OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

November 8, 2021 | San Jose, CA

RUNTIME MANAGEMENT OF THE COMPOSABLE MEMORY

Alexander Branover, AMD Boston
Nathan Kalyanasundharam, AMD Santa Clara

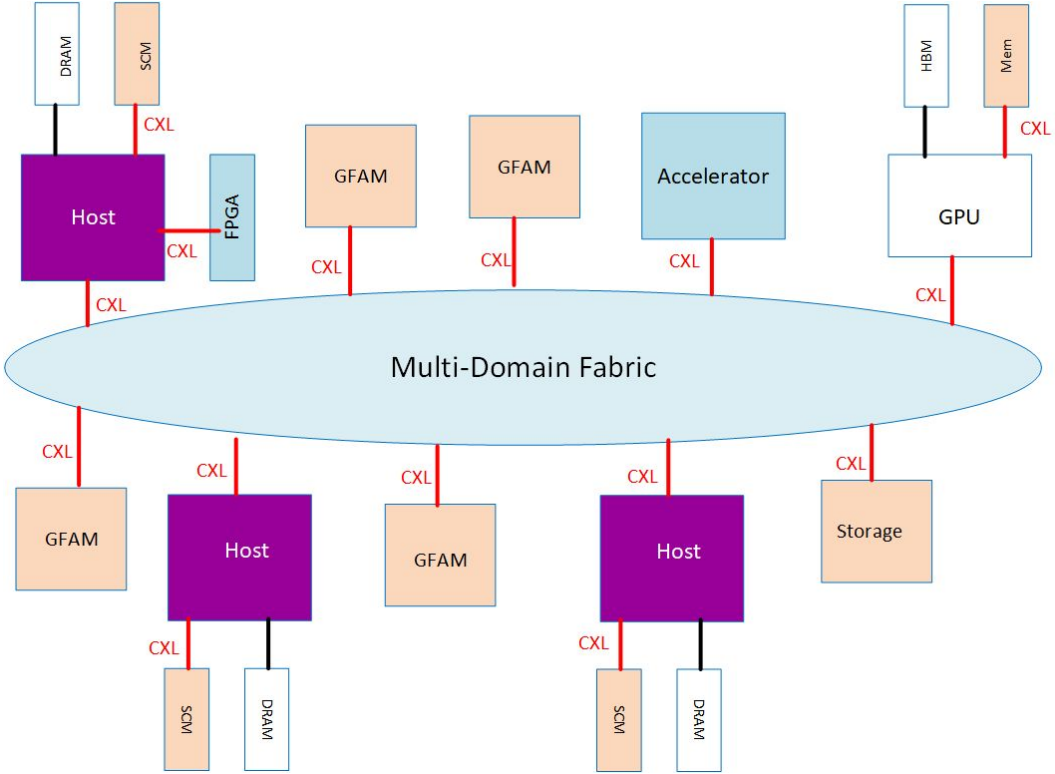
AGENDA

- Paradigm Shift and Memory Composability Progression
- Runtime Memory Management
- Tiered Memory
  - NUMA domains and Page Migration
- Multi-Type Memory Management
  - Persistent and Combined Memory/Storage operation
- Runtime Memory Pooling and Borrowing

PARADIGM SHIFT

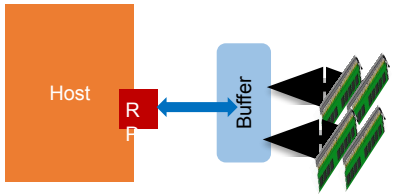
- Scalable, high-speed CXL™ Interconnect and PIM (Processing in Memory) contribute to the paradigm shift in memory intensive computations

- Efficiency Boost of the next generation data center
  - Management of the Host/Accelerator subsystems combined with the terabytes of the Fabric Attached Memory
  - Reduced complexity of the SW stack combined with direct access to multiple memory technologies



