An Open Source Memory-Centric Distributed Storage System

Haoyuan Li, Tachyon Nexus & UC Berkeley

November 19, 2015 @ AMPCamp 6

Outline

- **Open Source**
- Introduction to Tachyon (Before 2015)
- Deployments and New Features
- Getting Involved

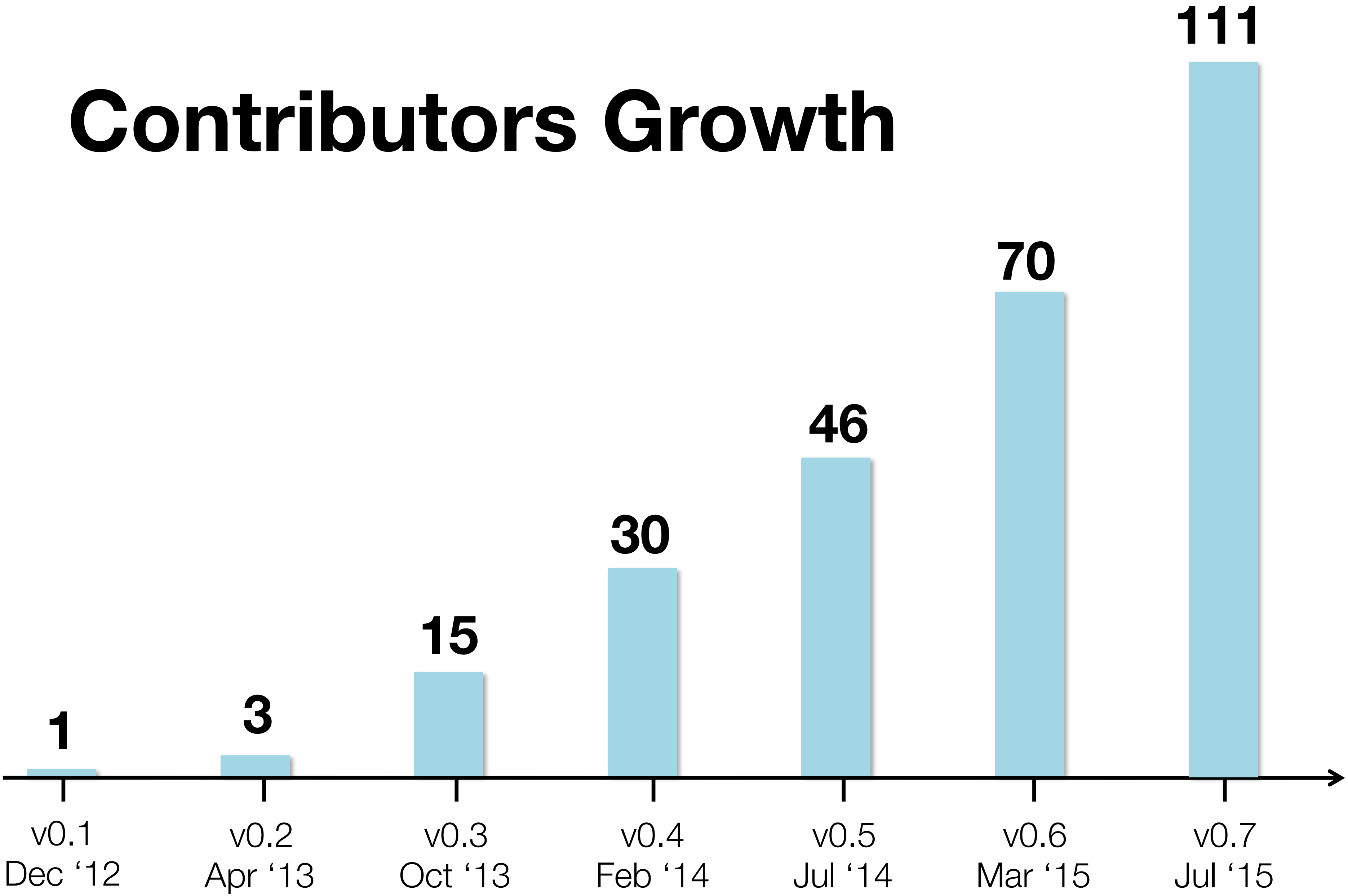
Background

- Started at UC Berkeley AMPLab
  - From summer 2012
- Open sourced
  - April 2013 (two and half years ago)
  - Apache License 2.0
  - Latest Release: Version 0.8.2 (November 2015)
- Deployed at > 100 companies

Tachyon: one of the Fastest Growing Big Data Open Source Project

TACHYON NEXUS

Contributors Growth



TACHYON NEXUS

amplab UC BERKELEY

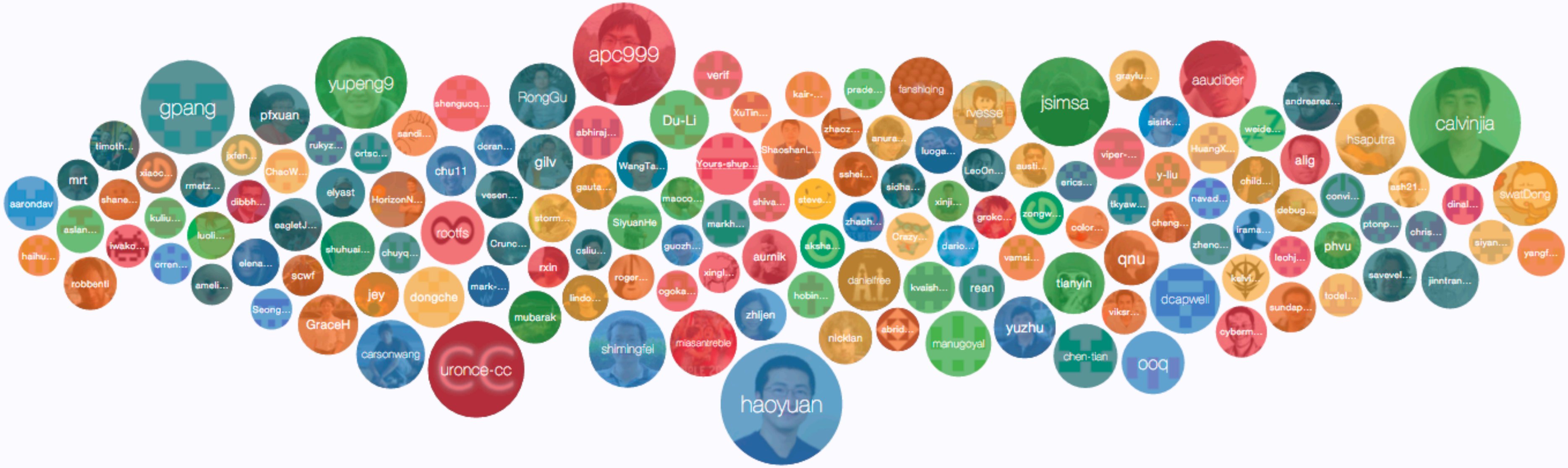
Contributors Growth

> 170 Contributors (V0.8)

(3x increment over the last AMPCamp)

> 50 Organizations

Thanks to Contributors and Users!



http://tachyon-project.org/community/

YAHOO!
intel
IBM
CLEMSON UNIVERSITY
SOUTH CAROLINA
1889
HUAWEI
Bai du 百度
NOKIA
databricks
CRAY
THE SUPERCOMPUTER COMPANY
redhat
Berkeley
UNIVERSITY OF CALIFORNIA
NANJING UNIVERSITY
1902
Atigeo
Palantir
Technologies
TACHYON
NEXUS
Mellanox
TECHNOLOGIES
Pivotal
TSINGHUA UNIVERSITY
清华
1911
ClearStory
DATA
Now You See It

One Tachyon Production Deployment Example

- Baidu (Dominant Search Engine in China, ~ 50 Billion USD Market Cap)
- Framework: SparkSQL
- Under Storage: Baidu's File System
- Storage Media: MEM + HDD
- 100+ nodes deployment
- 1PB+ managed space
- 30x Performance Improvement

Outline

- Open Source
- **Introduction to Tachyon (Before 2015)**
- New Features
- Getting Involved

Tachyon is an

Open Source

Memory-centric

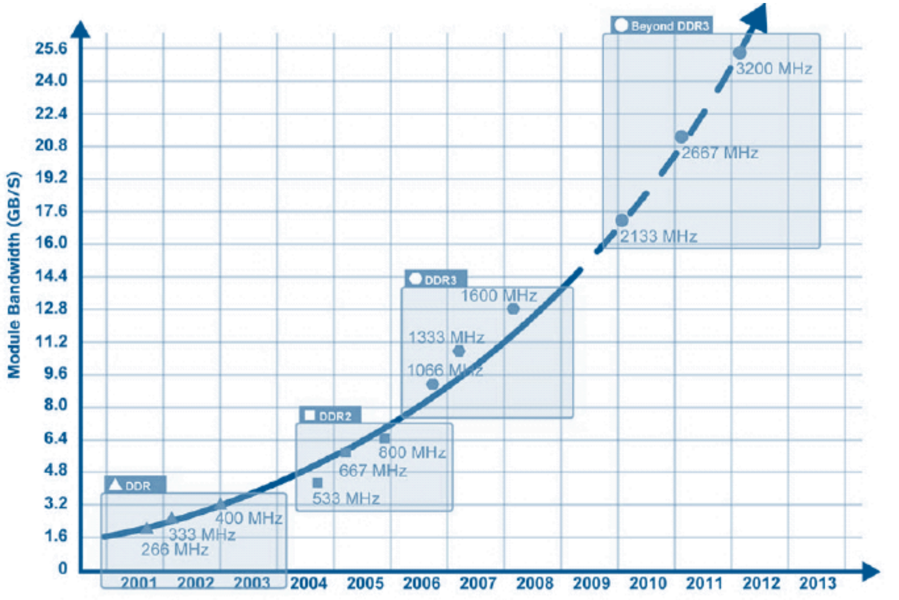
Distributed

Storage System

Why Tachyon?

