ALLUXIO

Spark Pipelines in the Cloud with Alluxio

Gene Pang, Alluxio, Inc.

Spark Summit EU - October 2017

About Me

- Gene Pang
- Software engineer @ Alluxio, Inc.
- Alluxio open source PMC member
- Ph.D. from AMPLab @ UC Berkeley
- Worked at Google before UC Berkeley
- Twitter: @unityxx
- Github: @gpang

Outline

1. Alluxio Overview
2. Data Pipelines
3. Experiments

History of Alluxio

Started at UC Berkeley AMPLab In Summer 2012

- Originally named as Tachyon
- Rebranded to Alluxio in early 2016

Open Sourced in 2013

- Apache License 2.0
- Latest Release: Alluxio 1.6.0

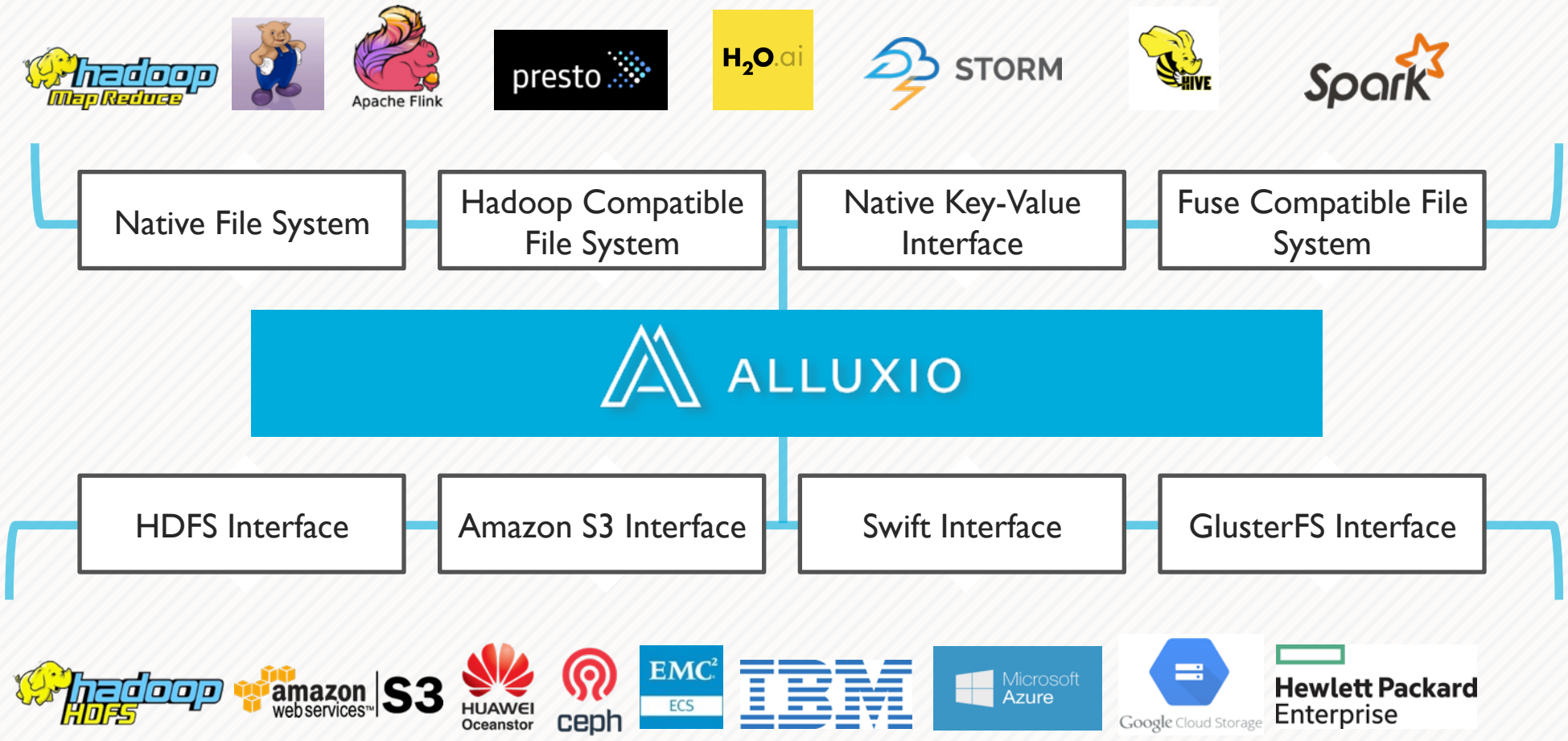
Alluxio: Unify Data at Memory Speed

Namespace Unification

Architecture Flexibility

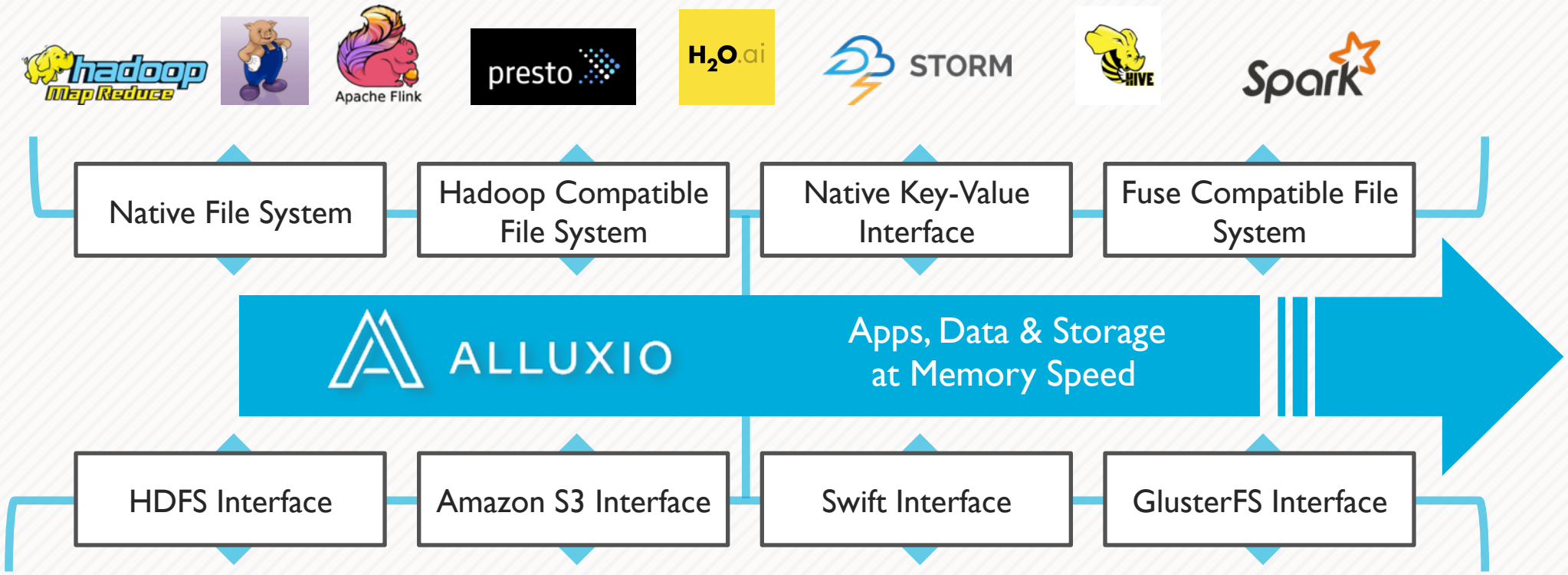
IO Performance

Data Ecosystem with Alluxio



- Apps only talk to Alluxio
- Simple Add/Remove
- No App Changes
- In-Memory Performance