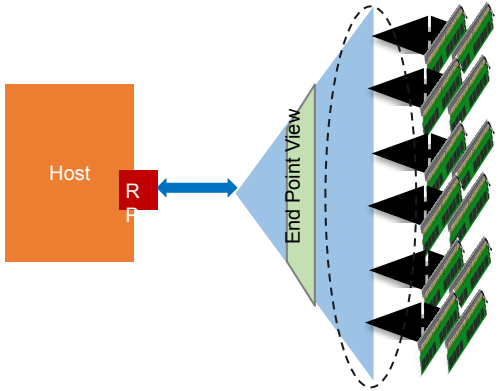
MEMORY COMPOSABILITY PROGRESSION

Mem Direct Attach

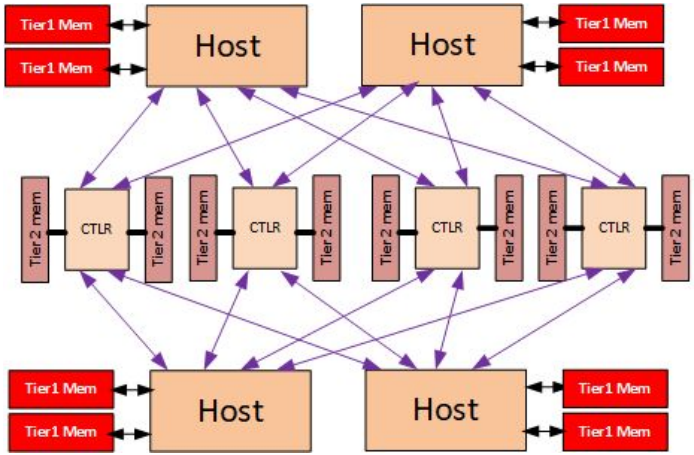


- Workloads/ applications benefiting from memory capacity
- Design optimization for {BW/$, Memory Capacity/$, BW/core}

