March 2021

Using Persistent Memory To Reduce Cloud Servers Costs

Jue Wang

Using Persistent Memory To Reduce Cloud Servers Costs

ABSTRACT

By virtue of its very high speed, DDR memory is an important component of cloud servers. Such memory is expensive and forms a substantial portion of the total cost of ownership (TCO) of cloud servers. This disclosure describes techniques to selectively use persistent memory (PMEM) as a cheaper alternative to DDR to reduce cost. A hypervisor is modified to include components that perform smart management of virtualization in a cloud environment to achieve significant memory savings by managing the data placement policy, e.g., demoting cold data to PMEM and promoting hot data to DDR RAM. Such operation is achieved with minimal performance impact on the virtual machine and with no modifications to the guest operating systems or customer applications.

KEYWORDS

- Memory cost
- DDR RAM
- Persistent memory
- Cloud computing
- Hardware counter
- Performance monitoring unit (PMU)
- Kernel idle pages
- Cold data
- Hot data
- Stale memory

BACKGROUND

By virtue of its very high speed, double data rate synchronous dynamic random access memory (DDR RAM) is an important component of cloud servers. DDR RAM is also expensive and can be a significant percentage of the total cost of ownership (TCO) of cloud servers. The price of DDR has gone up over generations, such that DDR4 is substantially more expensive than DDR3, and DDR5 is expected to be still more expensive than DDR4. However, DDR RAM

is generally not fully loaded; also, in the cloud context, a sizable fraction of data in server memory can be cold, e.g., infrequently accessed or never accessed after an initial write.

Persistent memory (PMEM) is an emerging memory technology that is a cheaper alternative to DDR RAM (50%-80% cost for the same capacity) at a somewhat lower performance (2x-4x higher latency and memory bandwidth lower by up to 10x). Although the price of PMEM is attractive, its performance characteristics prevent its use as a straightforward replacement of DDR.

Some current techniques deploy proprietary protocols to utilize PMEM, but suffer from low performance and from interference with DDR. Other techniques expose the PMEM as non-volatile memory to the customer's virtual machines (VM), which requires the customer workload to be rewritten to work with PMEM. This limits the utilization of PMEM significantly as it requires the guest operating system kernel (and driver support) and the customer applications to be redesigned and rebuilt from ground up to work with PMEM. Still other techniques achieve savings in VM guest memory, but those savings don't translate to dollar savings.

DESCRIPTION

This disclosure describes techniques to use persistent memory (PMEM) as a cheaper alternative to DDR memory to reduce the total cost of operations (TCO) of servers and virtual machines. Per the techniques, a hypervisor is used to smartly manage virtualization in a cloud environment to achieve significant memory savings by:

- hiding from the customer (the virtual machine guest) the hardware details of PMEM, thereby providing a transparent customer experience;
- managing the data placement policy, e.g., when and what data to be placed into PMEM versus DDR with minimal performance impact;

- maintaining and monitoring raw performance counters, and migrating the virtual machine to less occupied hosts when memory contention is detected based on the counters; etc.

The techniques enable a cloud service provider to provision virtual machine hosts with a portion of memory as PMEM, achieving substantial savings across their fleet of servers.

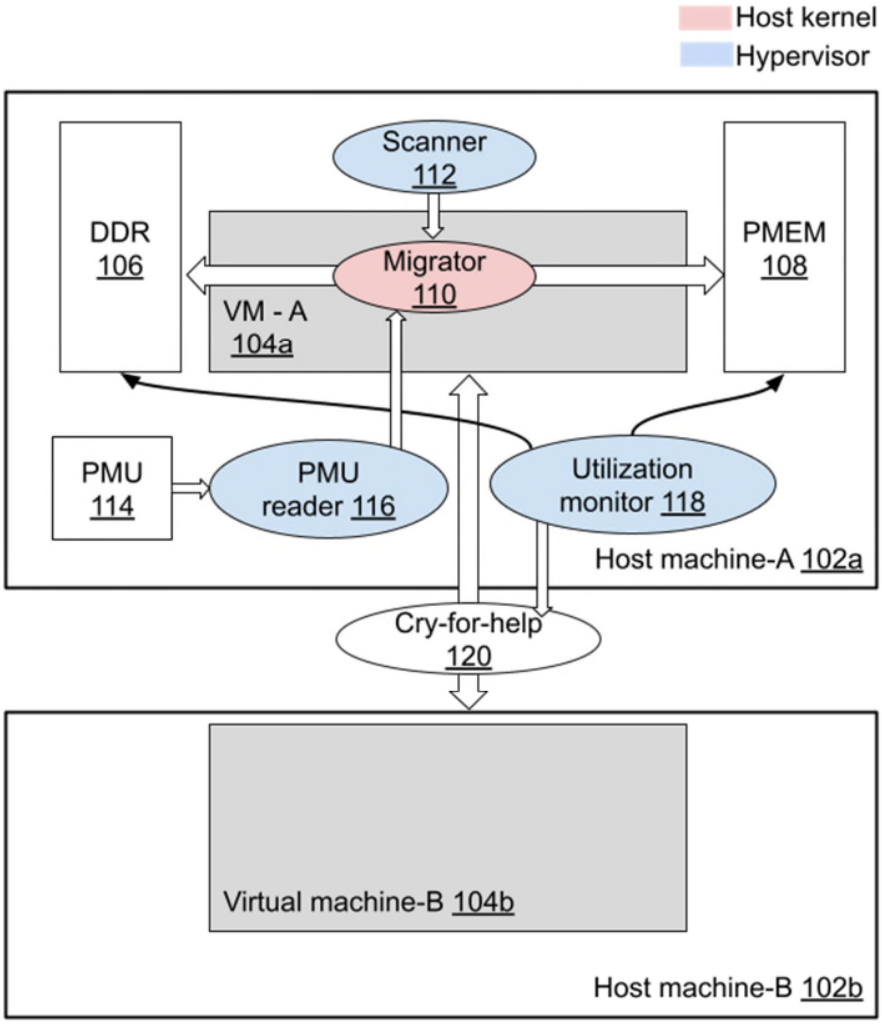

Fig. 1: Using persistent memory to achieve TCO-efficient cloud computing

Fig. 1 illustrates using persistent memory to achieve TCO-efficient cloud computing. A host machine A (102a) hosts a virtual machine A (104a). There are other hosts in the fleet, e.g., host machine B (102b), hosting virtual machine B (104b). Per the techniques of this disclosure, a

host machine is provisioned with both high-performance, expensive DDR (106) and lesser performing but cheaper PMEM (108).

A hypervisor and host-kernel framework migrate data between the DDR and the PMEM depending on the recency of its use. For example, cold data is moved (demoted) into PMEM while hot data is moved (promoted) into DDR. Specifically, the following components interact to achieve promotion or demotion of data based on recency of memory usage:

- A component in the host kernel, migrator (110), manages memory allocation and migration between DDR and PMEM in a manner transparent to the VM and its workload.
- A component in the hypervisor, scanner (112), periodically scans the VM guest memory looking for cold data (or stale memory pages) and invokes the migrator to demote such data or pages to PMEM. The scanner also looks for hot data or memory pages to promote to DDR.
- A performance monitoring unit (PMU, 114) is an on-CPU unit that includes raw hardware-performance counters exposing instruction execution, memory access, elapsed cycles, cache hits, cache misses, etc. A component in the hypervisor, PMU reader (116), collects PMU hardware counters and converts these into more granular memory page access patterns of VM guest memory. These fine-grained access patterns are used to guide fine-grained page promotion and demotion.
- A component in the hypervisor, utilization monitor (118), monitors the utilization of the DDR and the PMEM and triggers a cry-for-help procedure (120) if the PMEM bandwidth is fully saturated, and/or memory contention becomes unacceptably high, and/or the memory access latency is trending up beyond a certain threshold. The cry-for-help procedure selects the VMs that initiated the most accesses to the PMEM (in terms of

bandwidth) or suffered the most from the elevated memory access latency, and migrates such VMs to another host, e.g., host machine-B.

The activities of the above-described components, the decisions made, and their consequent actions (e.g., live VM migration) are transparent to the virtual machine. The above-described components ensure that the VM experience of customers is impacted only minimally by the introduction of the cheaper PMEM. As explained earlier, only infrequently used data are placed in the slower PMEM media; even such data is promoted back to DDR when accessed. A fixed portion, e.g., 25% to 40%, of the memory of a host can be provisioned with PMEM instead of DDR to accrue substantial savings.

CONCLUSION

By virtue of its very high speed, DDR memory is an important component of cloud servers. Such memory is expensive and forms a substantial portion of the total cost of ownership (TCO) of cloud servers. This disclosure describes techniques to selectively use persistent memory (PMEM) as a cheaper alternative to DDR to reduce cost. A hypervisor is modified to include components that perform smart management of virtualization in a cloud environment to achieve significant memory savings by managing the data placement policy, e.g., demoting cold data to PMEM and promoting hot data to DDR RAM. Such operation is achieved with minimal performance impact on the virtual machine and with no modifications to the guest operating systems or customer applications.

REFERENCES

1. "Analyzing the Performance of Intel Optane DC Persistent Memory in App Direct Mode in Lenovo ThinkSystem Servers" https://lenovopress.com/lp1083-analyzing-the-performance-of-dcpmm-appdirect-mode accessed on Mar. 8, 2021.
2. "What's Inside an Optane DIMM?" https://thememoryguy.com/whats-inside-an-optane-dimm/ accessed on Mar. 8, 2021.
3. "What are the benefits of Intel® Optane™ Persistent Memory?" https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html accessed on Mar. 8, 2021.
4. "Introduction to the Compute Express Link Standard | DesignWare Technical Bulletin" https://www.synopsys.com/designware-ip/technical-bulletin/compute-express-link-standard-2019q3.html accessed on Mar. 8, 2021.
5. "Next Generation SAP HANA Large Instances with Intel® Optane™ drive lower TCO" https://azure.microsoft.com/en-us/blog/next-generation-sap-hana-large-instances-with-intel-optane-drive-lower-tco/ accessed on Mar. 8, 2021.
6. Denis Bakhvalov, "PMU counters and profiling basics" https://easyperf.net/blog/2018/06/01/PMU-counters-and-profiling-basics accessed on Mar. 8, 2021.
7. "Idle page tracking / working set estimation," https://lwn.net/Articles/459269/ accessed on Mar. 8, 2021.