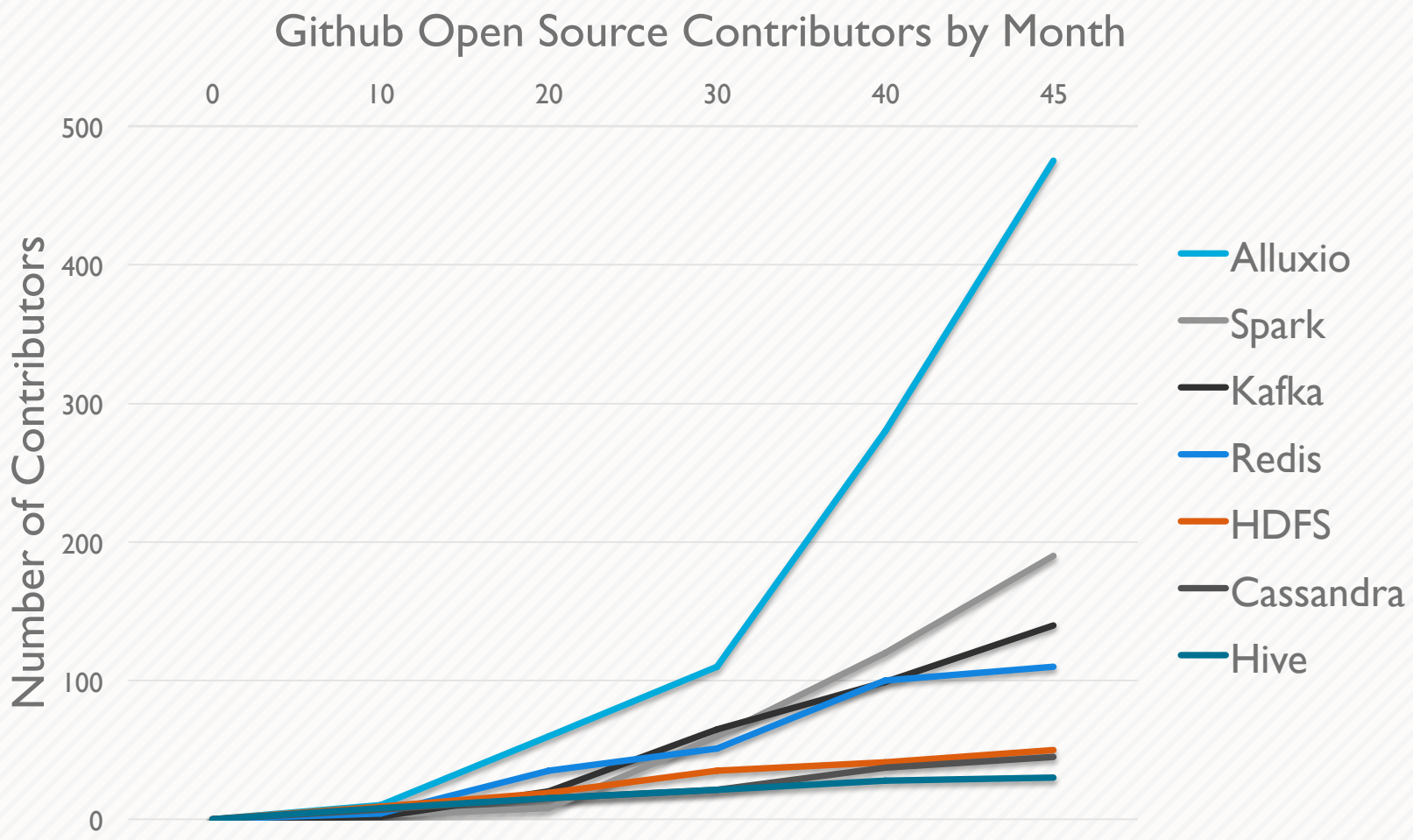
Next Gen Analytics with Alluxio



- ✓ Big Data/IoT
- ✓ AI/ML
- ✓ Deep Learning
- ✓ Cloud Migration
- ✓ Multi Platform
- ✓ Autonomous

hadoop HDFS · amazon web services S3 · HUAWEI Oceanstor · ceph · EMC ECS · IBM · Microsoft Azure · Google Cloud Storage · Hewlett Packard Enterprise