Memory Scale-Out



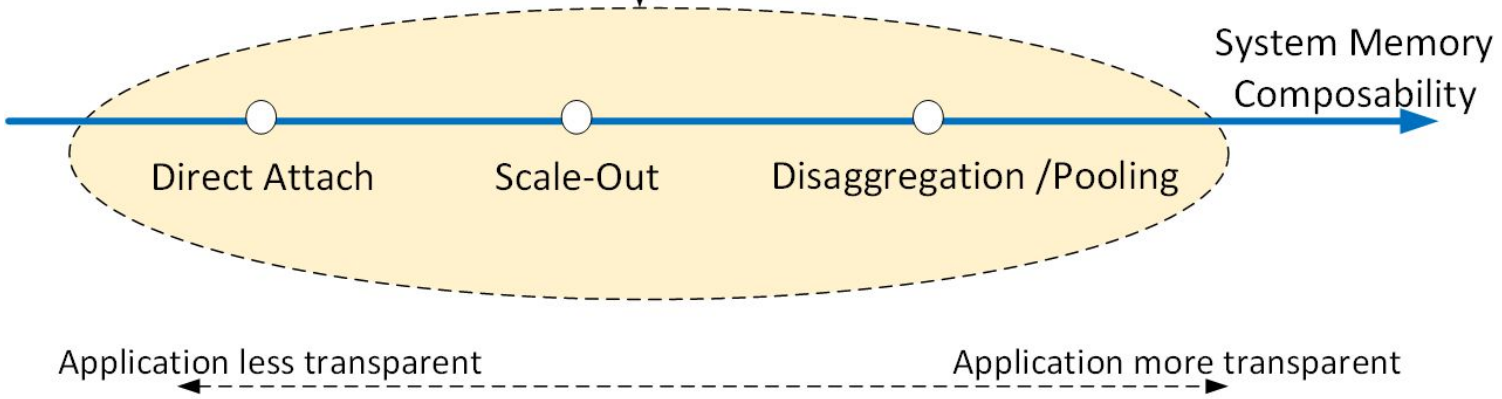
Mem Pooling & Disaggregation



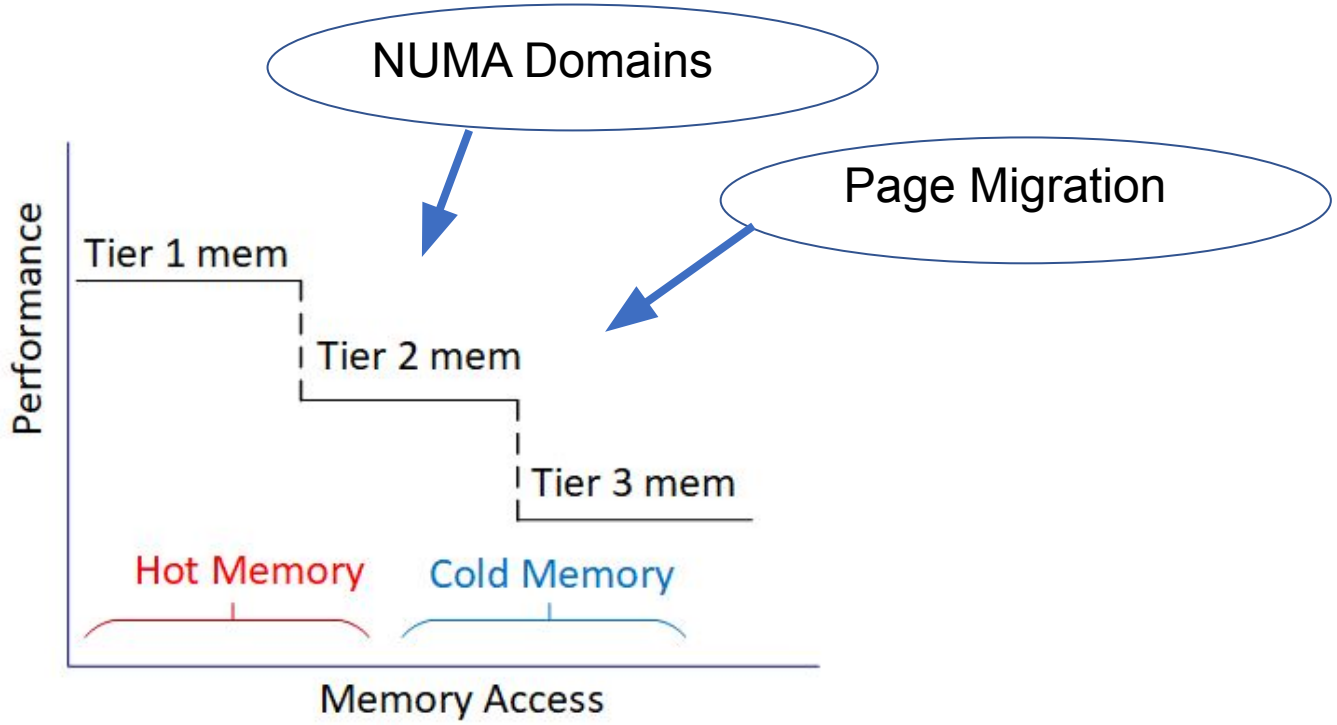
- Addresses the cost and underutilization of the memory
- Multi-domain Pooled Memory - memory in the pool is allocated/ released when required

RUNTIME MEMORY MANAGEMENT

- Runtime Management Approaches:
  - Memory Tiering / Page Migration
  - Multi-Type Memory Management
  - Runtime Memory Borrowing/Allocation



