Intel Optane Pmem Tier in data analytics

Xiaolei Ren Intel DCG Sales TSS

Agenda

- Intel Optane PMem introduction
- Alluxio introduction and Pmem enabling
- Alluxio performance Result
- Intel Optane Pmem usage in Data Analytics
- Summary

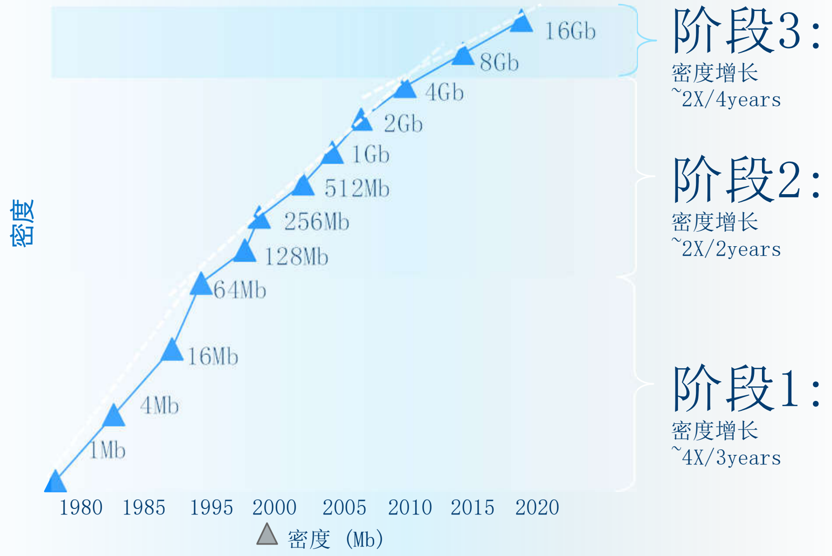
Intel Confidential

INTEL OPTANE PERSISTENT MEMORY

Intel Confidential

内存技术的扩展趋势

DRAM的密度增长率正在减慢

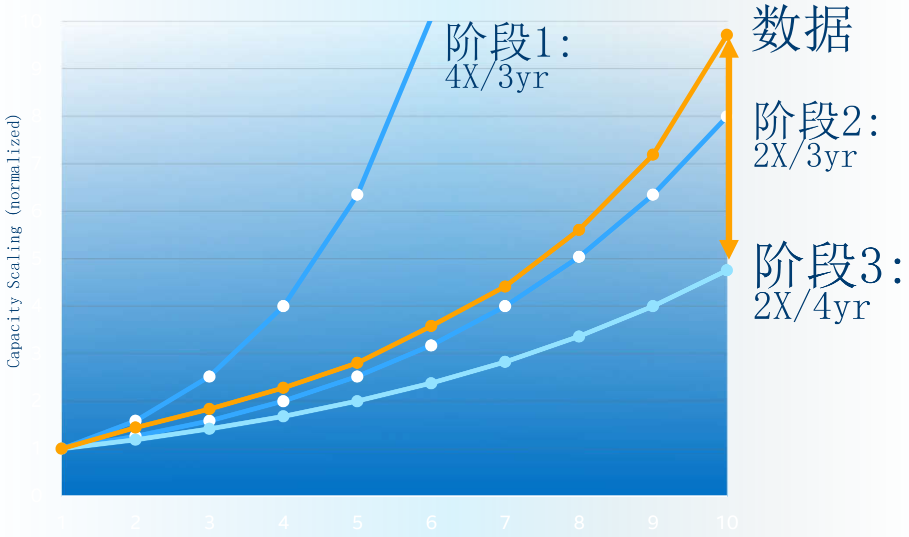


阶段3:
密度增长
~2X/4years

阶段2:
密度增长
~2X/2years

阶段1:
密度增长
~4X/3years

数据和内存容量之间的差距正在扩大



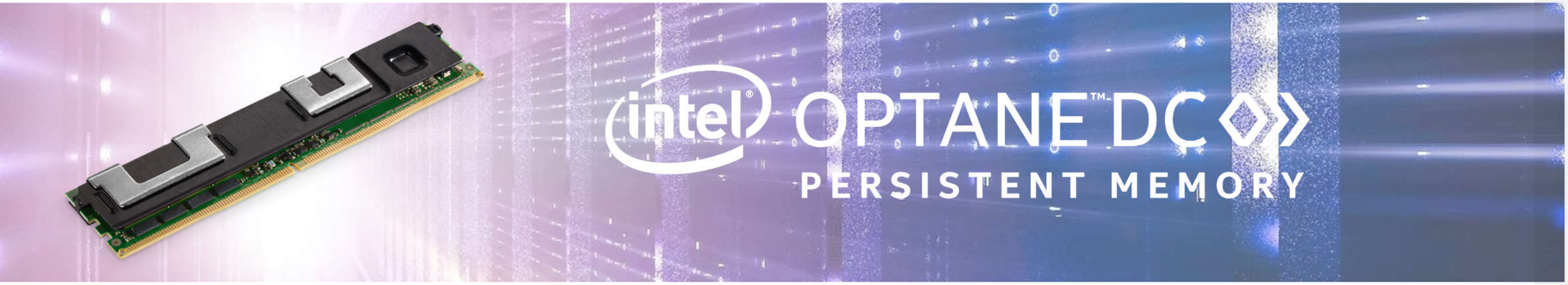
阶段1:
4X/3yr

数据

阶段2:
2X/3yr

阶段3:
2X/4yr

新硬件给客户带来新价值



Deliver **performance/TCO$** benefits on **large capacity systems**, to support **key workloads** to **save more, do more and go faster**

容量大

High capacity enabling systems with >3 TB of memory per CPU socket

128GB, 256GB, 512GB 三种规格

天然的非易失特性

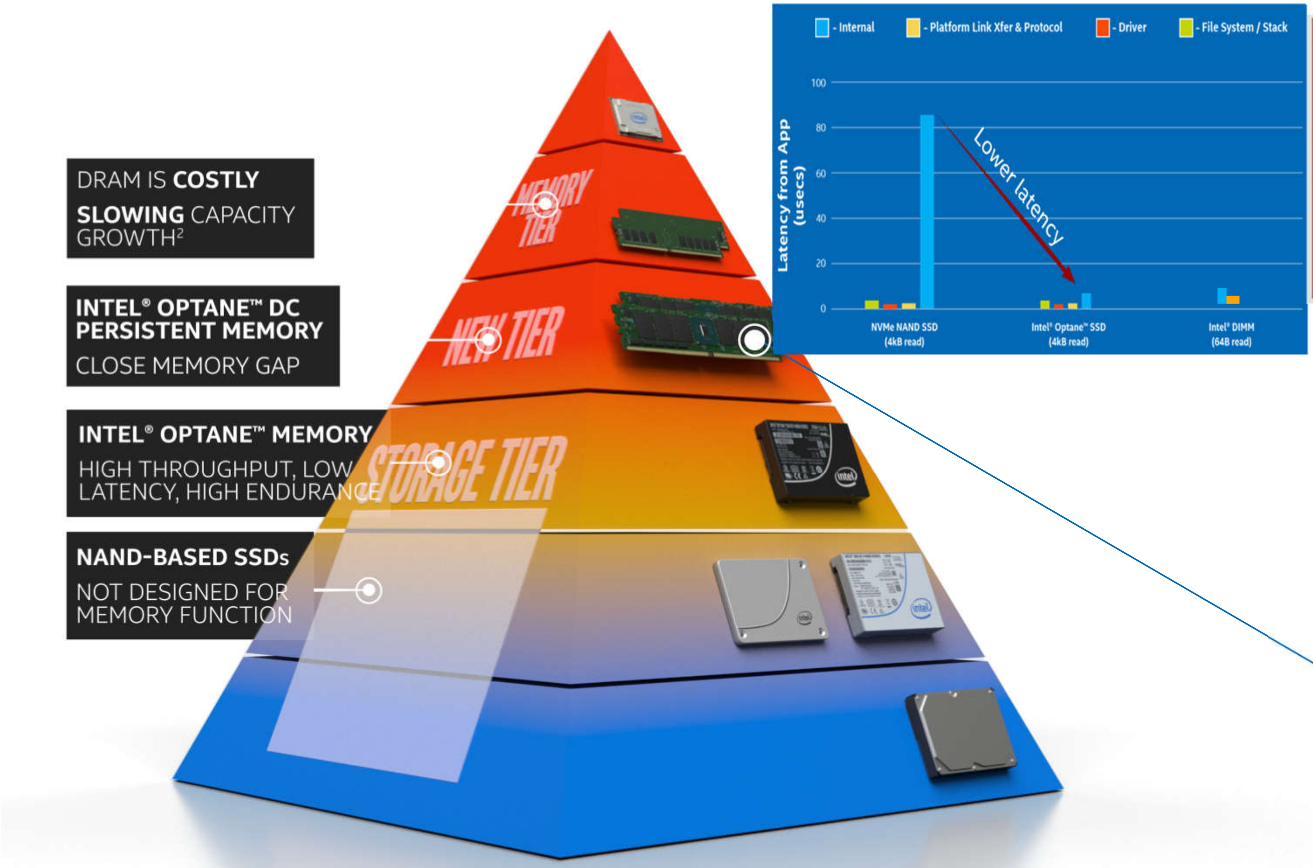
Persistent, with near-memory performance, HW encryption

高性价比

More capacity for your investment.

英特尔® 傲腾™ 持久内存的定位

INTEL OPTANE DC PERSISTENT MEMORY

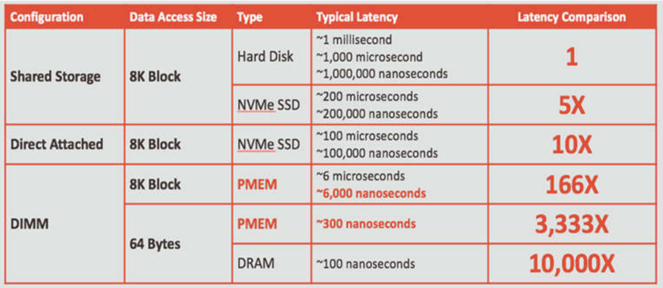


DRAM IS **COSTLY**
SLOWING CAPACITY GROWTH²

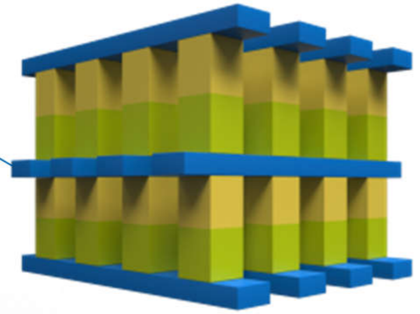
INTEL® OPTANE™ DC PERSISTENT MEMORY
CLOSE MEMORY GAP

INTEL® OPTANE™ MEMORY
HIGH THROUGHPUT, LOW LATENCY, HIGH ENDURANCE

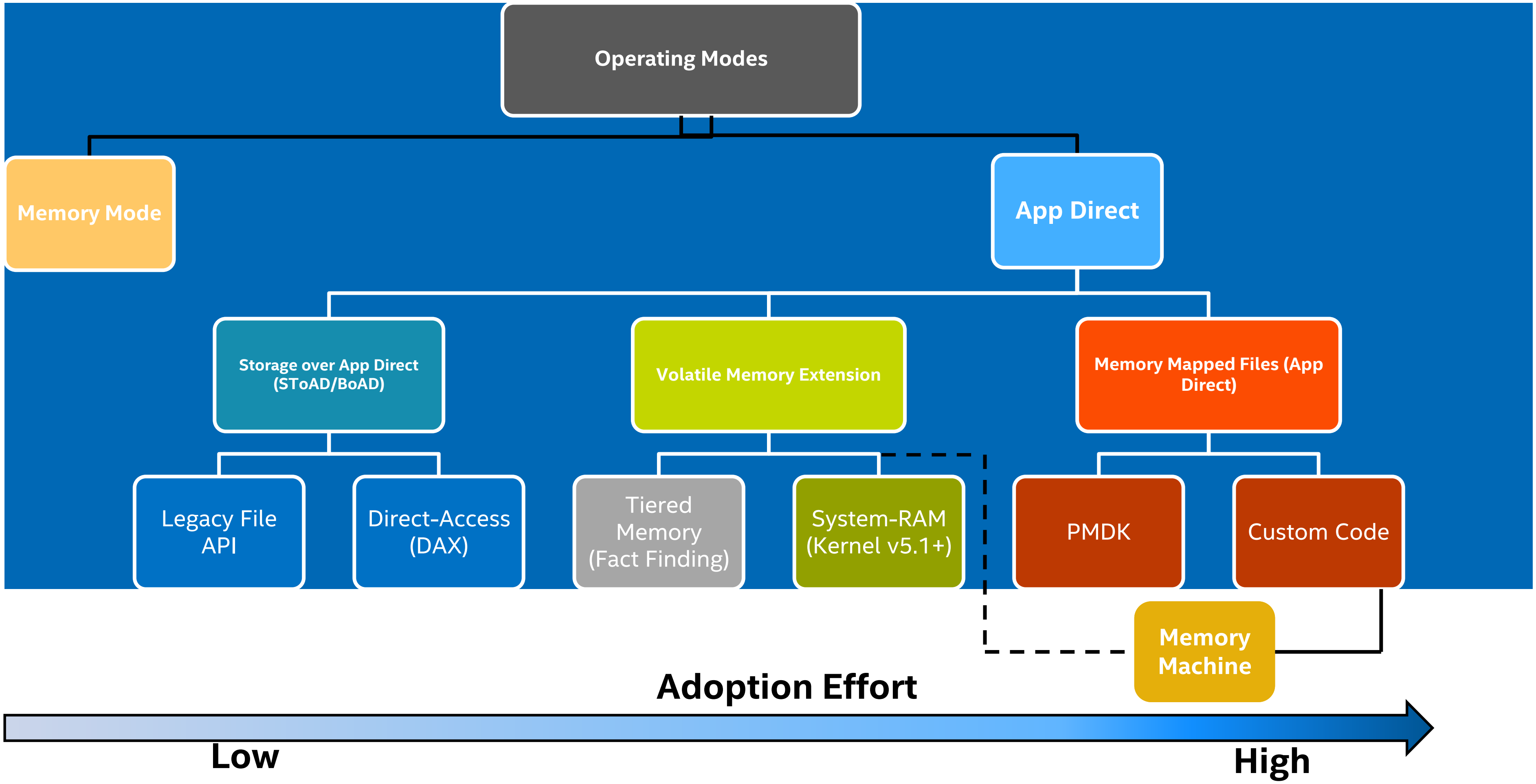
NAND-BASED SSDs
NOT DESIGNED FOR MEMORY FUNCTION



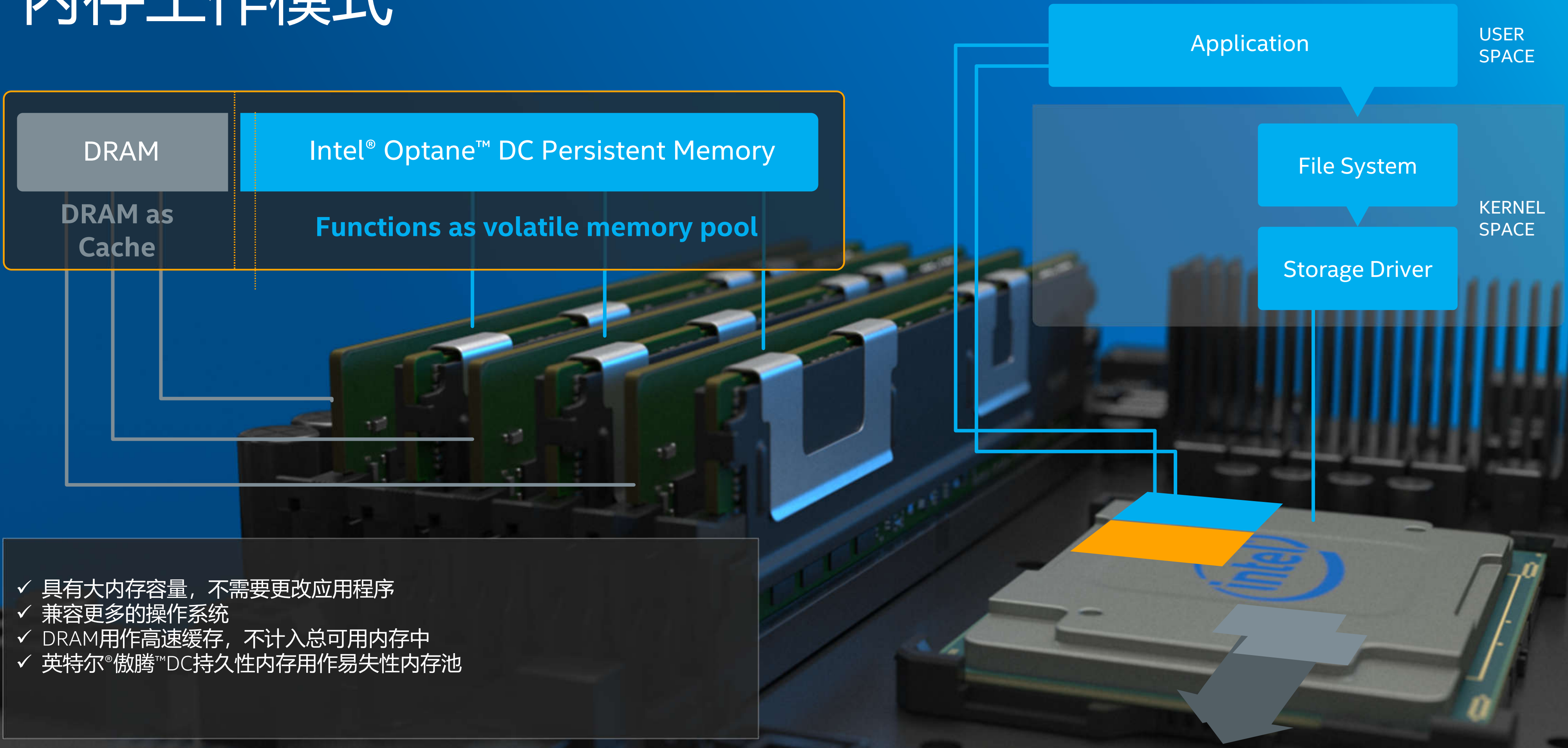
Configuration	Data Access Size	Type	Typical Latency	Latency Comparison
Shared Storage	8K Block	Hard Disk	~1 millisecond ~1,000 microsecond ~1,000,000 nanoseconds	1
		NVMe SSD	~200 microseconds ~200,000 nanoseconds	5X
Direct Attached	8K Block	NVMe SSD	~100 microseconds ~100,000 nanoseconds	10X
DIMM	8K Block	PMEM	~6 microseconds ~6,000 nanoseconds	166X
	64 Bytes	PMEM	~300 nanoseconds	3,333X
		DRAM	~100 nanoseconds	10,000X



傲腾持久内存的操作模式

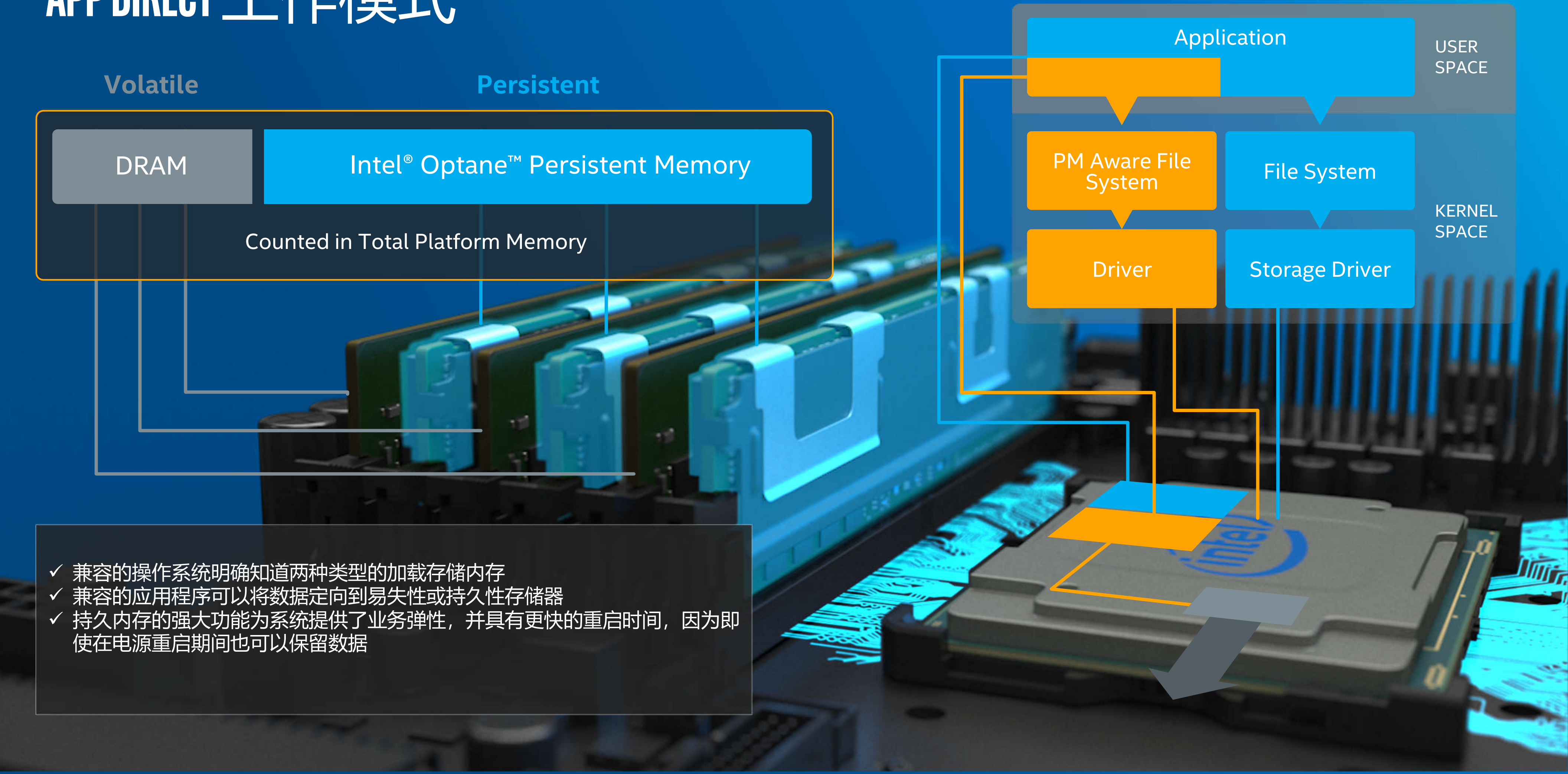

Operating Modes
Memory Mode
App Direct
Storage over App Direct (SToAD/BoAD)
Volatile Memory Extension
Memory Mapped Files (App Direct)
Legacy File API
Direct-Access (DAX)
Tiered Memory (Fact Finding)
System-RAM (Kernel v5.1+)
PMDK
Custom Code
Memory Machine
Adoption Effort
Low
High

内存工作模式



✓ 具有大内存容量，不需要更改应用程序
✓ 兼容更多的操作系统
✓ DRAM用作高速缓存，不计入总可用内存中
✓ 英特尔®傲腾™DC持久性内存用作易失性内存池

APP DIRECT 工作模式



Volatile

Persistent

DRAM

Intel® Optane™ Persistent Memory

Counted in Total Platform Memory

Application

USER SPACE

PM Aware File System

File System

KERNEL SPACE

Driver

Storage Driver

- ✓ 兼容的操作系统明确知道两种类型的加载存储内存
- ✓ 兼容的应用程序可以将数据定向到易失性或持久性存储器
- ✓ 持久内存的强大功能为系统提供了业务弹性，并具有更快的重启时间，因为即使在电源重启期间也可以保留数据

PMDK LIBRARIES

http://pmem.io

https://github.com/pmem/pmdk

High Level Interfaces
(in development)

Experimental C++ Persistent Containers

PCJ – Persistent Collection for Java

pmemkv

Language bindings

C++

C

JAVA

Python

Transaction Support

Interface to create a persistent memory resident log file

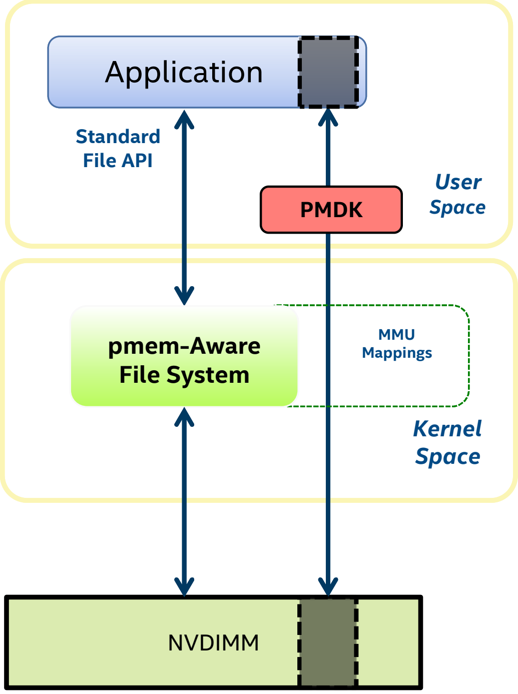
libpmemlog

Interface for persistent memory allocation, transactions and general facilities

libpmemobj

Interface to create arrays of pmem-resident blocks, of same size, atomically updated

libpmemblk



Support for **volatile** memory usage

memkind

vmemcache

Low level support for local persistent memory

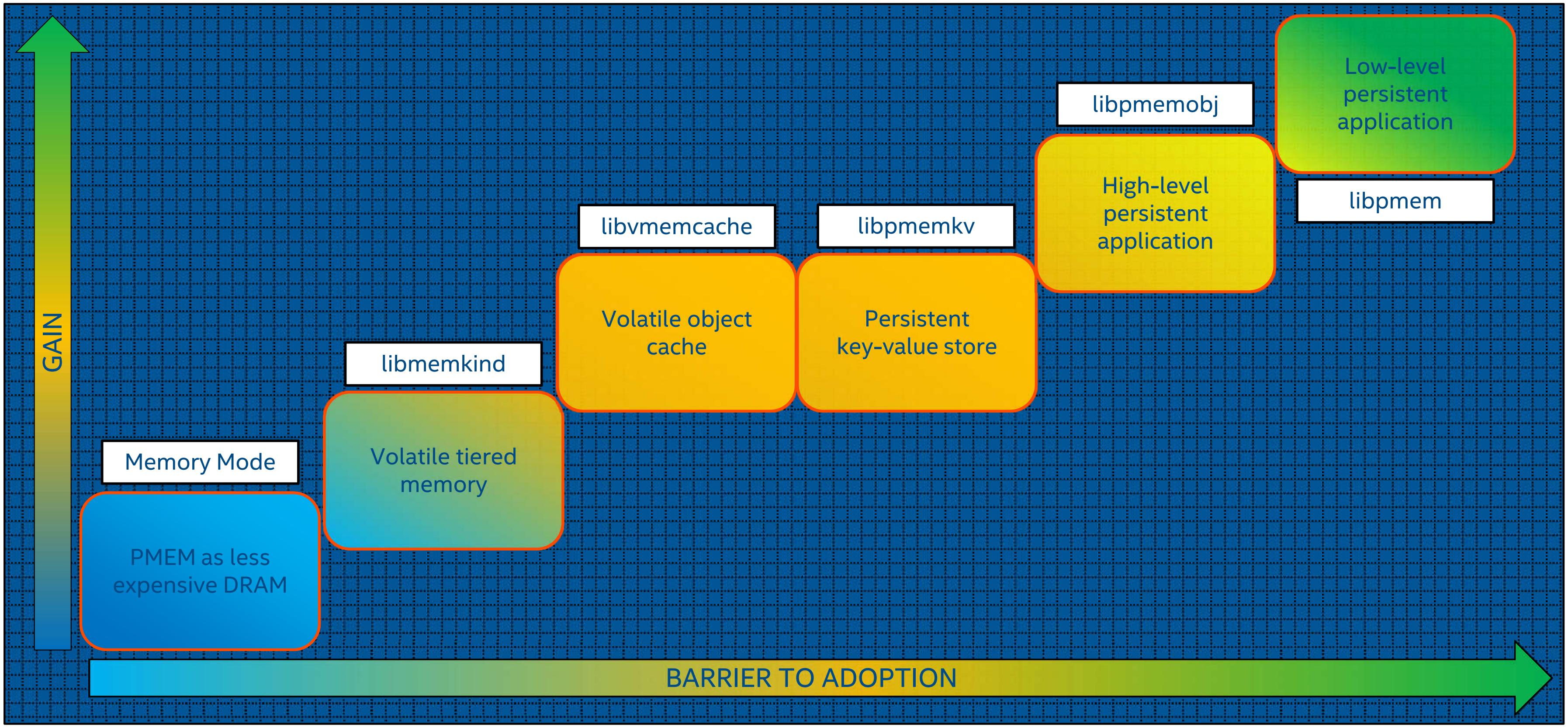
libpmem

Low level support for remote access to persistent memory

librpmem

Low-level support

DIFFERENT WAYS TO USE PERSISTENT MEMORY



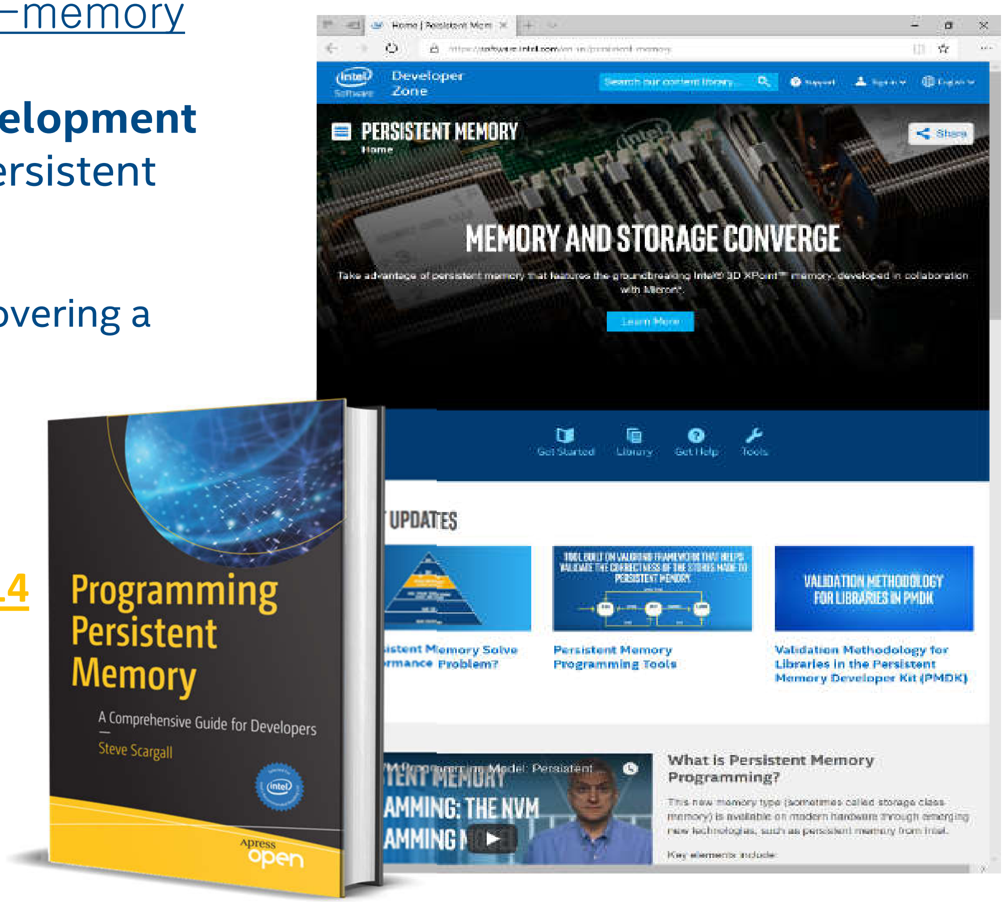
PERSISTENT MEMORY PROGRAMMING @ INTEL DEVELOPER ZONE

https://software.intel.com/en-us/persistent-memory

Learn how to use the **Persistent Memory Development Kit (PMDK)** to create or update apps to use persistent memory, with:

- 30+ technical articles, webinars and videos covering a broad range of related topics
- Code samples for C/C++, Java and Python

https://www.apress.com/us/book/9781484249314



PERSISTENT MEMORY

Home

MEMORY AND STORAGE CONVERGE

Take advantage of persistent memory that features the groundbreaking Intel® 3D XPoint™ memory, developed in collaboration with Micron®.

Learn More

UPDATES

VALIDATION METHODOLOGY FOR LIBRARIES IN PMDK

...istent Memory Solve ...rmance Problem?

Persistent Memory Programming Tools

Validation Methodology for Libraries in the Persistent Memory Developer Kit (PMDK)

What is Persistent Memory Programming?

This new memory type (sometimes called storage class memory) is available on modern hardware through emerging new technologies, such as persistent memory from Intel.

Key elements include:

傲腾持久内存的适用场景下的用户价值

用户的诉求

- DRAM 成本太高
- 纵向扩展太贵
- 内存容量不够
- 运行低效
- 系统性能不佳
- 存储太慢

USE INTEL ® OPTANE™ DC PERSISTENT MEMORY TO…

$ 降低成本

部分替代DRAM

Systems >512GB

改善TCO

Workloads that need large &/or persistent memory

增加负荷

增加内存容量

Large memory or SW license fees per core

合并负荷

High VMs, with low CPU utilization

运行更快

打破I/O瓶颈

3
High Disk I/O Traffic

高速存储

Tiered storage subsystem

傲腾持久内存的潜使用场景

最优先方向

云和虚拟化	数据库	存储	高性能计算	大数据分析	电信
扩展虚机内存容量	内存数据库	超快存储	大内存应用	堆外内存	NFVi
增加系统虚机和容器数量	大容量数据库缓存层	Meta-data管理	任务断点检查	实时分析	认知网络
	数据库日志	写缓存	PMoF	新兴分析平台	内容分发网络 (CDN)
OS 内存扩展	RDMA 复制	存储缓存层	文件系统交换	AI 数据分析	

云 & IaaS

大内存增加虚拟机密度，降本增效

案例
- 内存增加 改善在虚拟化的多租户环境中运行的工作负载
- 提高容器应用和云服务能力

内存数据库 / 超融合

大内存增加负荷，降成本提高性能

案例
- SAP HANA, Oracle Exadata, MSFT SQL
- Redis
- Cassandra, MongoDB

大数据分析和AI

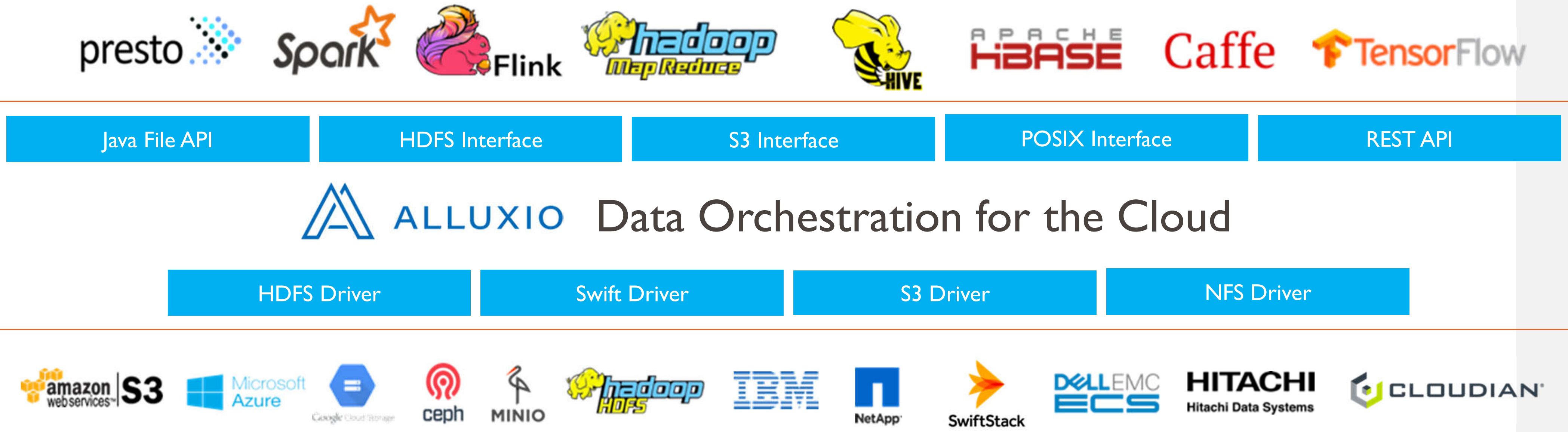
高性能，降成本，数据持久性

案例
- 企业和云存储的大容量高性能非易失缓存
- 网络存储的本地大容量缓存
- 高性能直连存储

ALLUXIO INTRODUCTION AND PMEM ENABLING

Intel Confidential

Data Orchestration for the Cloud



Java File API | HDFS Interface | S3 Interface | POSIX Interface | REST API

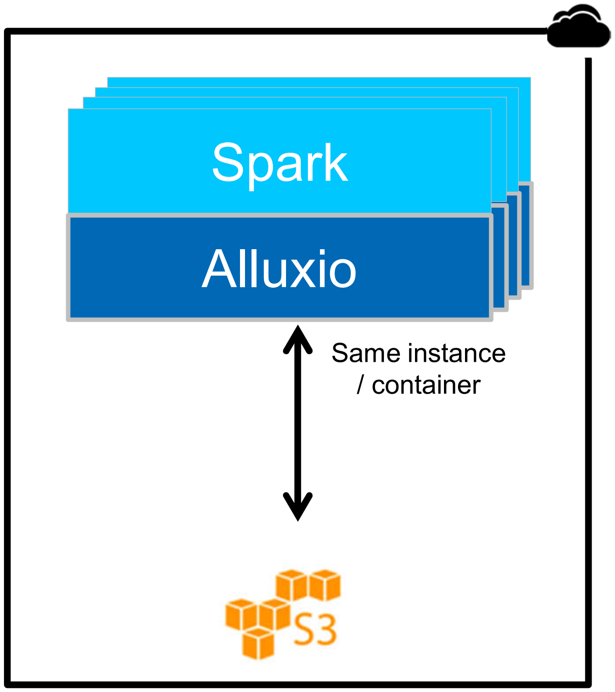
ALLUXIO Data Orchestration for the Cloud

HDFS Driver | Swift Driver | S3 Driver | NFS Driver

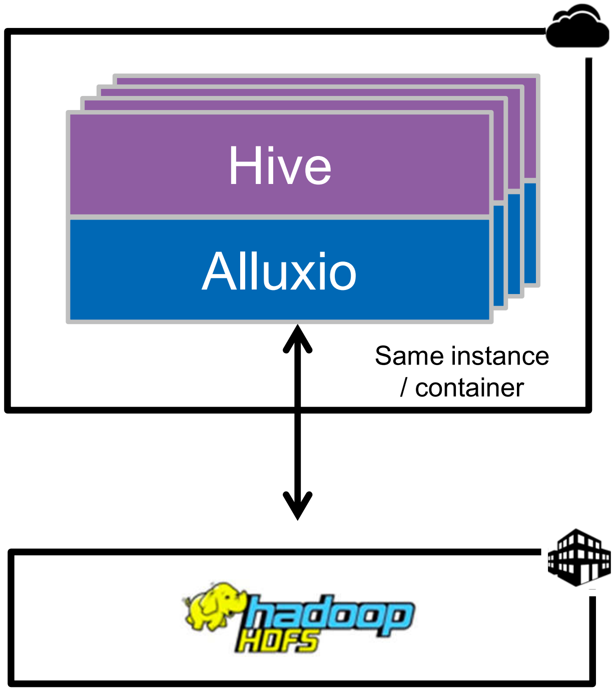
source: Alluxio

USE CASES ALLUXIO ENABLES

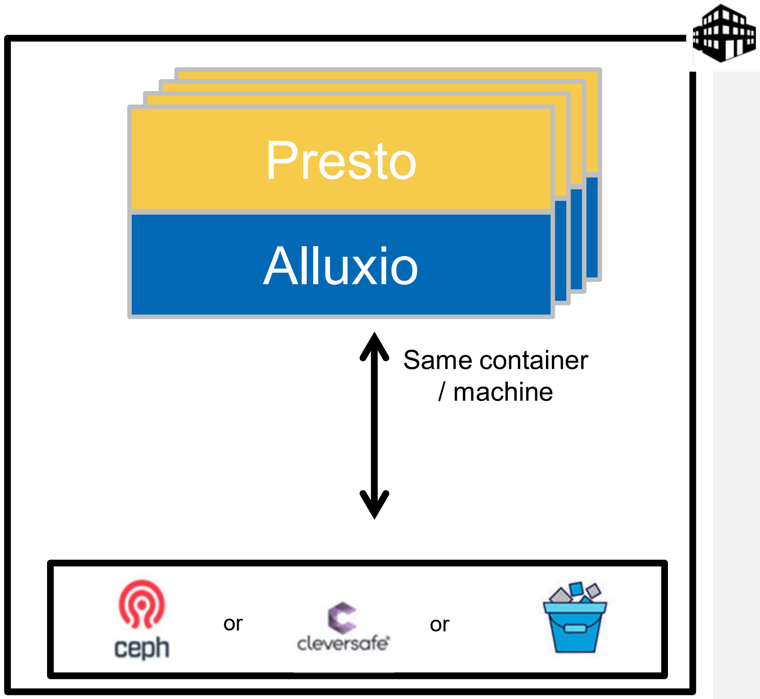
Accelerate big data frameworks on the public cloud



Burst big data workloads in hybrid cloud environments



Dramatically speed-up big data on object stores on premise



source: Alluxio

Alluxio – Key innovations

Data Locality
with Intelligent
Multi-tiering

Accelerate big data
workloads with transparent
tiered local data

Data Accessibility
for popular APIs &
API translation

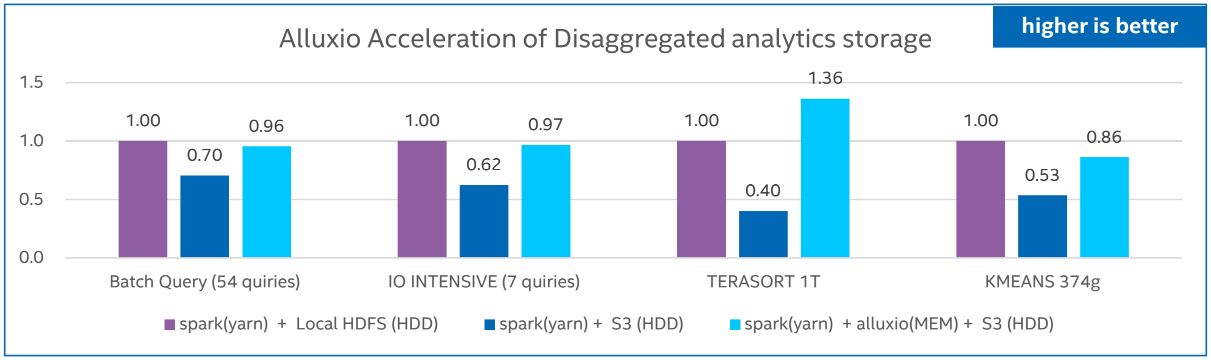
Run Spark, Hive, Presto, ML
workloads on your data
located anywhere

Data Elasticity
with a unified
namespace

Abstract data silos & storage
systems to independently scale
data on-demand with compute

source: Alluxio

ALLUXIO CACHE PERFORMANCE ACCELERATION FOR COMPUTE AND STORAGE DISAGGREGATED ARCHITECTURE



Using Alluxio IMDA as cache:

- For terasort, **3.4x** speedup over S3 object storage, **1.36x** speedup over local HDFS.
- For TPCDS test, up to **1.56x** performance speedup for IO intensive queries, slightly lower than local HDFS.
- For KMeans test, **1.62x** speedup over S3 object storage, 14% lower compared with local HDFS.

  https://www.alluxio.io/blog/speeding-big-data-analytics-on-the-cloud-with-in-memory-data-accelerator/

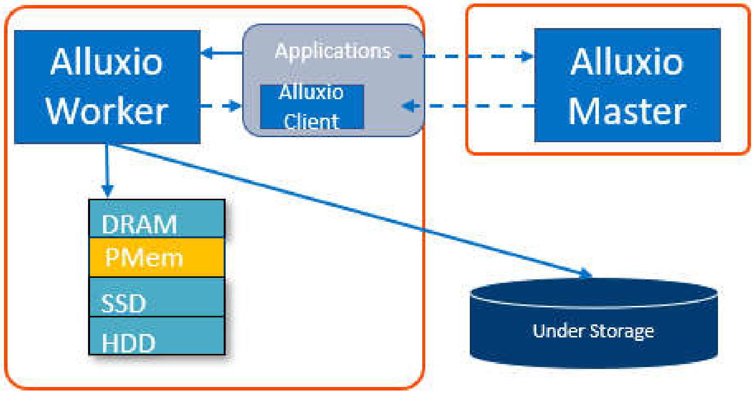
Using Alluxio IMDA cache improved in IO intensive workloads

ALLUXIO PMEM TIER ARCHITECTURE DESIGN

- Alluxio PMem tier
  - A new PMEM tier layer introduced to provide higher performance with lower cost
    - Large Capacity -> Cache more data
    - Higher performance compared with NVMe SSD
    - Deliver dedicated SLA to mission critical applications

PMem Mode

- AD mode: Leverage PMDK lib to bypass filesyster overhead and context switches, 90% POC code completed
- SoAD: No code changes, Performance evolution and value proposition working in progress
- 2LM mode: using Alluxio DRAM tier



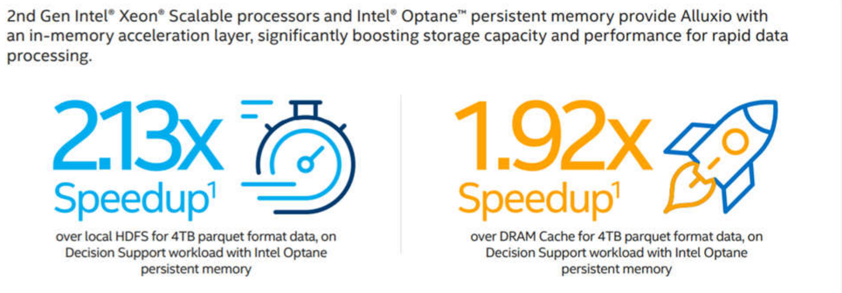
ALLUXIO & ALLUXIO PMEM TIER

Alluxio PMem tier demonstrated significant performance speedup over DRAM tier in ISO-cost configuration

- PMem tier (SoAD) performance with Decision support workload 4TB (parquet format)
  - 2.13x speedup over without cache configuration
  - 1.92x speedup over DRAM tier in ISO-cost configuration
- PMem tier (2LM) performance with Decision support workload 4TB (parquet format)
  - 1.24x speedup over without cache configuration
  - 1.12x speedup over DRAM tier in ISO-cost configuration

GTM

- Alluxio Launches Enhanced Hybrid Cloud Solution based on Intel Optane Persistent Memory
- Press rlease, blog, white paper, solution brief ready



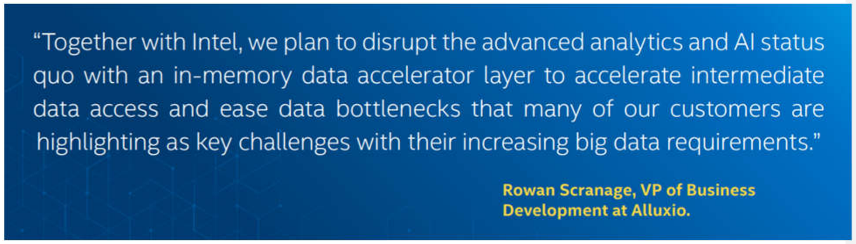
2nd Gen Intel® Xeon® Scalable processors and Intel® Optane™ persistent memory provide Alluxio with an in-memory acceleration layer, significantly boosting storage capacity and performance for rapid data processing.

2.13x Speedup[1]

over local HDFS for 4TB parquet format data, on Decision Support workload with Intel Optane persistent memory

1.92x Speedup[1]

over DRAM Cache for 4TB parquet format data, on Decision Support workload with Intel Optane persistent memory



"Together with Intel, we plan to disrupt the advanced analytics and AI status quo with an in-memory data accelerator layer to accelerate intermediate data access and ease data bottlenecks that many of our customers are highlighting as key challenges with their increasing big data requirements."

Rowan Scranage, VP of Business Development at Alluxio.

Alluxio launch PMem based enhanced hybrid cloud solutions: https://www.alluxio.io/press-releases/alluxio-launches-enhanced-hybrid-cloud-solution-based-on-intel-optane-persistent-memory/

Alluxio blog: https://www.alluxio.io/resources/whitepapers/accelerate-and-scale-big-data-analytics-with-alluxio-and-intel-optane-persistent-memory

White Paper: https://www.alluxio.io/app/uploads/2020/05/Intel-Alluxio-PMem-Whitepaper-200507.pdf

Solution brief: https://www.alluxio.io/app/uploads/2020/04/Alluxio-Intel_SolutionsBrief_Final1.pdf

Intel Confidential

ALLUXIO PMEM TIER PERFORMANCE

Intel Confidential

SYSTEM CONFIG

Red=Required
Black=Required if used
Blue=Internal Intel tracking

SVRINFO: Tool to capture hardware information and mitigation status; Software details will still need to be entered manually

https://github.intel.com/ssgcce/svrinfo

Attach svrinfo log file: use same name format as presentation.

PMem SoAD	PMem 2LM	DRAM
svr_info.html	svr_info.html	svr_info.html

Note on OS/Kernel for CLX:

CLX hardware mitigations require:

RHEL or CentOS requires 3.10.0-957.el7.x86_64 or newer kernel. Kernel.org version should be 4.19 or newer. Older kernels will only provide software mitigations which could mean lower performance

	PMem Tier	DRAM tier
Test by	Intel	Intel
Test date	12/06/2019	12/06/2019
SUT Setup		
Platform	S2600WF0	S2600WF0
# Nodes	2 compute, 3 storage	2 compute, 3 storage
# Sockets	2	2
CPU	6240 for compute, 6140 for storage	6240 for compute, 6140 for storage
Cores/socket, Threads/socket	18,36	18/36
Microcode	0x500002C	0x500002C
HT	On	On
Turbo	On	On
BIOS version	SE5C620.86B.0X.02.0094.102720191711	SE5C620.86B.0X.02.0094.102720191711
BKC version – E.g. ww47	ww08.2019	ww08.2019
PMem FW version – E.g. 5336	01.02.00.5410	-
System DDR Mem Config: slots / cap / speed	12 slots / 16GB / 2666	24 slots / 32 GB / 2666
System PMem Config: slots / cap /speed	8 slots / 128 GB /2666	-
Total Memory/Node (DDR, PMem)	192GB, 1024GB	768GB, 0
Storage - boot	1x INTEL 400GB SSD (SSDSC2BA400G3) OS Drive	1x INTEL 400GB SSD (SSDSC2BA400G3) OS Drive
Storage - application drives	11x 1TB HDD (ST1000NX0313) OSD for Ceph storage node	11x 1TB HDD (ST1000NX0313) OSD for Ceph storage node

Intel Confidential

SOFTWARE/WORKLOAD CONFIGURATION

Red=Required
Black=Required if used
Blue=Internal Intel tracking

Note on OS/Kernel for CLX:
CLX hardware mitigations require:
RHEL or CentOS requires 3.10.0-**957**.el7.x86_64 or newer kernel. Kernel.org version should be 4.19 or newer. Older kernels will only provide software mitigations which could mean lower performance

	Config1 - ISO Cost PMem	Config2 – ISO Cost DRAM	Config3 – Without Alluxio
OS	Fedora 29	Feodra 29	Feodra 29
Kernel	5.3.11-100.fc29.x86_64	5.3.11-100.fc29.x86_64	5.3.11-100.fc29.x86_64
Workload & version	Decision Support Workloads, 54 queries	Decision Support Workloads, 54 queries	Decision Support Workloads, 54 queries
Compiler	gcc 8.2.1	gcc 8.2.1	gcc 8.2.1
Libraries	N/A	N/A	N/A
Other SW	Hadoop 3.1.2	Hadoop 3.1.2	Hadoop 3.1.2
Other SW	Hive 3.1.1	Hive 3.1.1	Hive 3.1.1
Other SW	Spark 2.3.0	Spark 2.3.0	Spark 2.3.0
Other SW	Alluxio 2.0.0	Alluxio 2.0.0	Alluxio 2.0.0
Mitigation log attached (required for each config)	Yes	Yes	Yes
AEP mode: ex. 2LM or AD-volatile (replace DDR) or AD-persistent (replace NVME)	SoAD/2LM	N/A	N/A
Run Method: ex. cold (fresh-boot), warm (post-boot after few back to back iterations)	Warm	Warm	Warm
Iterations and result choice (median, average, min, max)	3 runs, median	3 runs, median	3 runs, median
Dataset size	**Decision support (54 SQL queries)**: 4TB parquet format	**Decision support (54 SQL queries)**: 4TB parquet format	**Decision support (54 SQL queries)**: 4TB parquet format

Why 54 queries

- The query was carefully selected based on previous joint work with Redhat and Silicon Valley Data Science to simulate common operations against object storage.
- https://www.redhat.com/en/blog/why-spark-ceph-part-1-3

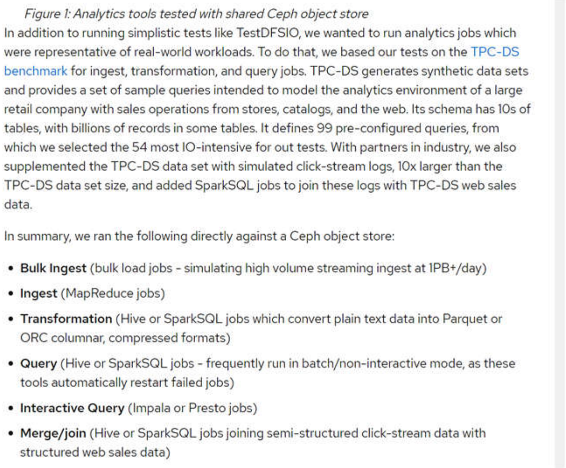


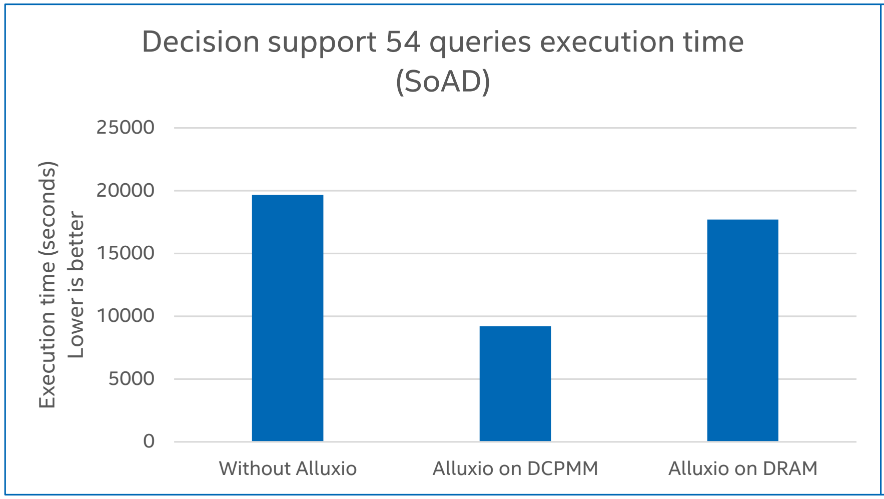
Figure 1: Analytics tools tested with shared Ceph object store

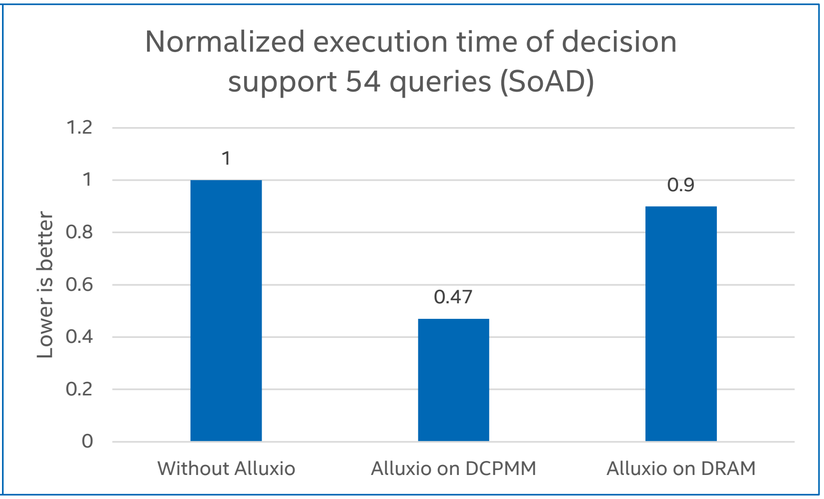
In addition to running simplistic tests like TestDFSIO, we wanted to run analytics jobs which were representative of real-world workloads. To do that, we based our tests on the TPC-DS benchmark for ingest, transformation, and query jobs. TPC-DS generates synthetic data sets and provides a set of sample queries intended to model the analytics environment of a large retail company with sales operations from stores, catalogs, and the web. Its schema has 10s of tables, with billions of records in some tables. It defines 99 pre-configured queries, from which we selected the 54 most IO-intensive for out tests. With partners in industry, we also supplemented the TPC-DS data set with simulated click-stream logs, 10x larger than the TPC-DS data set size, and added SparkSQL jobs to join these logs with TPC-DS web sales data.

In summary, we ran the following directly against a Ceph object store:

- **Bulk Ingest** (bulk load jobs - simulating high volume streaming ingest at 1PB+/day)
- **Ingest** (MapReduce jobs)
- **Transformation** (Hive or SparkSQL jobs which convert plain text data into Parquet or ORC columnar, compressed formats)
- **Query** (Hive or SparkSQL jobs - frequently run in batch/non-interactive mode, as these tools automatically restart failed jobs)
- **Interactive Query** (Impala or Presto jobs)
- **Merge/join** (Hive or SparkSQL jobs joining semi-structured click-stream data with structured web sales data)

ALLUXIO PMEM TIER PERFORMANCE OVERVIEW (SOAD)

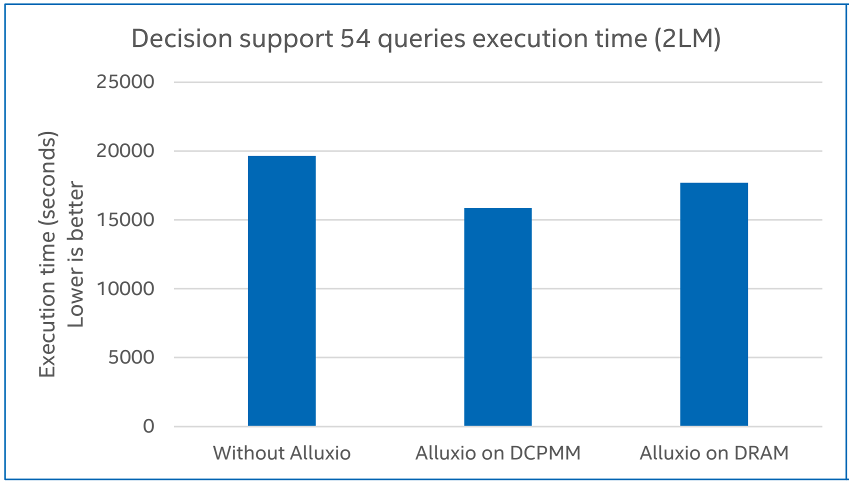

Decision support 54 queries execution time (SoAD)

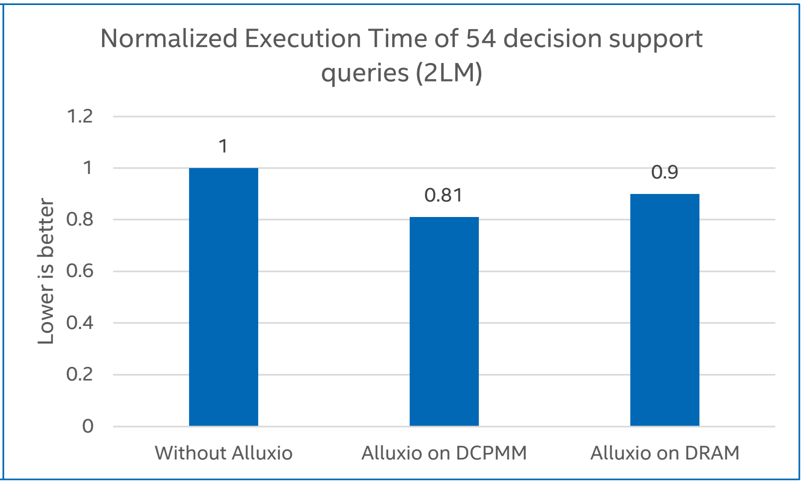

Normalized execution time of decision support 54 queries (SoAD)

- Raw capacity & data size
  - Alluxio: 900GB * 2 , DRAM 560GB * 2
  - 4TB parquet: 1547GB
- Alluxio PMem tier (SoAD) performance with decision support workload on 4TB parquet data
  - 2.13x speedup over without cache configuration
  - 1.92x speedup over DRAM tier in ISO-cost configuration.

Performance results are based on testing as of 12/06/2019 and may not reflect all publicly available security updates. See configuration disclosure on slide for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Configurations refer to page 31

ALLUXIO PMEM TIER PERFORMANCE OVERVIEW (2LM)


Decision support 54 queries execution time (2LM)


Normalized Execution Time of 54 decision support queries (2LM)

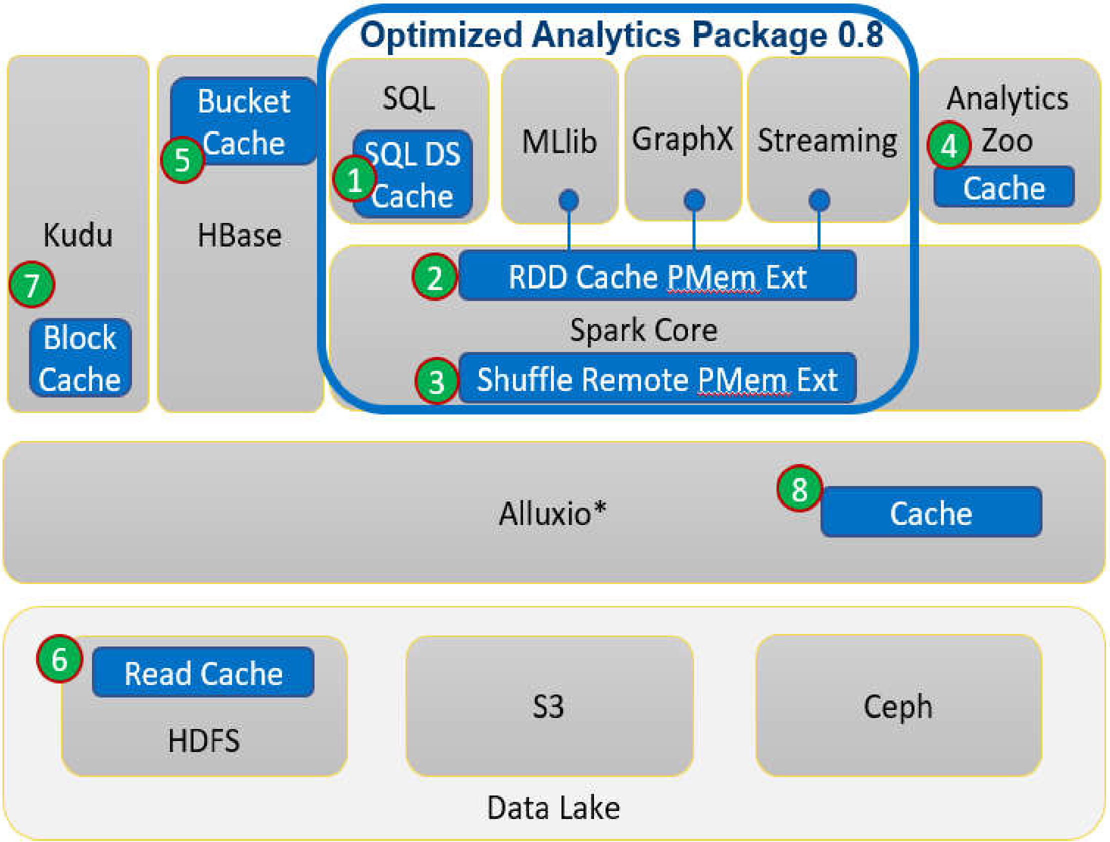
- Raw capacity & data size
  - Alluxio: 768GB * 2 , DRAM 560GB * 2
  - 4TB parquet: 1547GB
- Alluxio PMem tier (2LM) performance with decision support workload on 4TB parquet data
  - 1.24x speedup over without cache configuration
  - 1.12x speedup over DRAM tier in ISO-cost configuration.

Performance results are based on testing as of 12/06/2019 and may not reflect all publicly available security updates. See configuration disclosure on slide for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Configurations refer to page 31

INTEL OPTANE PMEM USAGE IN DATA ANALYTICS

Intel Confidential

PMEM USAGE IN BIG DATA ANALYTICS STACK



1 PMEM as Spark SQL data source cache, requiring to use Intel's Optimized Analytics Package (OAP) for Spark.

2 PMEM as another layer of storage media for RDD persistence

3 PMEM as more efficient storage for shuffle data (i.e., intermediate results)

4 PMEM as data cache for Intel Analytics Zoo on Spark.

5 PMEM as Hbase Bucket cache, speedup random read

6 PMEM as HDFS read cache, replace DRAM cache

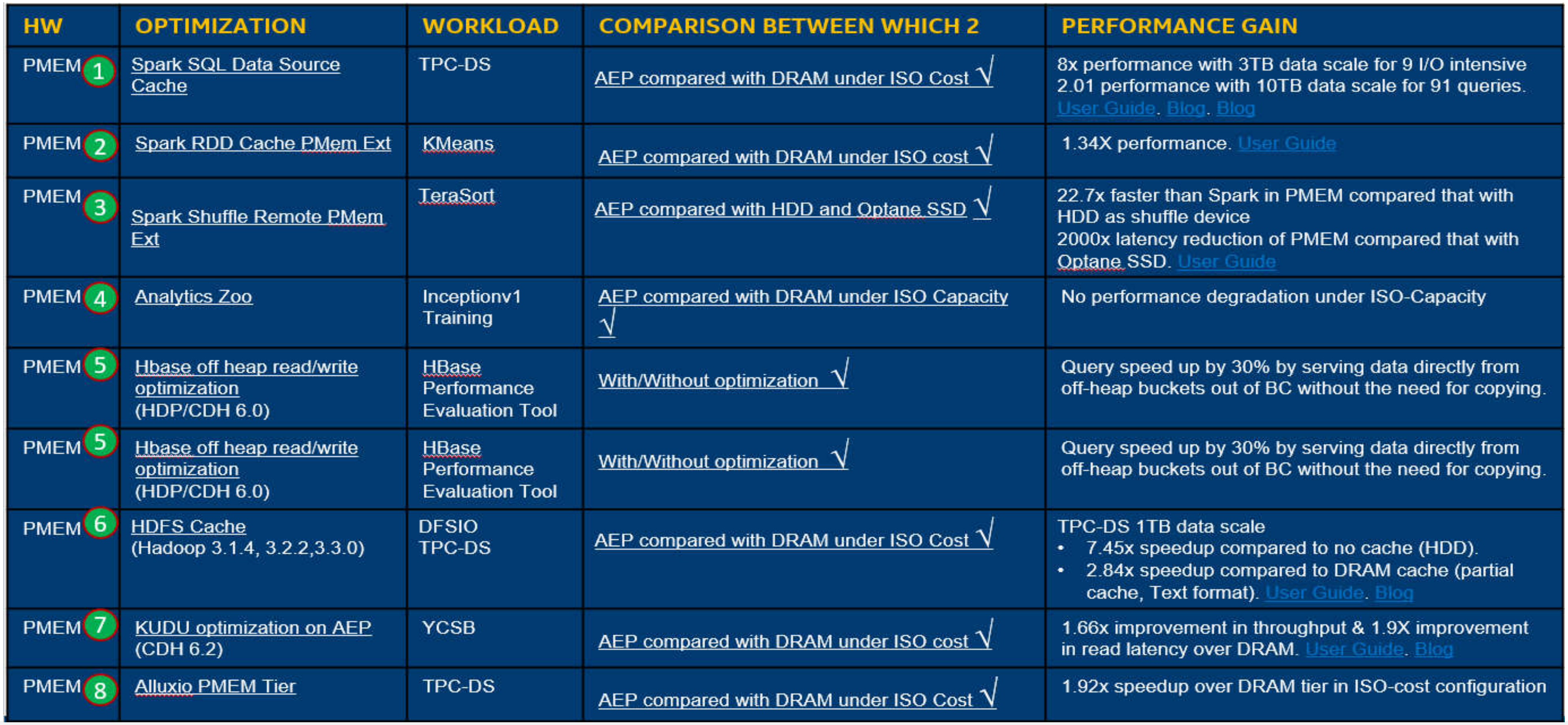
7 PMEM as Kudu block cache, reduce DRAM footprint

8 PMEM as Alluxio Pmem tier cache, reduce cost

● Ready for Adoption

5 + 7 Adopted by Cloudera in CDP

IA OPTIMIZATION – READY FOR ADOPTION



HW	OPTIMIZATION	WORKLOAD	COMPARISON BETWEEN WHICH 2	PERFORMANCE GAIN
PMEM 1	Spark SQL Data Source Cache	TPC-DS	AEP compared with DRAM under ISO Cost √	8x performance with 3TB data scale for 9 I/O intensive 2.01 performance with 10TB data scale for 91 queries. User Guide. Blog. Blog
PMEM 2	Spark RDD Cache PMem Ext	KMeans	AEP compared with DRAM under ISO cost √	1.34X performance. User Guide
PMEM 3	Spark Shuffle Remote PMem Ext	TeraSort	AEP compared with HDD and Optane SSD √	22.7x faster than Spark in PMEM compared that with HDD as shuffle device 2000x latency reduction of PMEM compared that with Optane SSD. User Guide
PMEM 4	Analytics Zoo	Inceptionv1 Training	AEP compared with DRAM under ISO Capacity √	No performance degradation under ISO-Capacity
PMEM 5	Hbase off heap read/write optimization (HDP/CDH 6.0)	HBase Performance Evaluation Tool	With/Without optimization √	Query speed up by 30% by serving data directly from off-heap buckets out of BC without the need for copying.
PMEM 5	Hbase off heap read/write optimization (HDP/CDH 6.0)	HBase Performance Evaluation Tool	With/Without optimization √	Query speed up by 30% by serving data directly from off-heap buckets out of BC without the need for copying.
PMEM 6	HDFS Cache (Hadoop 3.1.4, 3.2.2,3.3.0)	DFSIO TPC-DS	AEP compared with DRAM under ISO Cost √	TPC-DS 1TB data scale • 7.45x speedup compared to no cache (HDD). • 2.84x speedup compared to DRAM cache (partial cache, Text format). User Guide. Blog
PMEM 7	KUDU optimization on AEP (CDH 6.2)	YCSB	AEP compared with DRAM under ISO cost √	1.66x improvement in throughput & 1.9X improvement in read latency over DRAM. User Guide. Blog
PMEM 8	Alluxio PMEM Tier	TPC-DS	AEP compared with DRAM under ISO Cost √	1.92x speedup over DRAM tier in ISO-cost configuration

Yarn&Spark Configuration

Key	Value	Description
yarn.resourcemanager.scheduler.class	org.apache.hadoop.yarn.server.resourcemanager.scheduler.fair.FairScheduler	
yarn.nodemanager.resource.memory-mb	204800	
yarn.nodemanager.vmem-check-enabled	false	
yarn.scheduler.maximum-allocation-vcores	72	Must be greater than spark_exec_nums * spark_exec_cores
yarn.nodemanager.resource.cpu-vcores	72	
yarn.scheduler.minimum-allocation-mb	1024	

Key	Value	Description
spark.executor.instances	12	Total spark exec number, each node has 6 executors
spark.executor.cores	10	Per exec core number
spark.driver.memory	10g	
spark.executor.memory	24g	Spark exec mem total is 24+5 = 29G
spark.executor.memoryOverhead	5g	

Alluxio configurations

alluxio (PMem SoAD)		alluxio (PMem 2LM)		alluxio (DRAM)	
alluxio.worker.tieredstore.levels	1	alluxio.worker.tieredstore.levels	1	alluxio.worker.tieredstore.levels	1
alluxio.worker.ieredstore.level0.alias	SSD	alluxio.worker.ieredstore.level0.alias	MEM	alluxio.worker.ieredstore.level0.alias	MEM
alluxio.underfs.s3a.list.objects.v1	TRUE	alluxio.underfs.s3a.list.objects.v1	TRUE	alluxio.underfs.s3a.list.objects.v1	TRUE
alluxio.underfs.s3.threads.max	80	alluxio.underfs.s3.threads.max	80	alluxio.underfs.s3.threads.max	80
alluxio.underfs.s3.request.timeout	0	alluxio.underfs.s3.request.timeout	0	alluxio.underfs.s3.request.timeout	0
alluxio.worker.tieredstore.level0.dirs.path	/mnt/pmem0, /mnt/pmem1	alluxio.worker.tieredstore.level0.dirs.path	/mnt/ramdisk	alluxio.worker.tieredstore.level0.dirs.path	/mnt/ramdisk
alluxio.worker.tieredstore.level0.dirs.quota	450G,450G	alluxio.worker.tieredstore.level0.dirs.quota	786G	alluxio.worker.tieredstore.level0.dirs.quota	560G
alluxio.user.ufs.block.read.location.policy	alluxio.client.block.policy.DeterministicHashPolicy	alluxio.user.ufs.block.read.location.policy	alluxio.client.block.policy.DeterministicHashPolicy	alluxio.user.ufs.block.read.location.policy	alluxio.client.block.policy.DeterministicHashPolicy
alluxio.worker.network.async.cache.manager.threads.max	32	alluxio.worker.network.async.cache.manager.threads.max	32	alluxio.worker.network.async.cache.manager.threads.max	32
alluxio.worker.network.block.reader.threads.max	8192	alluxio.worker.network.block.reader.threads.max	8192	alluxio.worker.network.block.reader.threads.max	8192
alluxio.user.file.passive.cache.enabled	false	alluxio.user.file.passive.cache.enabled	false	alluxio.user.file.passive.cache.enabled	false

Comparison w/ NVMe

- Alluxio was typically deployed in data center for:
  - Simplify Hadoop for the hybrid cloud by making on-prem HDFS accessible to any compute in the cloud.
  - Get in-memory data access for Spark, Presto, or any analytics framework on AWS, Google Cloud Platform, or Microsoft Azure.
  - And the disaggregated environment evolving with diskless trend
  - Most of Alluxio customer is actually deployed w/ DRAM only

NVMe challenges

- Not enough PCIe slots, a PCIe raiser & NVMe combo hot swap bay module will be required
- Not enough performance, NVMe BW will be limited by PCIe riser

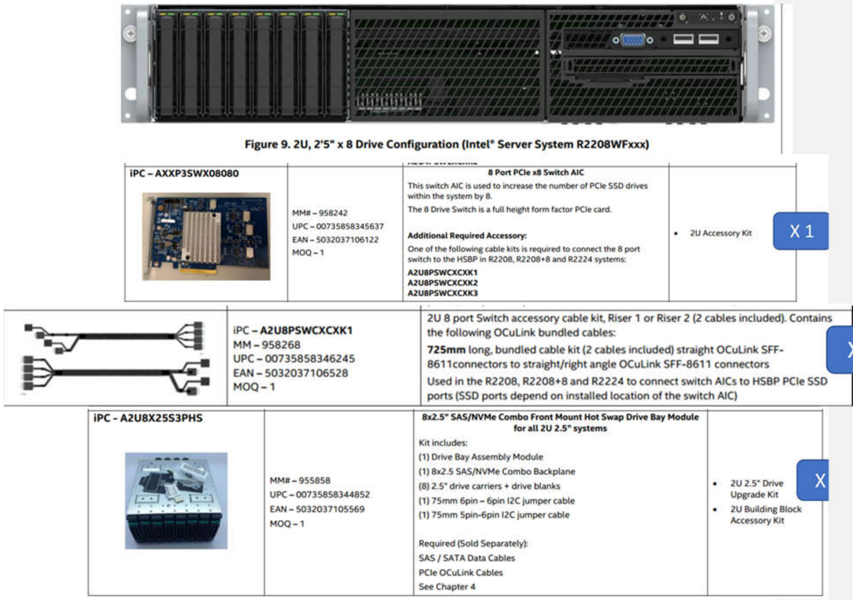


Figure 9. 2U, 2'5" x 8 Drive Configuration (Intel® Server System R2208WFxxx)

iPC – AXXP3SWX08080	MM# – 958242 UPC – 00735858345637 EAN – 5032037106122 MOQ – 1	**8 Port PCIe x8 Switch AIC** This switch AIC is used to increase the number of PCIe SSD drives within the system by 8. The 8 Drive Switch is a full height form factor PCIe card. **Additional Required Accessory:** One of the following cable kits is required to connect the 8 port switch to the HSBP in R2208, R2208+8 and R2224 systems: **A2U8PSWCXCXK1** **A2U8PSWCXCXK2** **A2U8PSWCXCXK3**	• 2U Accessory Kit

X 1

	iPC – **A2U8PSWCXCXK1** MM – 958268 UPC – 00735858346245 EAN – 5032037106528 MOQ – 1	2U 8 port Switch accessory cable kit, Riser 1 or Riser 2 (2 cables included). Contains the following OCuLink bundled cables: **725mm** long, bundled cable kit (2 cables included) straight OCuLink SFF-8611connectors to straight/right angle OCuLink SFF-8611 connectors Used in the R2208, R2208+8 and R2224 to connect switch AICs to HSBP PCIe SSD ports (SSD ports depend on installed location of the switch AIC)

X

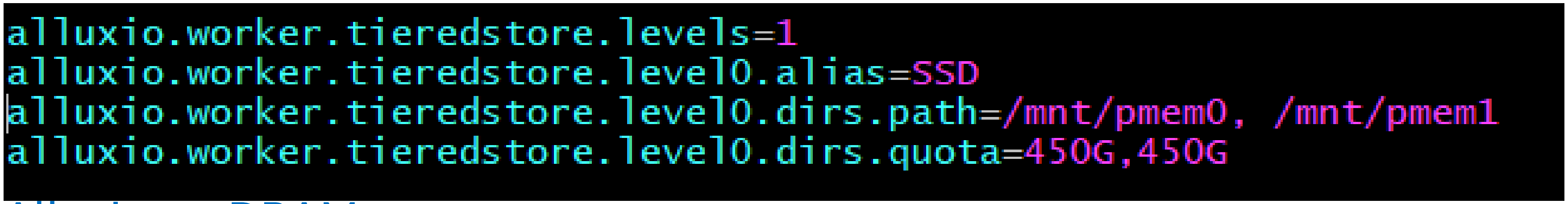
iPC - A2U8X25S3PHS	MM# – 955858 UPC – 00735858344852 EAN – 5032037105569 MOQ – 1	**8x2.5" SAS/NVMe Combo Front Mount Hot Swap Drive Bay Module for all 2U 2.5" systems** Kit includes: (1) Drive Bay Assembly Module (1) 8x2.5 SAS/NVMe Combo Backplane (8) 2.5" drive carriers + drive blanks (1) 75mm 6pin – 6pin I2C jumper cable (1) 75mm 5pin-6pin I2C jumper cable Required (Sold Separately): SAS / SATA Data Cables PCIe OCuLink Cables See Chapter 4	• 2U 2.5" Drive Upgrade Kit • 2U Building Block Accessory Kit

X

Micro workloads tests (fio)	Bandwidth	Comments
NVMe	7GB/s	Fio, 4x P4610
PMem numa binding_	42GB	Fio, via libpmemengine, 8x 128GB PMem
PMem non_numa	17GB	Fio, via libpmemengine
PMem snoopy_mode	42GB/s	Fio, via libpmemengine

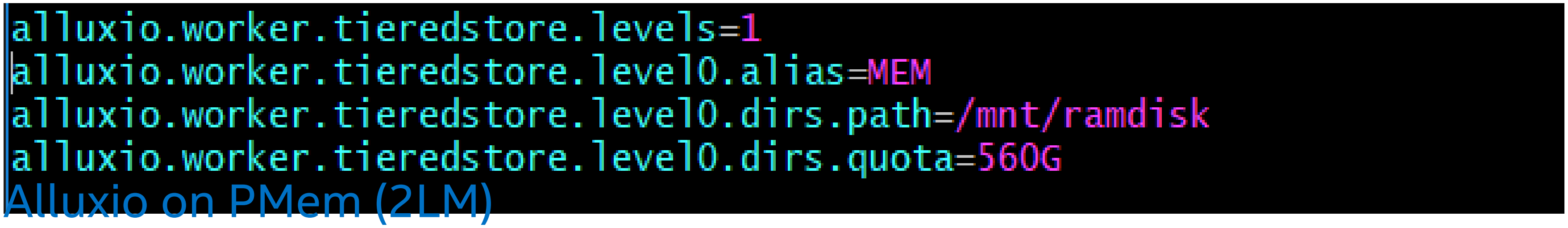
Alluxio PMem tier SoAD mode configuration

- Alluxio on PMem (SoAD)



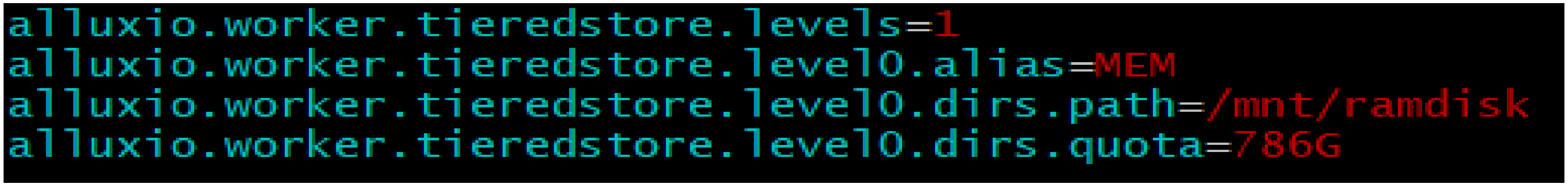
```
alluxio.worker.tieredstore.levels=1
alluxio.worker.tieredstore.level0.alias=SSD
alluxio.worker.tieredstore.level0.dirs.path=/mnt/pmem0, /mnt/pmem1
alluxio.worker.tieredstore.level0.dirs.quota=450G,450G
```

- Alluxio on DRAM



```
alluxio.worker.tieredstore.levels=1
alluxio.worker.tieredstore.level0.alias=MEM
alluxio.worker.tieredstore.level0.dirs.path=/mnt/ramdisk
alluxio.worker.tieredstore.level0.dirs.quota=560G
```

- Alluxio on PMem (2LM)



```
alluxio.worker.tieredstore.levels=1
alluxio.worker.tieredstore.level0.alias=MEM
alluxio.worker.tieredstore.level0.dirs.path=/mnt/ramdisk
alluxio.worker.tieredstore.level0.dirs.quota=786G
```

Alluxio enabling trouble-shooting

- Alluxio can’t be started using PMem
  - Set alluxio.worker.tieredstore.level0.alias to SSD
  - Or, use ramdisk to launch Alluxio and redirect ramdisk to PMem
- Ceph rgw domain name can’t resolved
  - alluxio.master.mount.table.root.option.alluxio.underfs.s3.disable.dns.buckets=tru
- Switch hive table location is too time-consuming using official recomme way
  - Use hive `metatool` to do the trick