Fastest Growing Big Data Open Source Projects



Fastest Growing open-source project in the big data ecosystem

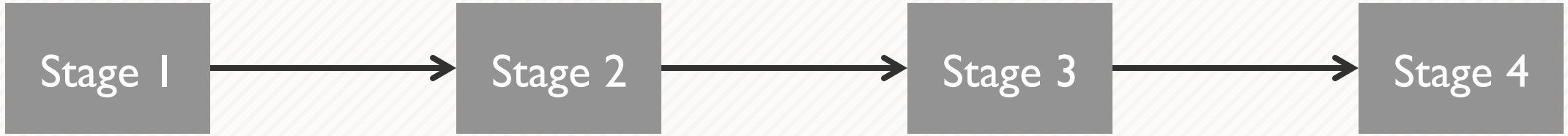
Running in large production clusters

Now: 600+ Contributors from 100+ organizations

Outline

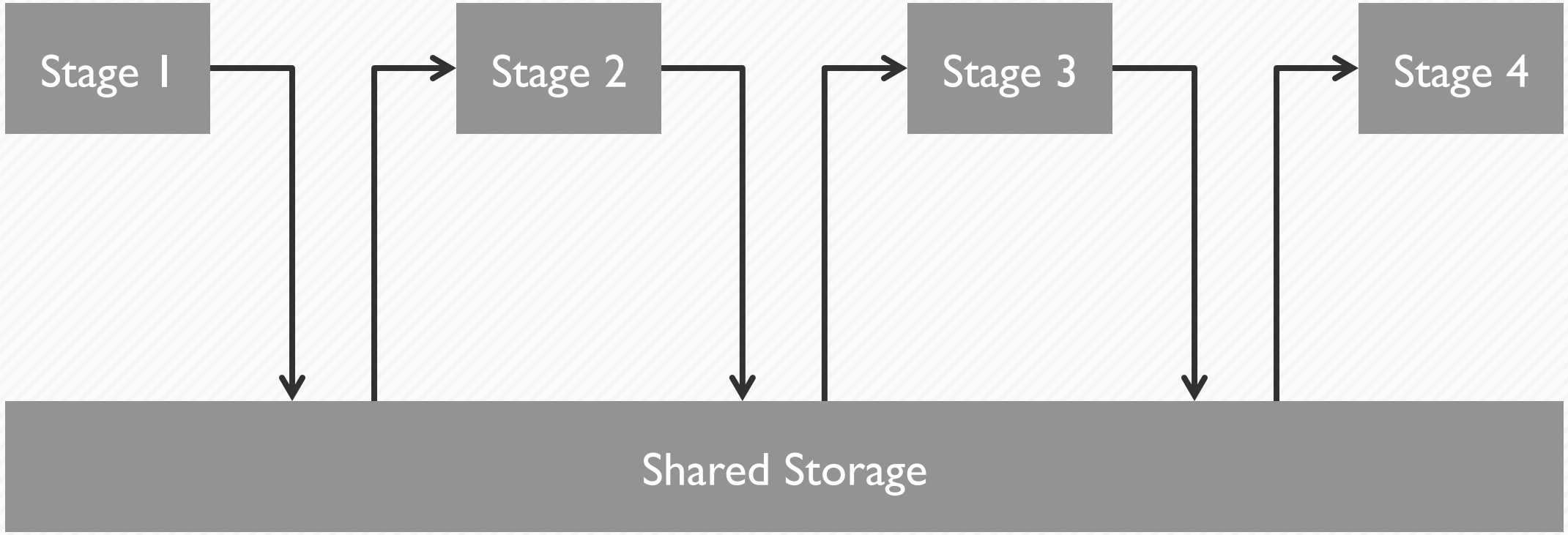
1. Alluxio Overview
2. Data Pipelines
3. Experiments