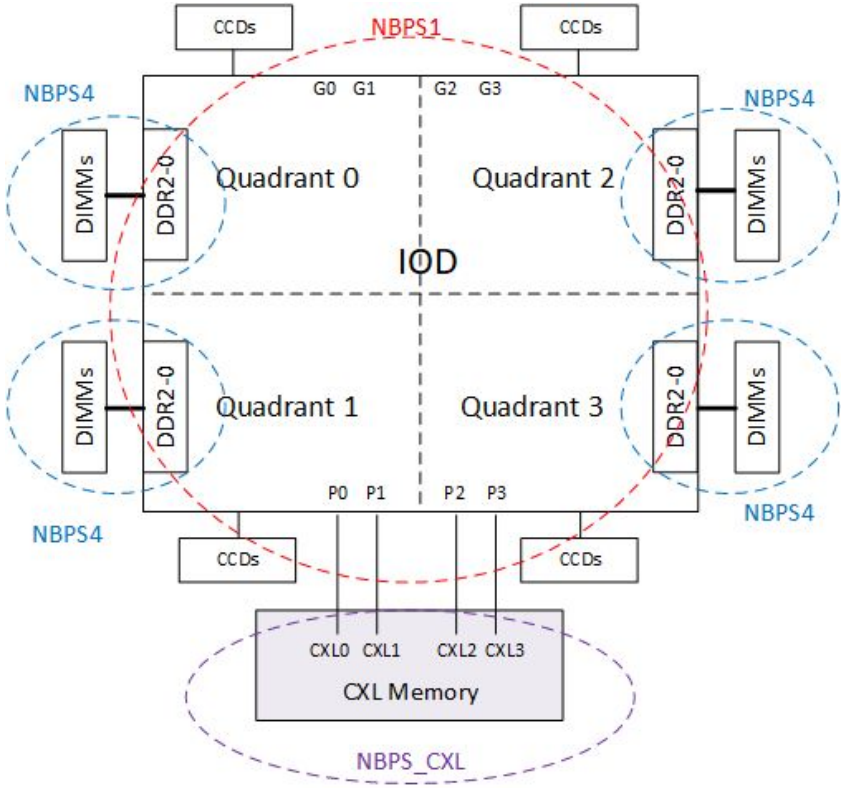
TIERED MEMORY



TIERED MEMORY

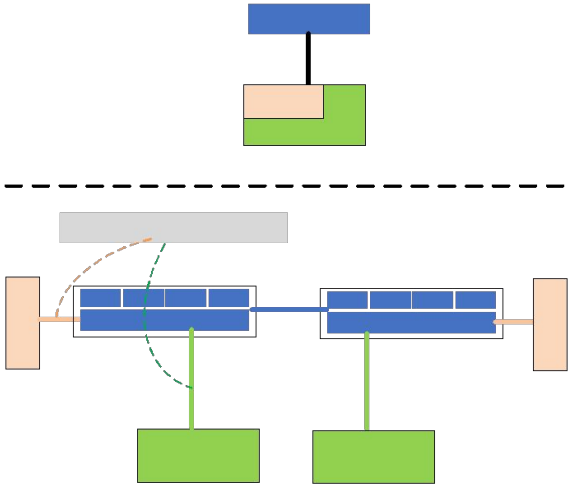
NUMA DOMAINS

- Exposed to the HV, Guest OS, Apps
- OS-assisted optimization of the memory subsystem
- Base on ACPI objects - SRAT/SLIT/HMAT



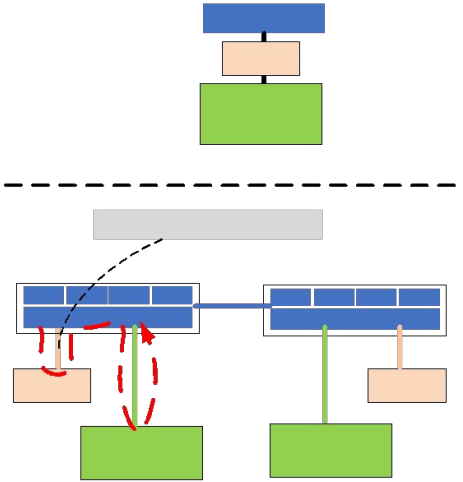
TIERED MEMORY

PAGE MIGRATION



SW Assisted Page Migration

- Active page migration between Far and Near memories
- HV/Guest migrates hot pages into Near Mem and retire cold pages into Far Mem
- Focused DMA to transfer required datasets from the Far to Near Mem



