Hypervisor-based Virtualization for Emerging Memory Devices

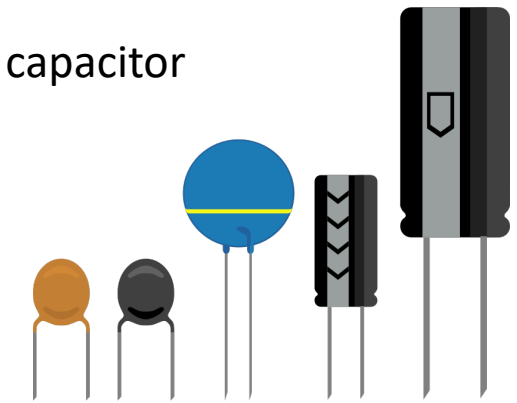
Takahiro Hirofuchi, Ph.D.

Senior Research Scientist, National Institute of Advanced Industrial Science and Technology (AIST)

France-Japan-Germany Trilateral Workshop, Nov 6-7, 2019

DRAM is not scalable anymore …

DRAM

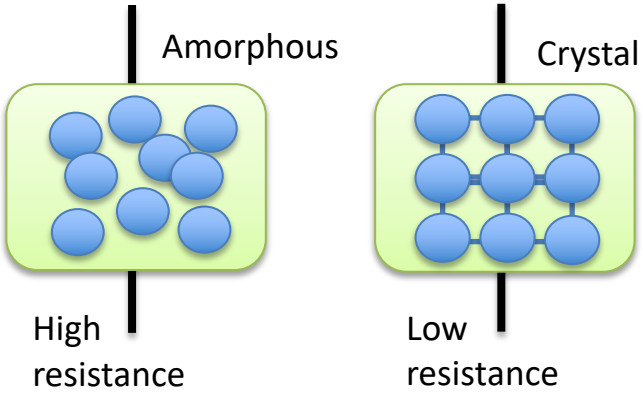


- Energy consuming
  - Need refresh energy
  - More than 30% of energy consumption of a data center
- Not scalable anymore
  - Serious energy dissipation

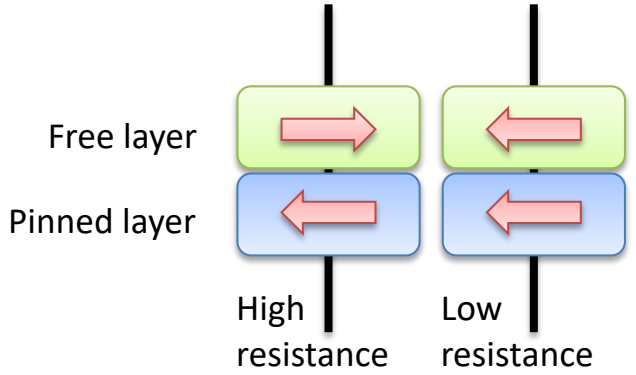
- Memory bandwidth wall
- Memory capacity wall

Emerging memory devices are promising.

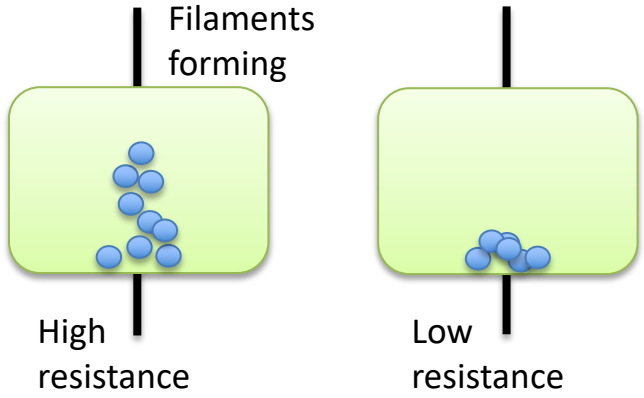
Phase Change Memory (PCM)



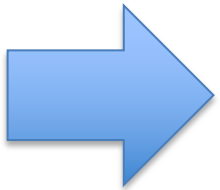
Magneto-resistive Random Access Memory (MRAM)



Resistive Random Access Memory (ReRAM)



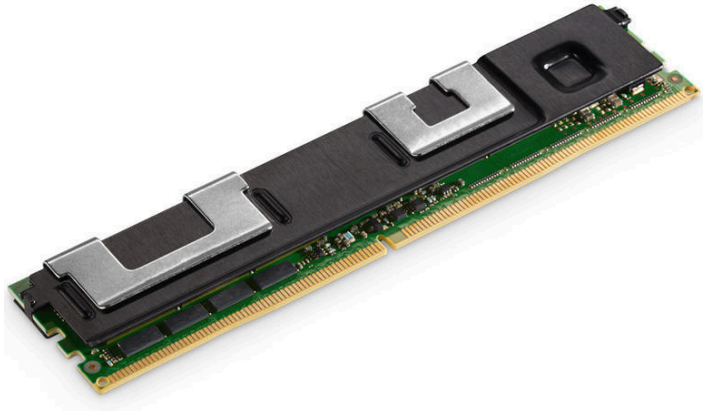
- Energy efficient
  - No refresh energy
  - Non-volatile
- Scalable in theory
  - Higher density will be possible

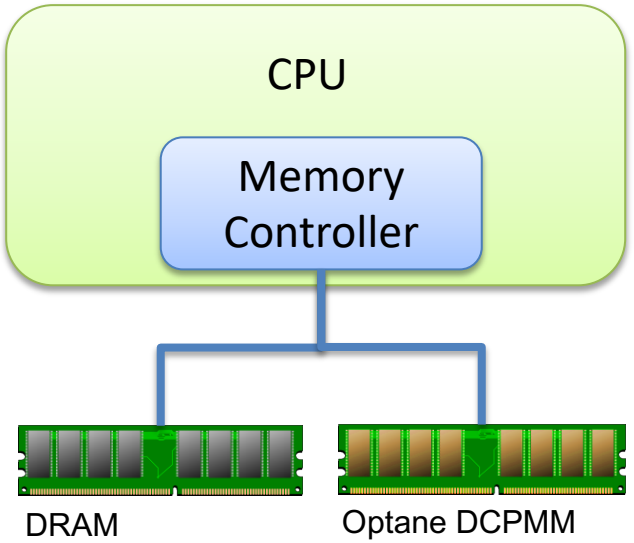


- Potential to drastically extend the capacity of main memory
- Potential for 3D-stack integration drastically increasing bandwidths
- Key technology for upcoming computing systems