Data Processing Pipeline



Output of stage is input of next stage

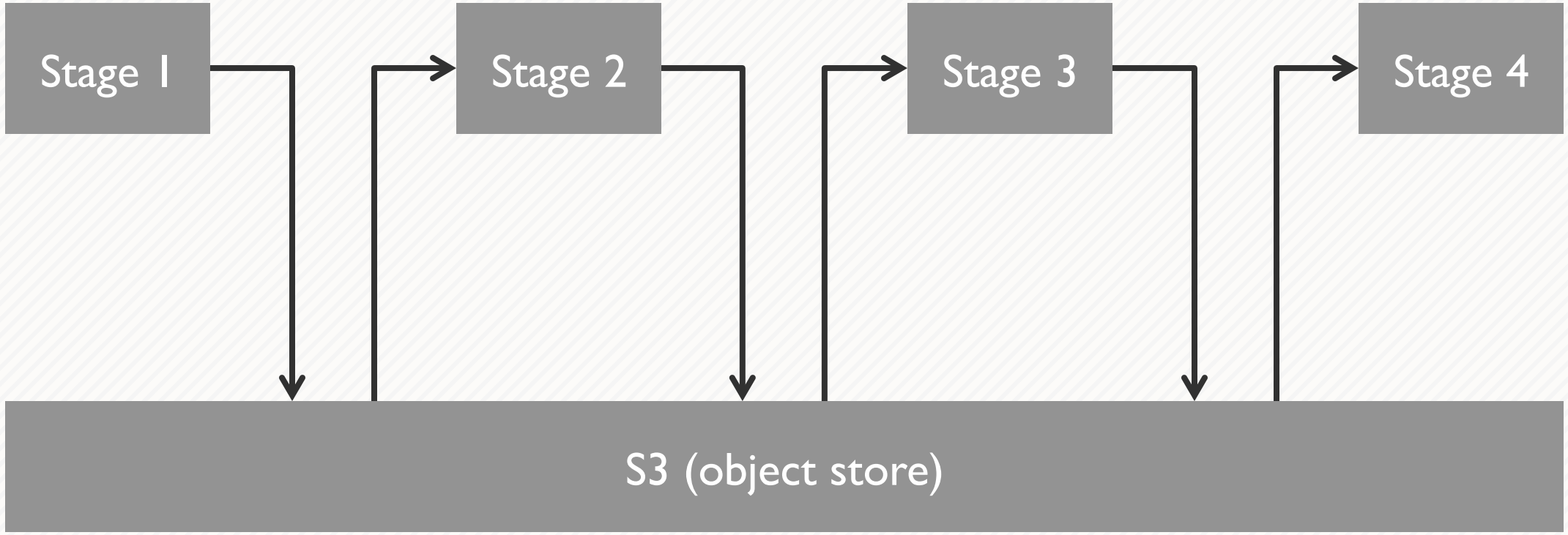
Data Processing Pipeline



Sharing via common storage

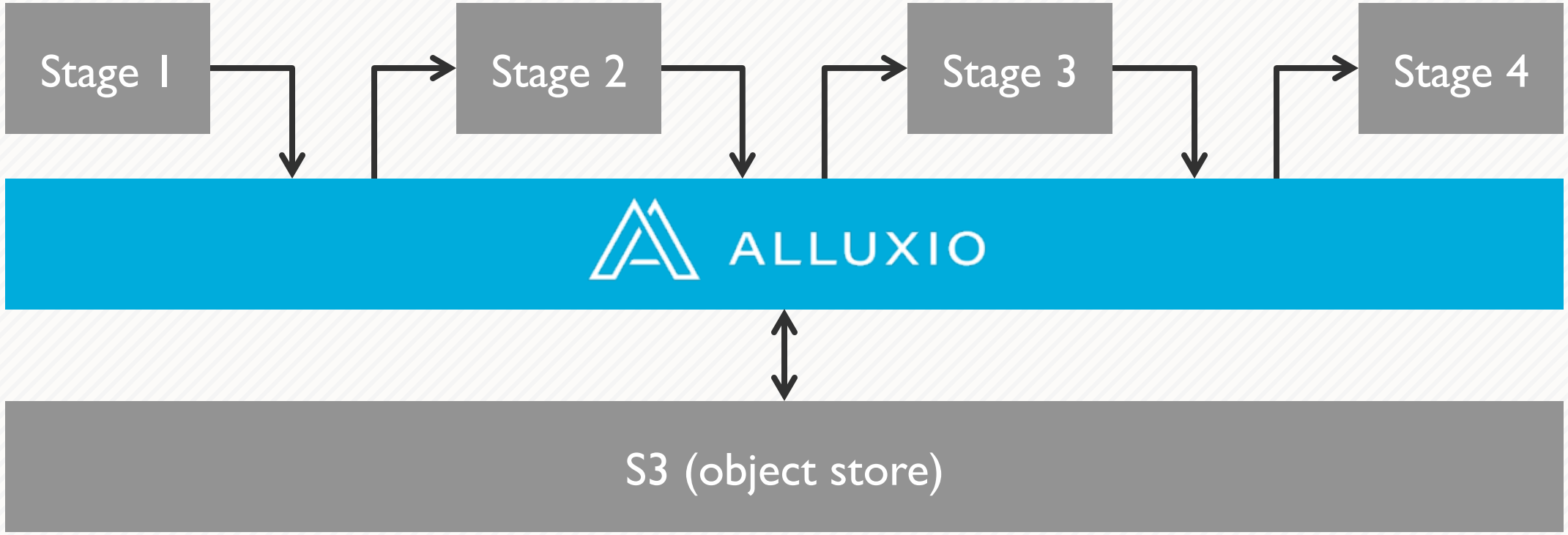

What about
pipelines in the cloud?

Data Processing Pipeline in the Cloud



Sharing data via cloud storage slows down performance

Cloud Pipeline with Alluxio



Sharing via Alluxio memory

Sharing Data in the Cloud

Previous stage writes output to storage

Next stage reads input from storage

…

Sharing Data in the Cloud with Alluxio

ALLUXIO

Previous stage writes output to ~~storage~~ **memory**

Next stage reads input from ~~storage~~ **memory**

…

Alluxio enables

in-memory data sharing

Faster pipeline performance

Alluxio – Fast Durable Writes

Improves write performance,

without sacrificing fault tolerance

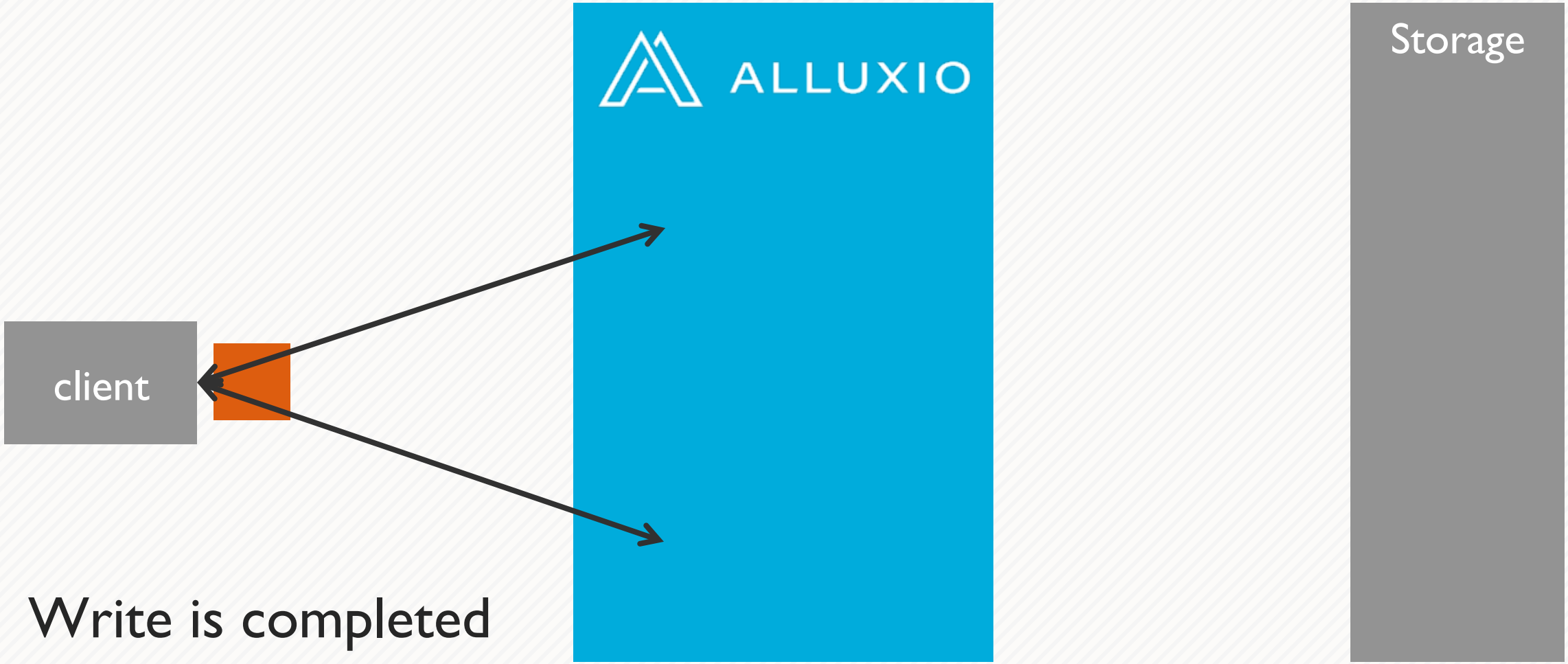
Alluxio – Fast Durable Writes

Synchronously write to replicas in Alluxio memory

+

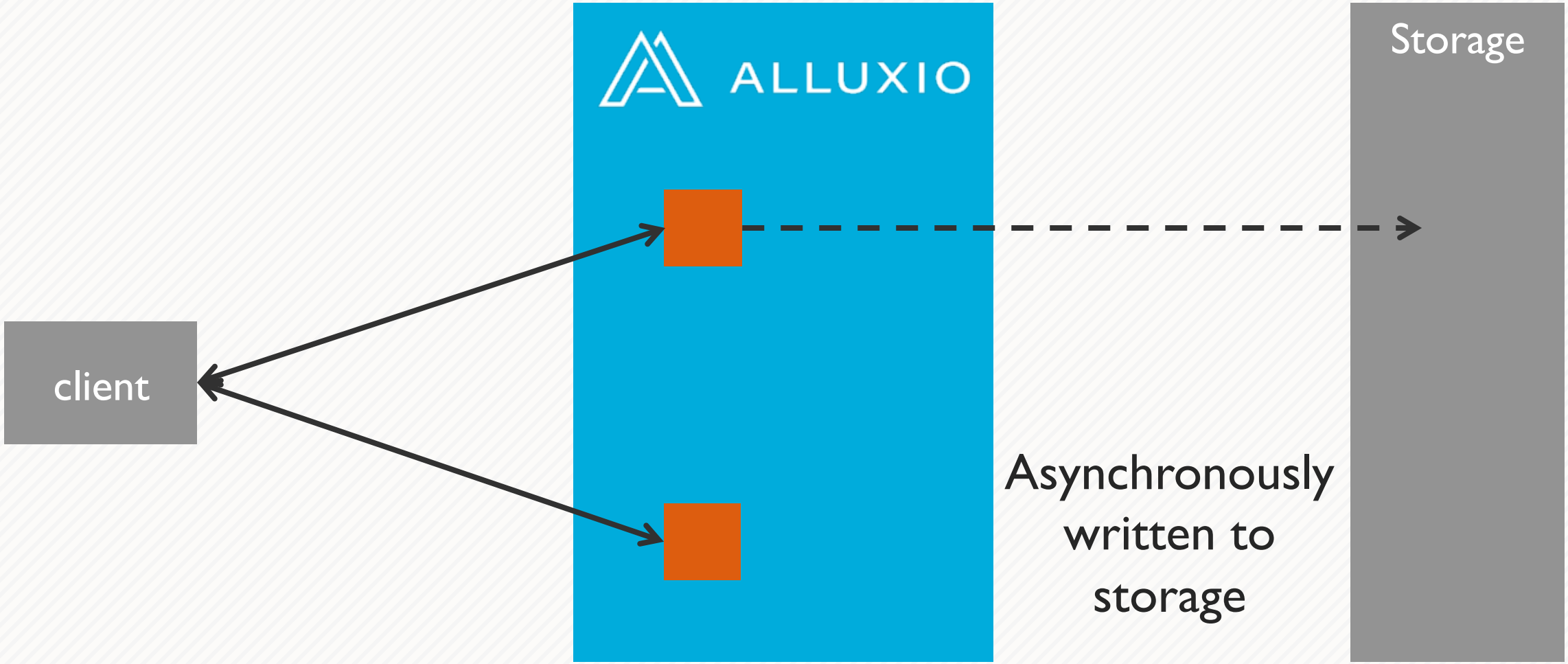
Asynchronously write to underlying storage

Alluxio – Fast Durable Writes



Write is completed

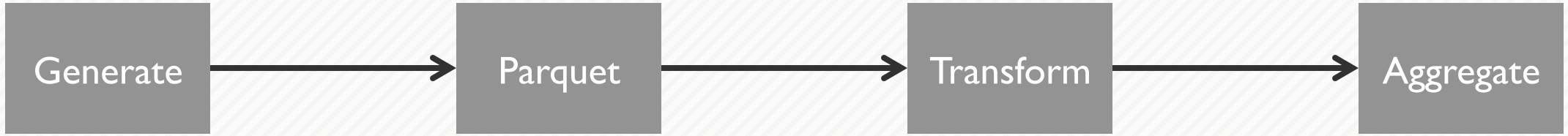
Alluxio – Fast Durable Writes



Outline

1. Alluxio Overview
2. Data Pipelines
3. Experiments

Log Pipeline in Amazon Web Services



Generate: [MapReduce] Create random csv log data

Parquet: [MapReduce] Convert csv to parquet format

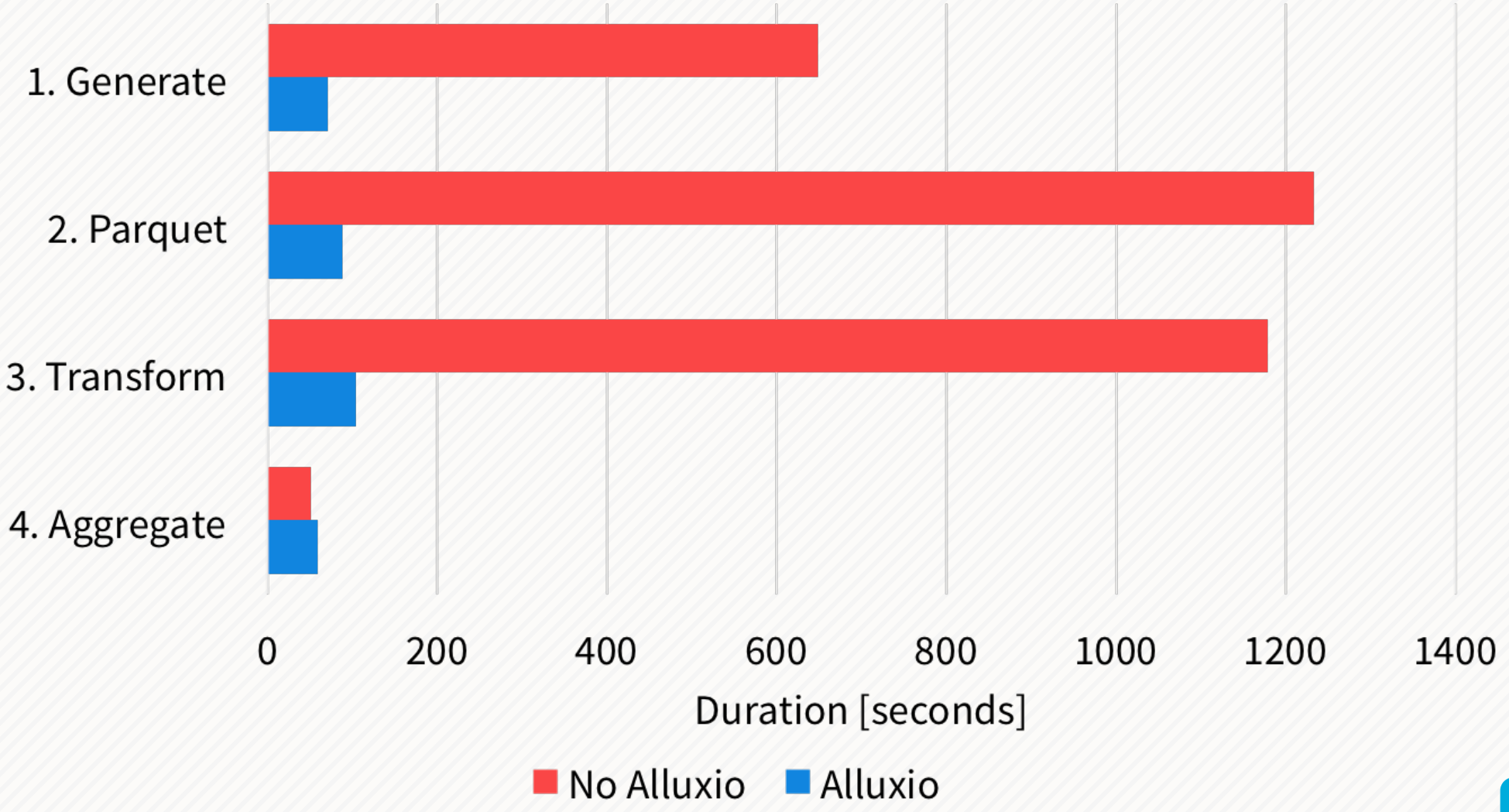
Transform: [Spark] Update column values

Aggregate: [Spark] Compute group by / aggregate

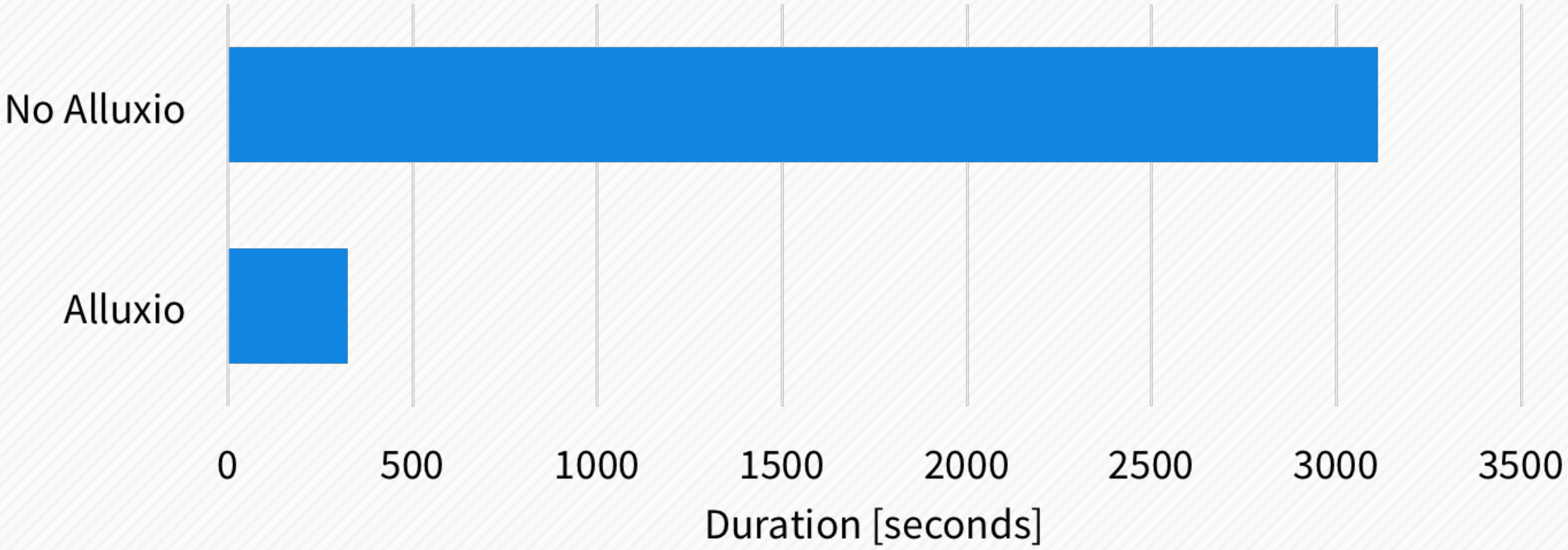
Log Pipeline Environment

- r4.2xlarge instances (61 GB ram, 8 CPUs)
  - 1 master, 3 workers
- Apache Spark 2.2.0
- Apache Hadoop 2.7.2
- Alluxio 1.6.0
- Generate 12 GB of logs
- Compare AWS S3 vs Alluxio w/ Fast Durable Writes

Average Stage Completion Time



Pipeline Completion Time



Over 9x speedup!

Alluxio and Pipelines in the Cloud

Alluxio enables in-memory sharing for data pipelines in the cloud

Alluxio's Fast Durable Write feature increases performance without sacrificing fault tolerance

Thank you!

Gene Pang
gene@alluxio.com
Twitter: @unityxx

Website

www.alluxio.com

E-mail

info@alluxio.com

Social Media

Twitter.com/alluxio

Linkedin.com/alluxio