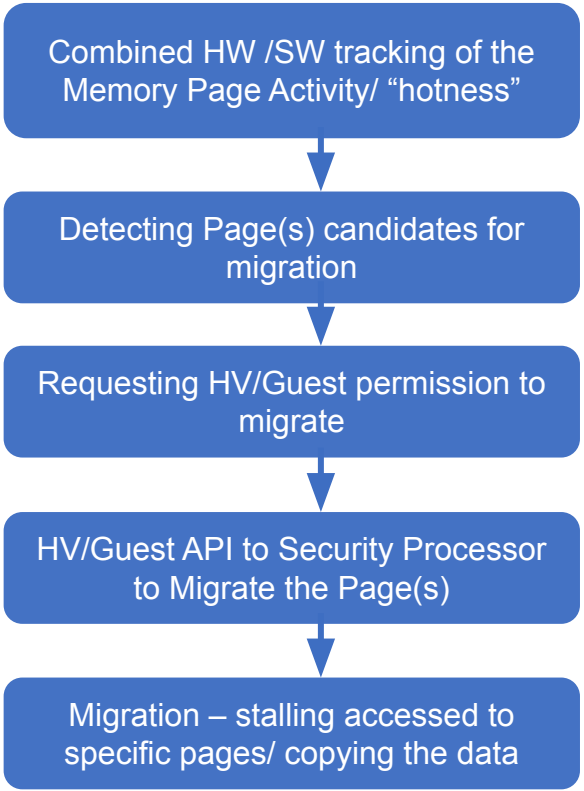
DRAM as a cache optimization

- HW managed Hot Dataset
- Near Mem Miss redirected to the Far Mem
- App/ HV unawareness

TIERED MEMORY

SW ASSISTED PAGE MIGRATION

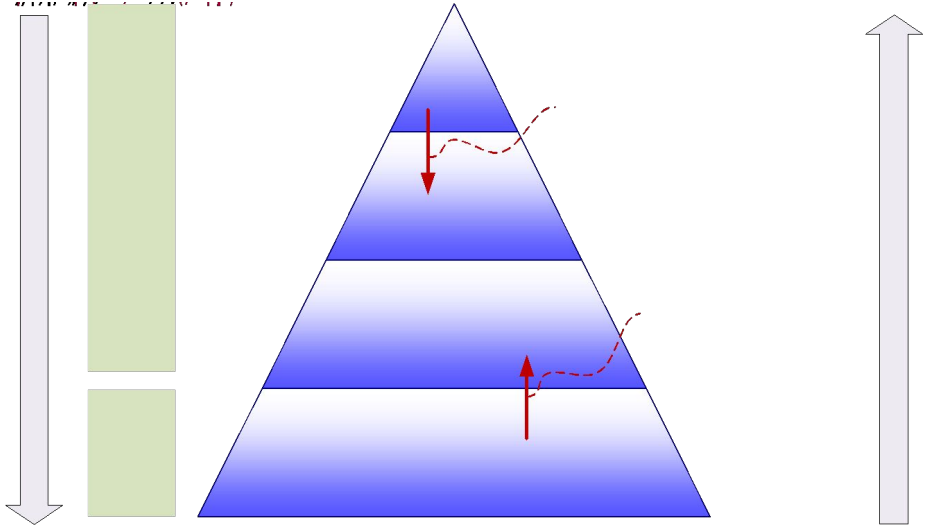
- Page “hotness” –combined action of the HW and SW tracking
- HV/Guest authorization of the migration
- Security Processor as a root of trust for performing the migration



