



Flash Memory Summit

HPC and Remote Persistent Memory

A Few Thoughts on HPC Usage

Jim Harrell



Flash Memory Summit

Agenda

- A few Observations on Persistent Memory
- High Level Architectures
- Programming Model Considerations
 - Flat Memory
 - Object Storage
- Applications
- Alternatives

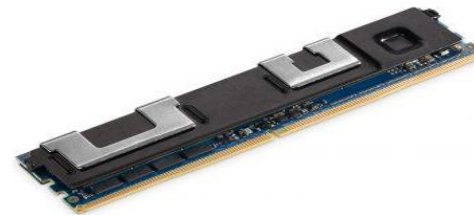


**Which came first,
the technology chicken?**

or the application egg?

Emerging Technologies

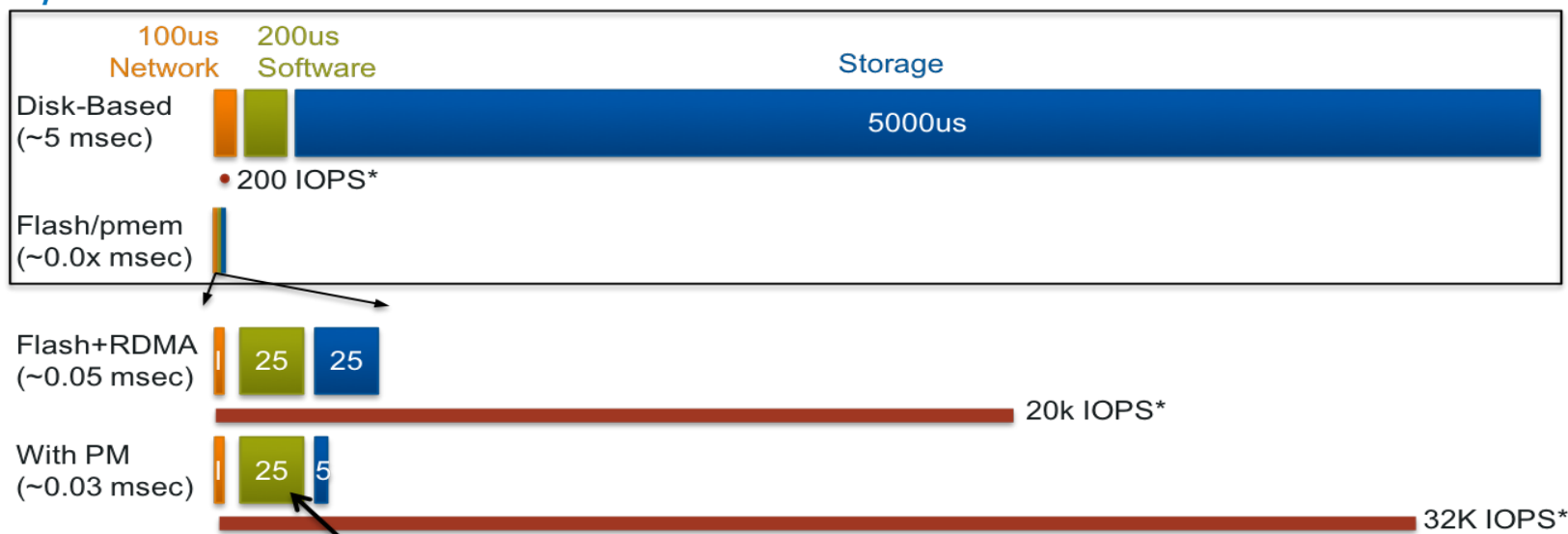
- Multiple NVRAM technologies nearing commercialization
 - Phase Change (PCM): a middle ground between DRAM and Flash
 - MRAM: DRAM replacement? density past 8 Gb, lower idle power
 - ReRAM: Flash replacement? High density, better endurance
 - CNTRAM: Carbon Nanotube based memory – another DRAM replacement?



