DDR4 DIMM Devices With Hybrid DRAM & NVM For Big Data Performances At Low Cost

Xiaobing Lee*, Florian Longnos**, Shaojie Chen**, Wei Yang**

*Storage Media Application Laboratory, Futurewei

**Shannon Laboratory, Huawei

Xiaobing.lee@huawei.com, florian.longnos@huawei.com

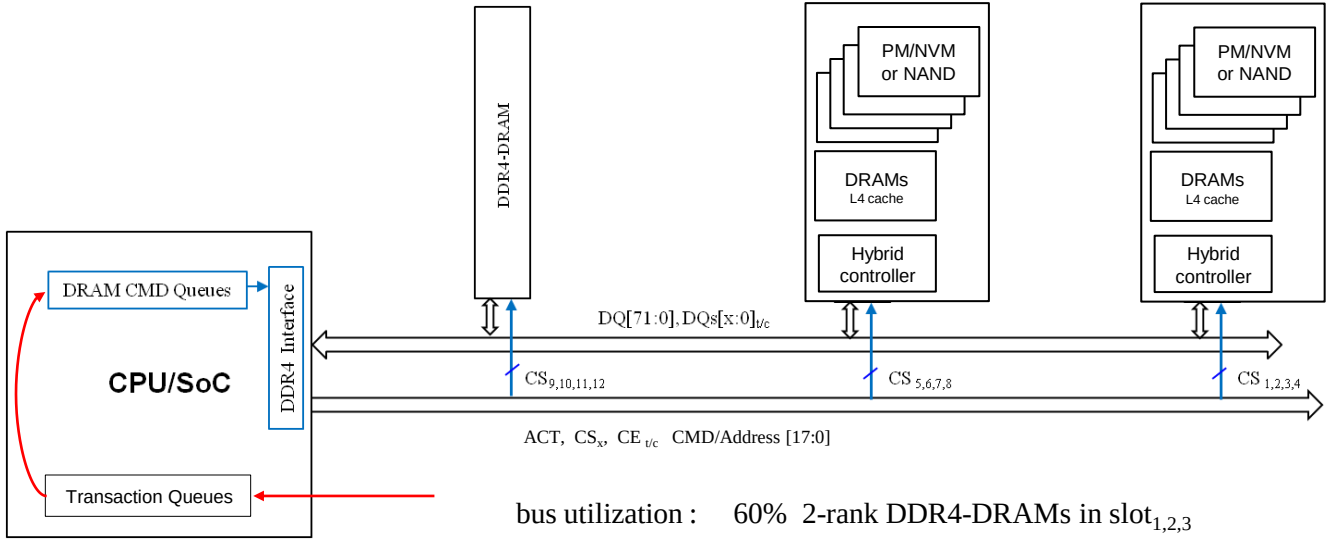
Contents

- Big data applications demands huge host memory
- Hybrid DIMM vs. separate DIMMs
- Optimizing hit rate for big data applications

Big Data Demands Huge Host Memory

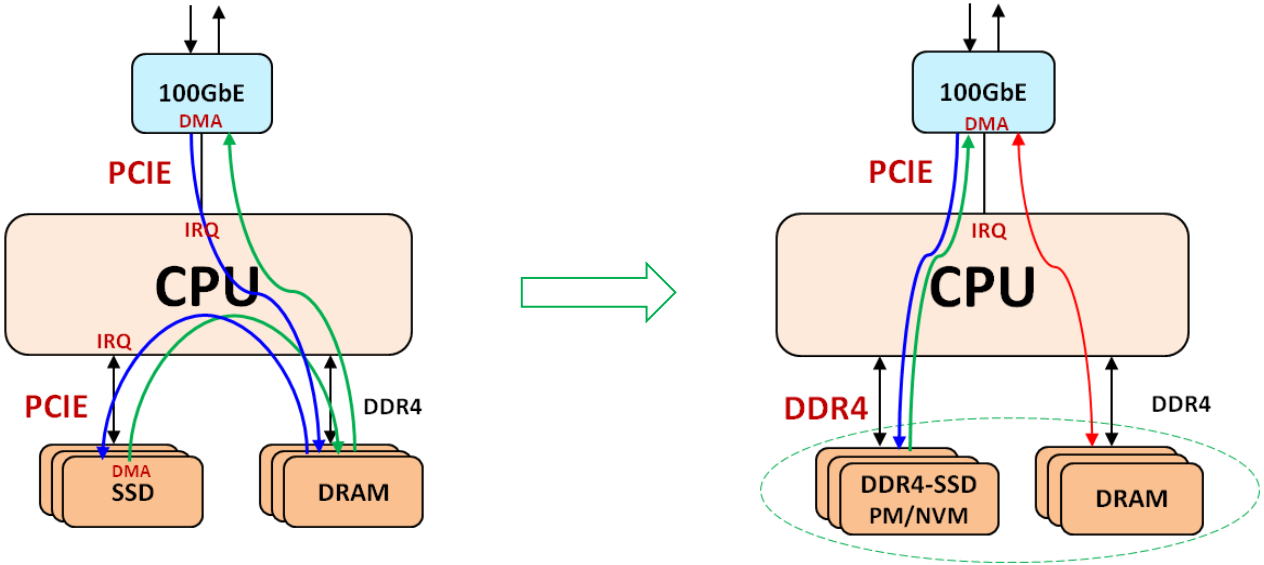
- **Big data applications demand huge host memory for high performance**
- **4-rank DDR4 or 8-rank 3DS DRAM DIMMs for higher bus utilization**

✓ **Some DIMM devices could be hybrid DRAM and PM/NVM to reduce costs**



bus utilization : 60% 2-rank DDR4-DRAMs in $slot_{1,2,3}$

NVMe Block Device Over DDR4 Bus



IOC to SSD storage runs 2x DMA-IRQ ops

- PCIe-SSD/SAS-SSD rd/wr ops needs CPU to process DMA-IRQ ops 2 times, using CPU DDR4 bus twice, and then reduce the CPU handle IOPs capacity
  Header: Data ~ 1 : 1000

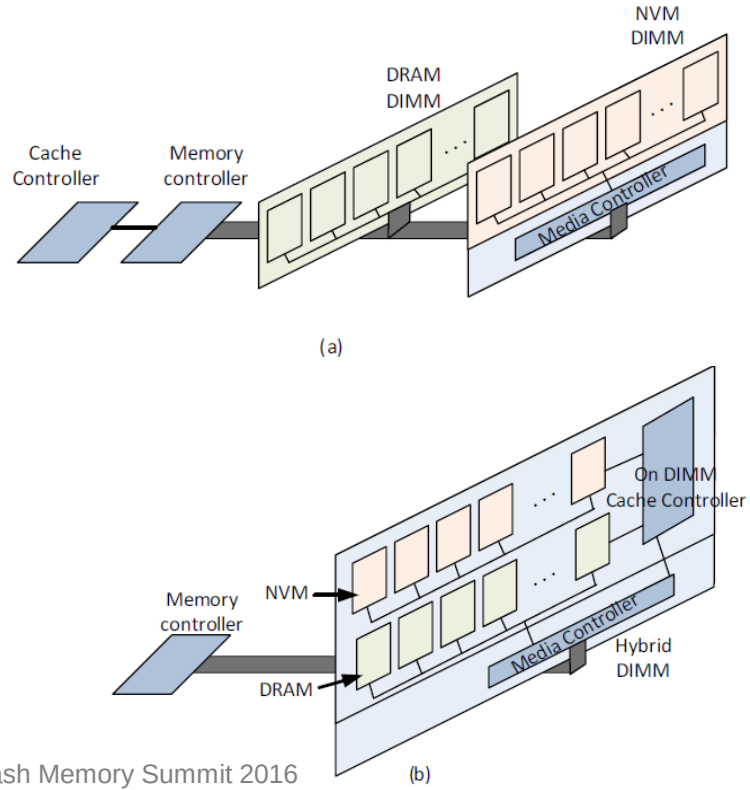
SSD devices are better, if put on DDR4 bus

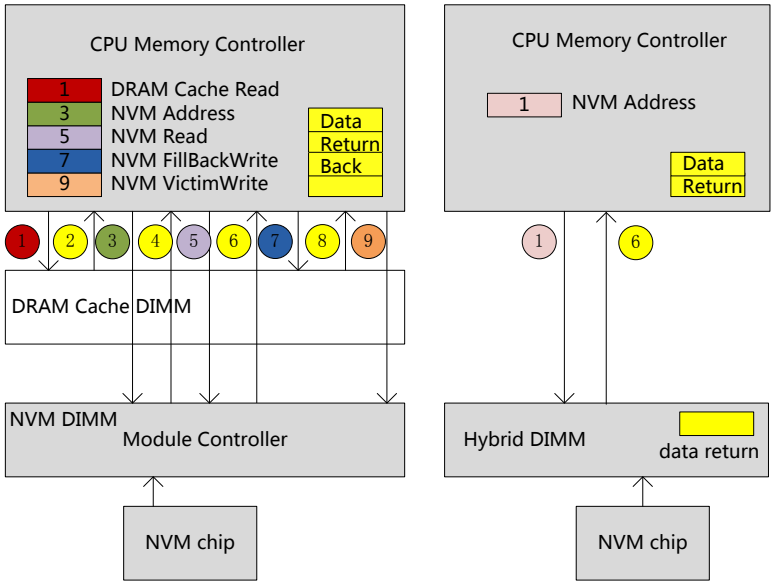
- ✓ CPU only process IOC DMA and IRQ ops once, could eliminate all SSD related PCIe DMA and IRQ ops
- ✓ Hybrid DIMM (**NVDIMM-P**) offloads DMA ops on-DIMM
- ✓ DDR4 data traffic is cut in half then could double IOPs

Contents

- Big data applications demands huge host memory
- Hybrid DIMM vs. separate DIMMs
- Optimizing hit rate for big data applications

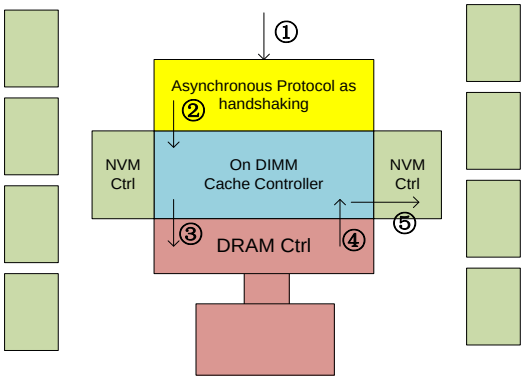
Separate vs. Hybrid DIMM





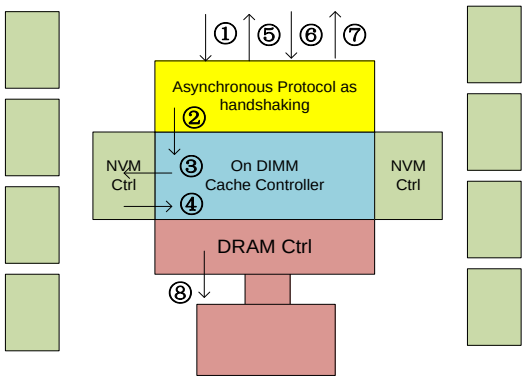
Cache management traffic is confined into the Hybrid DIMM, thus reducing transactions on the memory bus

Requests' Handling in Hybrid DIMM



Write dataflow

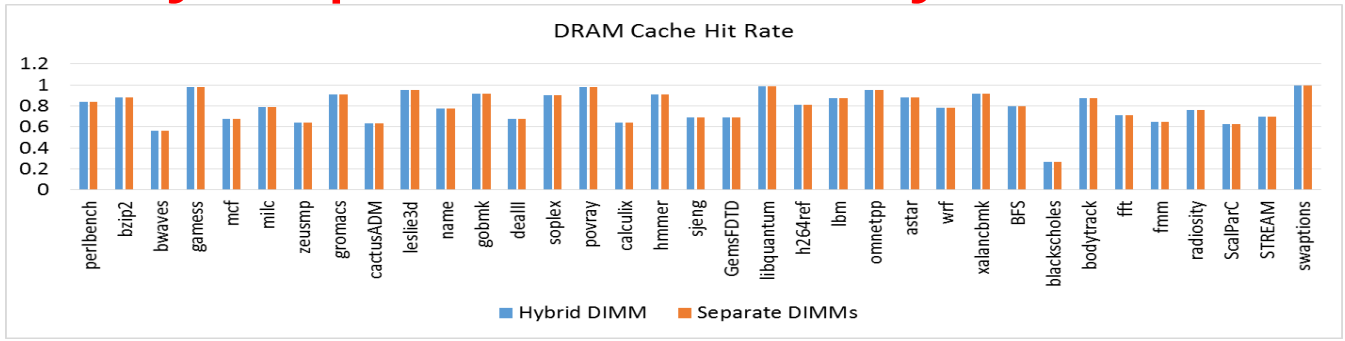
1. Host issues WRite (WR) request
2. WR request is transferred to on-DIMM cache controller
3. Cache controller schedules WR for DRAM cache
4. DRAM Ctrl transfers victim cacheline to cache controller
5. Cache controller schedules write-back to NVM

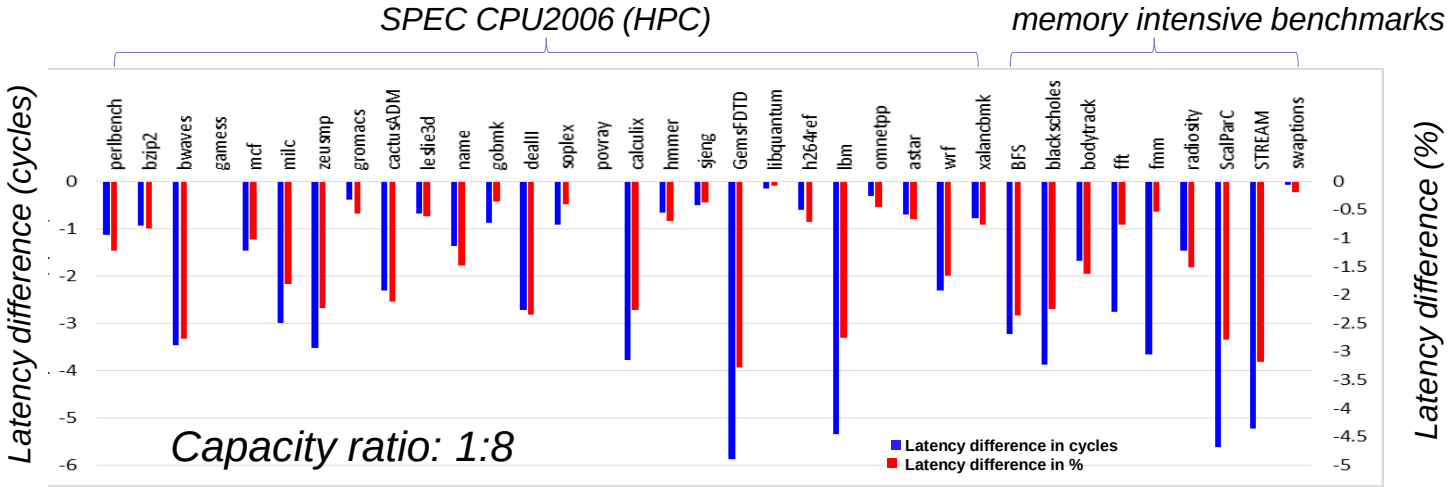


Read dataflow

1. Host issues ReaD (RD) request
2. RD request is transferred to on-DIMM cache controller
3. Cache controller schedules RD for NVM
4. NVM Ctrl returns RD data to cache controller
5. Hybrid DIMM notifies Host data is ready
6. Host grants the bus to hybrid DIMM
7. Hybrid DIMM returns RD data to Host
8. DRAM Ctrl fills RD data on DRAM

Latency Improvement In Hybrid DIMM





- Reduced bus competition
- Lower latency for most applications
- Higher ratio requires high hit rate for improvement

Contents

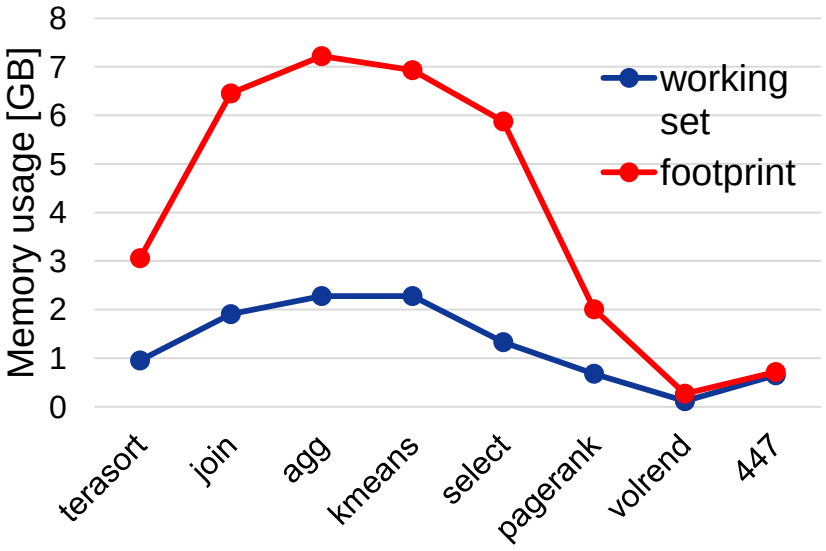
- Big data applications demands huge host memory
- Hybrid DIMM vs. separate DIMMs
- Optimizing hit rate for big data applications

Applications For Hybrid DIMM

Big Data Benchmarks (data traffics)

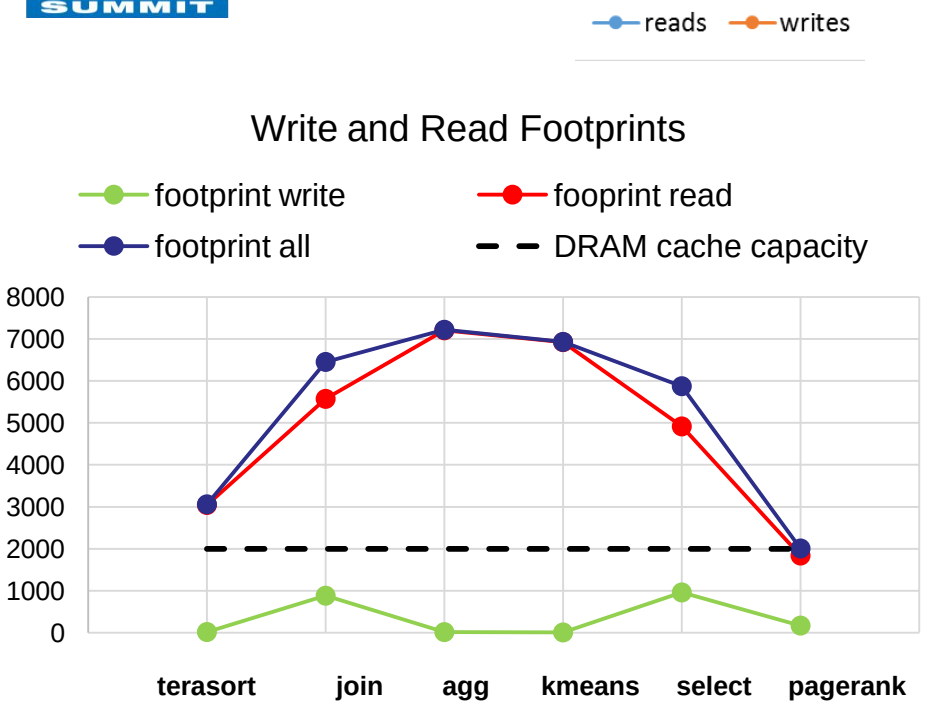
- Terasort input 1GB
- Hive Join Query
- Hive Aggregation Query
- Kmeans input 2GB
- Hive Select Query
- Pagerank

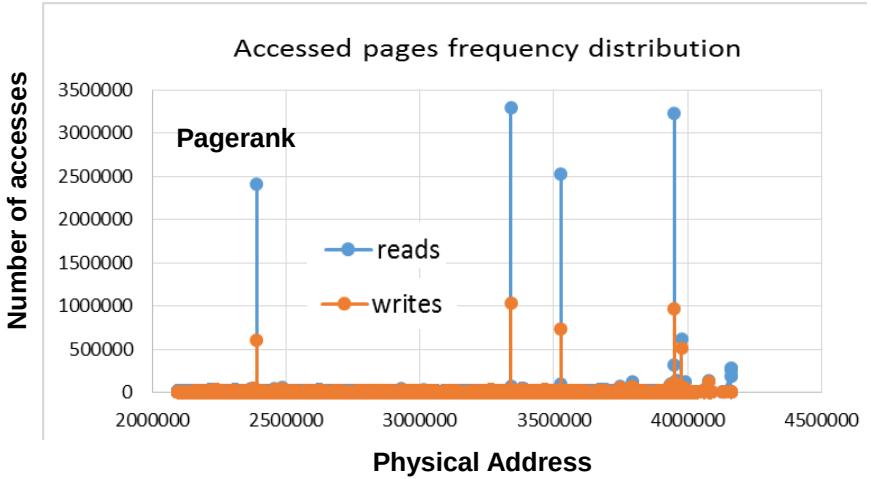
Hybrid DIMM will be beneficial for the applications that **have large footprint** and **high locality**, making most of the working set fit into DRAM cache

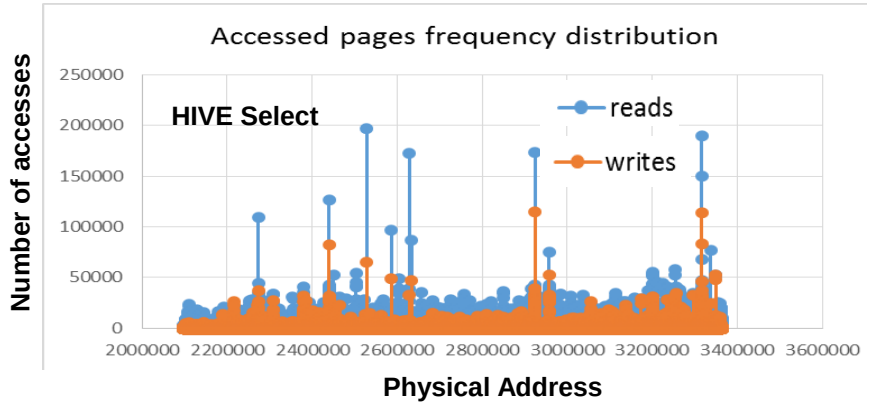


- Direct mapping or associative DRAM cache
- Replacement policy: LRU

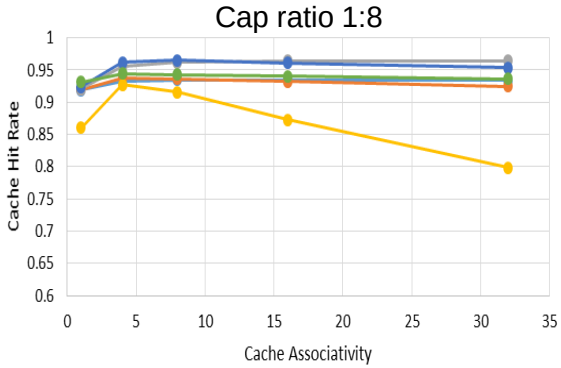
Read/Write Intensivity

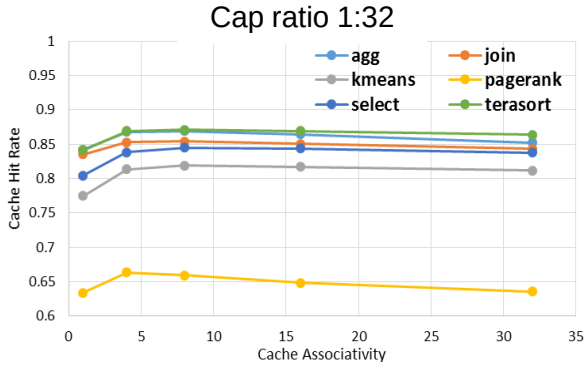


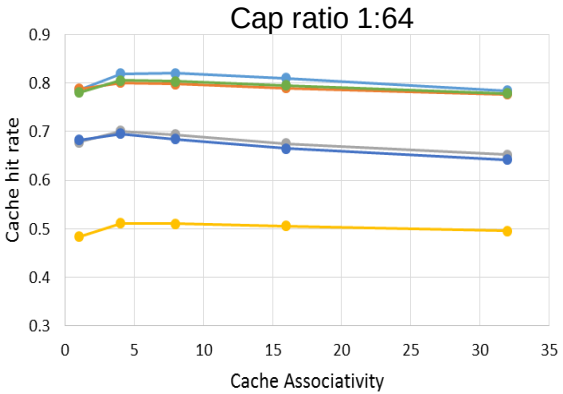


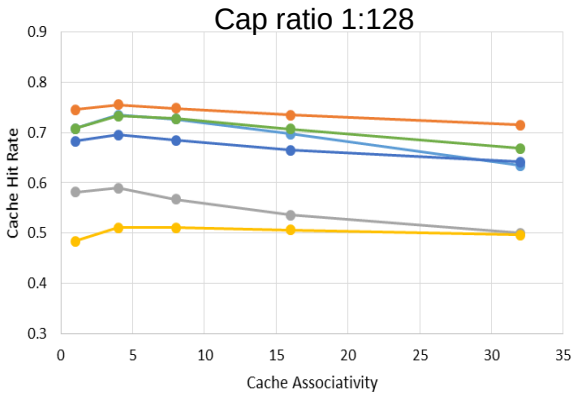


Impact Of Associativity On Cache Hit Rate



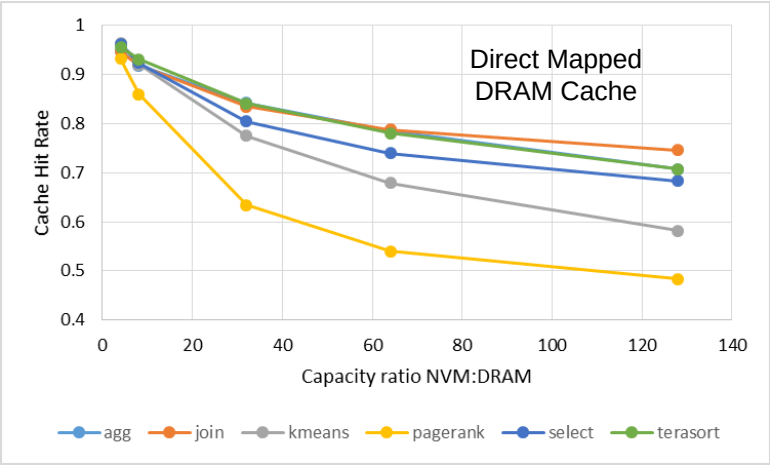


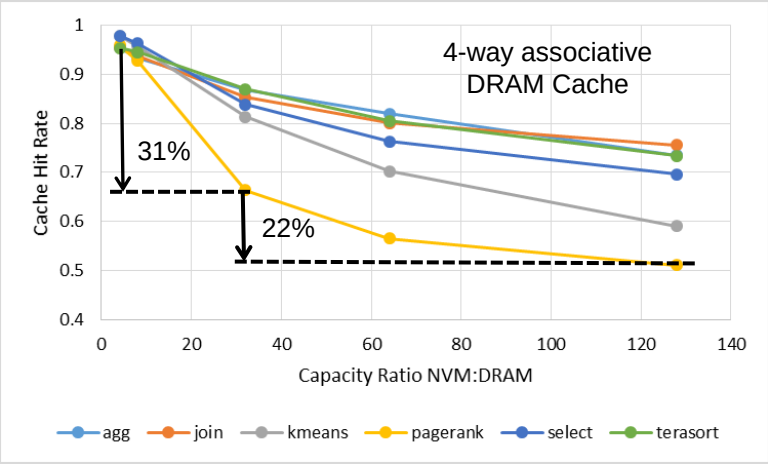




- Small capacity ratio & large DRAM: Associativity has little effect
- Above 1:32 ratio: Highest hit rate for associativity of 4 (+4% for 1:8)
- Further increase of associativity decreases hit rate
- PAGERANK hit rate is affected by associativity for all capacity ratios
- HIVE Agg's is significantly degraded for high associativity values (e.g. 64 or 128)

Impact Of Capacity Ratio On Cache Hit Rate





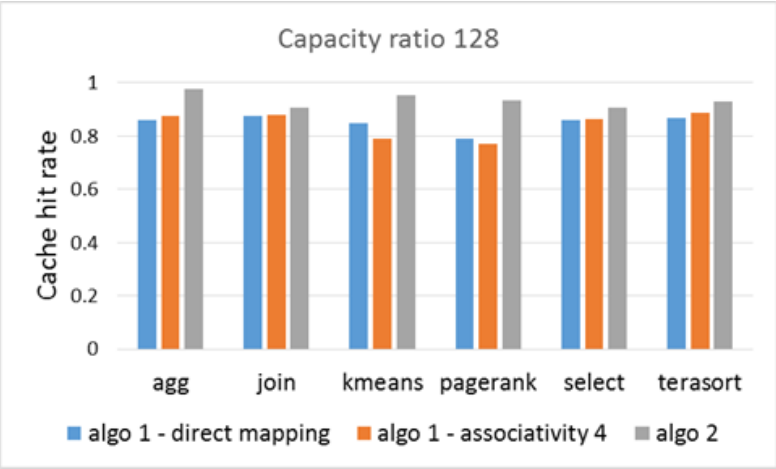
- TERASORT, SELECT, and JOIN have a regular hit rate decrease, almost linear
- JOIN and PAGERANK show a steep decrease in hit rate from 4:1 to 32:1 capacity ratio

- PAGERANK is the most affected by decrease in capacity ratio, with 46% decrease from 4:1 to 128:1 capacity ratio
- The behavior is unchanged whichever direct mapping or 4-way associative cache is used

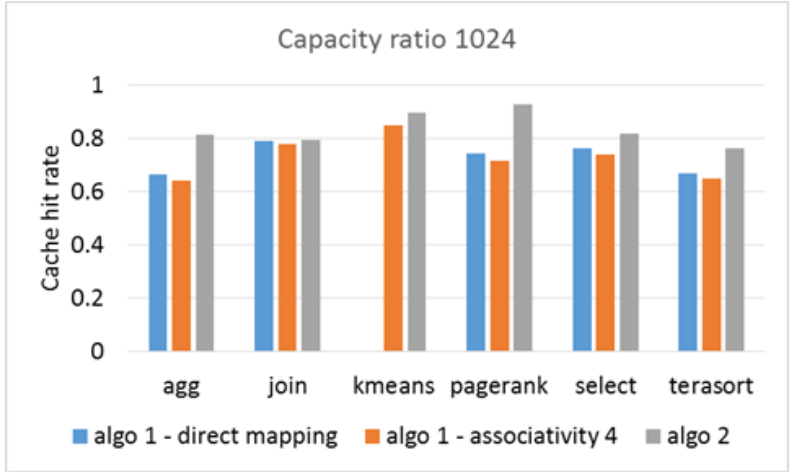
Improving Cache Traffic Management

Conventional cache management scheme vs. a different algorithm:

- We simulate the behavior of the two schemes over the first 10 million memory accesses
- Both solutions use the same prefetch scheme at the same granularity for fair comparison
- Both solutions use the same write/read access granularity



1) Hit rate could reach better than 90% for all studied big data benchmarks, when using algorithm 2 instead of algorithm 1
2) Hive AGGR and above all got hit rate performance boosted, PAGERANK hit rate increased 18% by using algorithm 2



1) Hit rate is increased for all studied big data benchmarks, when using algorithm 2 instead of algorithm 1
2) Hive AGG and above all got hit rate performance boosted. PAGERANK hit rate increased 25% by using algorithm 2

Conclusions

- Hybrid DIMM provides reduced latency due to less bus contention vs. separate DIMM design (best case –10 cycles; average –6 cycles)
- All big data applications have large footprint, but their frequently-accessed data, especially written pages, could potentially fit into DRAM L4-cache
- Optimizing cache hierarchy for writes can allow making the most of hybrid DIMM for such applications
- Improving hit rate can be done through cache traffic management algorithm, in various hybrid memory designs (25% hit rate increase)
- PM/NVM/NAND memory medias do not really affect the cache hit rates, but big data traffic patterns and caching algorithms affect hit performance, and their read latencies only marginally affect cache miss cases

Bibliography of authors

Dr. Xiaobing Lee is a Principal Engineer / Storage Architect in Huawei Storage Media Application Laboratory. He has over 25 years experience in ICT industries, including previous positions with AT&T labs, Sarnoff labs, SeaChange International, and Wegener Communications. His major interests include unified network and DRAM/NVM hybrid controllers, all-flash arrays SSD, LPDDR4-T / DDR4-T and NVDIMM-P to hybrid or hyper memories, emerging memories and applications.

Dr. Florian LONGNOS is a Research Engineer in Huawei Shannon Laboratory. He received his PhD Degree in microelectronics and nanotechnologies from Grenoble Institute of Technology, France, in 2014. He has been with Huawei Technologies Co. since 2015. His responsibilities and current research interests include computer architecture, memory sub-system design, emerging memory technologies.

Dr. Chen Shaojie is a research engineer at Huawei Shannon Lab since 2008. He received the Communication degree signal and information processing from Hangzhou Dianzi University, China, in 2008. His responsibilities are computer architecture, emerging memory technologies, memory controller design and memory protocol design. He attend the system architecture design and development of data-center server, such as High-Throughput Computer.