MULTI-TYPE MEMORY MANAGEMENT

PERSISTENT MEMORY

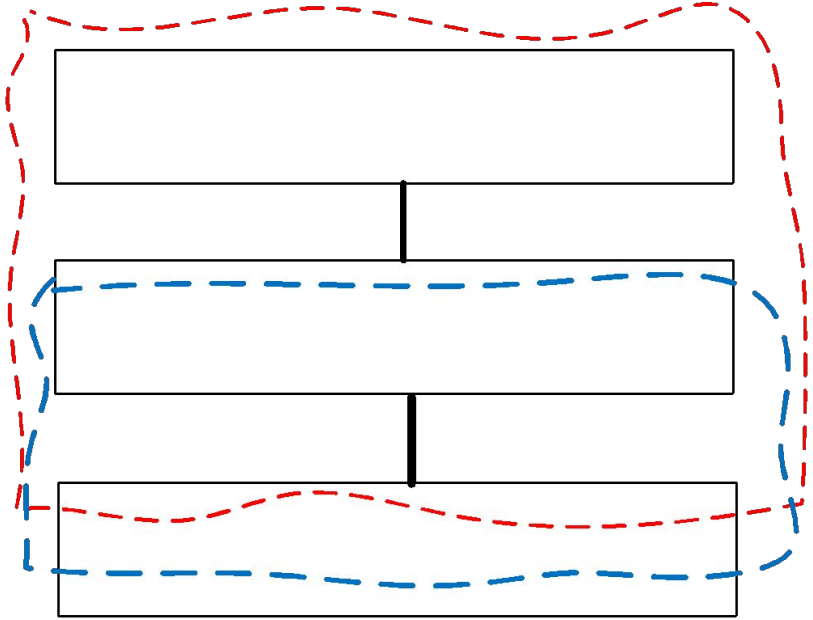
- PMEM aware apps (reduced data movement for power fail recovery, DB load time, etc.)
- Instant-on computer systems with persistent state
- Fully memory-mapped systems (no storage IO protocols)



MULTI-TYPE MEMORY MANAGEMENT

PERSISTENT MEMORY

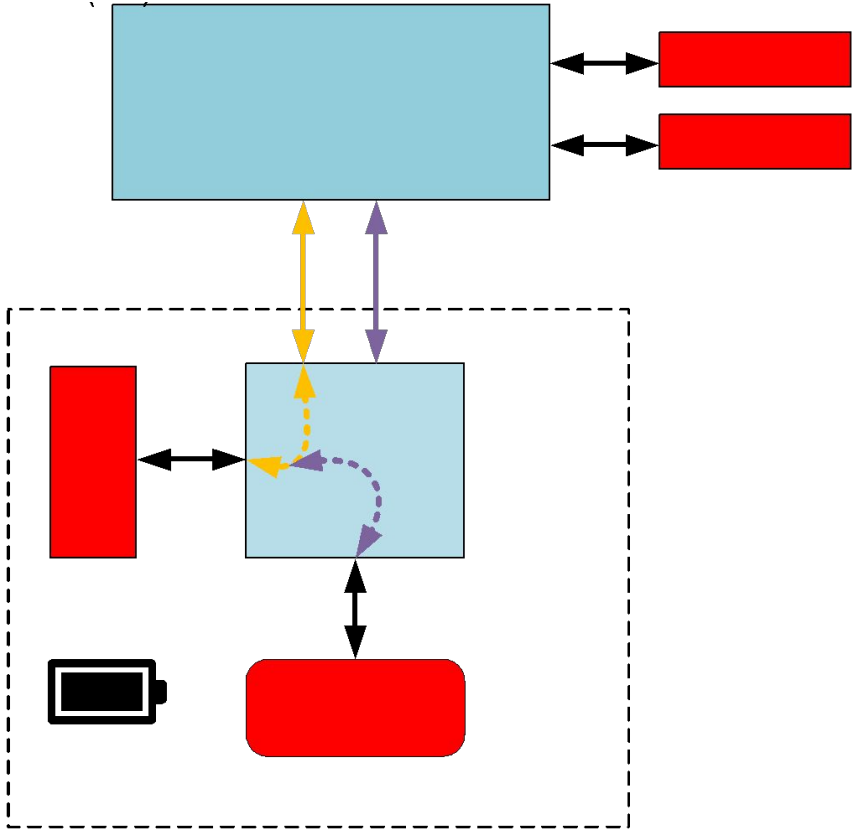
- Global Persistent Data Flush (GPF)
  - In the event of imminent power failure/hazardous events
  - Associated with Enhanced Endurance Region
  - No application awareness
- Basic Persistent Flush (BPF)
  - Limited to Basic Endurance Region
  - Applicable to the systems with limited hold-up
  - Requires SW involvement in periodic data flush to endurance