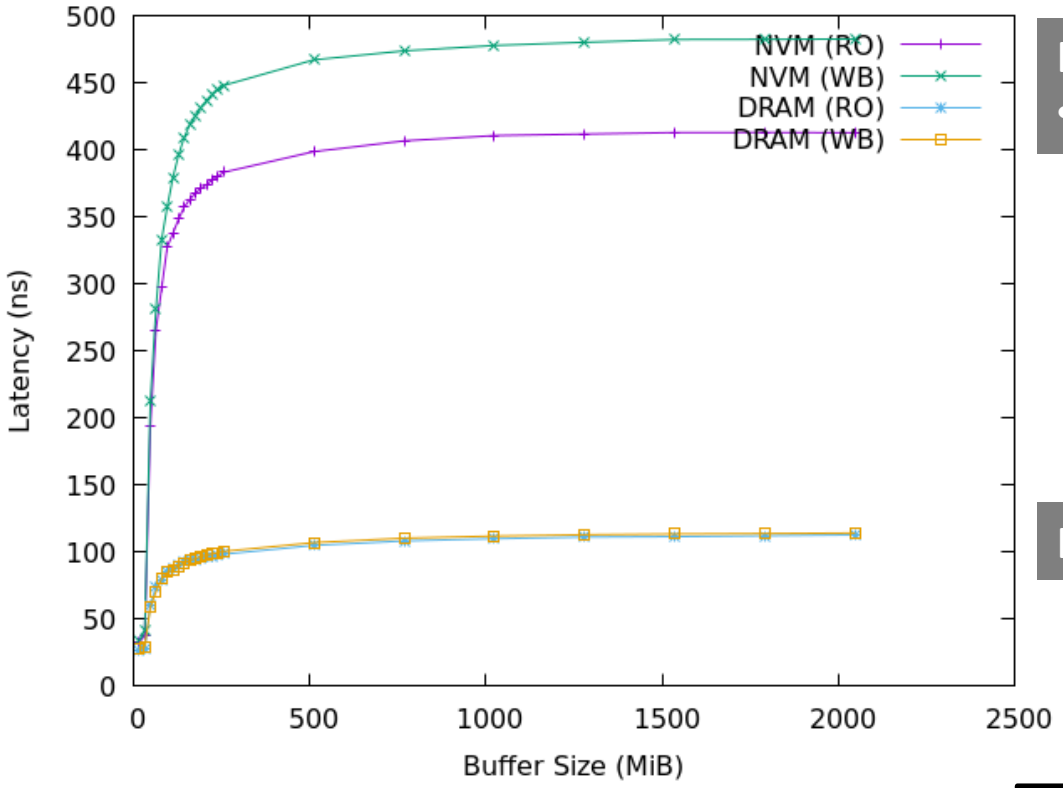
Intel Optane Data Center Persistent Memory (DCPMM)

- Released in April 2019
- First byte-addressable NVM based on a next-generation memory device
  - 3D-stacked resistive memory cells
  - (Possibly) PCM-based
- Connected to the memory bus of CPU via the DIMM interface
- 128-512 GB per memory module
  - A DRAM module is typically 8-32 GB.





Optane DCPMM's Latency is quite high.



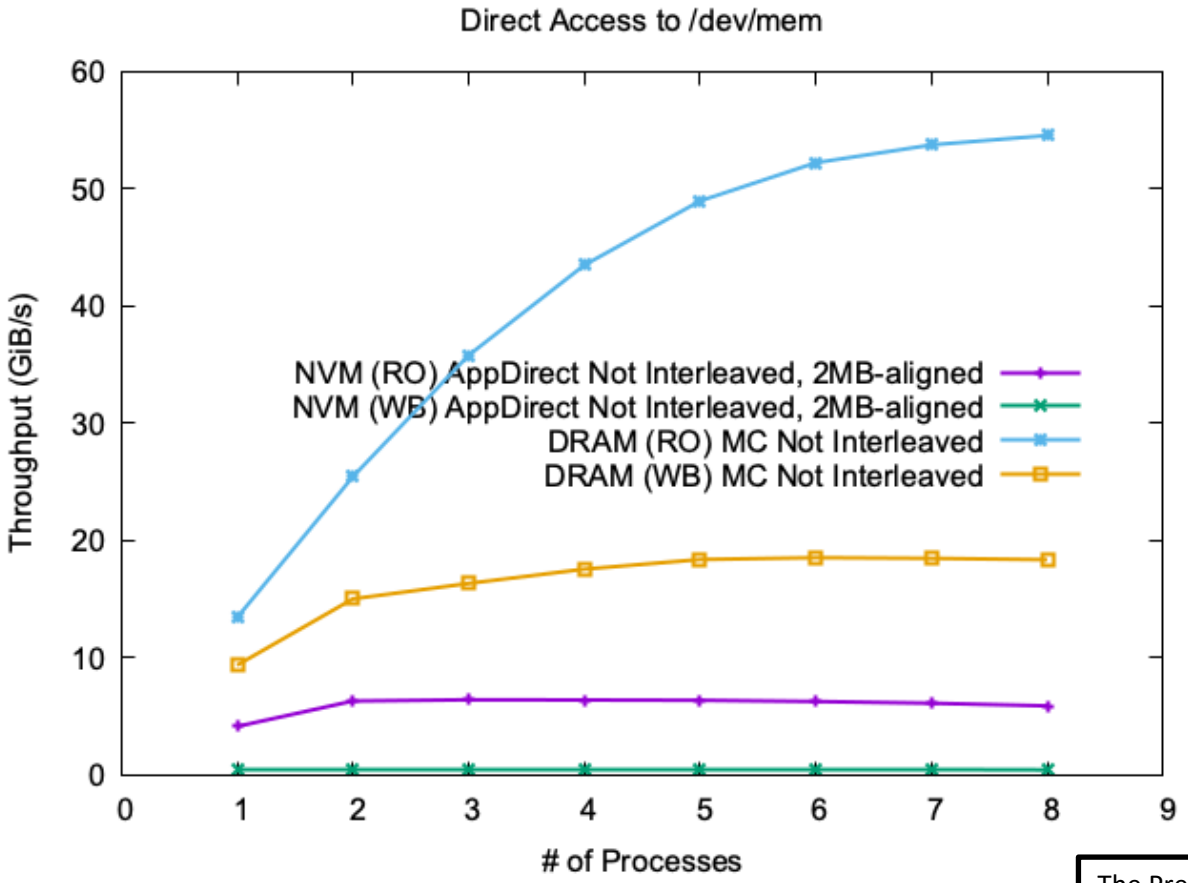
DCPMM latency is 400 ns.

- Write-back-involving latency is 480ns.

DRAM latency is at most 100 ns.

The Preliminary Evaluation of a Hypervisor-based Virtualization Mechanism for Intel Optane DC Persistent Memory Module, Takahiro Hirofuchi and Ryousei Takano, arXiv:1907.12014v1 [cs.OS] 28 Jul 2019

Optane DCPMM's Bandwidth is limited.



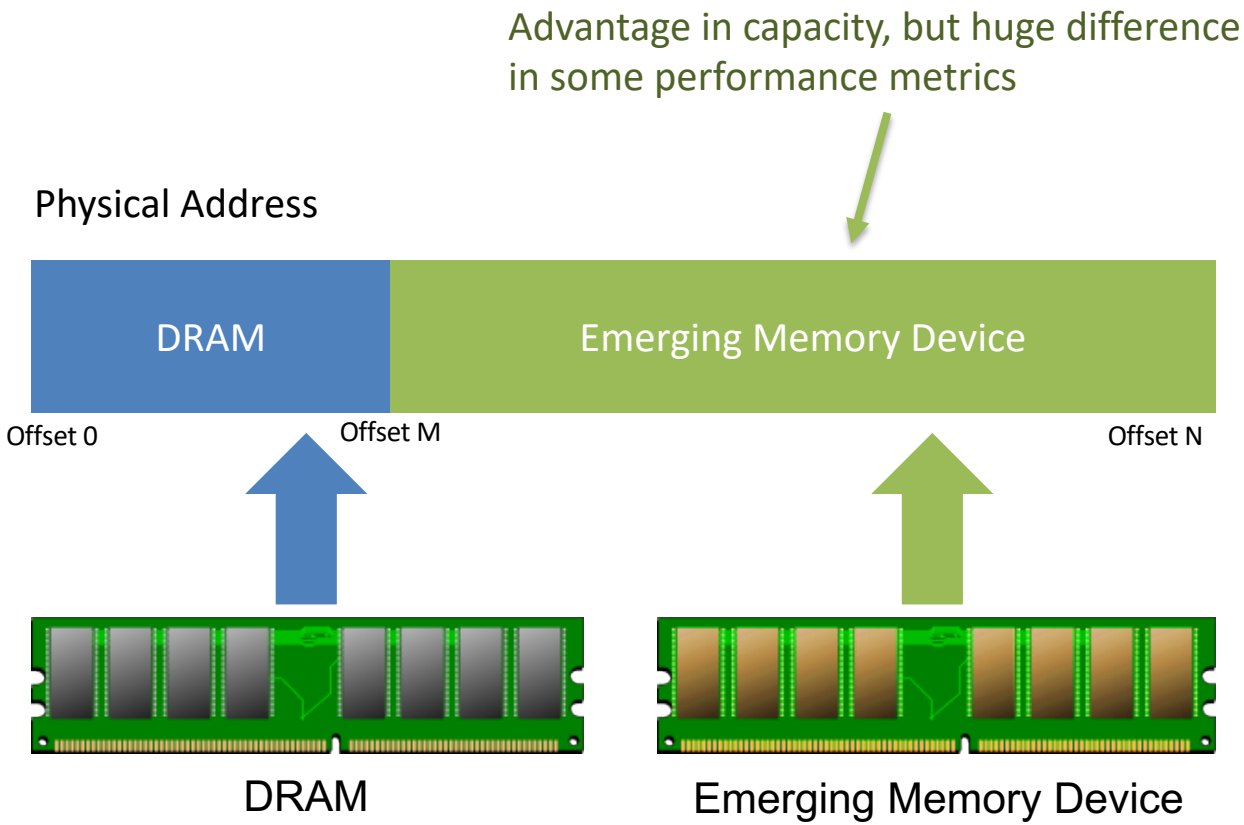
DRAM bandwidth is 55 GB/s, i.e., 110 GB/s per CPU socket.

DCPMM bandwidth is 5 GB/s, i.e., 10 GB/s per CPU socket.
- Write-back-involving bandwidth is 0.5 GB/s.

The Preliminary Evaluation of a Hypervisor-based Virtualization Mechanism for Intel Optane DC Persistent Memory Module, Takahiro Hirofuchi and Ryousei Takano, arXiv:1907.12014v1 [cs.OS] 28 Jul 2019

Huge Performance Gap in Main Memory

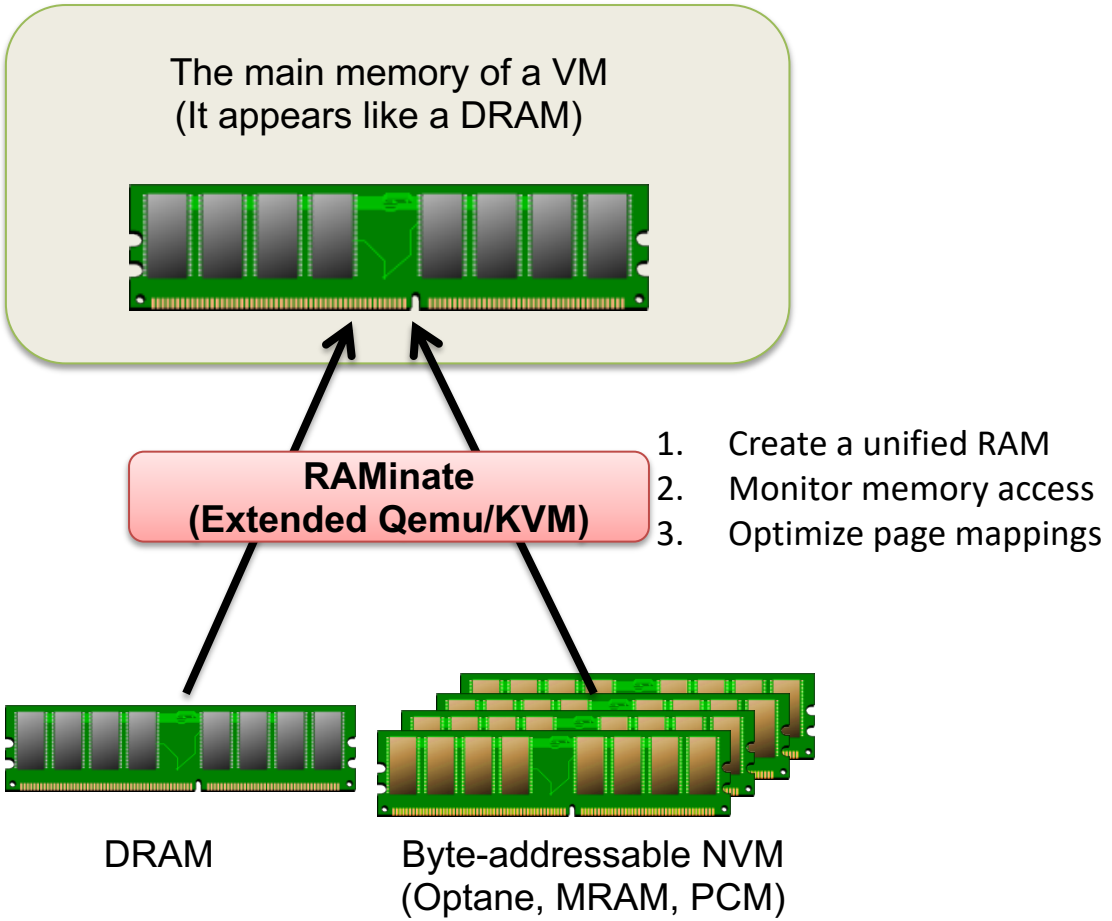
- Optane DCPMM
  - Capacity is 10 times larger.
  - Latency is 4 times higher.
  - Bandwidth is 10%.
- PCM and ReRAM
  - Slow write operations
  - High write energy
- STT-MRAM
  - High write energy



- Main memory now becomes hybrid.
- New system software studies are necessary.

