**
  - No application code refactoring is required to run on OS

Zero-Infinity: AI Model Hitting the Memory Wall



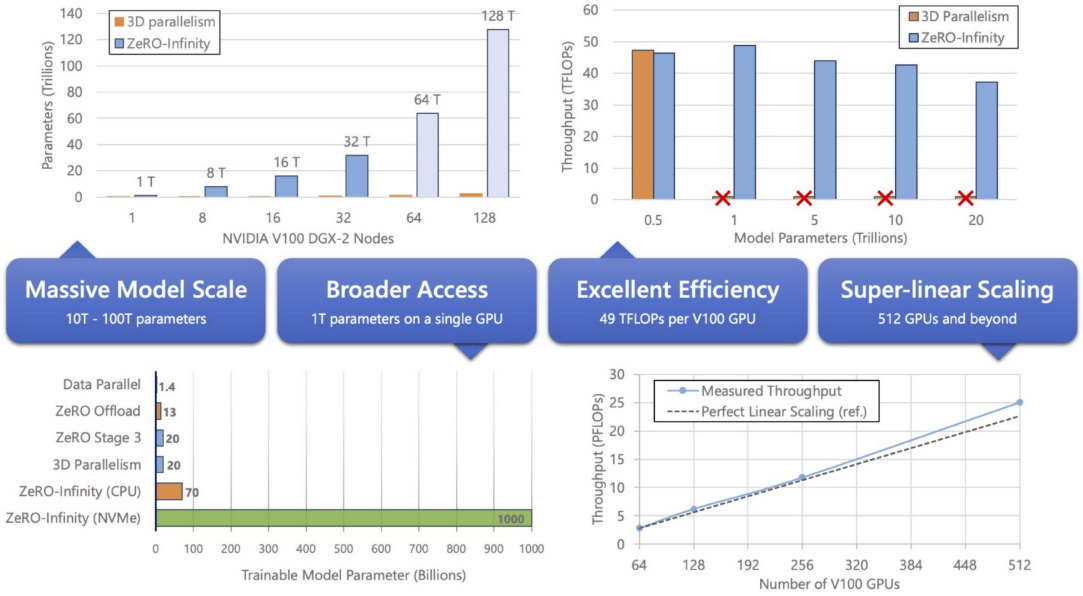
GPU memory size can't scale with AI model, especially for the natural language training. Opportunity to utilize another tier of ***larger media that can be slower and cheaper*** *[Gholami et. al]*

Zero-Infinity: High-Level Architecture



Zero-Infinity breaks down training jobs, schedules and moves data across different classes of media (GPU memory, CPU memory, and NVMe). No model code refactoring is required.

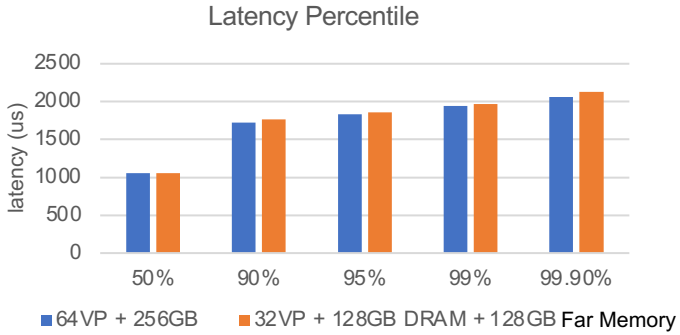
Zero-Infinity: Scale AI Model Linearly



*There is potential to use **far memory** as additional class(es) of media in Zero-Infinity*

Far Memory OS Pagefile

- Better COGs by replacing DRAM with far memory (cheaper & slower)
- Sample workload is a pure memory application for Machine Learning
  - Heavily rely on memory caching
- Use Windows OS to swap cold memory pages in DRAM to far memory
  - Transparent to application (no software change)
- Workload Profile
  - 32 threads; memory committed ~**150** GB
- System Configuration
  - 32 VCPU 128GB DRAM + 128GB far-memory OS Pagefile
    - **17%** working set on pagefile
  - 64 VCPU 256GB DRAM



Summary

- Systems with multi-tier memory are coming
- Work is needed to take advantage of this new hardware trend
- Azure is interested in using multi-tier memory to optimize existing services and create new services for our customers


OCP
TOTAL SOLUTIONS

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA