White Paper

Intel® Optane™ persistent memory
Cloud-Native Software

Fault Tolerance and Agility for Long-Running, Stateful Apps in the Cloud

MemVerge Memory Machine™ Cloud Edition enables long-running, stateful applications to use the data persistence capabilities of Intel® Optane™ persistent memory (PMem) without modifications to the software. Certification of this solution for Red Hat® OpenShift® container platform reduces risk and improves time to adoption, helping universalize cloud-native approaches across workloads. The PMem CSI Operator for Kubernetes and OpenShift provides deployment agility, streamlining the path to enterprise adoption.

Table of Contents

Linux containers are instrumental to digital transformation across industries, providing the workload portability and agility needed to take advantage of multi-cloud infrastructures. Packaging microservices with their dependencies in lightweight containers makes it possible for workloads to scale by passing freely among private, public, and hybrid clouds, with central orchestration in real time. Software-defined infrastructure reduces the need for idle system headroom, optimizing resource efficiency and reducing both capital expense (CapEx) and operating expense (OpEx). Containers are also a prerequisite for evolving technical teams toward modern working models such as DevOps, DevSecOps, and continuous integration/continuous delivery (CI/CD).

These forces will continue to expand the role of containers in software architectures for the foreseeable future. The first wave of container adoption focused on relatively simple cases, namely short-running, stateless applications that don't require storage on the host. Extending containerization to support stateful, long-running applications introduces challenges. Keeping track of session data on the host requires persistent storage, and traditional block storage is too slow to provide an excellent user experience while supporting the inherent requirements of containers to be portable and to quickly spin up and down on demand.

Software accesses Intel Optane PMem, shown in Figure 1, as persistent storage using the Intel Optane PMem container storage interface (CSI) driver, with dramatically lower latency than block storage. MemVerge Memory Machine builds on that ability by giving unmodified applications access to memory persistence using the Memory Machine Transparent Memory Service. In addition, the MemVerge platform allows instant snapshots of an application's runtime state to be copied among containers for fast restart or recovery. Both the Intel Optane PMem CSI driver and MemVerge Memory Machine Operator are certified for Red Hat OpenShift, streamlining the path to enterprise adoption.

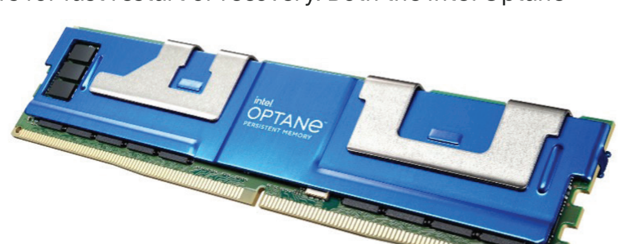

Figure 1. Intel® Optane™ PMem module.

Refactored Memory Hierarchy with Intel® Optane™ PMem

As businesses strive to generate value from dramatically growing data sets, they are limited by how quickly their algorithms can access that data. A key solution to this challenge has been to hold larger amounts of data in active system memory, where compute resources can access it on the memory bus, with far lower latency than reading from and writing to block storage. While this dramatic reduction in I/O overhead can accelerate workloads, increasing DRAM capacities becomes cost-prohibitive with terabyte-scale data. In addition, a system restart can easily lead to a delay of ten minutes or more as the full dataset is read into memory.

To address these challenges, Intel Optane PMem provides a more cost-effective approach to building memory capacity than DRAM, with support for data persistence. Socket-compatible with conventional DRAM, Intel Optane PMem 200 Series is available in 128, 256, and 512 GB DIMM modules. Intel Optane PMem and DRAM deployed side-by-side in a system based on 3rd Generation Intel® Xeon® Scalable processors enable total system memory of up to 6TB per socket.[1] The Intel Optane PMem 200 series provides an average of up to 32 percent more memory bandwidth than the previous generation.[1] In addition to being fast like memory, it is also persistent like storage. Intel Optane PMem spans both the memory and the storage tier, as shown in Figure 2, creating a new paradigm.

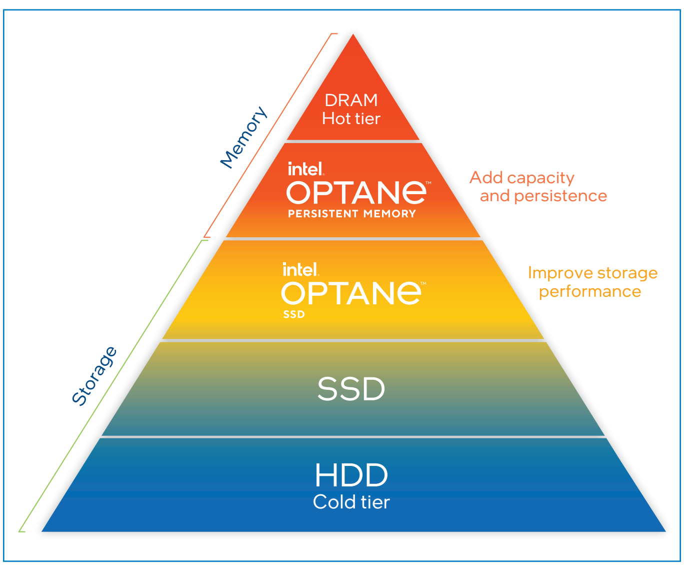

Figure 2. Hierarchical model of memory and storage.

Data is protected at rest in Intel Optane PMem by AES-256 encryption based on hardware-based cipher key processing, which requires no enablement in software. This feature also supports cryptographic erasure of data contents (by changing the key) for protection during module re-provisioning or at end-of-life.

Intel Optane PMem Operating Modes

In order to provide both more cost-effective system memory capacity than DRAM and lower-latency persistent storage than local drives, Intel Optane PMem provides two operating modes, as illustrated in Figure 3. This flexibility enables data center operators to match workloads to specific memory-storage architectures, while modules can be redeployed in one mode or the other as needs change.

- **Memory Mode** enables Intel Optane PMem to act as volatile system memory (without data persistence) to add capacity at significantly lower cost than DRAM, while requiring no changes to applications. Intel Optane PMem and DRAM are installed side by side, with the DRAM acting as a caching tier to optimize performance.
- **App Direct Mode** enables data persistence on Intel Optane PMem for enabled applications, at far lower latency than block storage. MemVerge Memory Machine Operator streamlines deployment of Intel Optane PMem as persistent container storage by removing the need for applications to be modified.

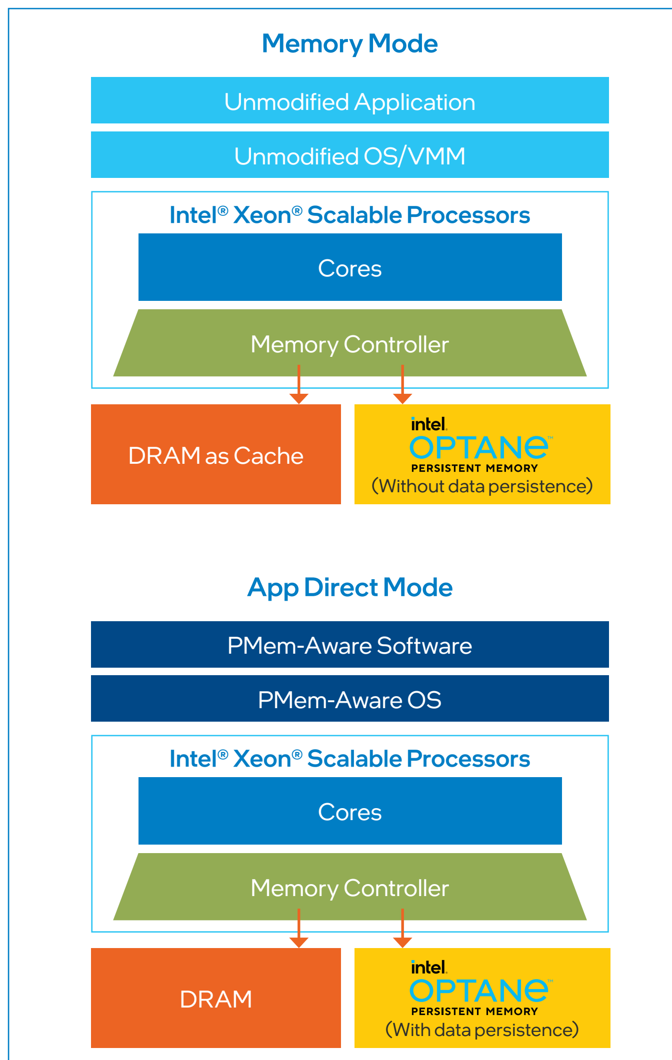

Figure 3. Intel® Optane™ PMem operating modes.

Implementing PMem as Container Storage

The Intel Optane PMem CSI driver exposes Intel Optane PMem to containerized applications in App Direct Mode as filesystem storage. The open source driver can either be deployed manually using YAML files or consumed as a Kubernetes operator. The operator is a software extension to Kubernetes that automates deployment and maintenance of the Intel Optane PMem CSI driver. Applications can then take advantage of high-capacity persistent storage with latency similar to that of DRAM. Capacity-aware scheduling places Kubernetes pods only on nodes with the storage capacity needed. The Intel Optane PMem CSI operator supports three types of volumes:

- **Persistent volumes**. Volumes are independent of the applications that use them, and an application must run on the same node as the volume to make use of it. Data is retained until the volume is deleted.
- **CSI ephemeral volumes**. A new volume is created at application start-up, on the same node as the application. The volume cannot be shared with other applications, and it is deleted when the application stops running.
- **Raw block volumes**. Intel Optane PMem CSI provisions volumes as raw block devices, and pages mapped on the raw block device go through the Linux page cache.

Optimized Support for Big, Fast Data with MemVerge Memory Machine™

MemVerge Memory Machine extends the Intel Optane PMem value proposition by virtualizing DRAM and PMem to create a two-tiered pool of software-defined memory. Memory Machine intelligently and dynamically assigns memory resources to applications, optimizing workload performance, availability, and agility.

The platform acts as a compatibility layer so unmodified applications can immediately benefit from new processors, memory components, interconnects, and other technologies. The service architecture of Memory Machine Cloud Edition is illustrated in Figure 4.

Snapshots and In-Memory Data Services

Memory Machine ZeroIO™ In-memory Snapshots capture application state in memory, with minimal resource and performance overhead. The lightweight nature of this functionality makes it feasible to capture and update snapshots frequently. Snapshots enable application data and state to be stored in PMem devices, supporting high availability in the memory tier, along with the following key software-defined memory services:

- **Time Rollback.** Automatic snapshots taken at intervals enable an application to be rolled back to its state at a previous point in time.
- **AutoSave.** Successive snapshots can protect applications from the impact of a crash, recovering automatically to pre-crash status with minimal disruption.
- **Thin Clone.** Memory Machine can clone an in-memory database for access by DevOps or other applications, with minimal time and resource requirements.
- **App Migration.** Moving a snapshot to another server enables new application instances to be started instantly, complete with existing application state.

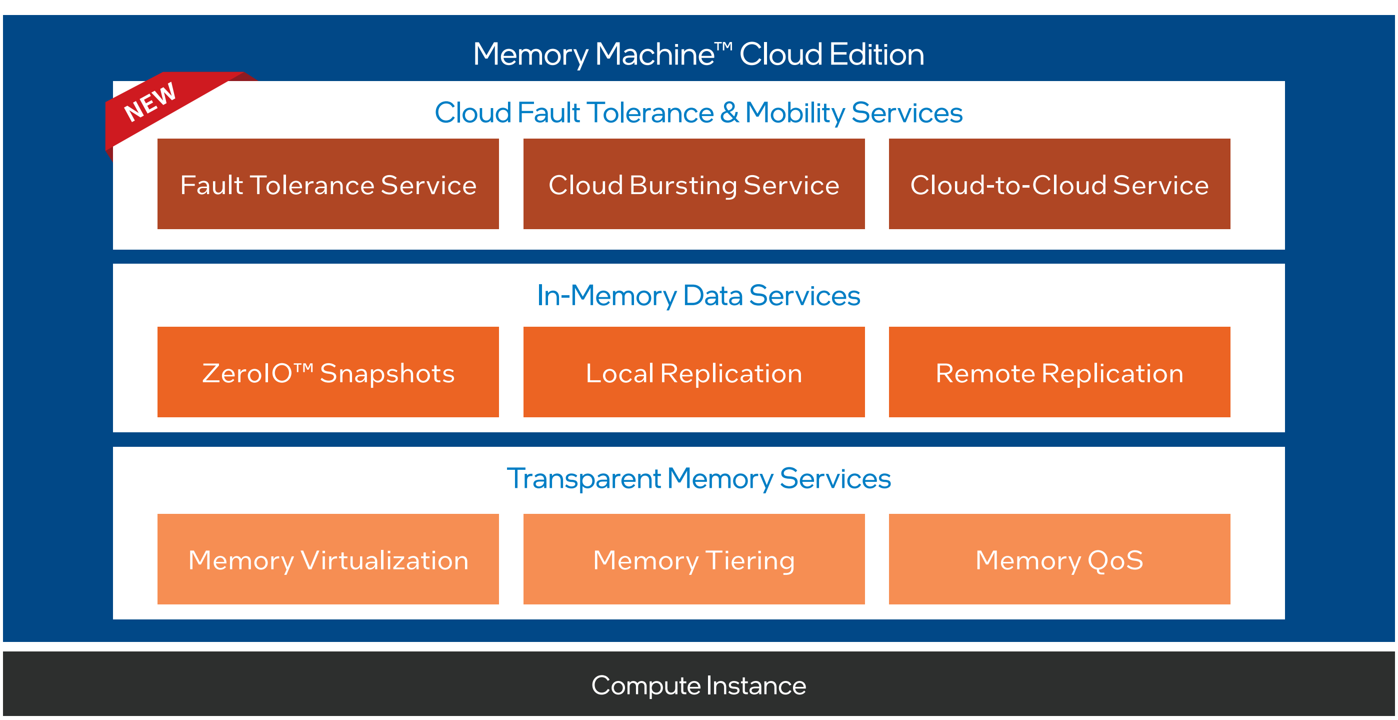

Figure 4. MemVerge Memory Machine™ Cloud Edition service architecture.

Example Use Case: Enabling Long-Running Apps for Low-Cost Spot Instances

Historically, it has been difficult to realize the agility and flexibility advantages of cloud deployment for certain classes of workloads. This has been particularly true for non-fault-tolerant, long-running applications that hold large amounts of stateful data in memory, such as large-scale analytics, HPC simulation and modeling, and genomic sequencing. It has been prohibitively complex and time-consuming with this class of workloads to burst capacity from on-premises resources to public clouds or to move them between clouds, for example.

Moreover, it is untenable for critical applications—especially in industries such as financial services or healthcare—to encounter delays on the order of ten minutes or more to recover after a restart. Memory Machine can load, replicate, recover, and transport workloads at in-memory speeds, as shown in Figure 5. An AppCapsule is created by using ZeroIO snapshot to capture all the application data needed to restart from a given point in time. When the cloud provider reclaims the spot instance, the application automatically avoids the impact of the unplanned instance termination by gracefully recovering and seamlessly resuming on a new instance.

Kubernetes Use Cases with Memory Machine and Intel Optane PMem

In Kubernetes implementations, the combination of Memory Machine and Intel Optane PMem enables cost-effective deployment of more memory per container, accelerating data-intensive workloads such as analytics and machine learning, improving ROI.

Specifying memory characteristics on a per-container basis allows DRAM-to-PMem ratios to be configured according to workload SLA requirements, optimizing the use of available memory resources. Point-in-time container snapshots can provide for instant crash recovery and failure analysis, rolling the production application back to a known good state and then using snapshots to replay and debug the failure.

Big Memory and Memory Machine with Red Hat® OpenShift® Platform

Red Hat OpenShift is built on top of Kubernetes, providing a fully supported, enterprise-ready version of the popular open source container environment for building cloud-native software architectures. The platform is distributed as open source and integrated with Red Hat's software development tools to optimize productivity and support for modern development practices such as DevOps and CI/CD. OpenShift incorporates components from other Red Hat technologies, including Red Hat Enterprise Linux, which help expand functionality, improve usability, and harden the platform for the enterprise. It provides a consistent container deployment and orchestration platform across on-premises, hosted, and public cloud infrastructures.

The Memory Machine Operator will deploy and configure Memory Machine in an OpenShift environment. PMem on each node is configured using the PMem CSI driver. All the benefits that the combination of Memory Machine and PMem provides in a bare-metal or virtual machine scenario are now available in an automated, containerized deployment.

To manage containers, Memory Machine runs in a pod on each container node (see Figure 6) and manages memory in a local PMem-based persistent volume (PV). A DaemonSet is used to create one Memory Machine pod on each node. This pod runs the Memory Machine memory manager. A local, shared PV is created for each application pod. These shared PVs mount the same directory on the host node, so all pods on the node can access Memory Machine-mediated PMem. For more information on the Memory Machine Operator implementation, see https://github.com/MemVerge/memory-machine-operator.

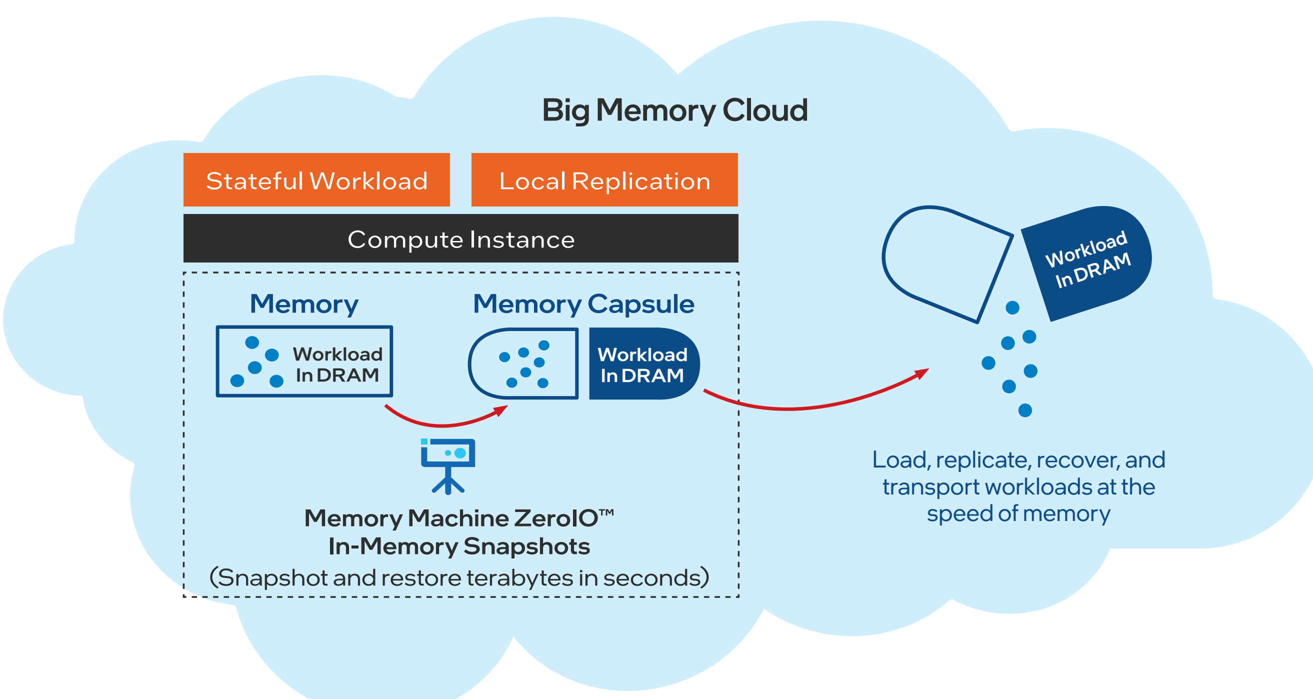

Figure 5. MemVerge ZeroIO™ in-memory snapshots can be used to create AppCapsules.

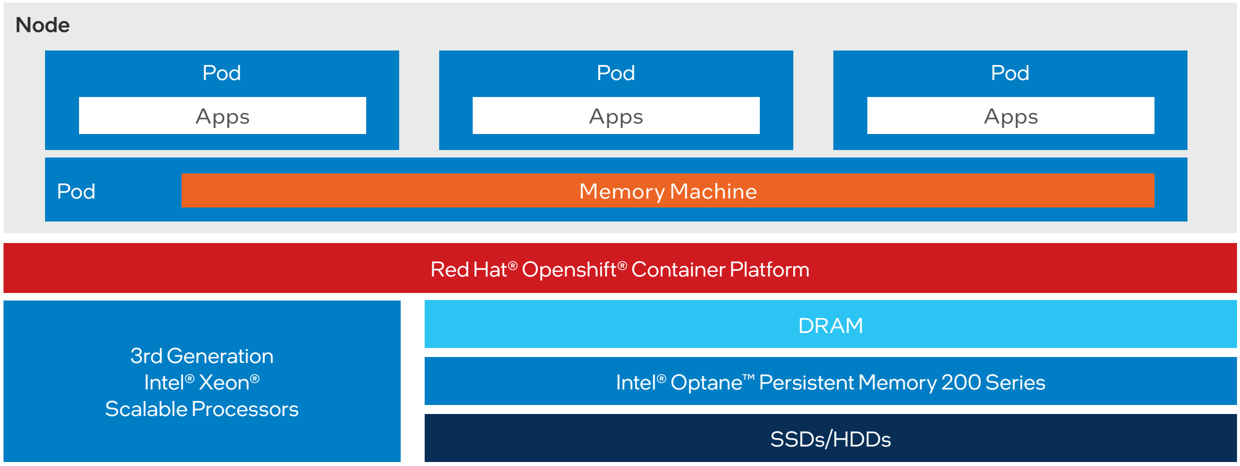

Figure 6. Memory Machine™ deployment model for Red Hat® OpenShift® container platform.

To facilitate building integrated enterprise solution stacks from compatible elements, Red Hat maintains an Ecosystem Catalog of hardware, software, and cloud service providers that are certified for specific Red Hat technologies, including OpenShift. Certification for software may include capabilities such as running on OpenShift as a target platform, being fully containerized, and being covered by a collaborative support agreement with Red Hat.

Enterprise architects can use this information to provide compatibility assurances about the combination of OpenShift and other solution elements. Both the MemVerge Memory Machine Operator and the Intel Optane PMem CSI Operator are certified for OpenShift and are delivered as SaaS offerings through the Red Hat Ecosystem Catalog, providing confidence for developers and architects as they deploy next-generation container environments.

Intel Optane PMem CSI has also passed badging certification. Certification badges extend the Red Hat OpenShift Operator certification into specific functional areas or infrastructure services that are key in cloud-native deployments. Built on the foundations of container and operator certification, badges indicate that a certified product delivers additional capabilities, such as conformance with Kubernetes plugin APIs including CSI or Container Networking Interface (CNI). These assurances build even better confidence for developers, architects, and customers as they deploy next-generation container environments.

Conclusion

Enterprise architects seeking to extend the benefits of containerization to long-running, stateful, non-fault-tolerant applications need efficient, dependable approaches to providing data persistence. Intel Optane PMem provides that capability, made available to containers using Intel Optane PMem CSI. MemVerge Memory Machine simplifies access to data persistence on Intel Optane PMem, without changes to applications, and virtualizes the pool of DRAM and PMem, allocating it as byte-addressable memory when needed by containerized applications. Certification of both Intel Optane CSI and Memory Machine for Red Hat OpenShift offers a proven foundation for extending containerization to these workloads.

More Information

Intel® Optane™ PMem: intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html

Intel® Optane™ PMem CSI operator (open source community): operatorhub.io/operator/pmem-csi-operator

Intel® Optane™ PMem CSI operator (Red Hat Ecosystem Catalog): catalog.redhat.com/software/operators/detail/612cd31535f8020a3f8bd5f2

Intel® Optane™ PMem CSI storage driver GitHub Readme file: intel.github.io/pmem-csi/1.0/README.html

MemVerge Memory Machine™ Cloud Edition: memverge.com/memory-machine-cloud-edition-data-sheet

MemVerge Memory Machine™ v1 container stack (Red Hat Ecosystem Catalog): catalog.redhat.com/software/container-stacks/detail/6020a5bd42f828f1b08f0567

Persistent Memory Programming open source project: pmem.io

Red Hat® OpenShift®: redhat.com/en/technologies/cloud-computing/openshift

[1] Intel® Optane™ Persistent Memory 200 Series Product Brief. https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html.

0422/CDH/MESH/346444-001US