An Optane DIMM prototype. Source: Intel



New Devices – Challenging DRAM and Flash



Software becomes a significant fraction of latency using Solid state storage, even with 4x improved efficiency

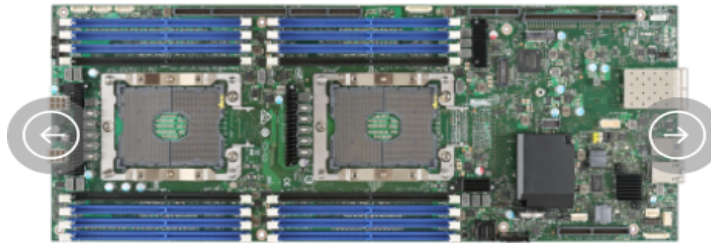
* Max potential 1-thread random sector



Flash Memory Summit

Local Server Configuration

- Simplest configuration is to add Persistent Memory (PM) to a server
 - Limits of about 1.5TB DDR per socket (max)
 - Expect <3TB with a mix of DDR and PM
 - Limits of about 8 sockets to a server – max PM <24TB
 - Is PM a memory extender or a new hierarchy?



- What is your problem size?



Flash Memory Summit

Clusters

- Customers could use clusters with Persistent Memory
 - Would probably require IB or OPA for features – RDMA and Atomics
 - Atomics are mostly a future capability
 - Scalability is limited
 - Slower access than to local PM
 - Cluster PM preferred by a DoE lab
 - Sharing issues across jobs
 - similar to HDFS
 - Security of data between jobs?
- Who is the data for?





Flash Memory Summit

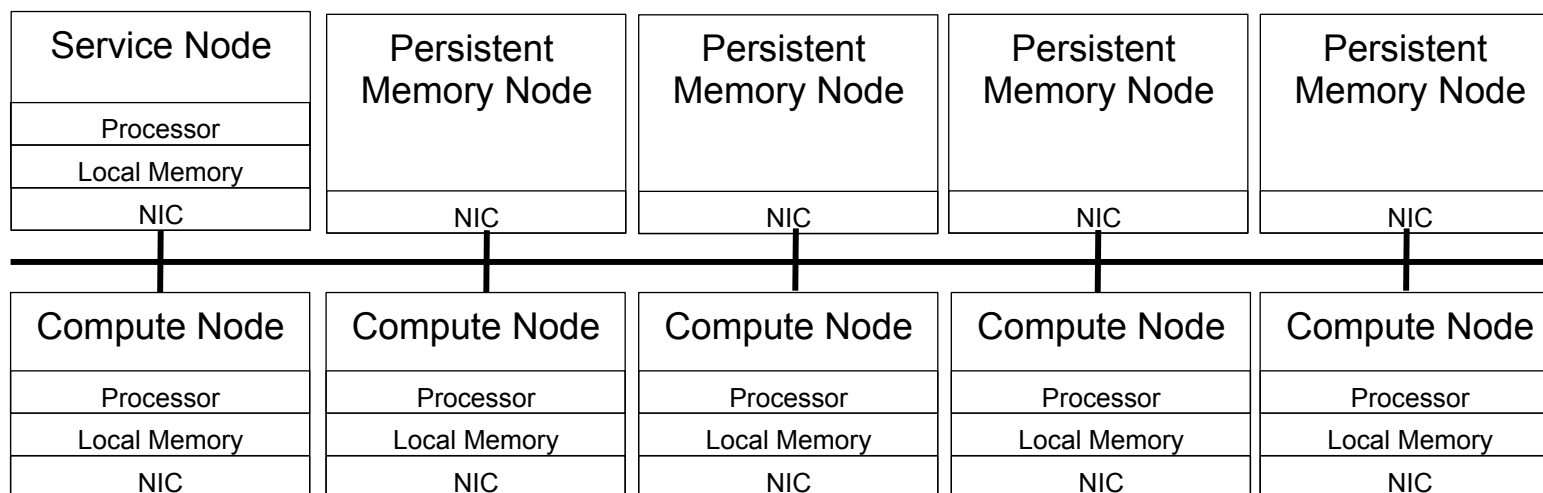
Persistent Memory as a System Resource

- Use servers or nodes as memory stores
- Model is the same as a disk server
- Several vendors using this model for flash including Cray
 - Block Stores, Filesystems, Object Stores, Key/Value Stores
 - Sharing is an issue – often like remote SCSI instead of shared resource (NVMeoF)
- Hardware can scale to fit problem size
- Works for Clusters and HPC systems
- Persistent Global Memory (PGM)
- Performance limited by *the network*
 - Network latency
 - Network bandwidth
 - PM latency – mostly hidden?
 - Software?



