APACHE HBASE × MI

HBASECON ASIA2019

THE COMMUNITY EVENT FOR APACHE HBASE™

HBASECON ASIA2019

HBase Bucket Cache On Persistent Memory

Anoop Sam John, Ramkrishna S Vasudevan, Xu Kai

intel

Notices and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks.

Performance results are based on testing as of 06 24, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. Configuration: See slide 9
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at [intel.com].

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request. No product or component can be absolutely secure.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel, the Intel logo, 3D XPoint, Optane, Xeon, Xeon logos, and Intel Optane logo are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others

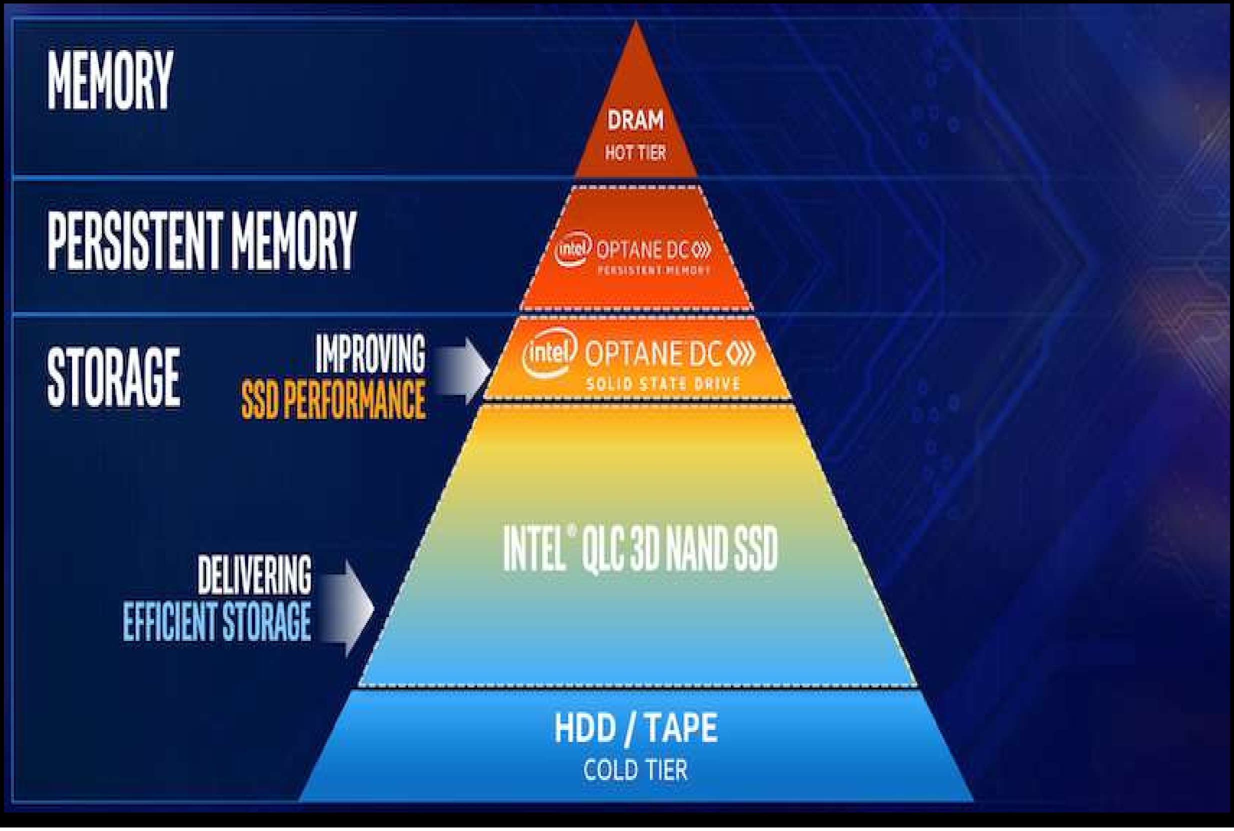
Persistent Memory Technology

Operation modes

Bucket cache On Persistent Memory

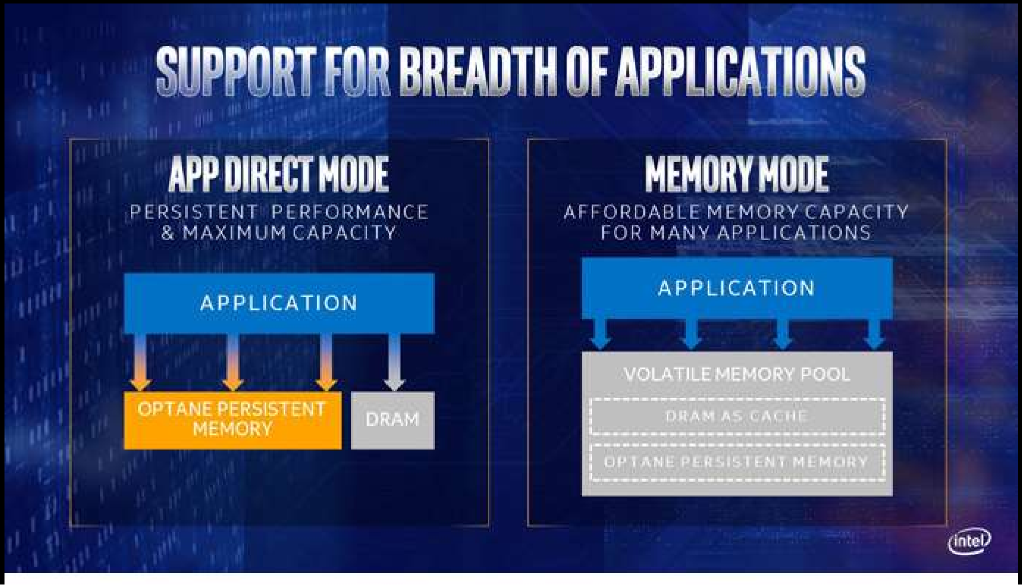
Performance Numbers

Persistent memory Technology



Memory modes

2 LM – Memory mode	App direct mode
Transparent to applications	Application is aware of Pmem and DRAM
DRAM acts as first level cache	Application decides whether to use DRAM or Pmem (HBase Bucket cache)
No persistence available, huge memory is made available for applications	Persistence available



SUPPORT FOR BREADTH OF APPLICATIONS

APP DIRECT MODE

PERSISTENT PERFORMANCE & MAXIMUM CAPACITY

APPLICATION

OPTANE PERSISTENT MEMORY

DRAM

MEMORY MODE

AFFORDABLE MEMORY CAPACITY FOR MANY APPLICATIONS

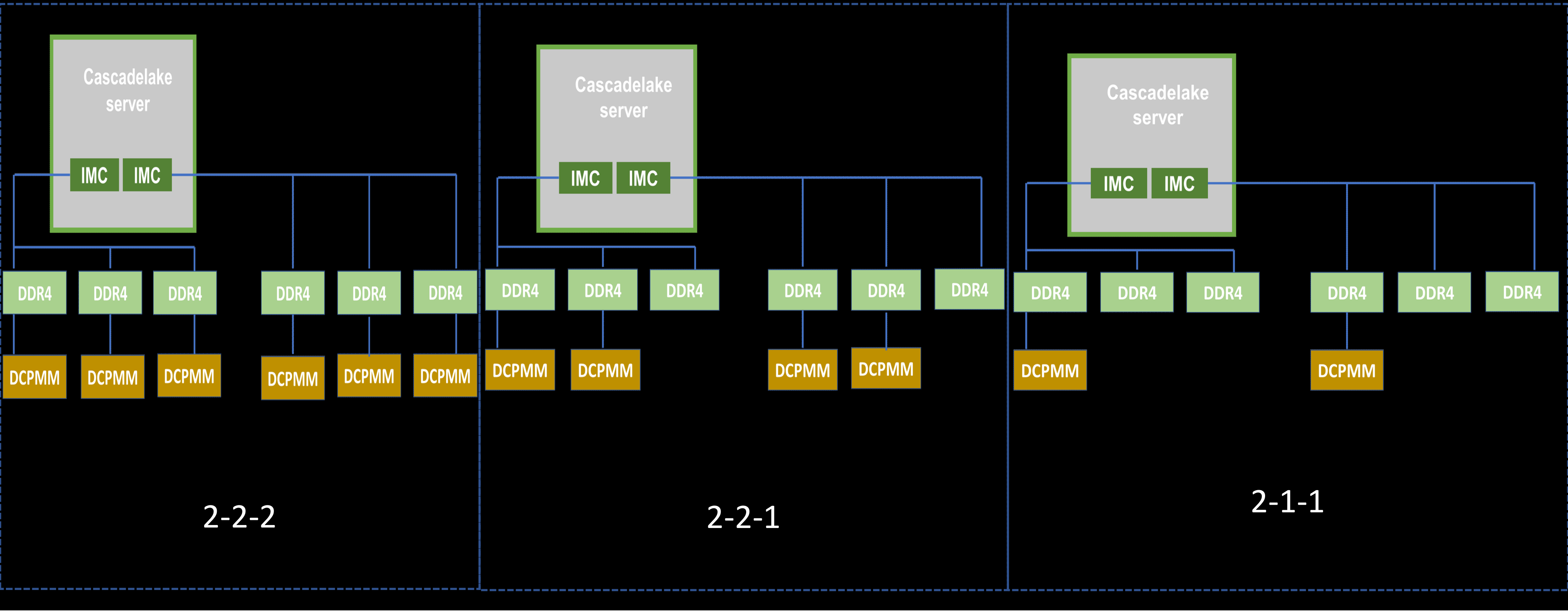
APPLICATION

VOLATILE MEMORY POOL

DRAM AS CACHE

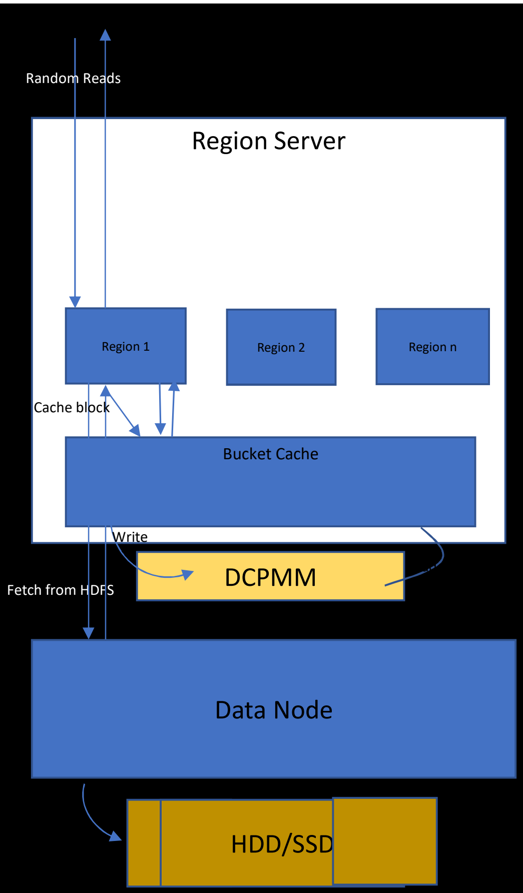
OPTANE PERSISTENT MEMORY

Configuration modes



BUCKET CACHE ON PERSISTENT MEMORY

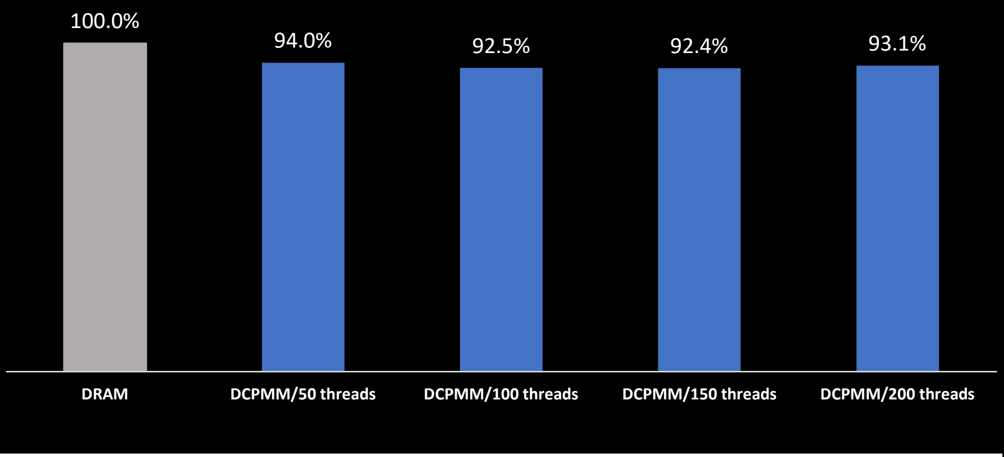
- HBase Bucket Cache overview:
  - Data read from HDFS is cached in BlockCache
  - HBase has various implementations of BlockCache
  - BucketCache is one implementation of Block Cache
    - BucketCache is allocated on DCPMM/DRAM using Java DirectByteBuffer mechanism
    - Modes: offheap (DRAM), file, mmap
    - New Mode : pmem (HBASE-21874), included in CDH6.2.0
    - Supports large BlockCache for high performance
    - Large BlockCache -> low latency and higher throughput
- This case study is with Bucket Cache in Offheap(DRAM) vs Pmem(DCPMM)
  - Equivalent capacity, DCPMM can be much cheaper than DRAM with minor performance drop
  - Same/Similar cost, DCPMM gives a larger size compared to DRAM, which means more data in cache and better latency/throughput
  - *Though the server with DCPMM has DRAM also, note that in DCPMM tests the amount of DRAM has no role to play in the bucket cache experiment.*



HBase Bucket Cache with Intel® Optane™ DC Persistent Memory – Similar Capacity

up to *94.0% * performance of DRAM when DCPMM/DRAM have similar capacity and all the data can fit within DCPMM/DRAM**

	DRAM Cluster	DCPMM Cluster
# of workers	1	
Processors	2nd Gen Intel Xeon Gold 6240 (dual sockets)(Casacade lake)	
DRAM	1.5TB (24 * 64GB)	192 GB (12 * 16GB)
DCPMM	N/A	1.5TB (12 * 128GB)
DCPMM Config	N/A	2-2-2 (AppDirect mode)
Storage	7.68 TB – 8 * 960GB SATA3 SSD	
Network	10Gb Ethernet	



HBase Random-Read Normalized Performance (x)

DCPMM vs. DRAM

TPS(Transaction per Second), Higher is Better

Benchmark kit	HBase performance evaluation tool
Dataset Size	1200GB (Data can fit within both DRAM and DCPMM)
SUT (system under test)	CDH 6.2.0

Performance results are based on testing as of 06 24, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks.

HBASECON ASIA2019

HBase Bucket Cache with Intel® Optane™ DC Persistent Memory – Similar Cost

up to *13.2X performance of DRAM when all of the data can fit within DCPMM, about ⅔ of the data can fit within DRAM**

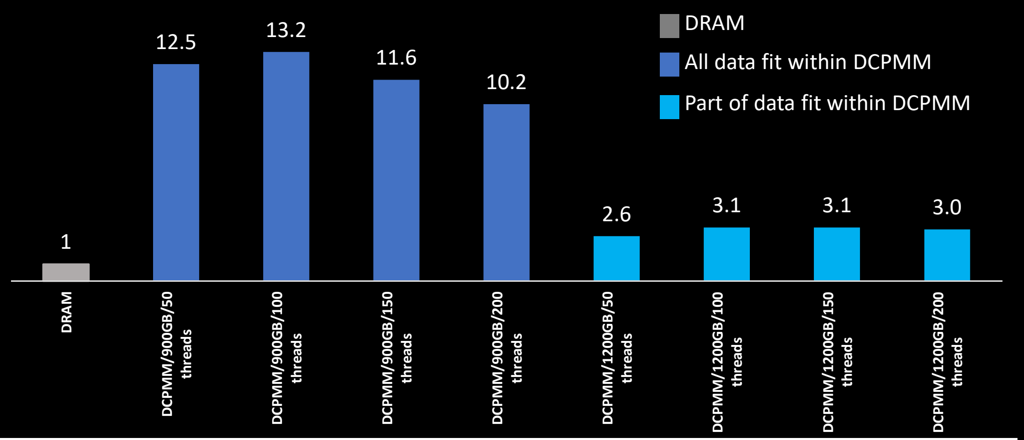
up to *3.1X performance of DRAM when all data can not fit within DCPMM/DRAM, but DCPMM can hold more(DCPMM holds about ¾ , DRAM holds about ½)**

	DRAM Cluster	DCPMM Cluster
# of workers	1	
Processors	2nd Gen Intel Xeon Gold 6240 (dual sockets)(Cascade lake)	
DRAM	768GB (12 * 64GB)	192 GB (12 * 16GB)
DCPMM	N/A	1TB (8 * 128GB)
DCPMM Config	N/A	2-2-1 (AppDirect mode)
Storage	7.68 TB – 8 * 960GB SATA3 SSD	
Network	10Gb Ethernet	

HBase Random-Read Normalized Performance (x)

DCPMM vs. DRAM

TPS(Transaction per Second), Higher is Better



Benchmark kit	HBase performance evaluation tool
Dataset Size	900GB (All data can fit within DCPMM, but part fit within DRAM) 1200GB (Data can not all fit within DCPMM/DRAM, but DCPMM can hold more)
SUT (system under test)	CDH 6.2.0

Performance results are based on testing as of 06 24, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks.

HBase release and JIRA

https://issues.apache.org/jira/browse/HBASE-21874 - Bucket cache on Persistent memory

Available in CDH 6.2.0 release.

Future Work

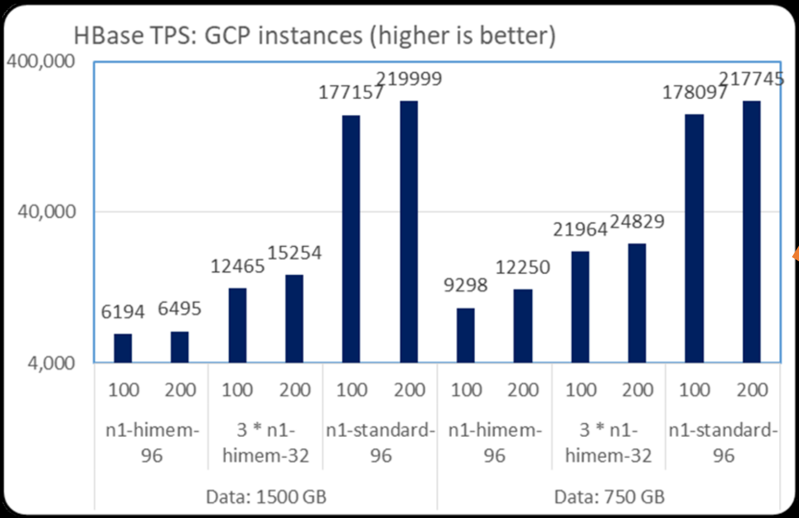
- Support Tiered cache – with cache residing on both DRAM and DCPMM.
- JDK support for DCPMM with old gen objects residing on DCPMM.
- Support DCPMM in write path (WALLess HBase).

Thanks！

Backup

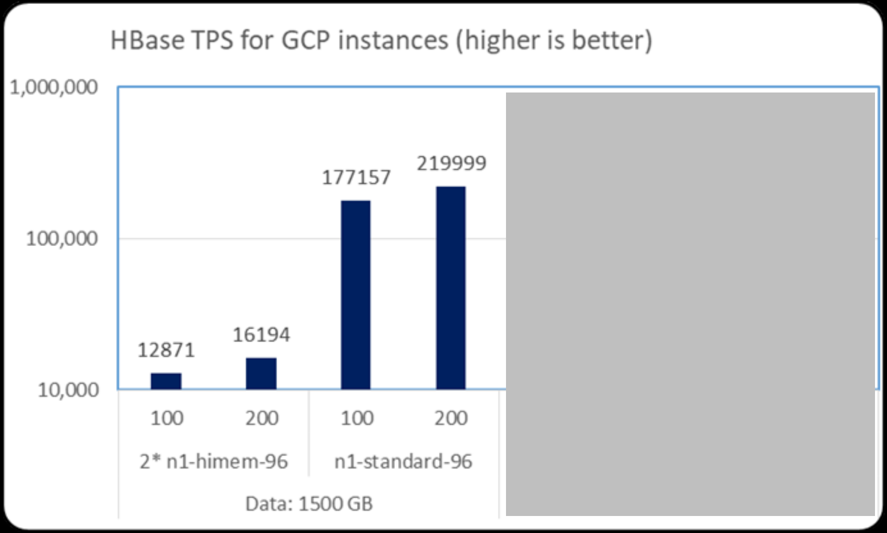
Tests on GCP

Name	n1-himem-96(DRAM only)	n1-himem-32(DRAM only)	n1-standard-96 (DRAM + AEP)
# of instances	Varies by test	3	1
CPU vCores Freq: Base/Turbo Cache	Xeon 96 2.0 GHz/false 55 MB	Xeon 32	Xeon 96 2.0 GHz/false 55 MB
DDR4 Memory	624 GB	208 GB	192 GB
HBase Bucket Cache	550 GB	180 GB	1500 GB
DCPMM Memory	NA	NA	1.6 TB (AD mode)
Storage	1 * 2 TB “SSD persistent disk”	1 * 2 TB “SSD persistent disk”	1 * 2 TB “SSD persistent disk”



- 28x (177157 vs. 6194) to 33x (219999 vs. 6495) TPS* speedup using one DCPMM-based instance compared with DRAM-only instances.

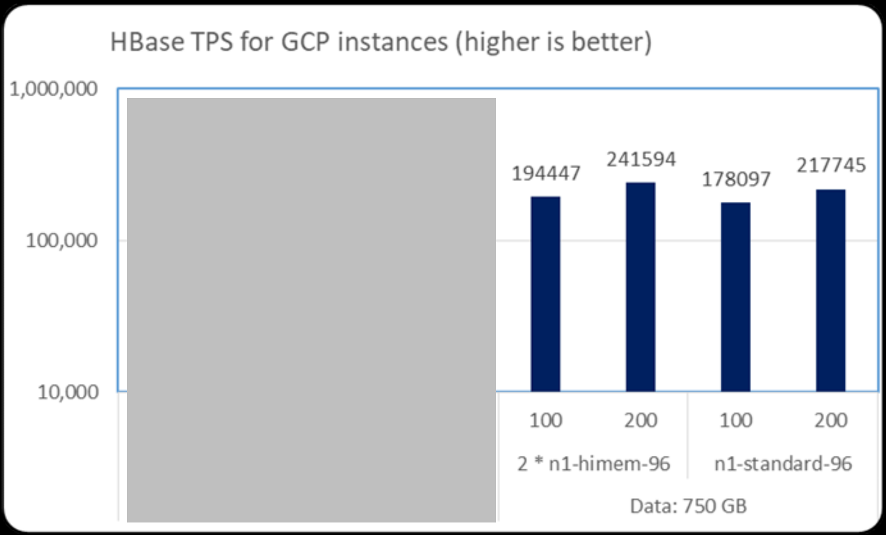
- 13.5x (219999 vs. 16194) to 13.7x (177157 vs. 12871) TPS* speedup using one DCPMM-based instance compared with two DRAM-only instances.



Scenario:
HBase data exceeds DRAM bucket-cache (550 GB), HBase data fits within DCPMM bucket-cache (1.5 TB)

Scenario:
HBase data exceeds DRAM bucket-cache (2 * 550 GB), HBase data fits within DCPMM bucket-cache (1.5 TB)

* TPS: Transactions Per Second

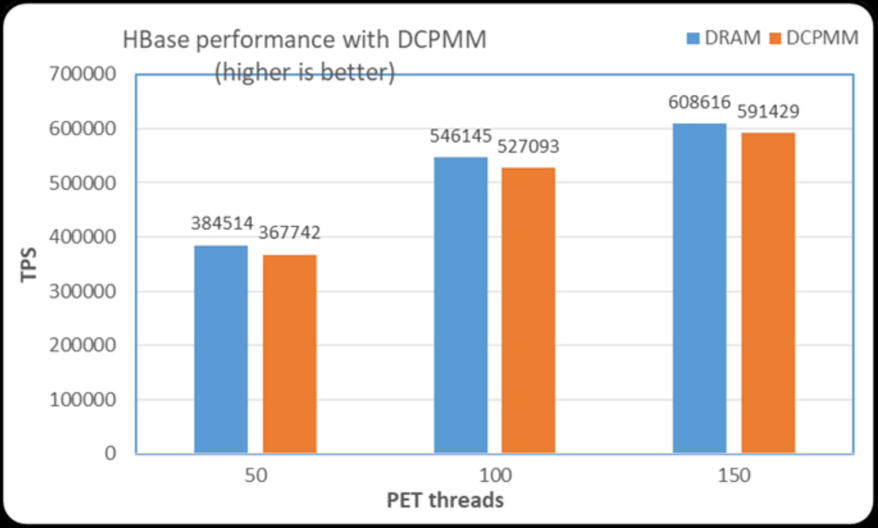


- TPS* using one DCPMM-based instance is 90.1% (217745 vs. 241594) to 91.5% (178097 vs. 194447) of the TPS using two DRAM-only instances.

Scenario:
HBase data fits within DRAM bucket-cache (2 * 550 GB),
HBase data fits within DCPMM bucket-cache (1.5 TB)

HBASECON ASIA2019

Tests on HPE infrastructure



* TPS: Transactions Per Second

Scenario:
HBase hot data (1.2 TB) fits within DCPMM bucket-cache (1.2 TB)
HBase hot data (1.2 TB) fits within DRAM bucket-cache (1.2 TB)