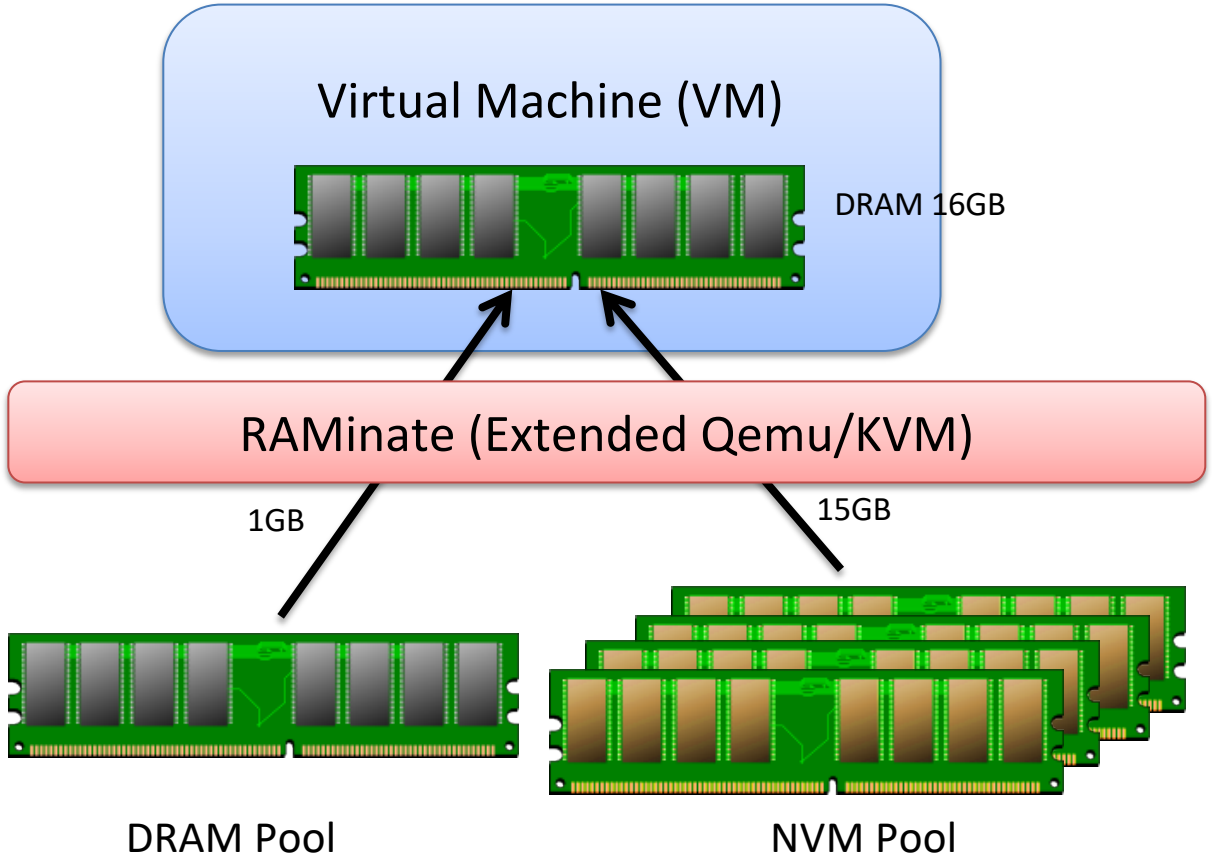
RAMinate: Hypervisor-based Virtualization for Hybrid Main Memory

- Virtually create a unified memory combined with DRAM and byte-addressable NVM (e.g., Optane DCPMM)
  - Make it appear like one memory device to an OS and applications
- Everything is done at the hypervisor layer
  - Support any OSes and applications without any modification to them
- Maximize the performance of a VM even with a small mixed ratio of DRAM
  - Hot memory pages are automatically relocated to DRAM (i.e., fast memory)
  - Cold memory pages are automatically relocated to NVM (i.e., slow memory)
- https://github.com/takahiro-hirofuchi/raminate
- **ACM SoCC 2016 Best Paper Award!**



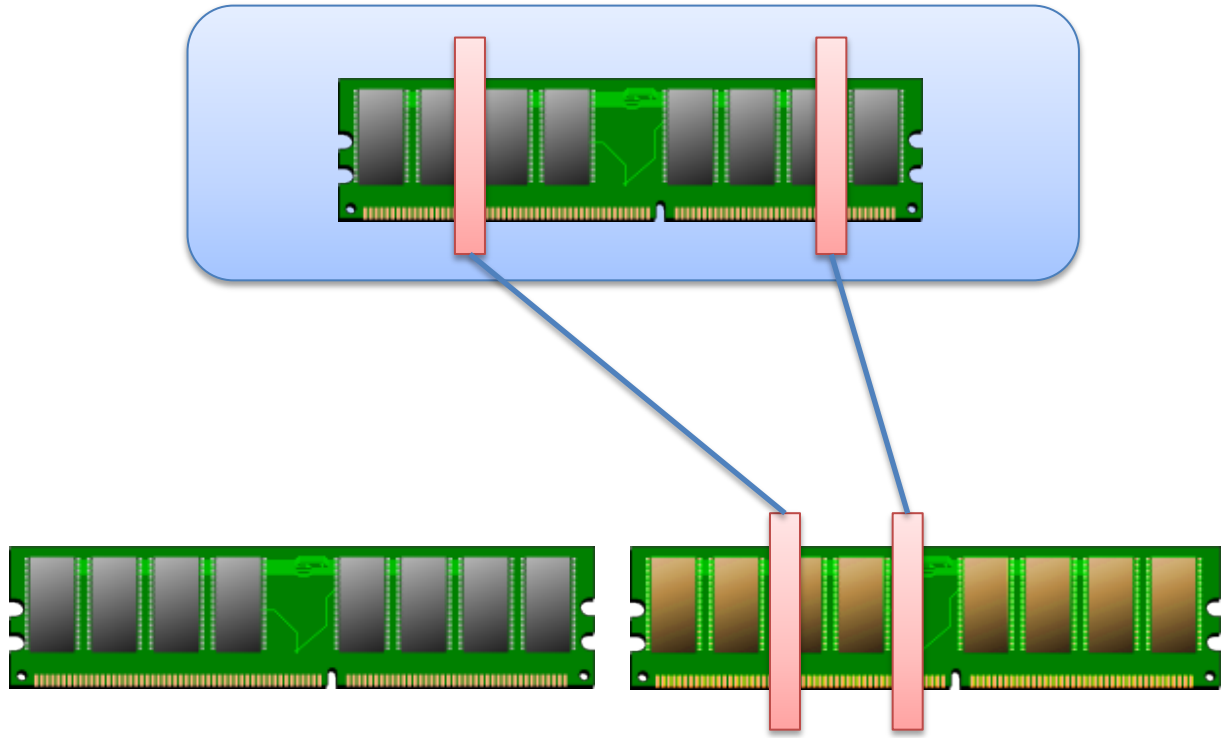
Overview of RAMinate (1)

- Allocate the RAM of a VM from the DRAM and NVM pools
- Guest OS sees a uniform RAM composed of one memory device



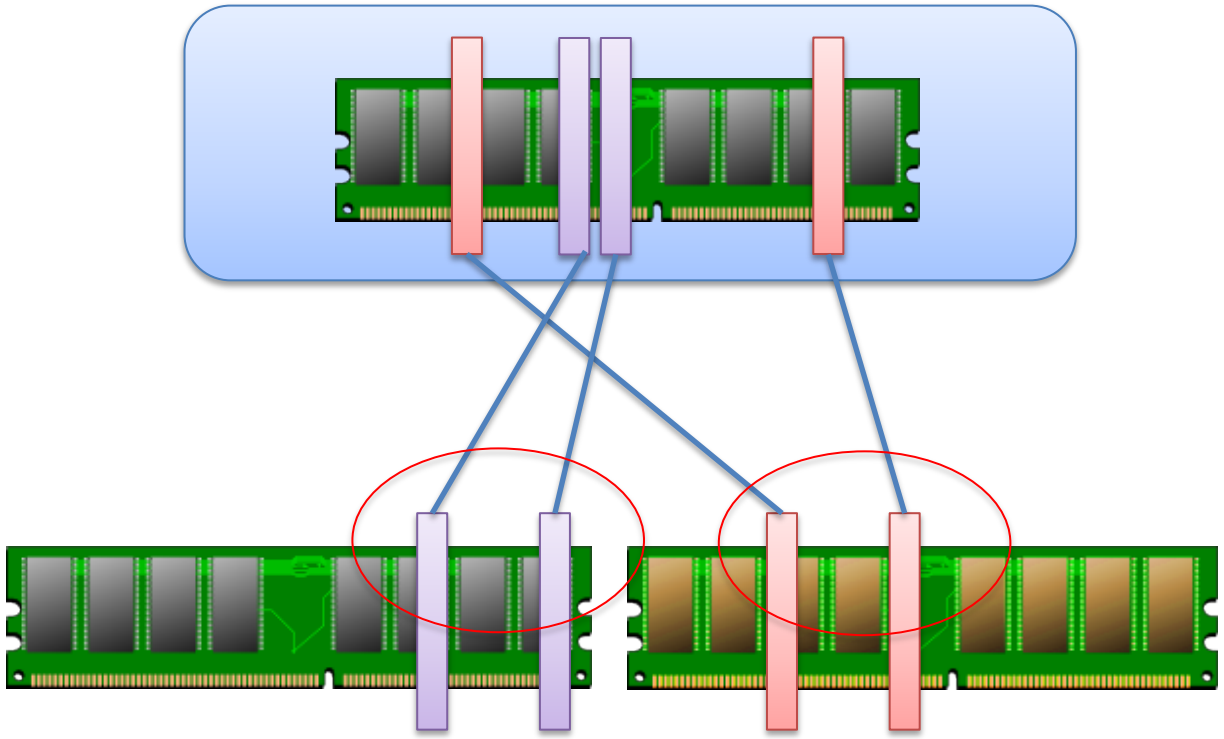
Overview of RAMinate (2)

1a. Detect read/write-intensive guest physical pages on NVM



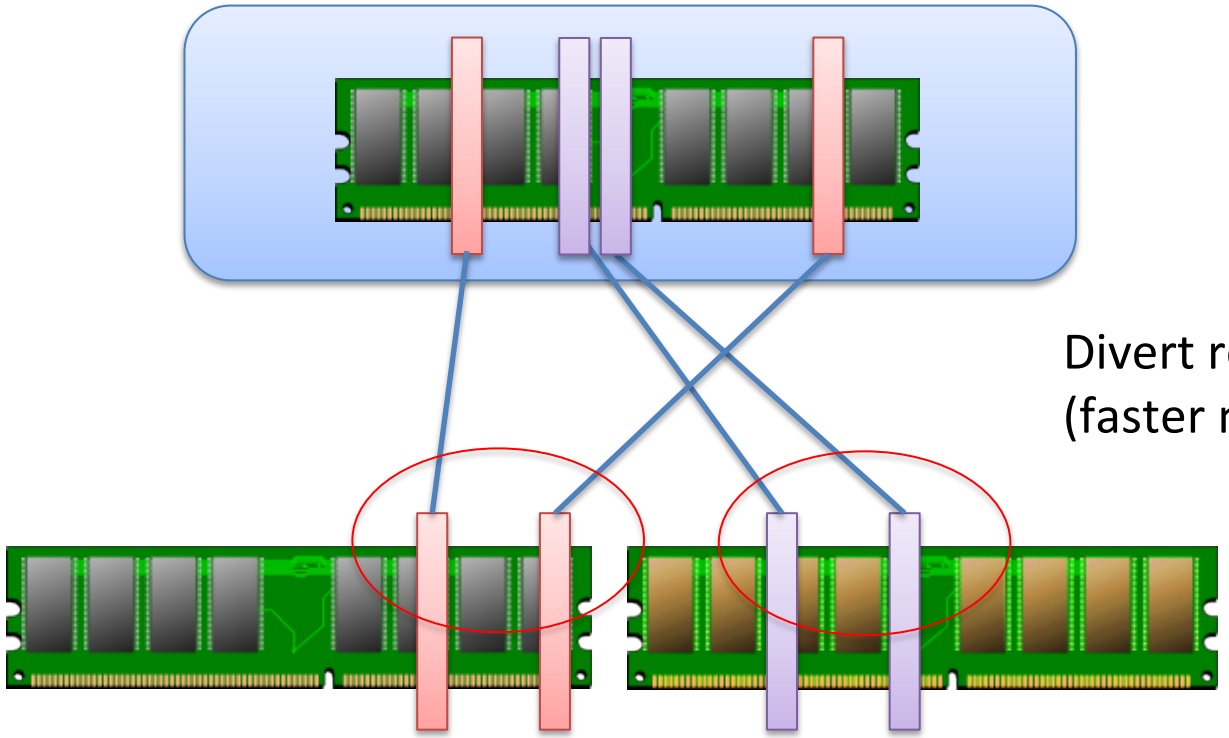
Overview of RAMinate (3)

1b. Detect guest physical pages on DRAM
that have few read/write operations



Overview of RAMinate (4)