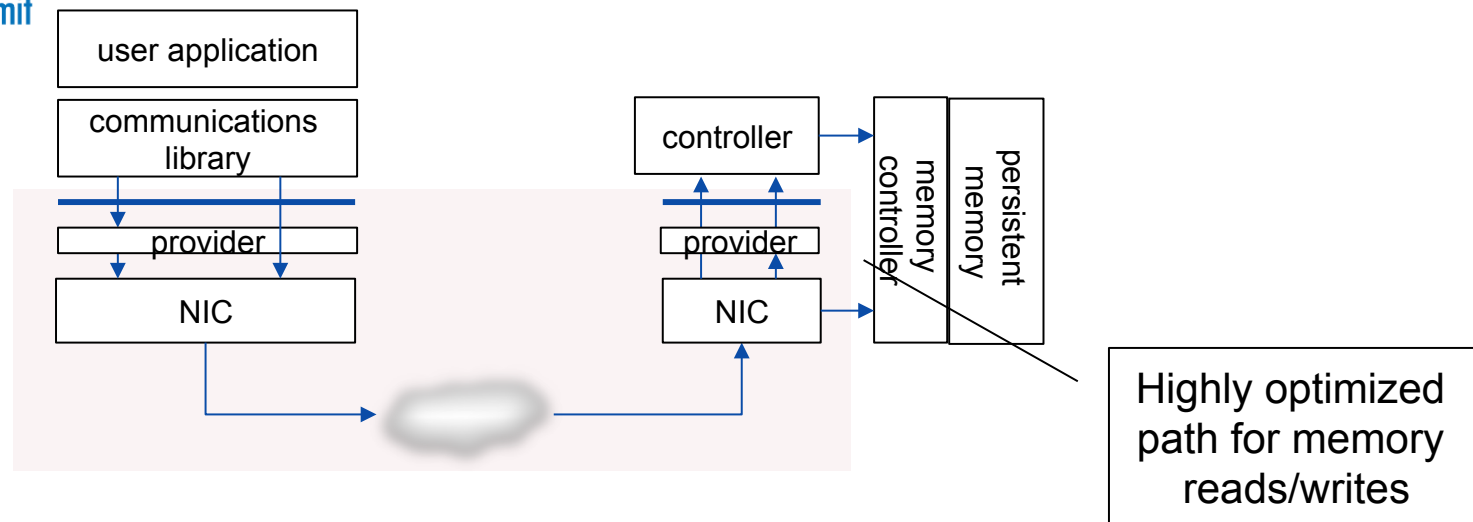


System Architecture



- This basic architecture is the same HPC architecture used since MPPs
- Here persistent memory nodes are part of the system and are managed as a system resource – fits nicely as a persistent resource
- Scalable, ..