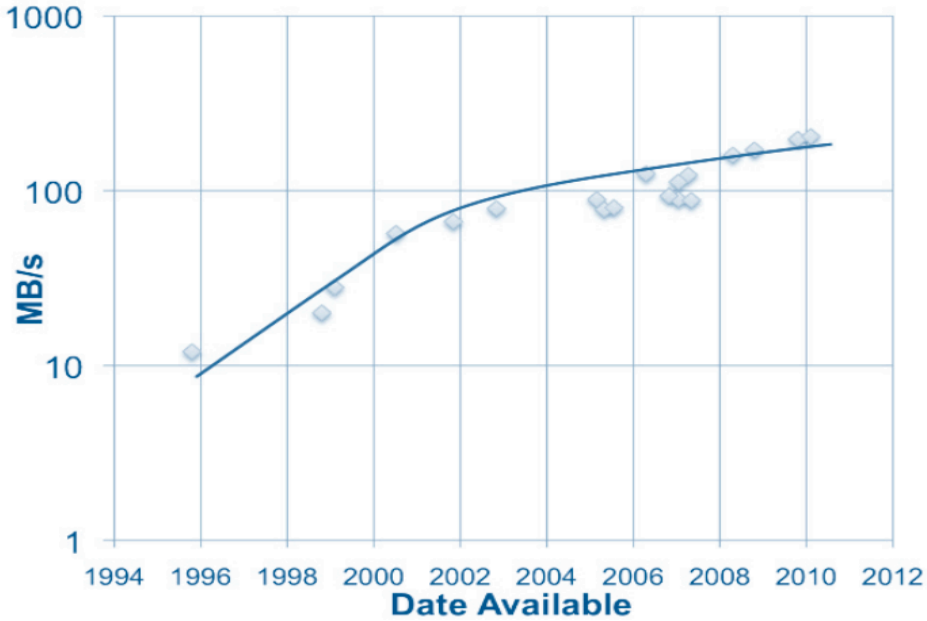
Performance Trend:
Memory is Fast

- RAM throughput increasing **exponentially**



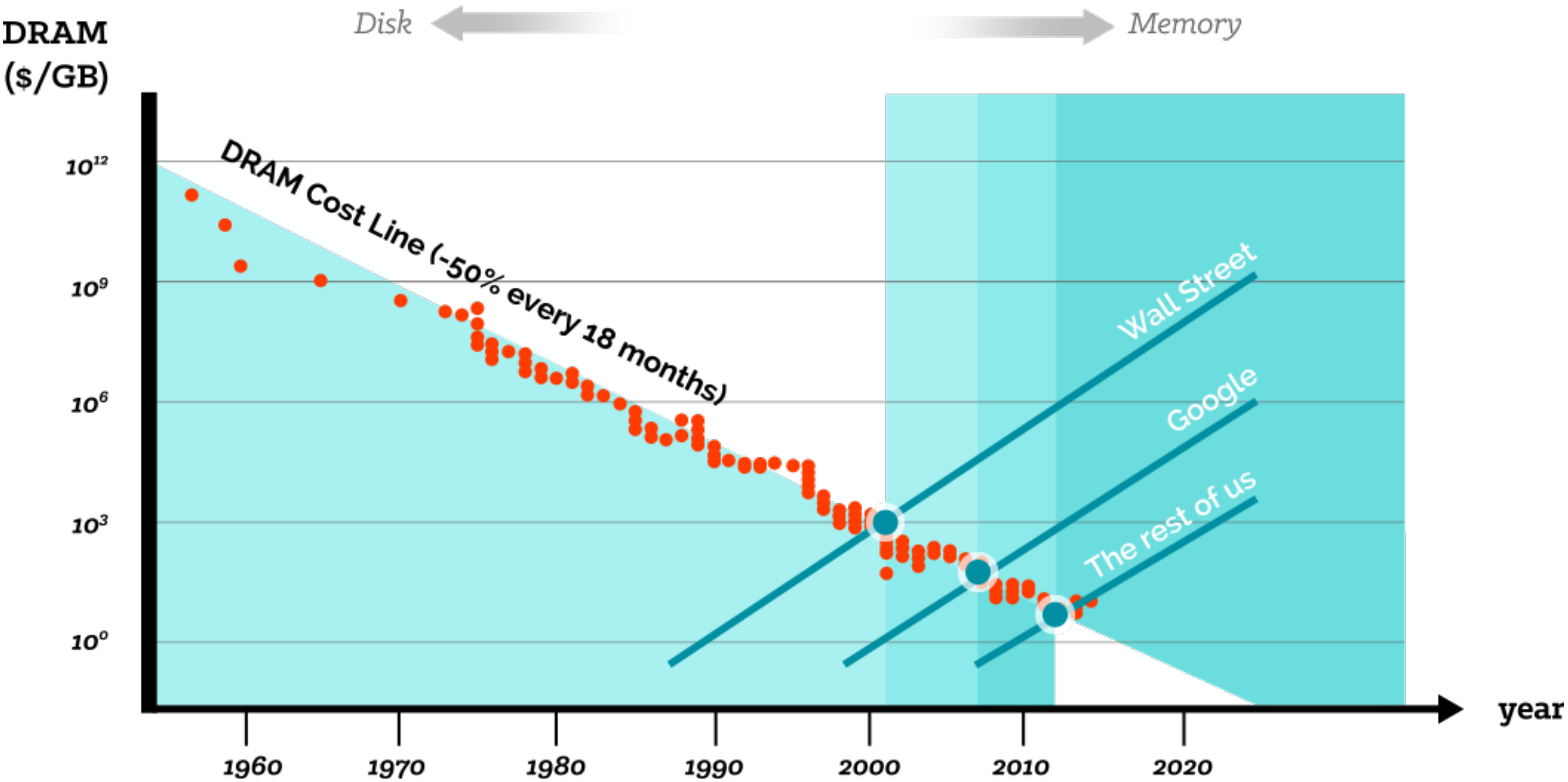
Bandwidths shown for 64-bit memory module. Date indicates approximate industry product introduction.

- Disk throughput increasing **slowly**



Memory-locality key to interactive response times

Price Trend: Memory is Cheaper



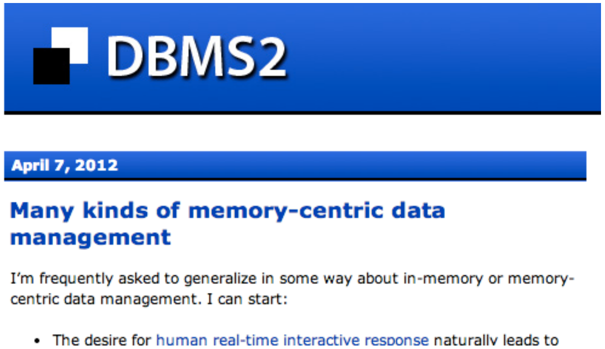
source: jcmit.com

TACHYON NEXUS

amplab UC BERKELEY

Realized by many…

Spark



DBMS2

April 7, 2012

Many kinds of memory-centric data management

I'm frequently asked to generalize in some way about in-memory or memory-centric data management. I can start:

- The desire for human real-time interactive response naturally leads to

cloudera
IMPALA

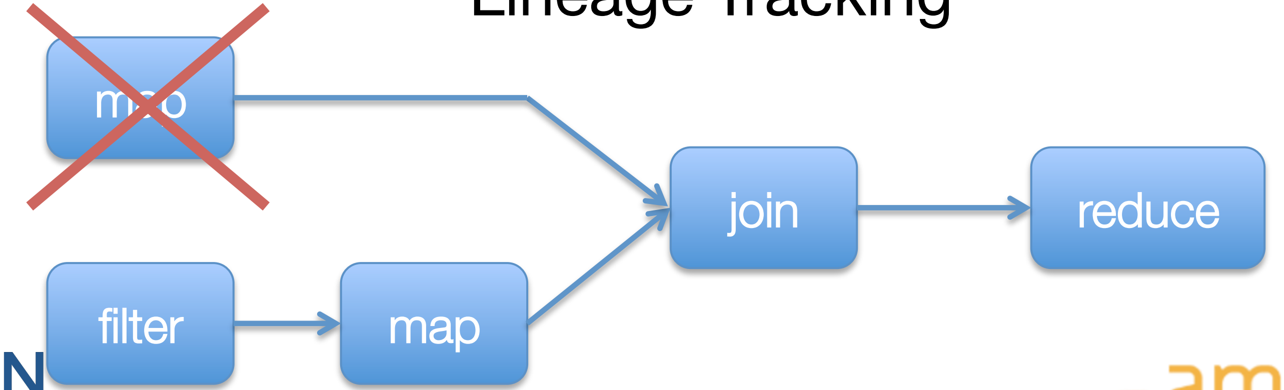
Is the Problem Solved?

Missing a Solution for the Storage Layer

A Use Case Example with Spark

- Fast, in-memory data processing framework
  - Keep **one in-memory** copy inside JVM
  - Track **lineage** of operations used to derive data
  - Upon failure, use lineage to recompute data

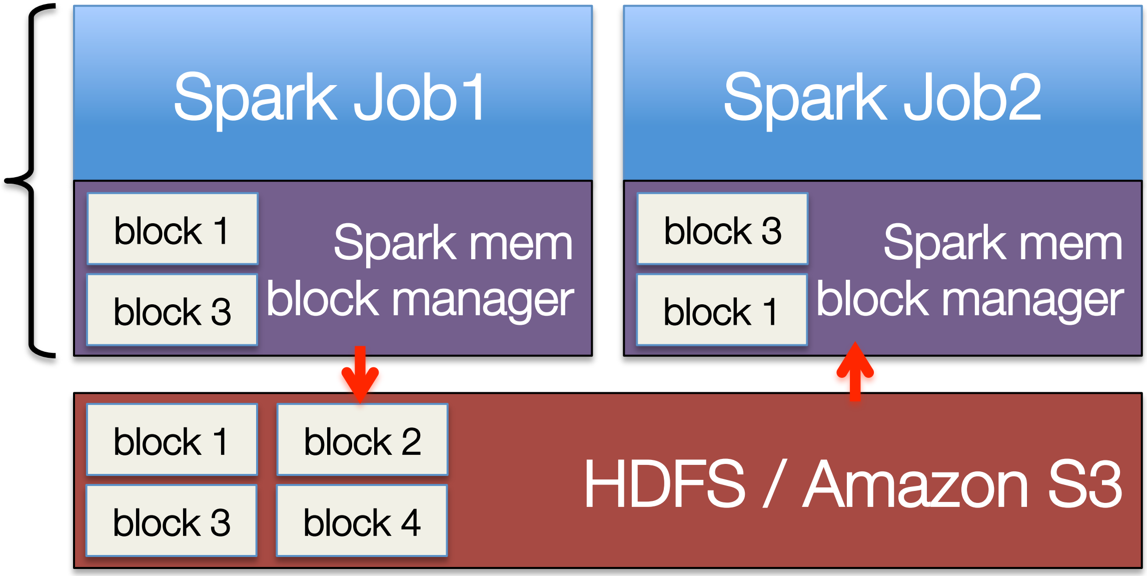
Lineage Tracking



Issue 1

Data Sharing is the bottleneck in analytics pipeline: Slow writes to disk

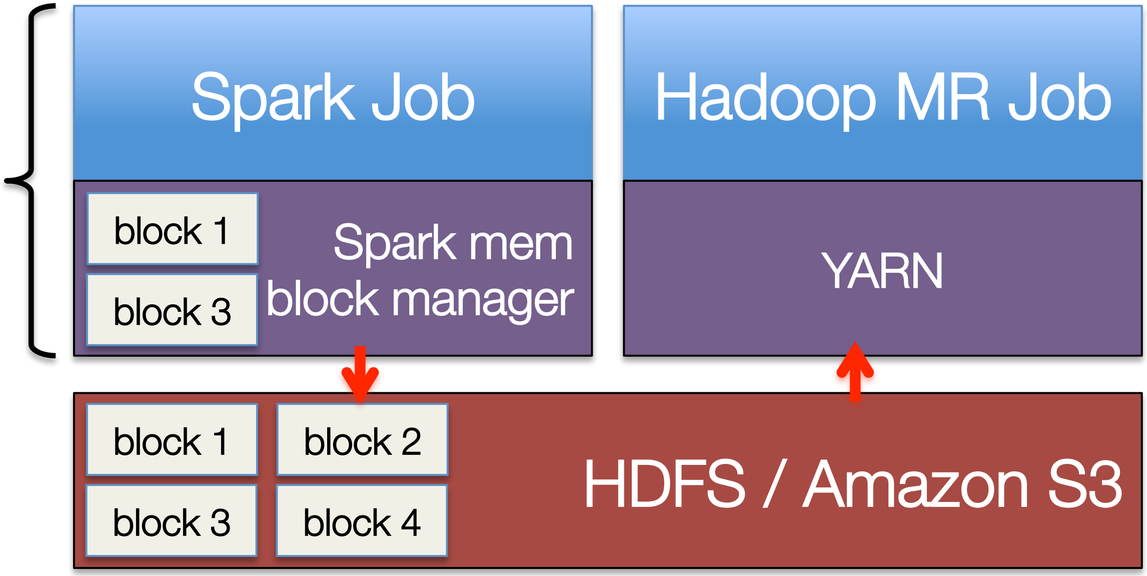
storage engine & execution engine same process (slow writes)



Issue 1

Data Sharing is the bottleneck in analytics pipeline: Slow writes to disk

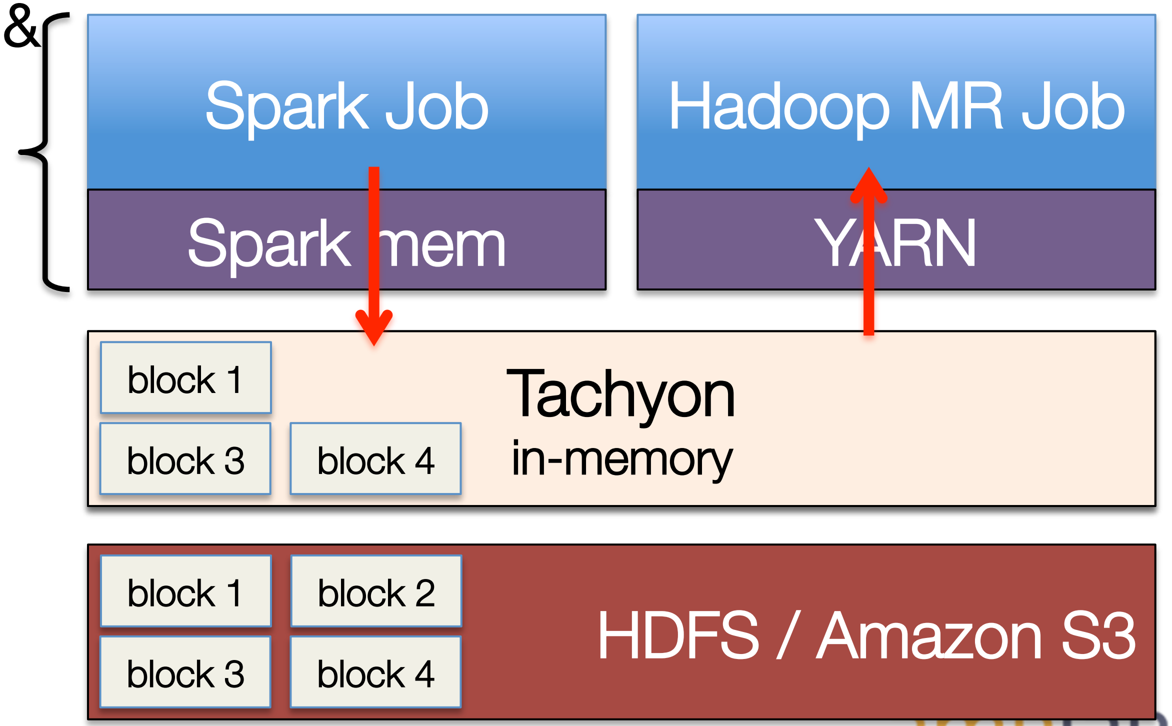
storage engine &
execution engine
same process
(slow writes)



Issue 1 resolved with Tachyon

Memory-speed data sharing among jobs in different frameworks

execution engine &
storage engine
same process
(fast writes)



Spark Job

Spark mem

Hadoop MR Job

YARN

block 1

block 3

block 4

Tachyon
in-memory

block 1

block 2

block 3

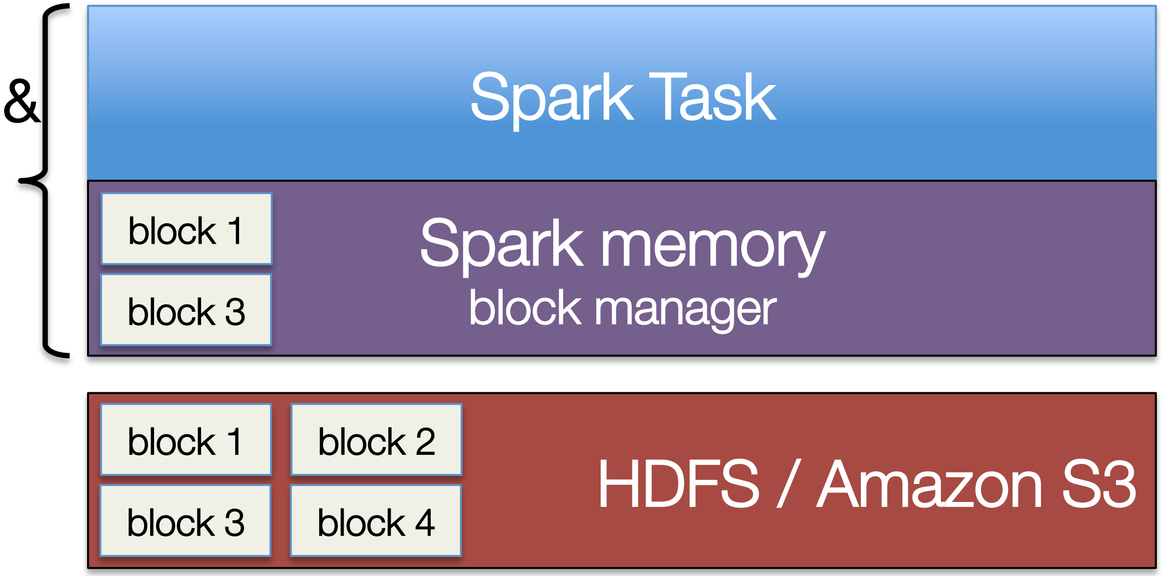
block 4

HDFS / Amazon S3

Issue 2

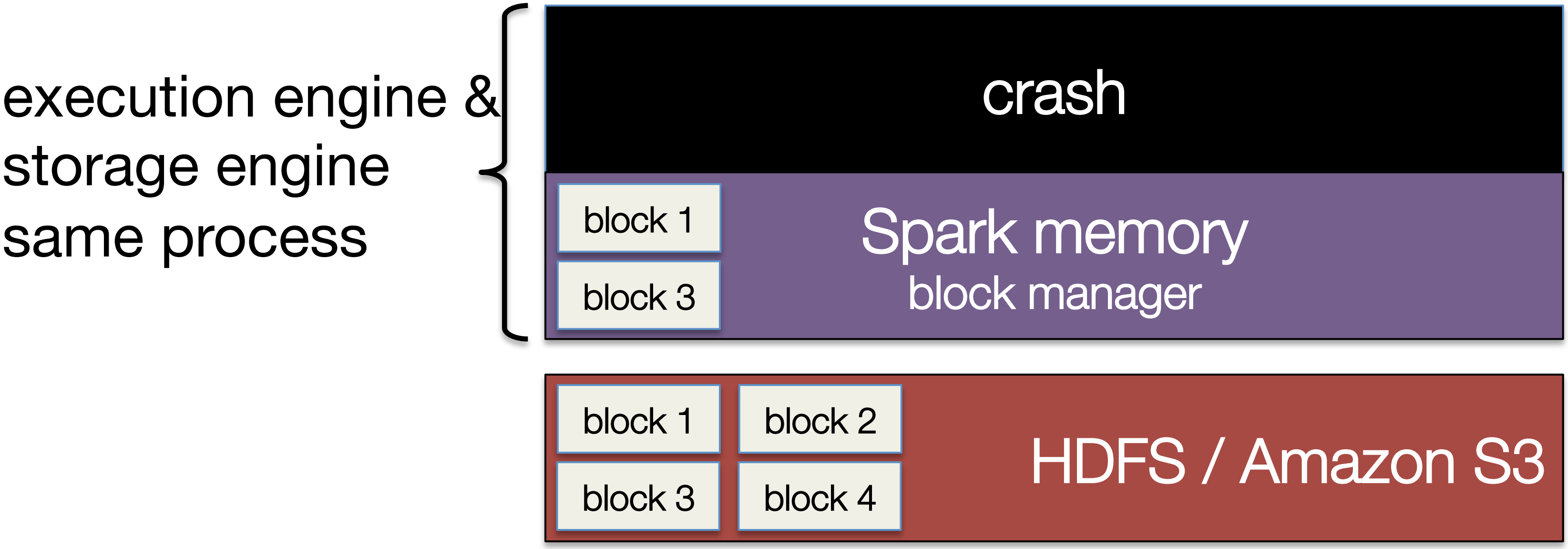
Cache loss when process crashes

execution engine & storage engine same process



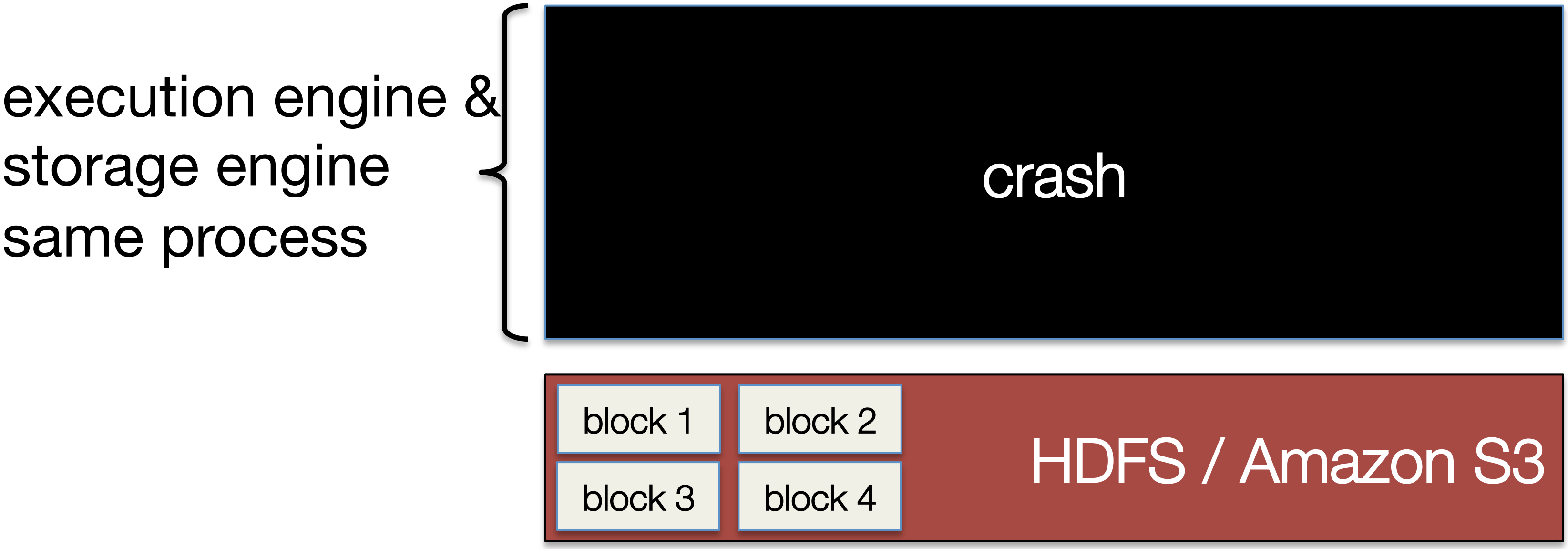
Issue 2

Cache loss when process crashes



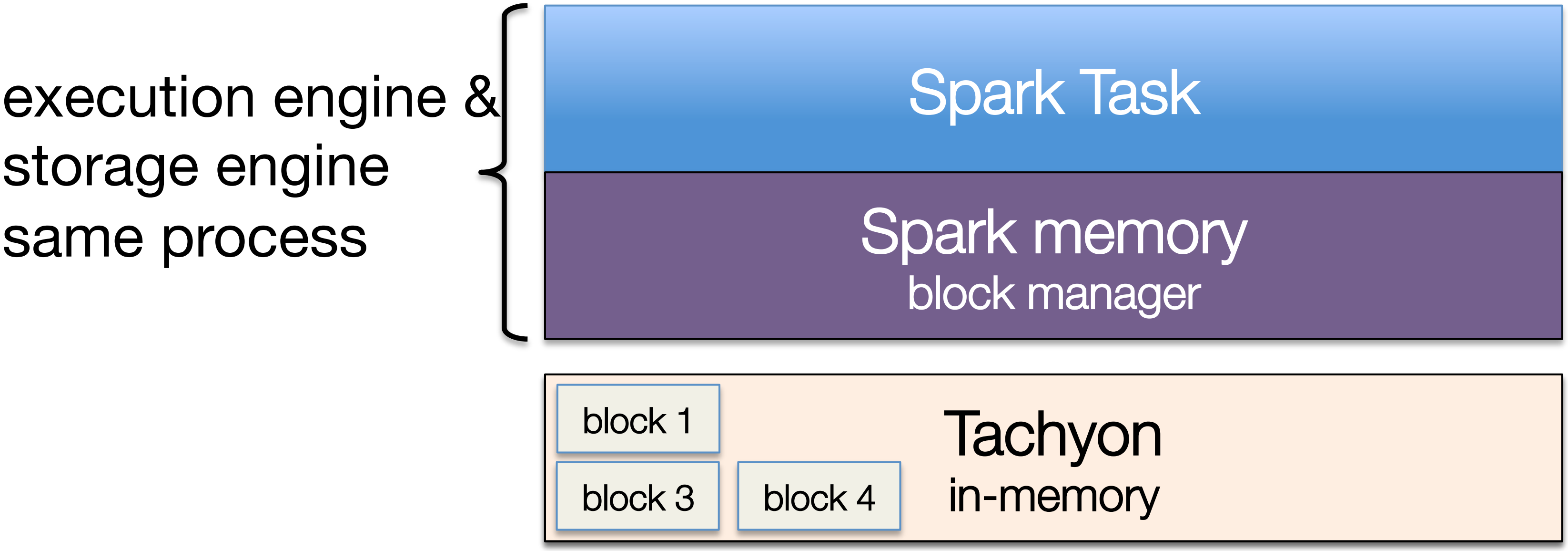
Issue 2

Cache loss when process crashes



Issue 2 resolved with Tachyon

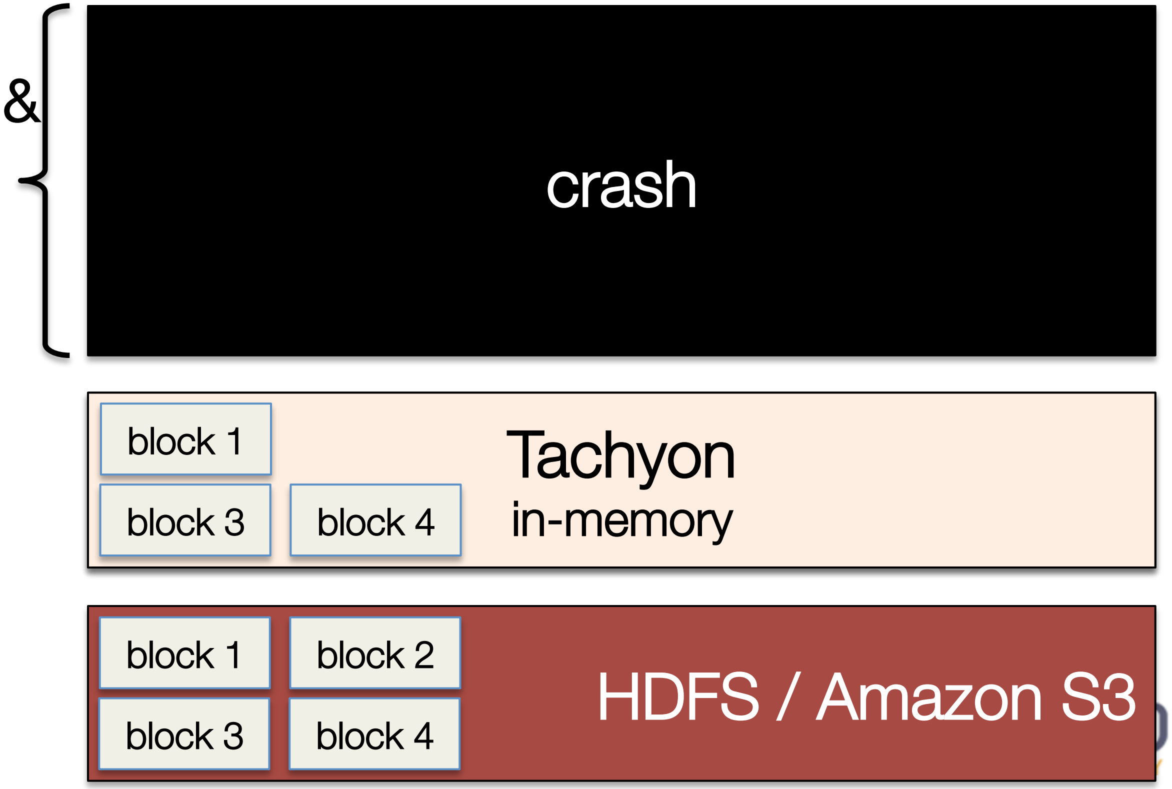
Keep in-memory data safe, even when a job crashes.



Issue 2 resolved with Tachyon

Keep in-memory data safe, even when a job crashes.

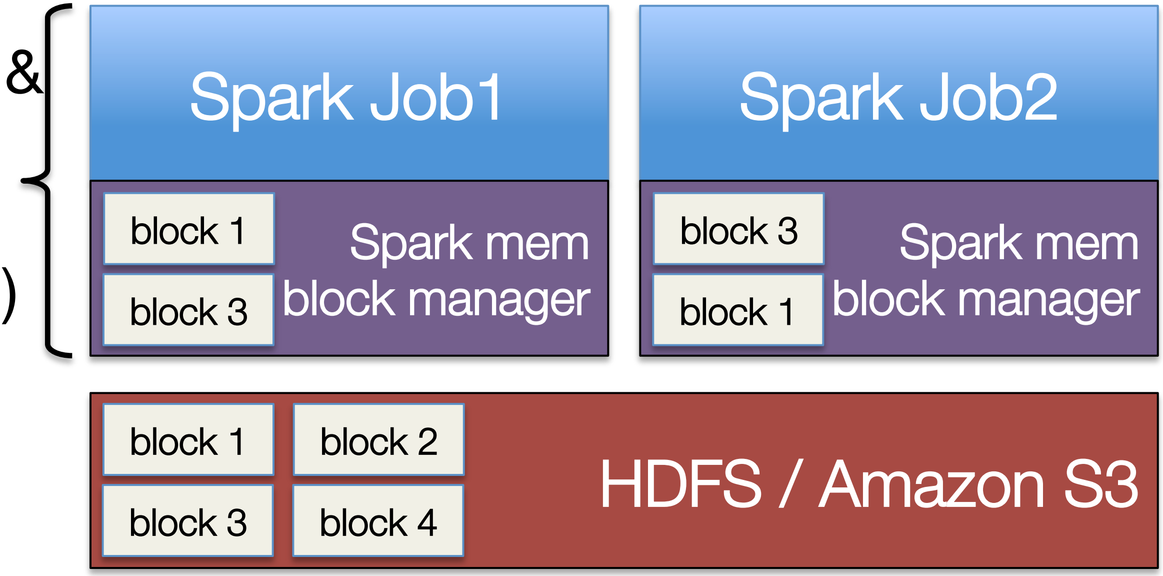
execution engine &
storage engine
same process



Issue 3

In-memory Data Duplication & Java Garbage Collection

execution engine & storage engine same process (duplication & GC)

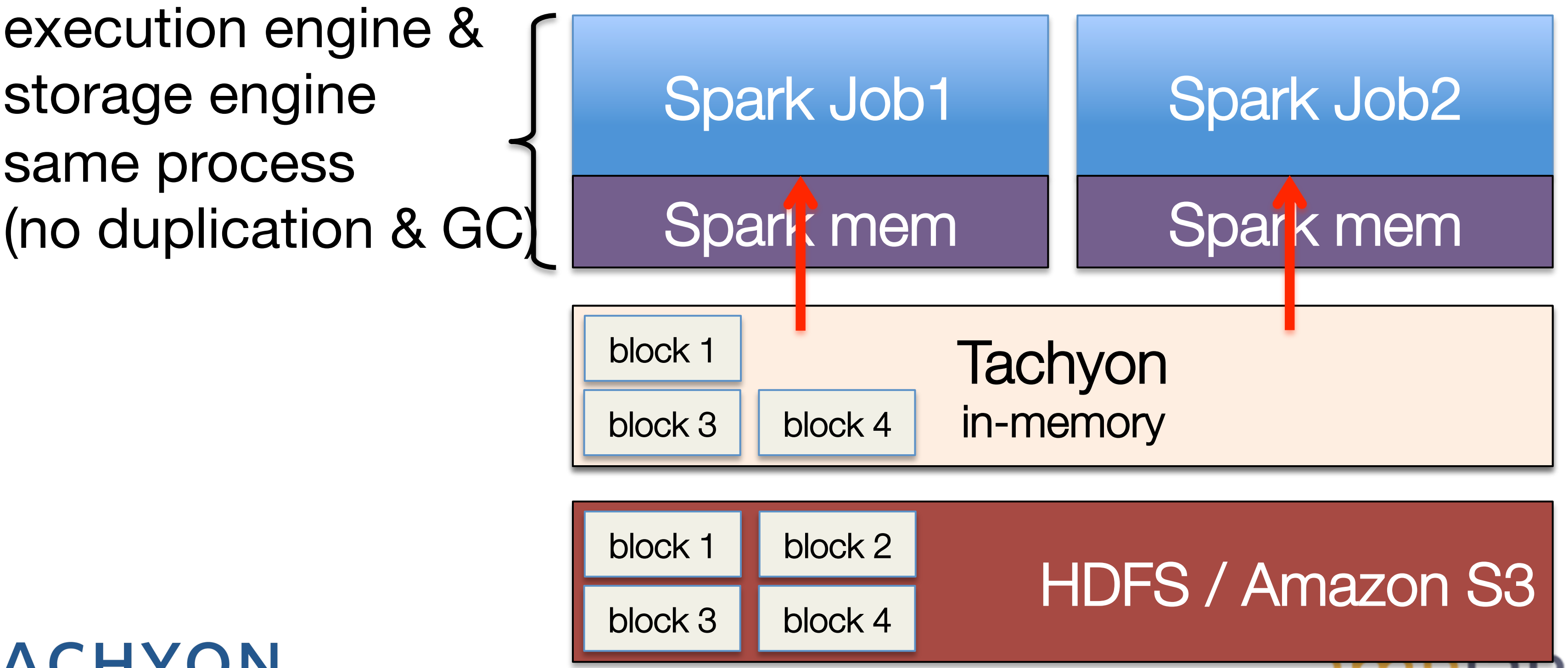


TACHYON NEXUS

amplab UC BERKELEY

Issue 3 resolved with Tachyon

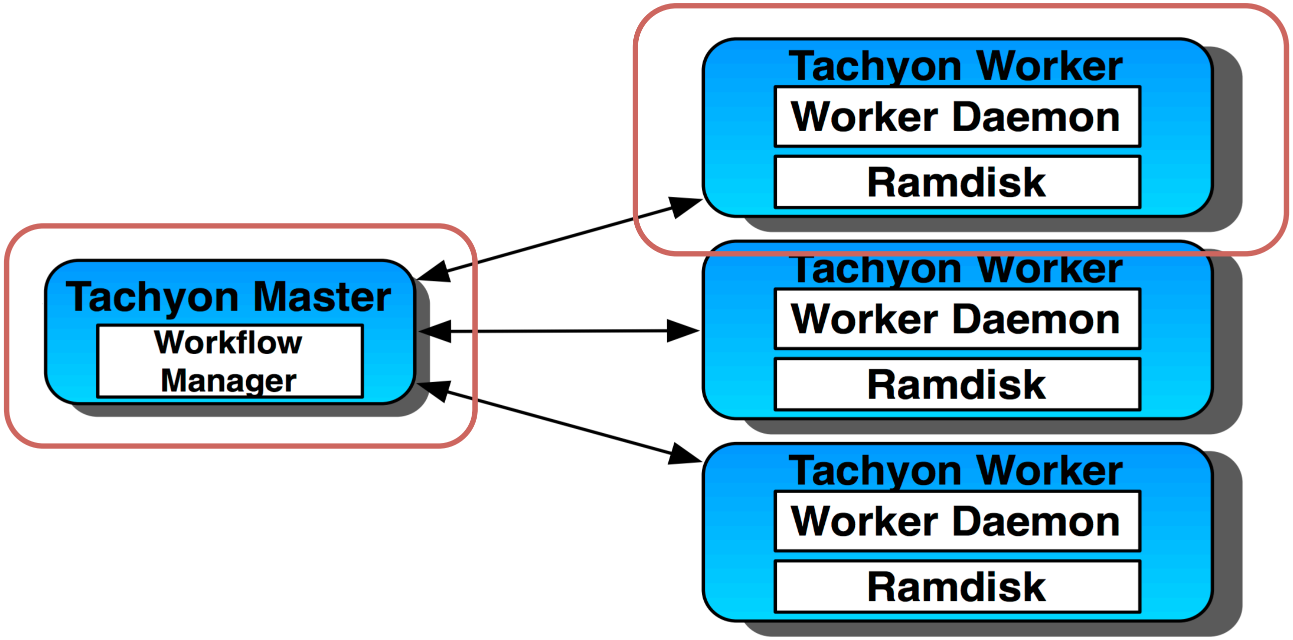
No in-memory data duplication, much less GC



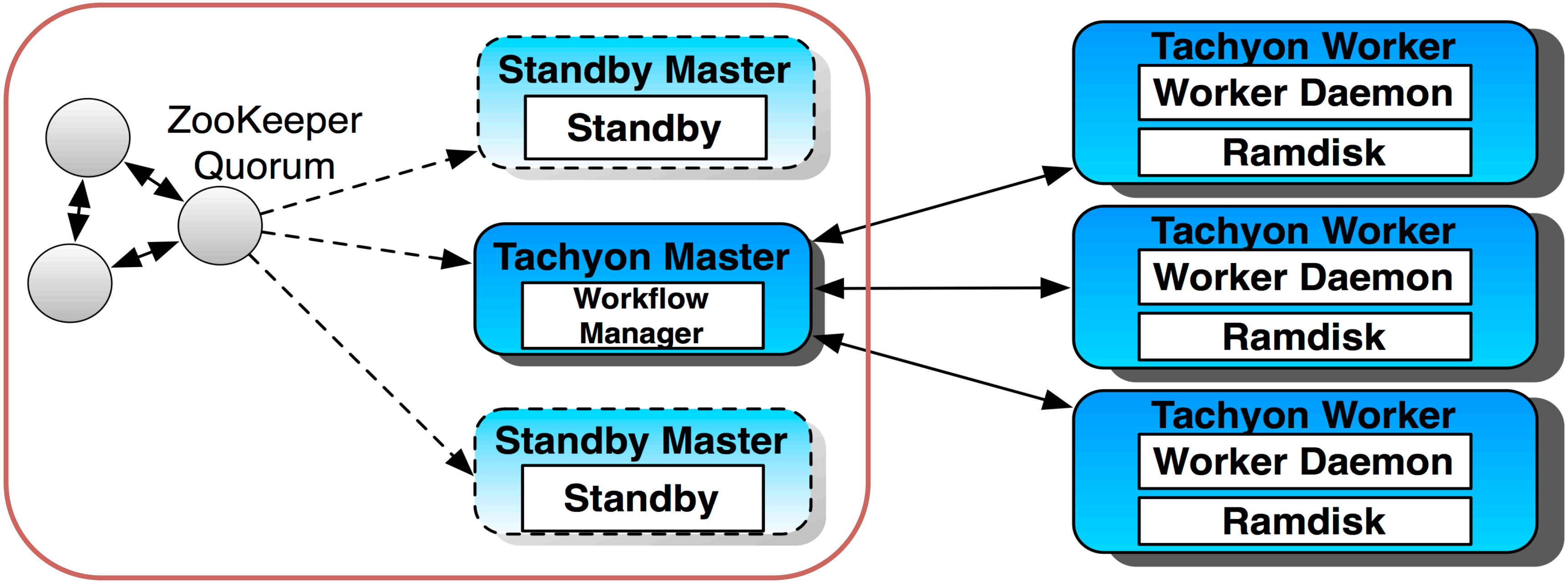
Previously Mentioned

- A memory-centric storage architecture
- Push lineage down to storage layer

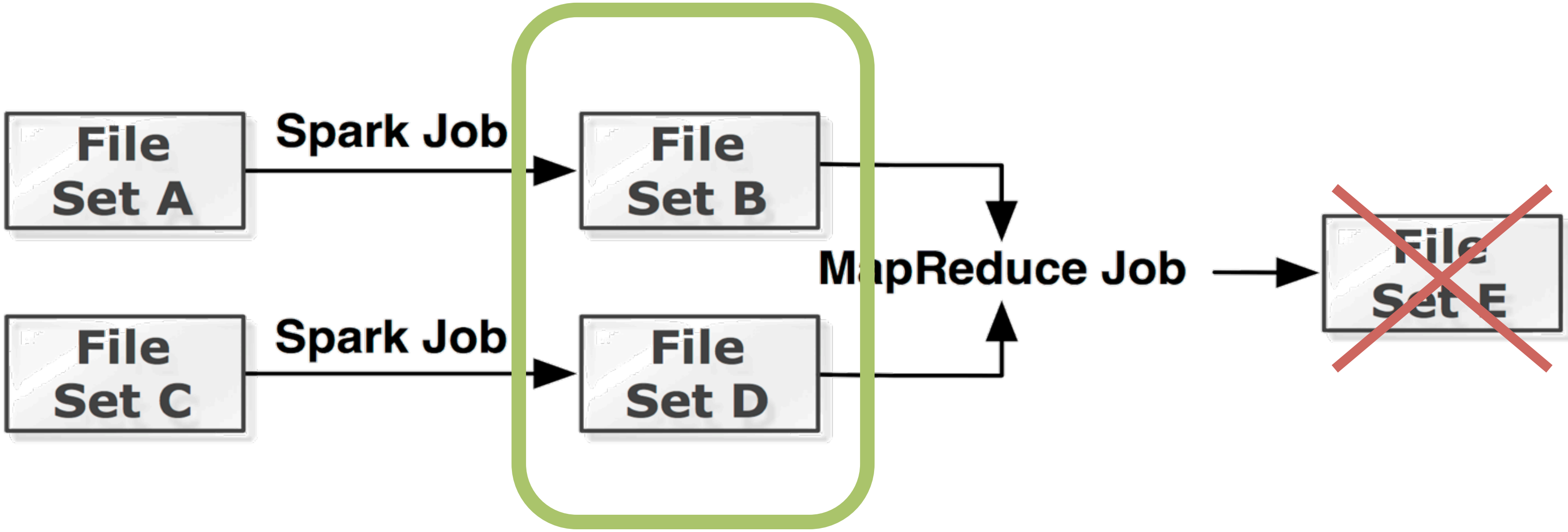
Tachyon Memory-Centric Architecture



Tachyon Memory-Centric Architecture



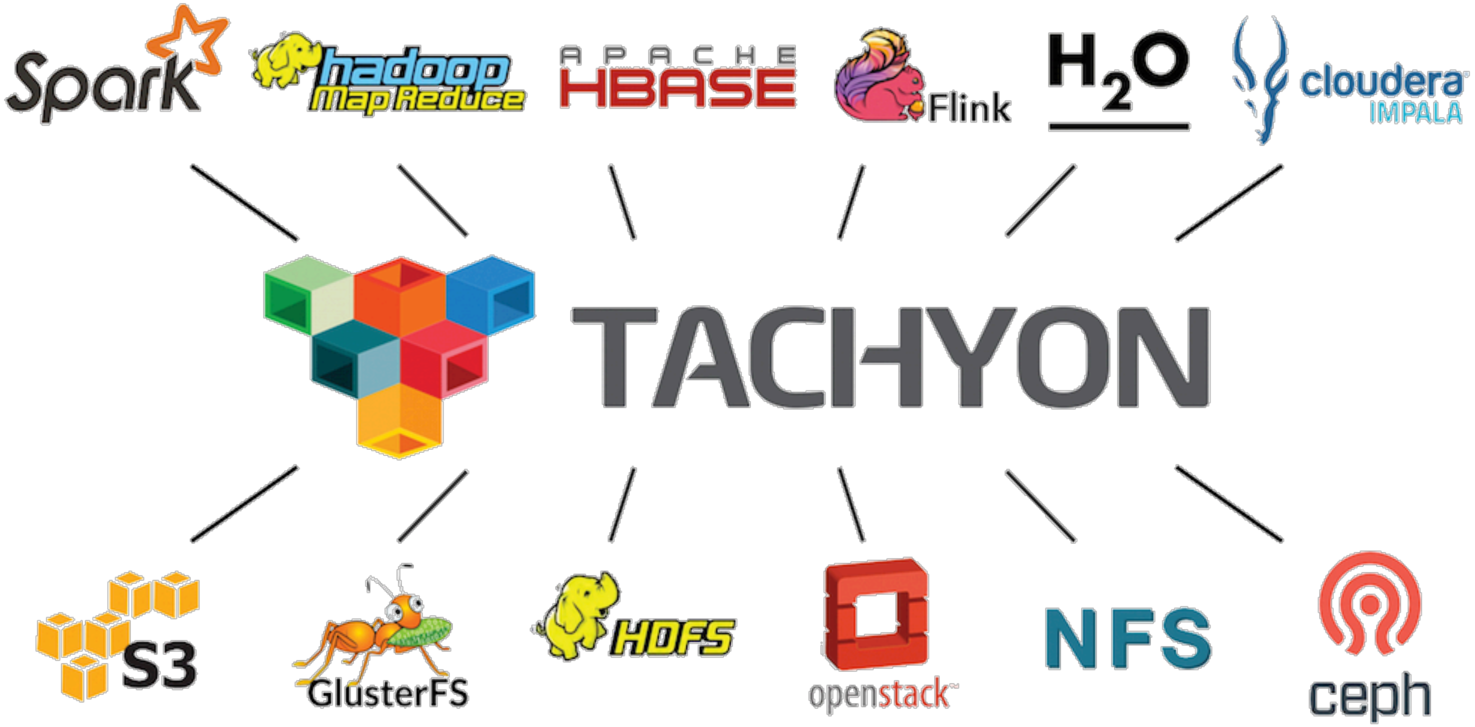
Lineage in Tachyon



Outline

- Open Source
- Introduction to Tachyon (Before 2015)
- **Deployments and New Features**
- Getting Involved

1) Eco-system:
Enable new workload in any storage;
Work with the framework of your choice;



2) Tachyon running in production environments, both in the Cloud and on Premise.

Use Case: Baidu

- Framework: SparkSQL
- Under Storage: Baidu's File System
- Storage Media: MEM + HDD
- 100+ nodes deployment
- 1PB+ managed space
- 30x Performance Improvement

Use Case: an Oil Company

- Framework: Spark
- Under Storage: GlusterFS
- Storage Media: MEM only
- Analyzing data in traditional storage

Use Case: a SAAS Company

- Framework: Impala
- Under Storage: S3
- Storage Media: MEM + SSD
- 15x Performance Improvement

Use Case: a Biotechnology Company

- Framework: Spark & MapReduce
- Under Storage: GlusterFS
- Storage Media: MEM and SSD

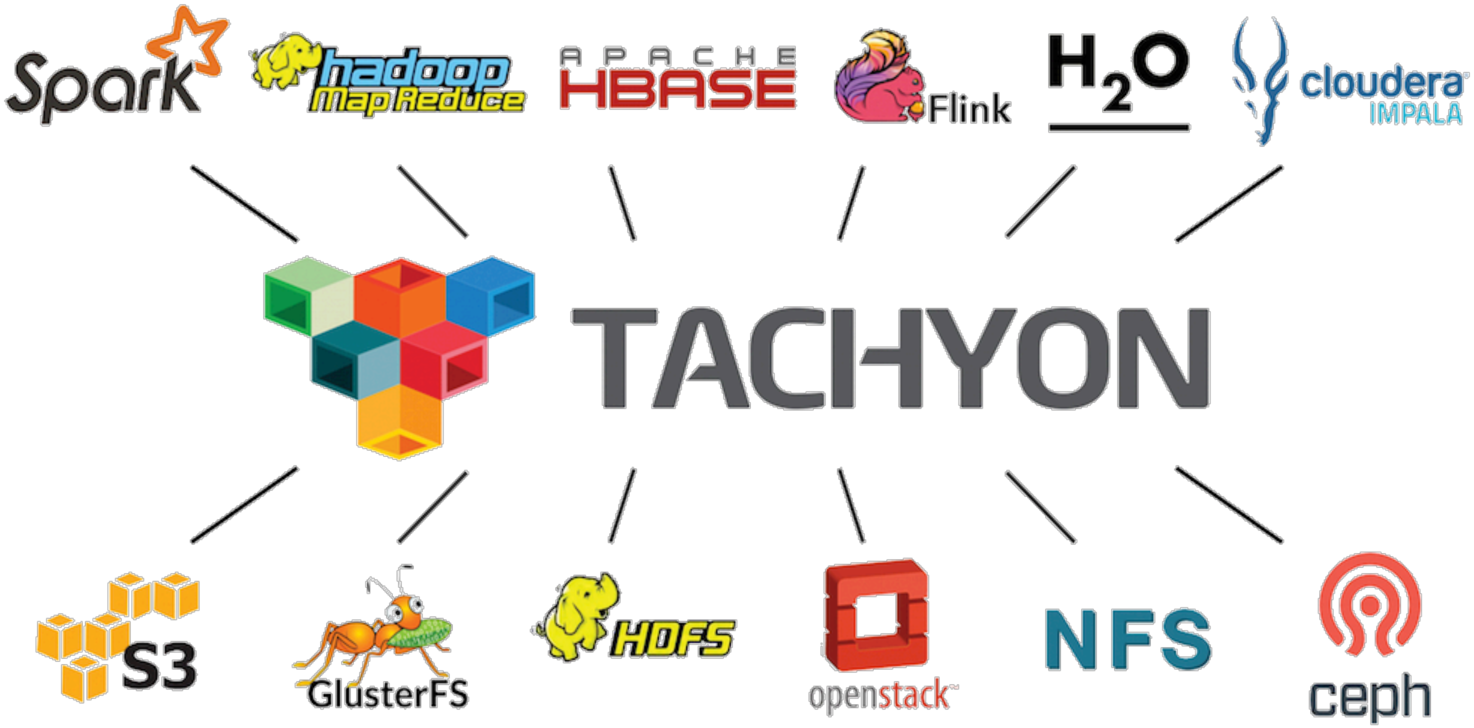
Use Case: a SAAS Company

- Framework: Spark
- Under Storage: S3
- Storage Media: SSD only
- Elastic Tachyon deployment

Use Case: a Retail Company

- Framework: Spark & MapReduce
- Under Storage: HDFS
- Storage Media: MEM

Run Everywhere

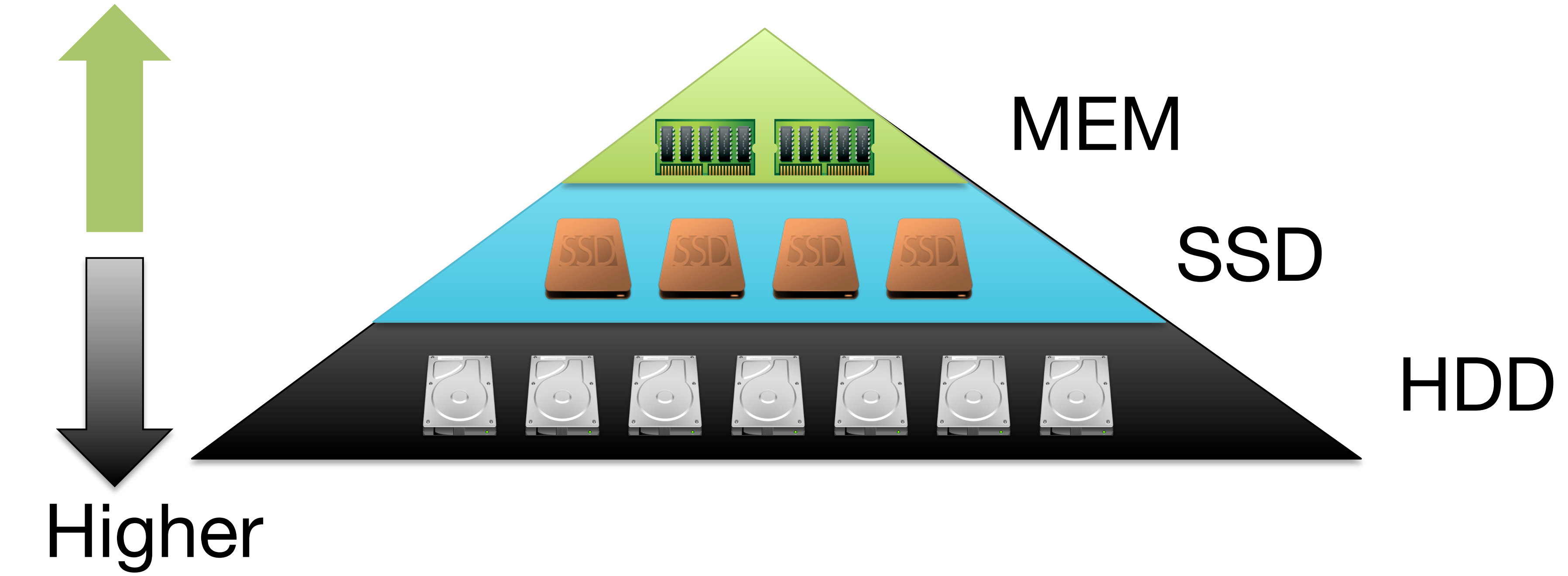


Enable Faster Innovation in Storage Layer

What if
data size exceeds
memory capacity?

3) Tiered Storage: Tachyon Manages More Than DRAM

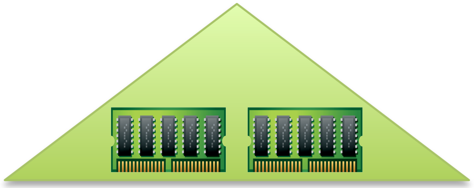
TACHYON



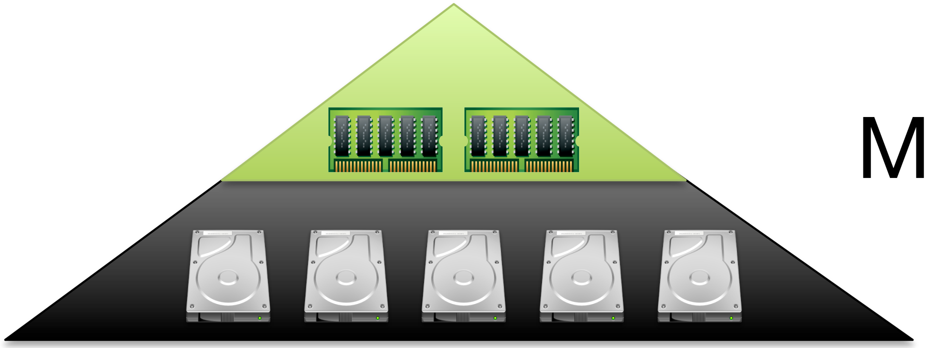
TACHYON NEXUS

amplab UC BERKELEY

Configurable Storage Tiers



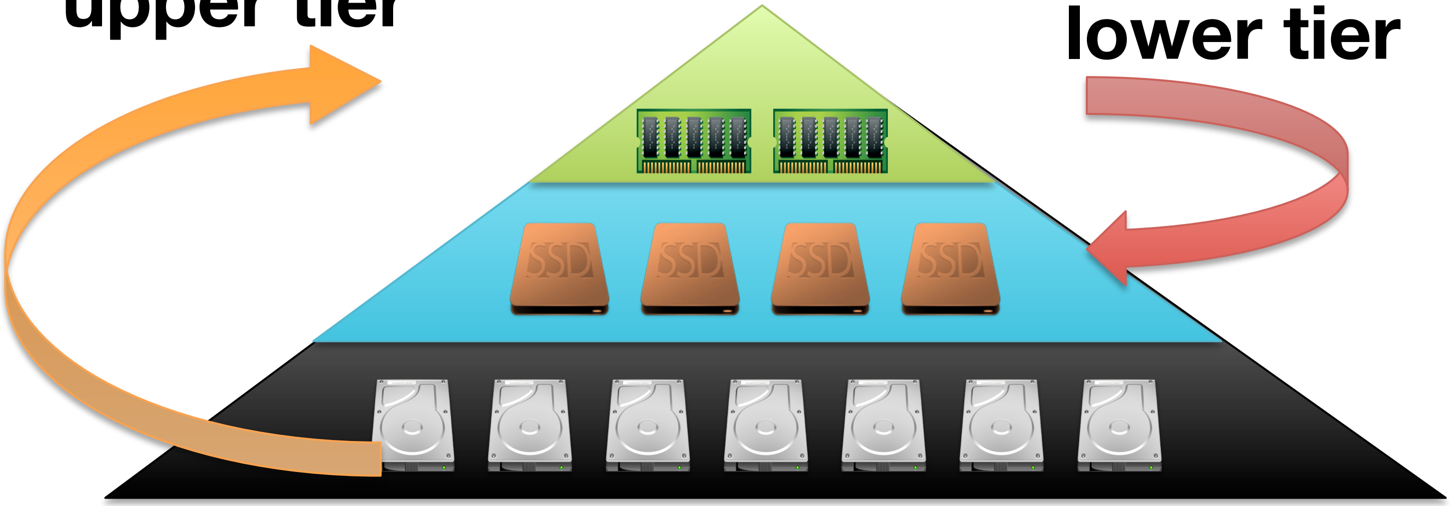
MEM only



MEM + HHD

SSD only

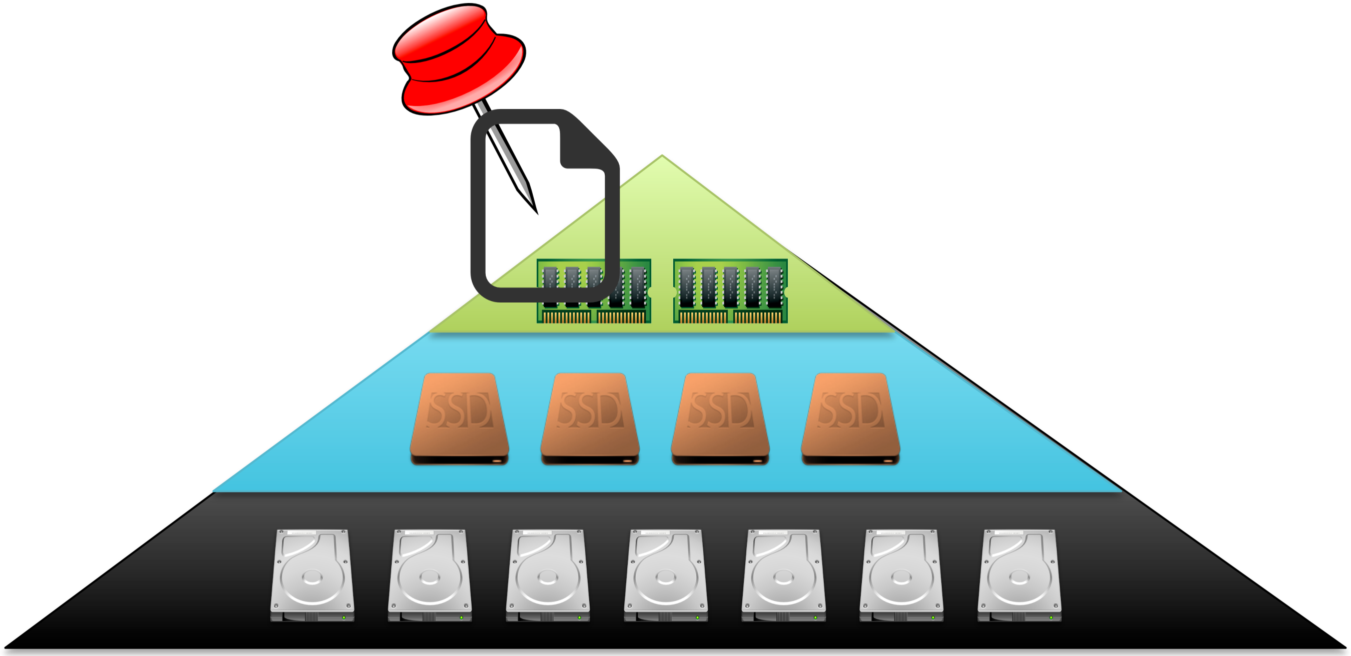
4) Pluggable Data Management Policy



Promote hot data to upper tier

Evict stale data to lower tier

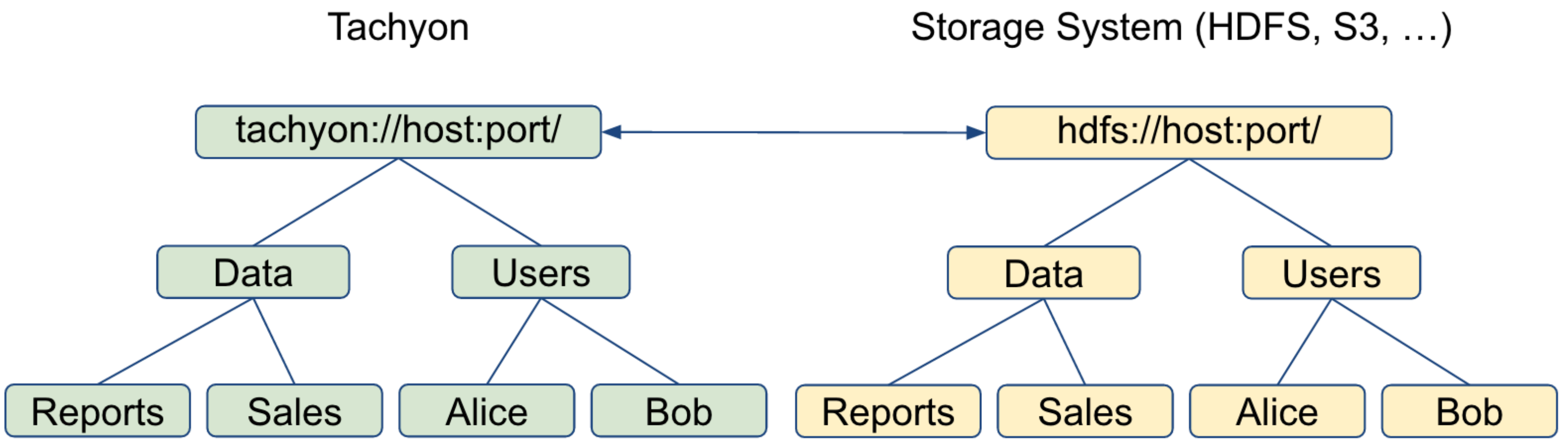
Pin Data in Memory



TACHYON NEXUS

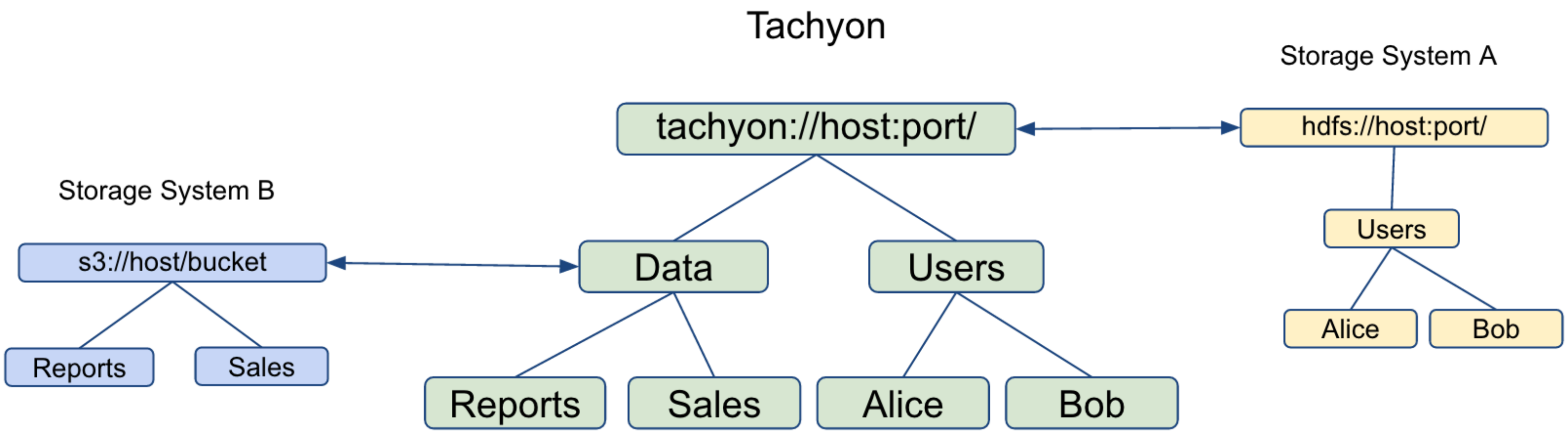
amplab UC BERKELEY

5) Transparent Naming



- Creation, renaming, and deletion of persisted Tachyon objects mapped to storage layer
- Tachyon paths are preserved in storage layer

6) Unified Namespace



- Unified namespace for multiple data sources
- Sharing of data across storage systems
- API for on-the-fly mounting / unmounting

More Features

- 7) Remote Write Support
- 8) Easy deployment with Mesos and Yarn
- 9) Initial Security Support
- 10) One Command Cluster Deployment
- 11) Metrics Reporting for Clients, Workers, and Master

12) More Under Storage Supports

HDFS

Gluster®
NON STOP STORAGE

amazon web services™ | S3

openstack
CLOUD SOFTWARE

ceph

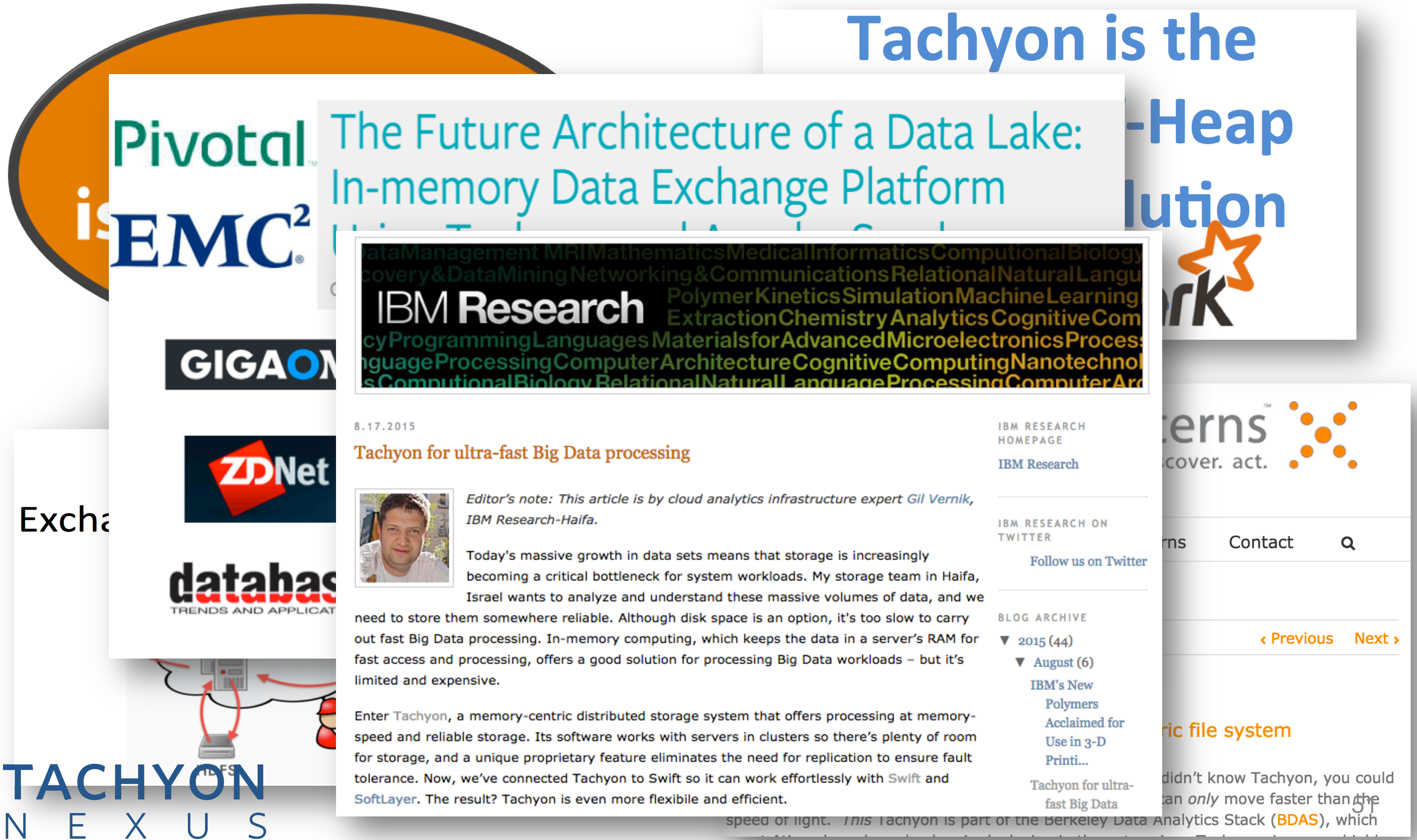
NFS
NETWORK FILE SYSTEM

OrangeFS™

lustre®

IBM
GPFS

Reported Tachyon Usage



is
Tachyon is the
-Heap
lution
rk
Pivotal
EMC²
GIGAOM
ZDNet
databas
TRENDS AND APPLICAT
Excha
HDFS
IBM Research
ataManagementMRIMathematicsMedicalInformaticsComputationalBiology
covery&DataMiningNetworking&CommunicationsRelationalNaturalLangu
PolymerKineticsSimulationMachineLearning
ExtractionChemistryAnalyticsCognitiveCom
cyProgrammingLanguagesMaterialsforAdvancedMicroelectronicsProcess
nguageProcessingComputerArchitectureCognitiveComputingNanotechnol
terns™
cover. act.
rns
Contact
‹ Previous
Next ›

The Future Architecture of a Data Lake: In-memory Data Exchange Platform

8.17.2015

Tachyon for ultra-fast Big Data processing

Editor's note: This article is by cloud analytics infrastructure expert Gil Vernik, IBM Research-Haifa.

Today's massive growth in data sets means that storage is increasingly becoming a critical bottleneck for system workloads. My storage team in Haifa, Israel wants to analyze and understand these massive volumes of data, and we need to store them somewhere reliable. Although disk space is an option, it's too slow to carry out fast Big Data processing. In-memory computing, which keeps the data in a server's RAM for fast access and processing, offers a good solution for processing Big Data workloads – but it's limited and expensive.

Enter Tachyon, a memory-centric distributed storage system that offers processing at memory-speed and reliable storage. Its software works with servers in clusters so there's plenty of room for storage, and a unique proprietary feature eliminates the need for replication to ensure fault tolerance. Now, we've connected Tachyon to Swift so it can work effortlessly with Swift and SoftLayer. The result? Tachyon is even more flexibile and efficient.

IBM RESEARCH HOMEPAGE

IBM Research

IBM RESEARCH ON TWITTER

Follow us on Twitter

BLOG ARCHIVE

- ▼ 2015 (44)
  - ▼ August (6)
    - IBM's New Polymers Acclaimed for Use in 3-D Printi...
    - Tachyon for ultra-fast Big Data

ric file system

didn't know Tachyon, you could ...an *only* move faster than the speed of light. *This* Tachyon is part of the Berkeley Data Analytics Stack (BDAS), which

TACHYON
NEXUS

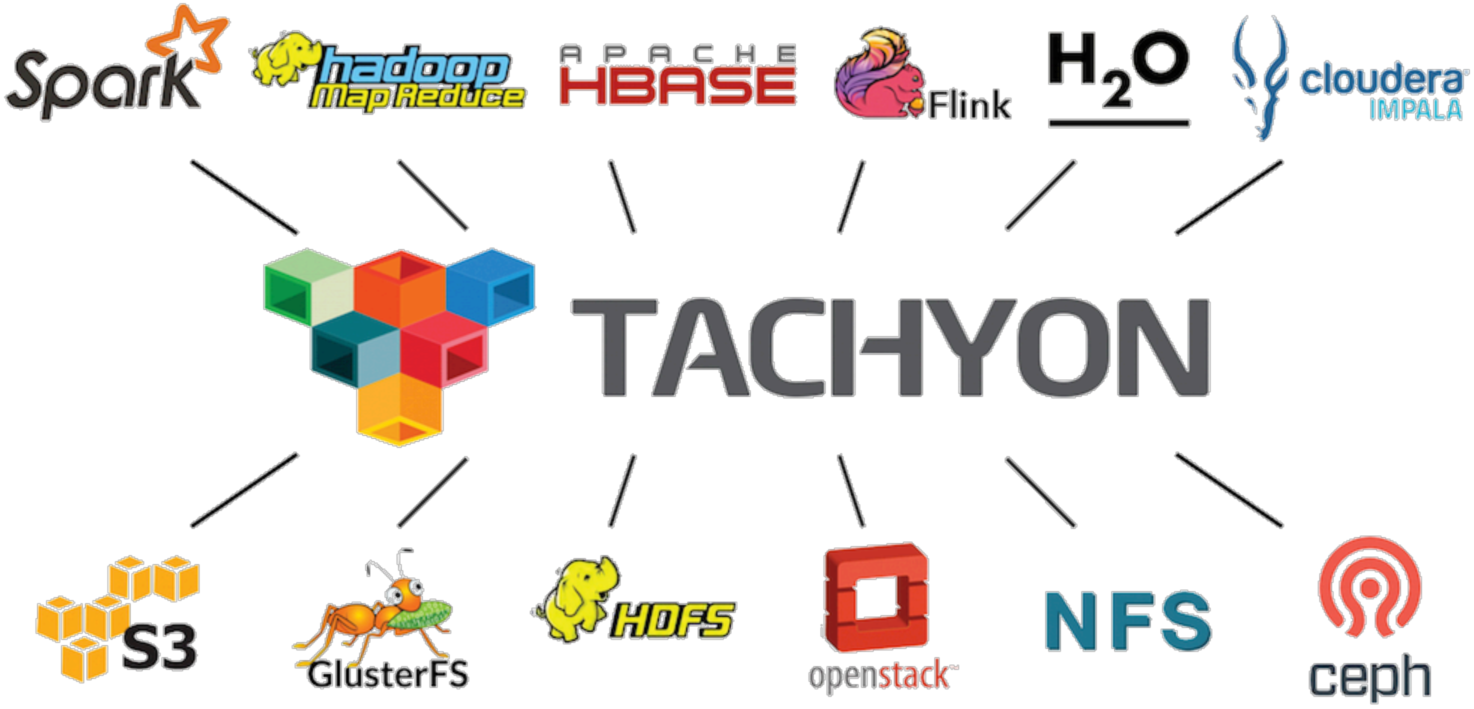
TACHYON NEXUS

- Team consists of Tachyon creators, top contributors
- Series A ($7.5 million) from Andreessen Horowitz
- Committed to Tachyon Open Source
- http://www.tachyonnexus.com

Outline

- Open Source
- Introduction to Tachyon (Before 2015)
- Deployments and New Features
- **Getting Involved**

Memory-Centric Distributed Storage



Welcome to try, contact, and collaborate!

JIRA New Contributor Tasks

TACHYON

- Try Tachyon: http://tachyon-project.org
- Develop Tachyon: https://github.com/amplab/tachyon
- Meet Friends: http://www.meetup.com/Tachyon
- Get News: http://goo.gl/mwB2sX
- Tachyon Nexus: http://www.tachyonnexus.com
- Contact us: haoyuan@tachyonnexus.com