MULTI-TYPE MEMORY MANAGEMENT

COMBINED MEMORY/STORAGE OPERATION

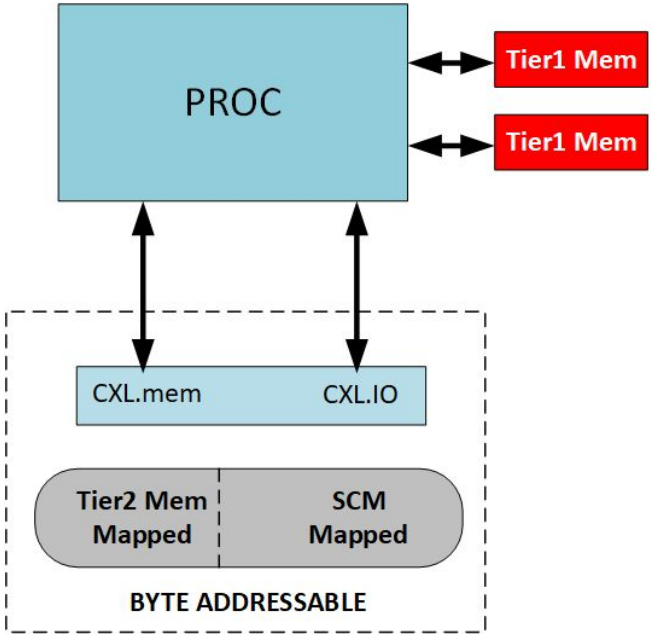
- Tier2 to NVM Data Transfer on power failure under battery power
- No data loss in the event of power failure while achieving full memory throughout (with Tier1/ Tier2) during normal operation
- Battery hold-up control to minimize the system power
- Dirty page tracking to speed up the Tier 2-> NVM data transfer



MULTI-TYPE MEMORY MANAGEMENT

COMBINED MEMORY / STORAGE OPERATION

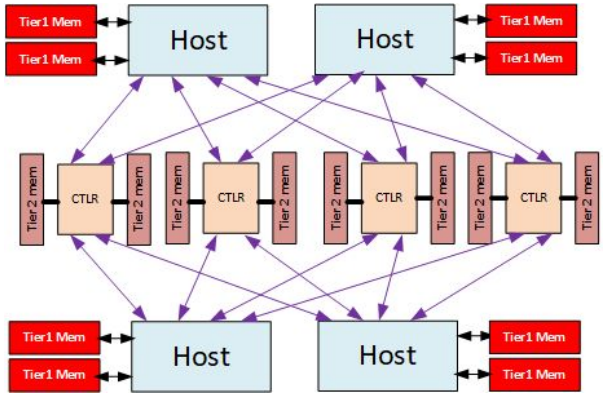
- High-performance storage on memory bus with flexible access (64B to 512B)
- Design scalability for {density, power, latency, pricing} through SW based memory tiering selection
- Diskless Server with Storage disaggregation
- Leverages existing software paradigm with application transparent integration (does not require persistent memory ecosystem)
- On-demand page re-mapping between Storage and Memory with zero DMA