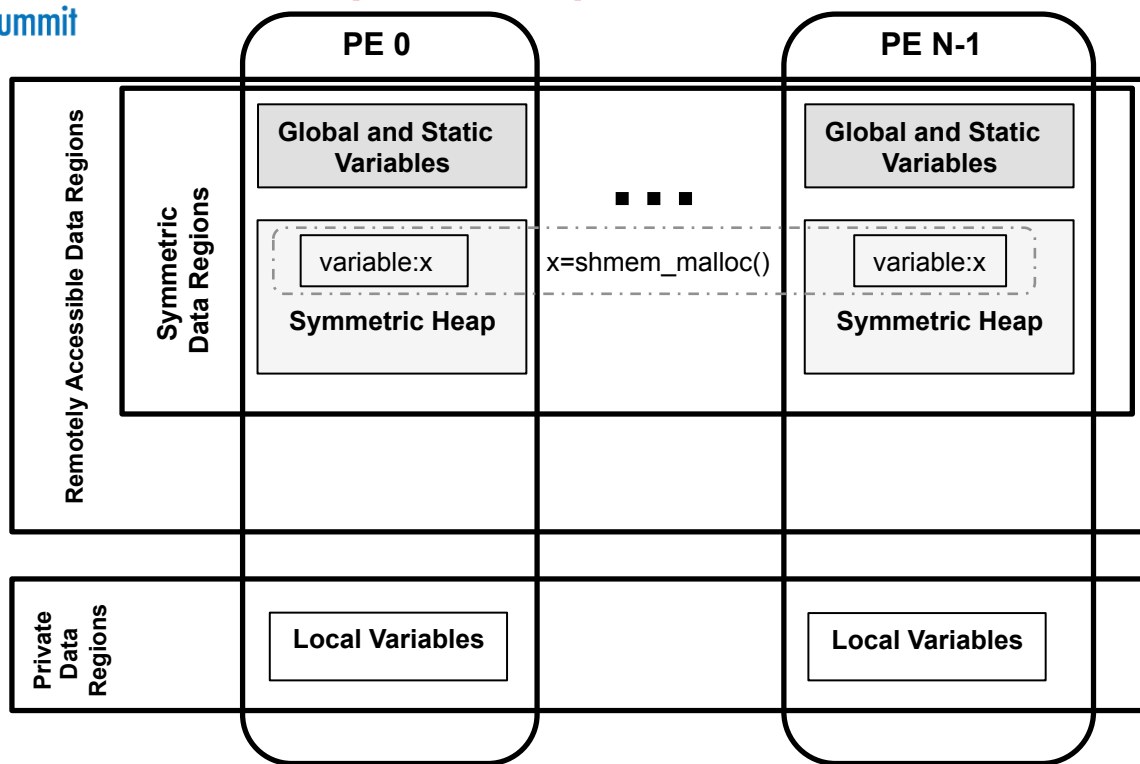
Application View of Persistent Global Memory



- No software interference with application remote memory access
- Applications have high performance PGM as “flat memory”