2. Swap read/write-intensive NVM pages
   with DRAM pages having few read/write requests



Divert read/write traffic to DRAM
(faster memory) behind the guest OS

DRAM Pool

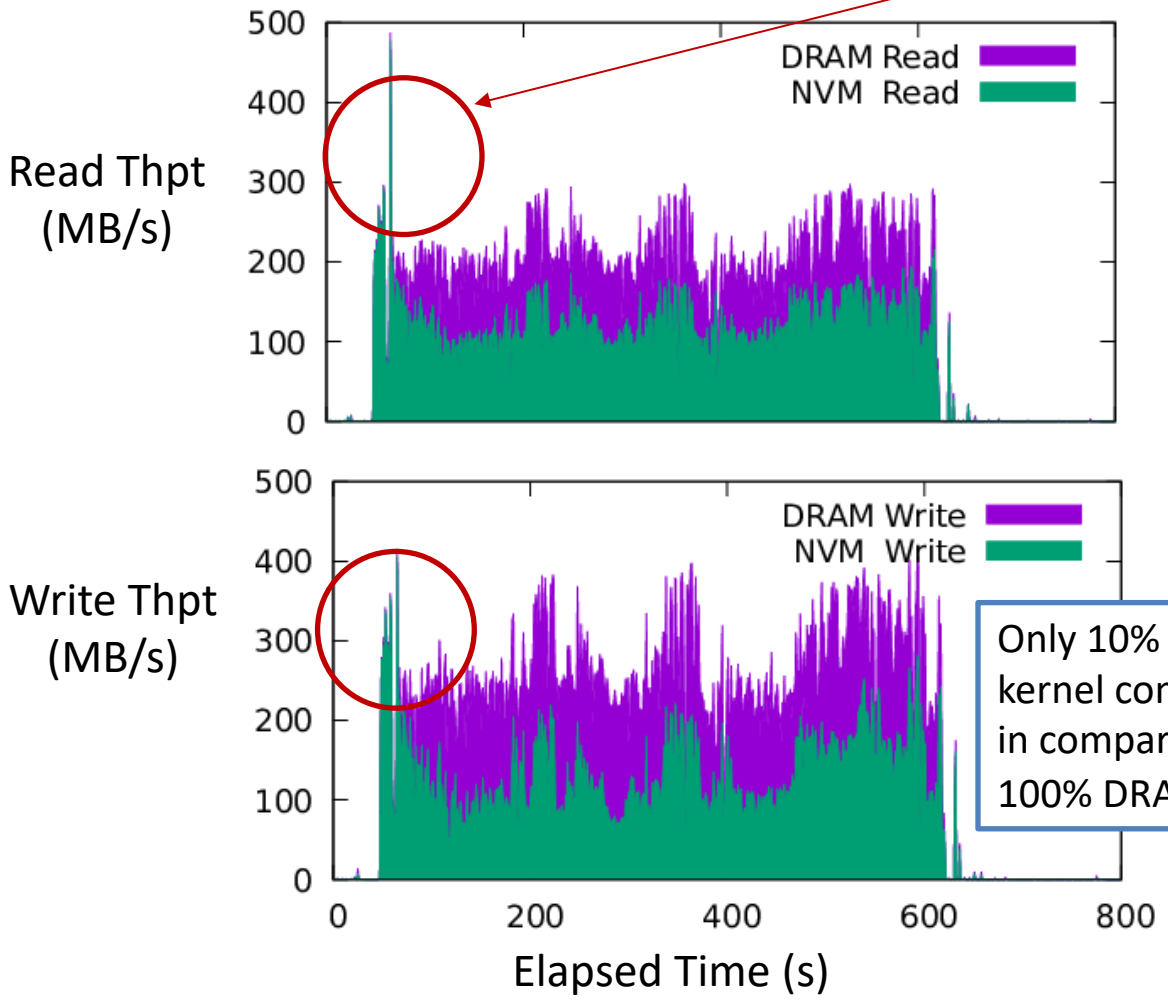
NVM Pool

A Use-case of RAMinate for Optane DCPMM

A VM is set up with 4 GB RAM of the mixed ratio of 1% DRAM and 99% DCPMM.

1. Just after kernel build started, most memory traffic was from/to the DCPMM side. But, after RAMinate optimized page locations, the memory traffic of DCPMM was reduced to 50%.

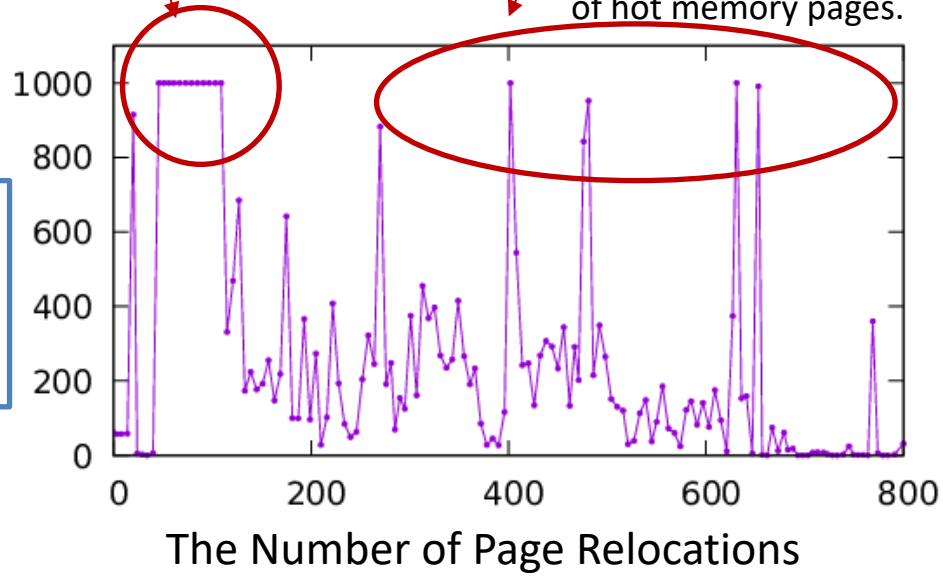
2. RAMinate detected hot memory pages and moved them to the DRAM side. It also moved cold memory pages to the Optane side.



Only 10% increase in kernel compile time in comparison to the 100% DRAM case

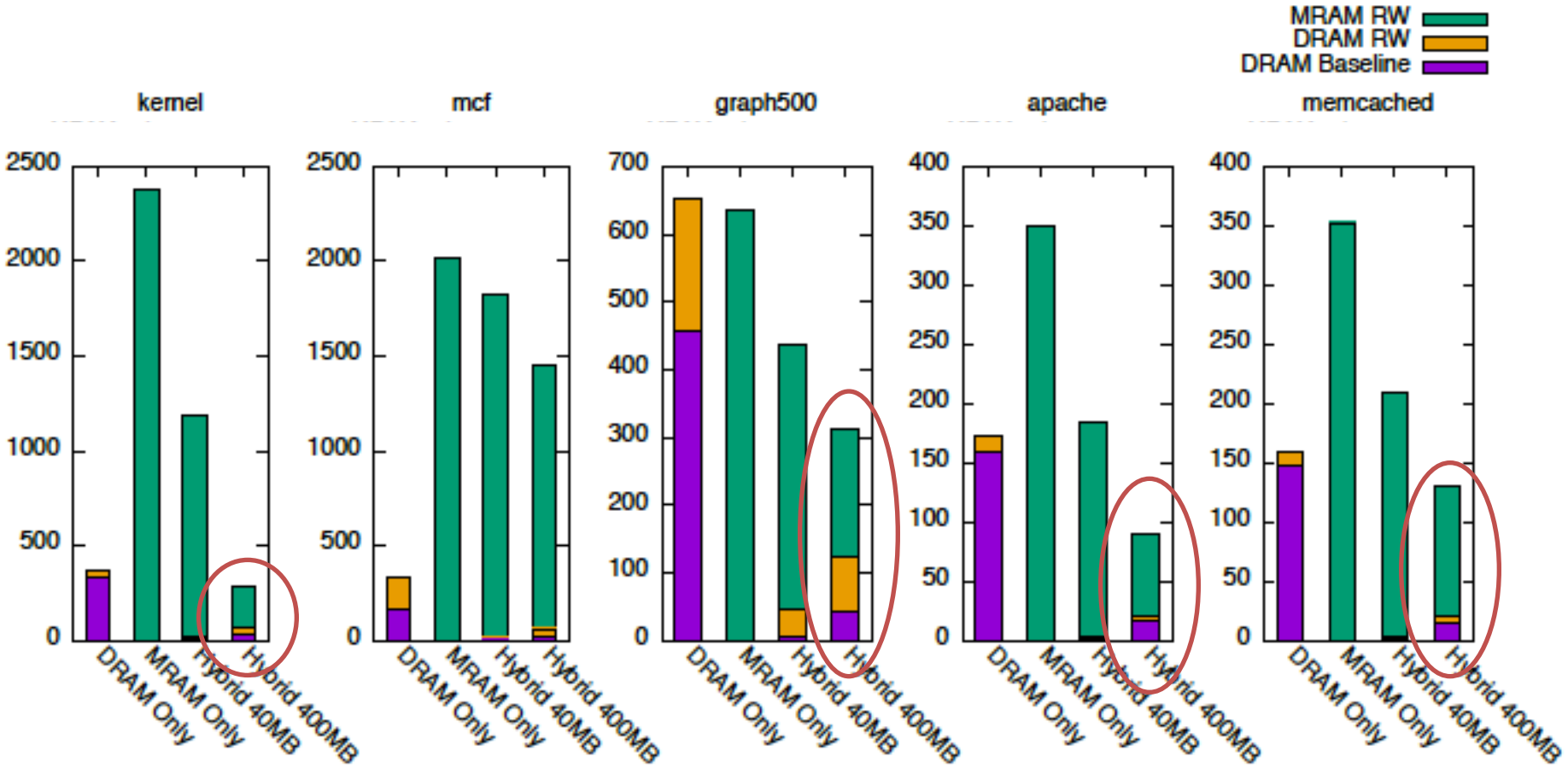
2a. Initially, it optimized page locations intensively.

2b. It continuously updated memory mappings in response to the change of hot memory pages.



A Use-case of RAMinate for Future STT-MRAM

Energy Consumption of Main Memory



ACM SoCC 2016, Hirofuchi et. al.

- Assumption:
  STT-MRAM is fast like DRAM. However, its write energy is 10^2 time larger than that of DRAM.
- Hybrid memory with 10% DRAM outperformed DRAM-only memory in terms of energy consumption.

A Software-based Emulator for Non-volatile Main Memory

- Virtually show a target program a slow NVM, by adjusting the execution speed of the program running on a DRAM-equipped machine
- Fast and accurate
- Aware of asymmetric read/write latencies

Table 4 NVM latencies configured by our prototype and measured with wbbench/robench.

Configured read/write lat.	wbbench		robench	
	Measured lat.	error	Measured lat.	error
122 ns/200 ns	202.1 ns	1.1 %	125.3 ns	2.7 %
122 ns/300 ns	300.4 ns	0.1 %	125.6 ns	3.0 %
122 ns/400 ns	399.2 ns	-0.2 %	125.8 ns	3.1 %
122 ns/500 ns	497.8 ns	-0.4 %	126.4 ns	3.6 %
122 ns/1000 ns	988.7 ns	-1.1 %	128.6 ns	5.4 %

A Software-based NVM Emulator Supporting Read/Write Asymmetric Latencies, IEICE Trans., Dec, 2019 (To appear)

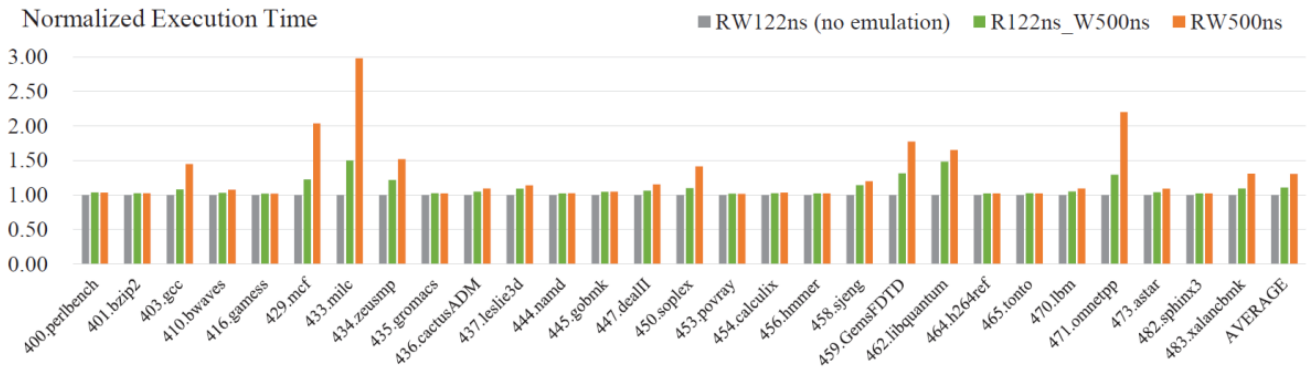
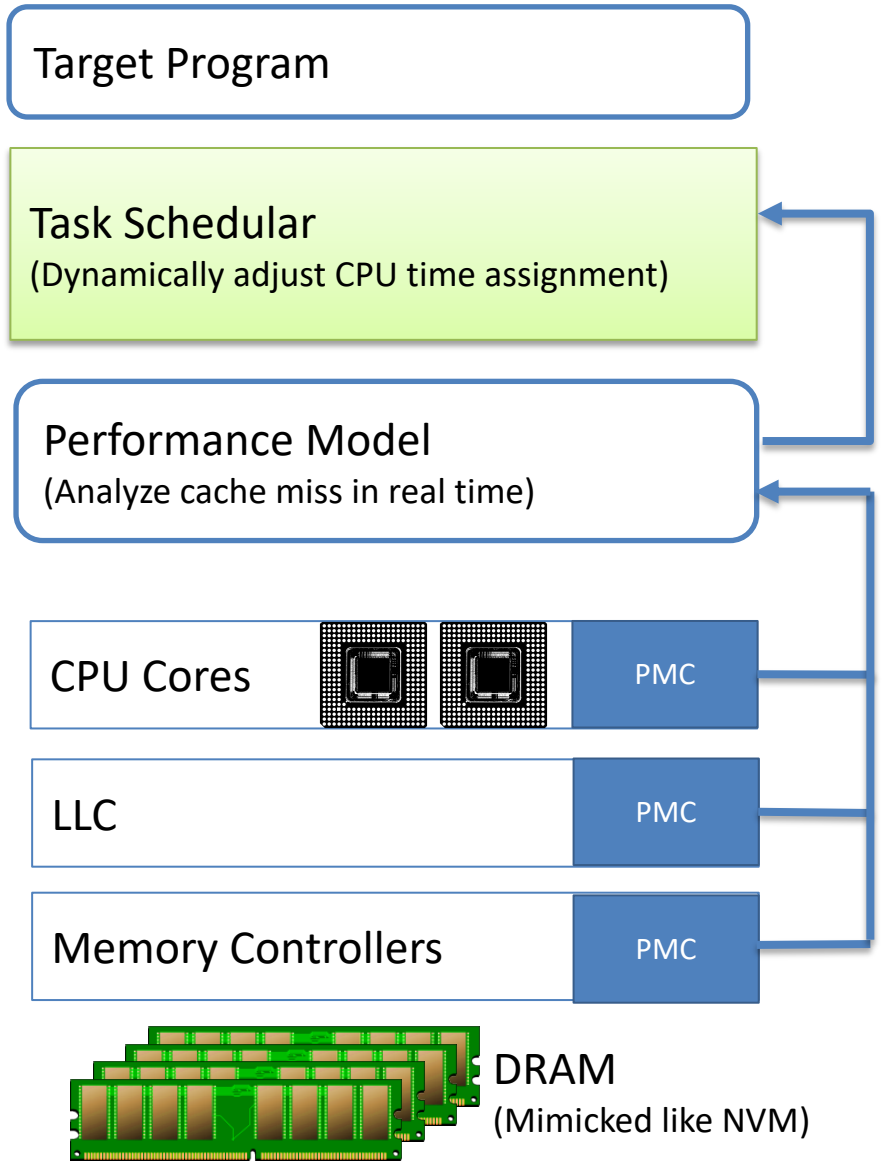
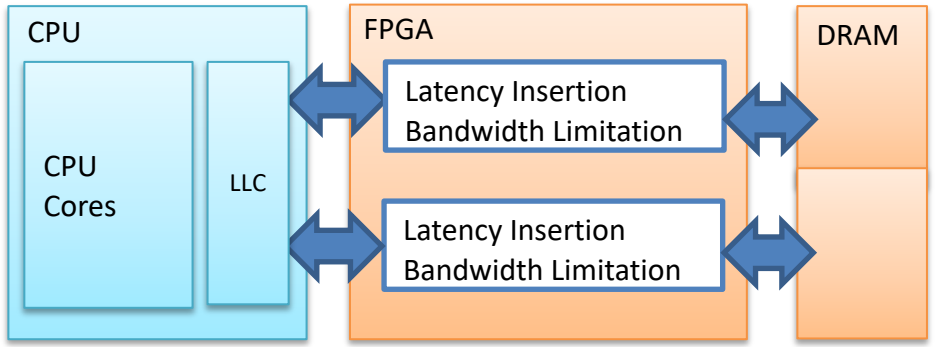


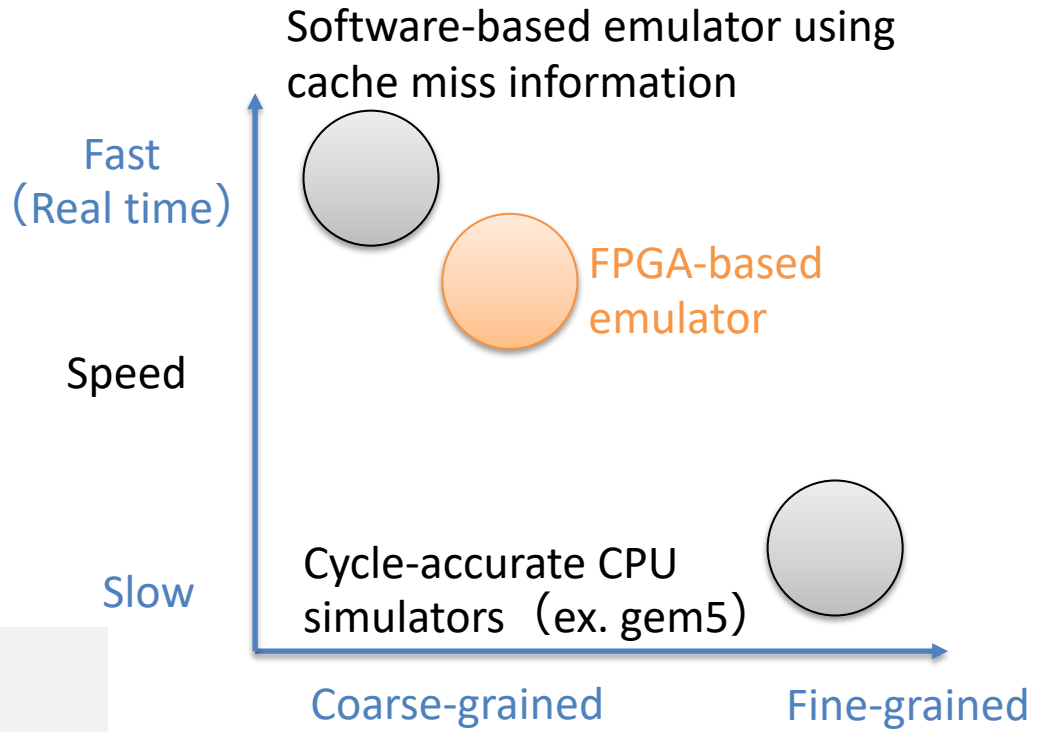
Fig. 10 Execution time of SPECCPU 2006 benchmarks when setting the emulated NVM read/write latencies to 500 ns. The results are normalized to *no emulation*.



An FPGA-based Emulator for Hybrid Main Memory

- Emulate hybrid main memory systems in the hardware level
  - Set latency, bandwidth and bit-flips in each physical address range
- Enable detailed performance evaluation of new system software mechanisms





Summary

- Emerging memory devices
  - PCM, MRAM, ReRAM
  - Optane DCPMM
- Main memory now becomes hybrid
  - New system software studies are necessary
- RAMinate
  - Hypervisor-based virtualization for hybrid main memory systems
- Software-based emulator using cache miss information
  - Support asymmetric read/write latencies
- FPGA-based emulator
  - Enable system software studies on emerging memory devices