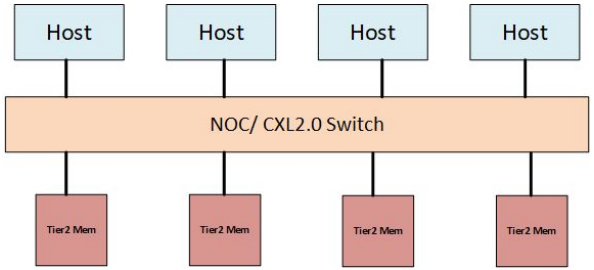


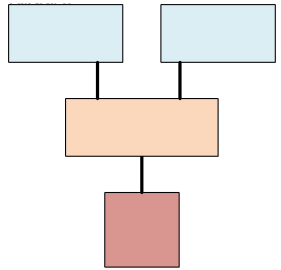
RUNTIME MEMORY ALLOCATION /POOLING

FABRIC ATTACHED MEMORY

- Multiple structures serve for fabric level memory pooling/ borrowing
- Combination of the private (dedicated to specific host) and shareable memory ranges
- Protection of the memory regions from unauthorized guests and hypervisor
- Borrowing/ Lending of the memory ranges between Hosts is regulated by the fabric aware SW layer (i.e., Fabric Manager)



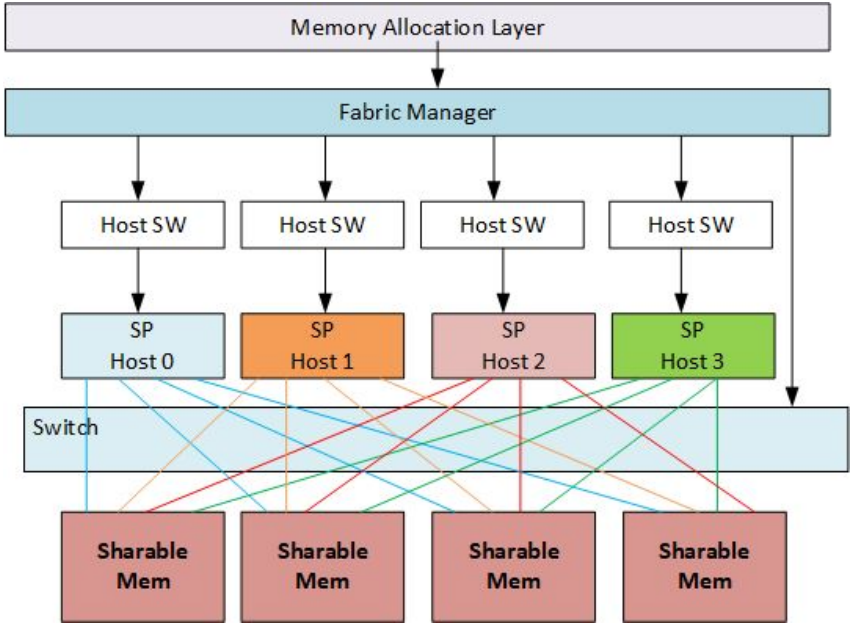




RUNTIME MEMORY ALLOCATION /POOLING

FABRIC ATTACHED MEMORY

- Memory Allocation Layer – communicates <new memory allocation per Host> based on the system/apps needs
- Fabric Manager – adjusts the fabric settings and communicates new memory allocations to the Host SW
- Host SW - Invokes Hot Add/Hot Removal method to increase/ reduce (or offline) an amount of memory allocated to the Host
  - In some instances, Host SW can directly invoke SP to adjust the memory size allocated to the Host
- On–die Security Processor (Root of Trust) is involved in securing an exclusive access to the memory range



SUMMARY

- Composable Disaggregated Memory is the key approach to address the cost and underutilization of the System Memory
- Further investment in the Runtime Management of the Composable & Multi-Type memory structures is required to maximize the system level performance across multiple use-cases
- Application Transparency is another goal of efficient Runtime Management by abstracting away an underlying fabric/memory infrastructure

Disclaimer & Attribution

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION


OCP
TOTAL SOLUTIONS

OCP FUTURE TECHNOLOGIES SYMPOSIUM

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


AMD