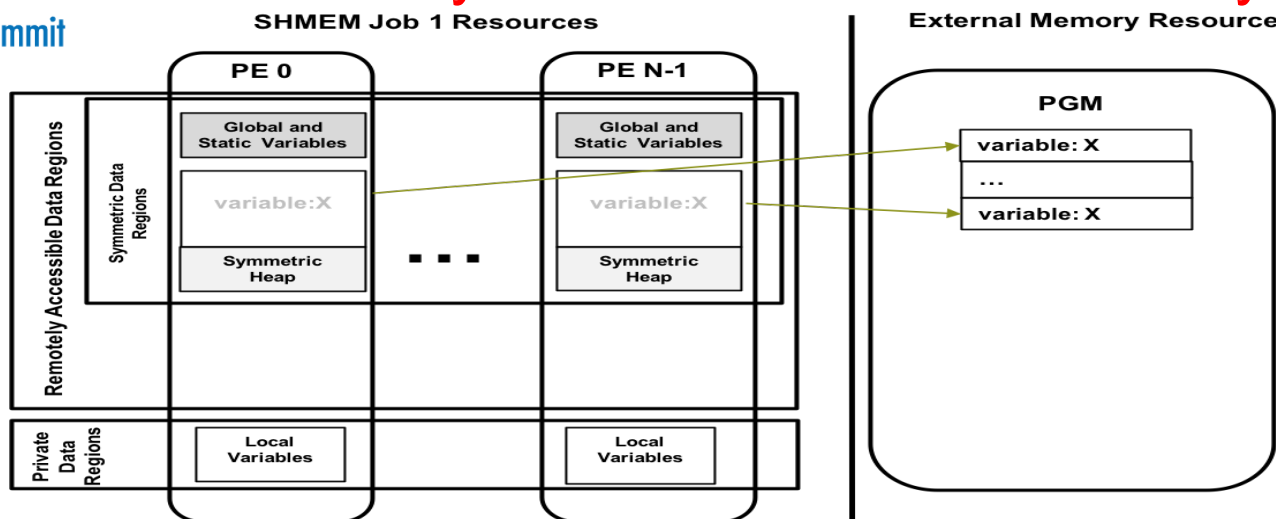


Example - OpenSHMEM Memory Model





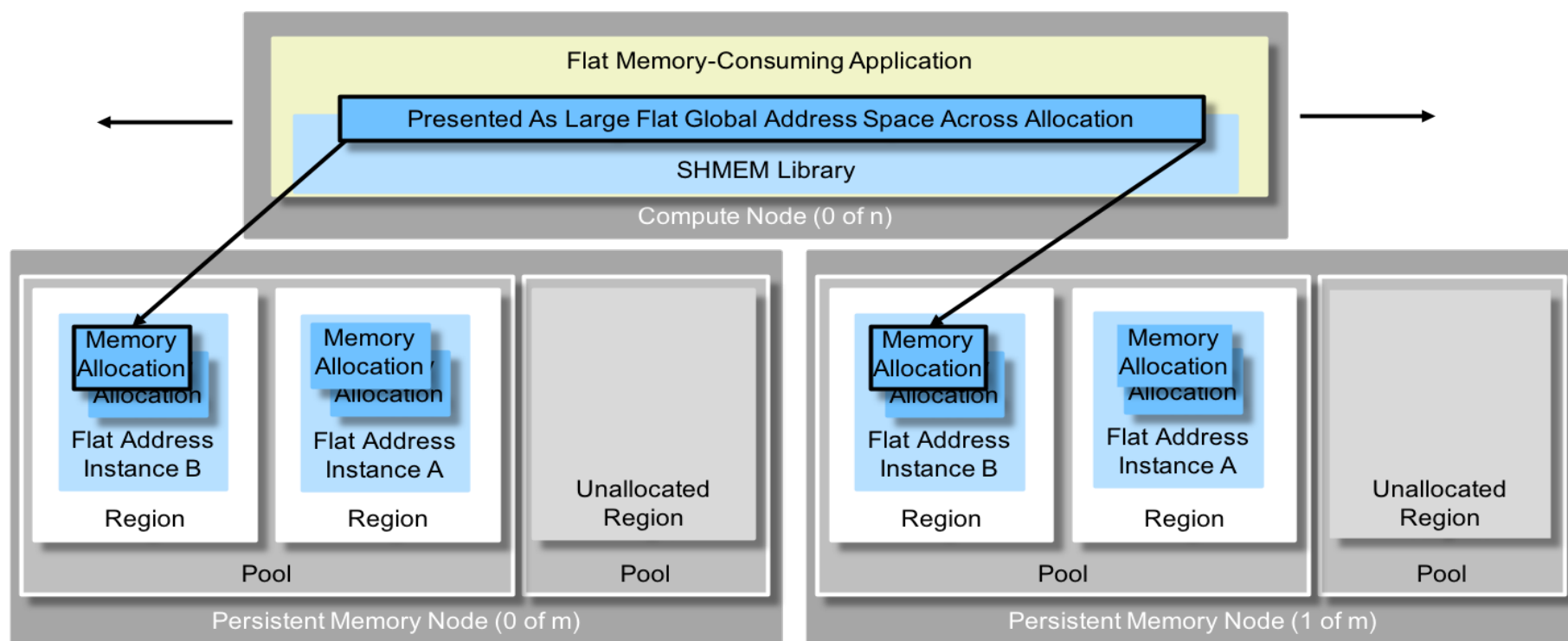
PGM - Cray SHMEM Extended Memory Model



- Symmetric heap partition backed by PGM
- Shared PGM regions
- SHMEM resiliency, enhanced error handling, and error reporting
- Loosely coupled jobs (SHMEM/non-SHMEM)



High Level View of Flat Memory Model





Flash Memory Summit

Productivity and PGM

- Using Python and Persistent Global Memory for Network Intrusion Detection
- Compared against a standard SQLAlchemy and PostgreSQL database implementation
- Python implementation was 5 to 11X faster
- Python showed pointer chasing performance similar to UPC

Is the World of Memory Just Flat?





Flash Memory Summit

Rationale/Motivation for Alternatives to Flat Memory

- Driving Needs
 - Persistent memory needs to be named, to be shared between applications
 - Data structures that are shared need to be secured
 - Persistence at scale requires resiliency and recovery
 - High-level language programmers prefer structures over flat memory spaces
- Point of convergence for memory and storage concepts
 - Memory throughput, latency, and access granularity
 - Storage-like robustness and data management
 - Highly concurrent access from many uncoordinated application instances
 - Eventual consistency via versioning coupled with triggered snapshots
- On-ramp to future programming models
 - Build complex, sharable data structures
 - Users shape data like their science, not the system
 - Extensible for data flows, streams, triggers, procedures, tiering to cloud

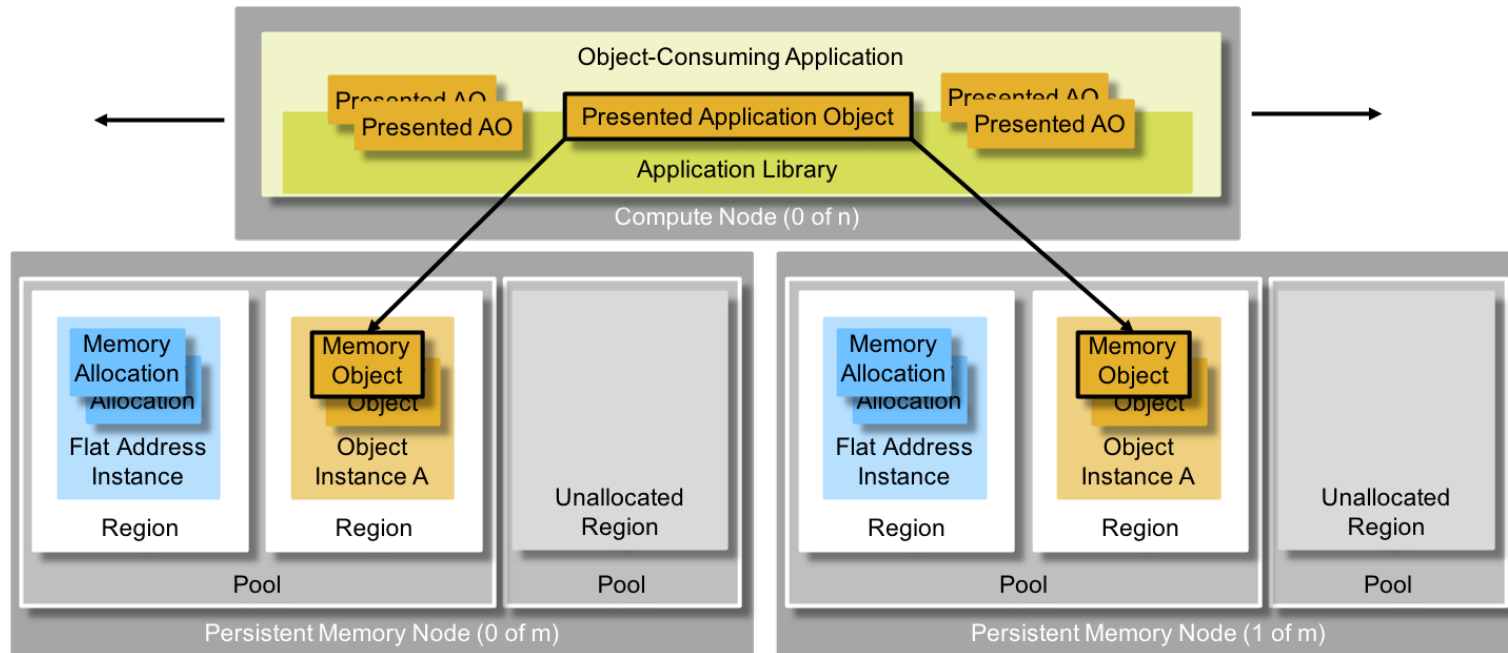


Flash Memory Summit

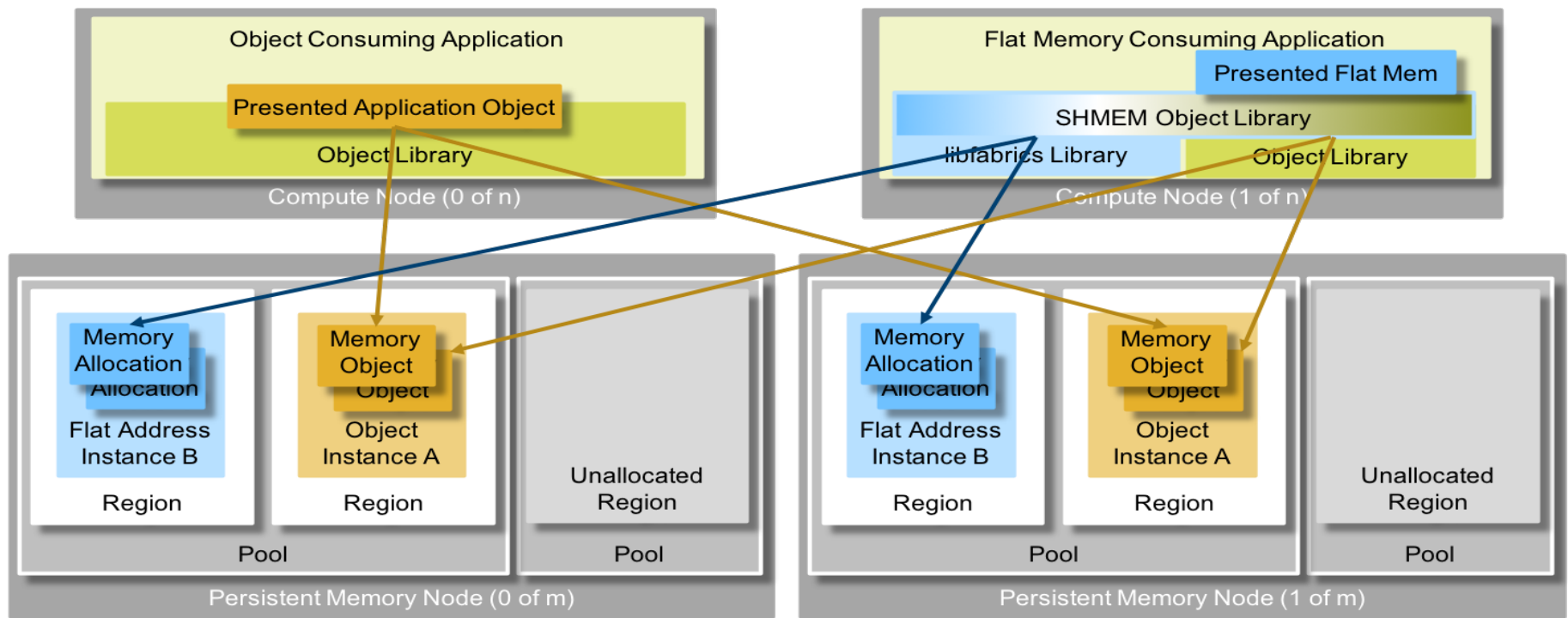
Potential Use Cases

- Analysts querying a common database
- Data trickling into pool of objects
- Data being expunged based on age or low priority
- Triggers / alerts / events on some key operations
- Association of related data through collections
- Express complex structures, groups of objects
- Optional use as base Key/Value store
- Optional use as simplified files

Memory Objects and Application Objects



Memory & Application Objects, Emulating Flat Memory





Flash Memory Summit

What Applications will use Persistent Memory?

- Applications that have huge memory needs
 - Where scaling compute nodes to add memory is not economical
 - Applications that do lots of small memory accesses - GUPS
 - Scaling memory to avoid disk access
- In Memory Databases
 - Currently application driven
 - Look for newer DB conversions to use PM in 2020++
- Productivity Improvements
 - Use the same tools on laptops, clusters, and HPC systems
 - SHMEM – extensible scaling, resiliency in SHMEM...






Flash Memory Summit

Alternatives to Persistent Global Memory?

- Not DDR – too expensive and no persistence – no real case for remote DDR memory server
- Flash – cheaper, available, ... , It will be in all systems anyway
- Why isn't anyone else working on PGM in HPC?
 - It isn't available yet and will roll out slowly – realistically 2020++
 - Current products aimed at single servers – now
 - Flash is helping most applications
 - Flash provides big improvements for small I/Os especially coupled with caching
 - Flash products are aimed at single Key/Value and Object stores – like NVMeoF
 - Nobody is thinking about many nodes of shared memory because they don't have the network atomics in their applications or they don't have Graph DB style applications that are embarrassingly parallel
- By 2021/2022 there will be applications looking for PGM – mostly newer applications and newer databases



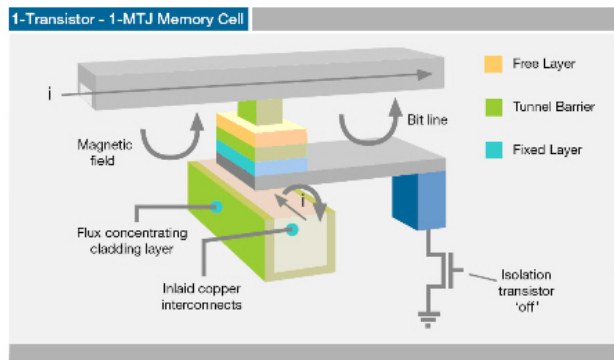
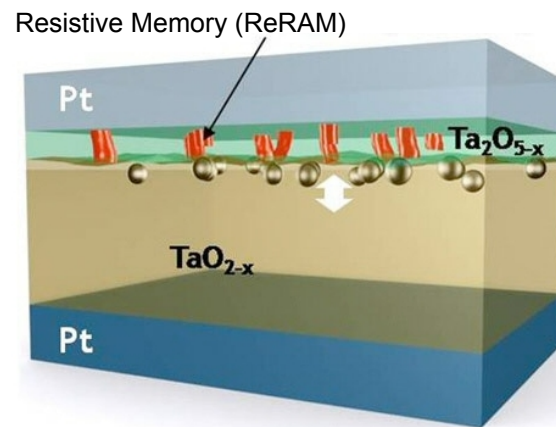
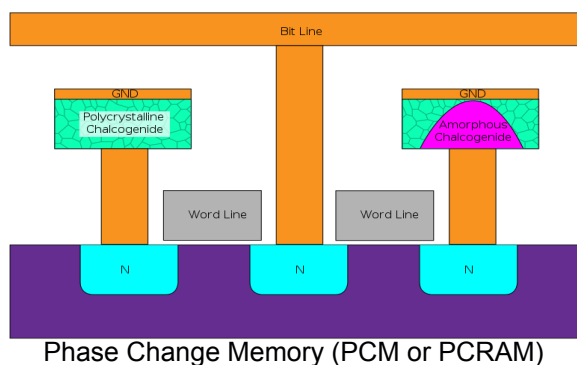
Flash Memory Summit 2018
Persistent Memory Track

FMS Persistent Memory Track Presented by:   

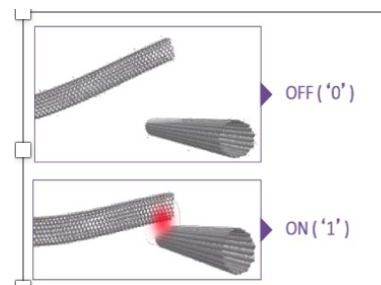


Flash Memory Summit

New Devices – Challenging DRAM and Flash



Spin-Transfer Torque Magneto-resistive Memory (STT-MRAM)



Flash Memory Summit 2018
Santa Clara, CA

FMS Persistent Memory Track Presented by:

