Tencent 腾讯

Running Hadoop Jobs on Typhoon System

Huican Zhu

Talk Overview

Tencent 腾讯

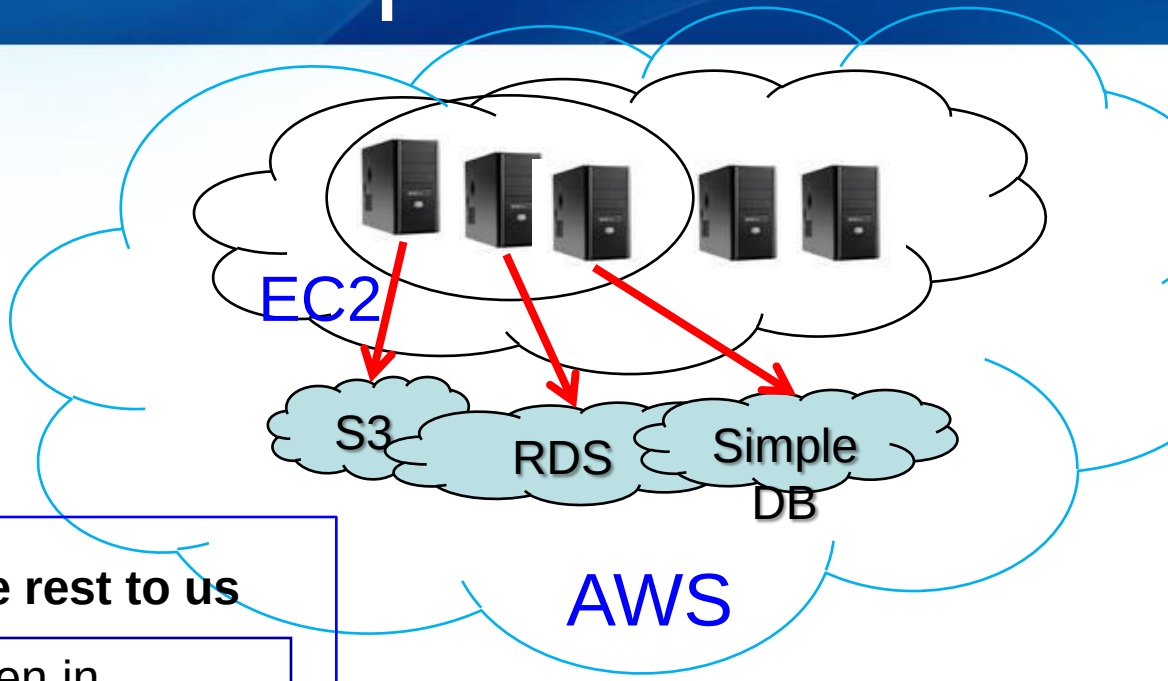
- Introduction To Typhoon
- Scheduling Hadoop Jobs to Typhoon
- Mapping HDFS to XFS

Typhoon System

Tencent 腾讯

- Typhoon is Tencent's cloud computing platform.
  - Focusing on IaaS + PaaS
- Purpose:
  - Manage storage and computing resources.
    - free programmers from maintenance work
  - Resource sharing:
    - Better resource utilization
  - Uniform management :
    - uniform monitoring, maintenance and security

Platform Comparison



Focus on your app, leave the rest to us

Google App Engine

- Apps written in Java, Python
- Access to HRD
- Run in sandboxed env.
-Quota, Limits

More Platforms

Condor
High Throughput Computing

- Heterogeneous Environment
- Harness idle resources
- Flexible matching algorithm
- High throughput

-Private cloud
-HDFS, MapReduce, Hbase
- Batch processing

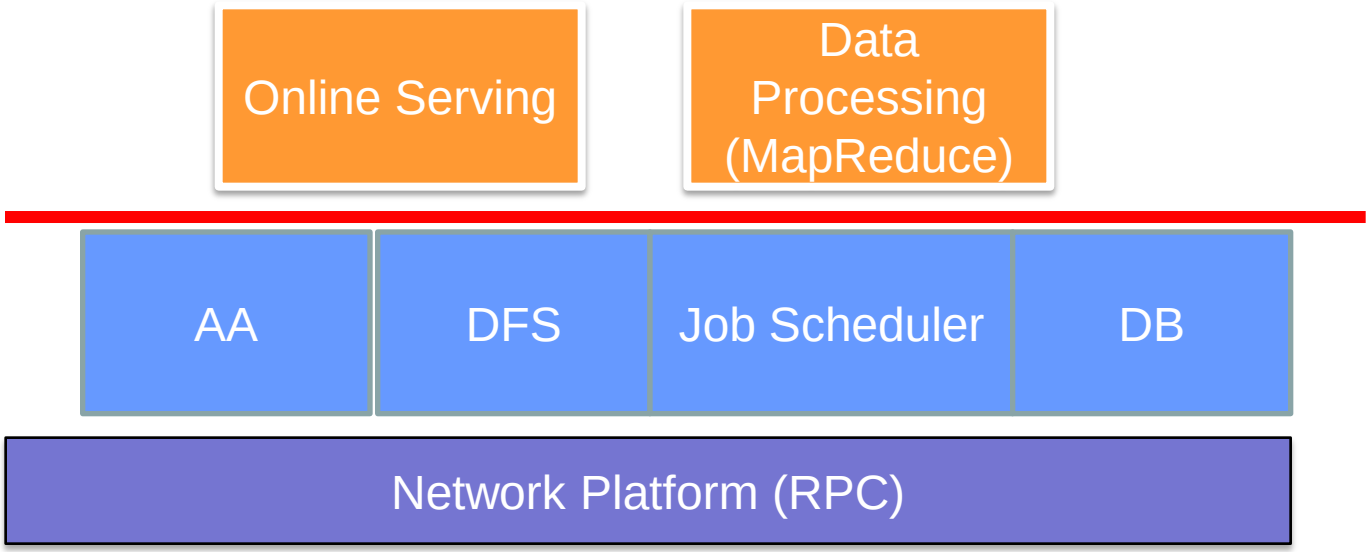
Cloud Computing as OS

Tencent 腾讯

- File system (Storage)
- Process and Resource management(CPU, memory, etc)
- Authentication and Authorization(accounts, permission, quota, etc)
- System software(database etc)
- APIs for software development （networking, threads, etc)

Typhoon Layers

Tencent 腾讯

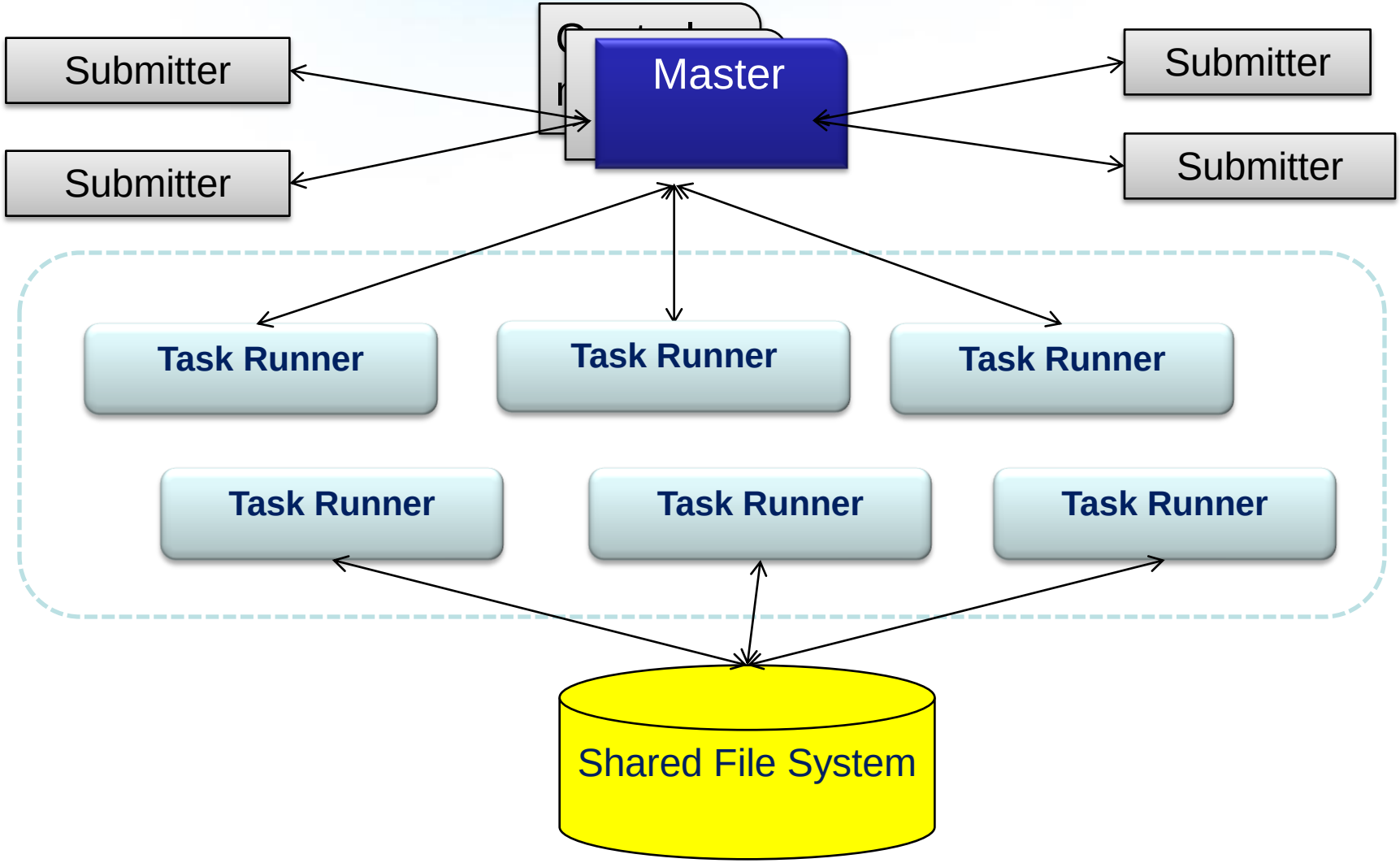


Typhoon Properties

Tencent 腾讯

- Private cloud
- Computing nodes are more uniform: Linux OS, central management
- Standalone job scheduler
- Written in C++

Scheduler Architecture



Scheduler Properties

Tencent 腾讯

- Support diverse binary types:
  - MapReduce programs
  - “Hello world” binary
  - Java programs
- Support diverse job types:
  - Online serving:
    - Latency sensitive, resource guarantee.
  - Offline processing:
    - Batch oriented. Throughput more important
- Resource guarantee: Quota

Talk Overview

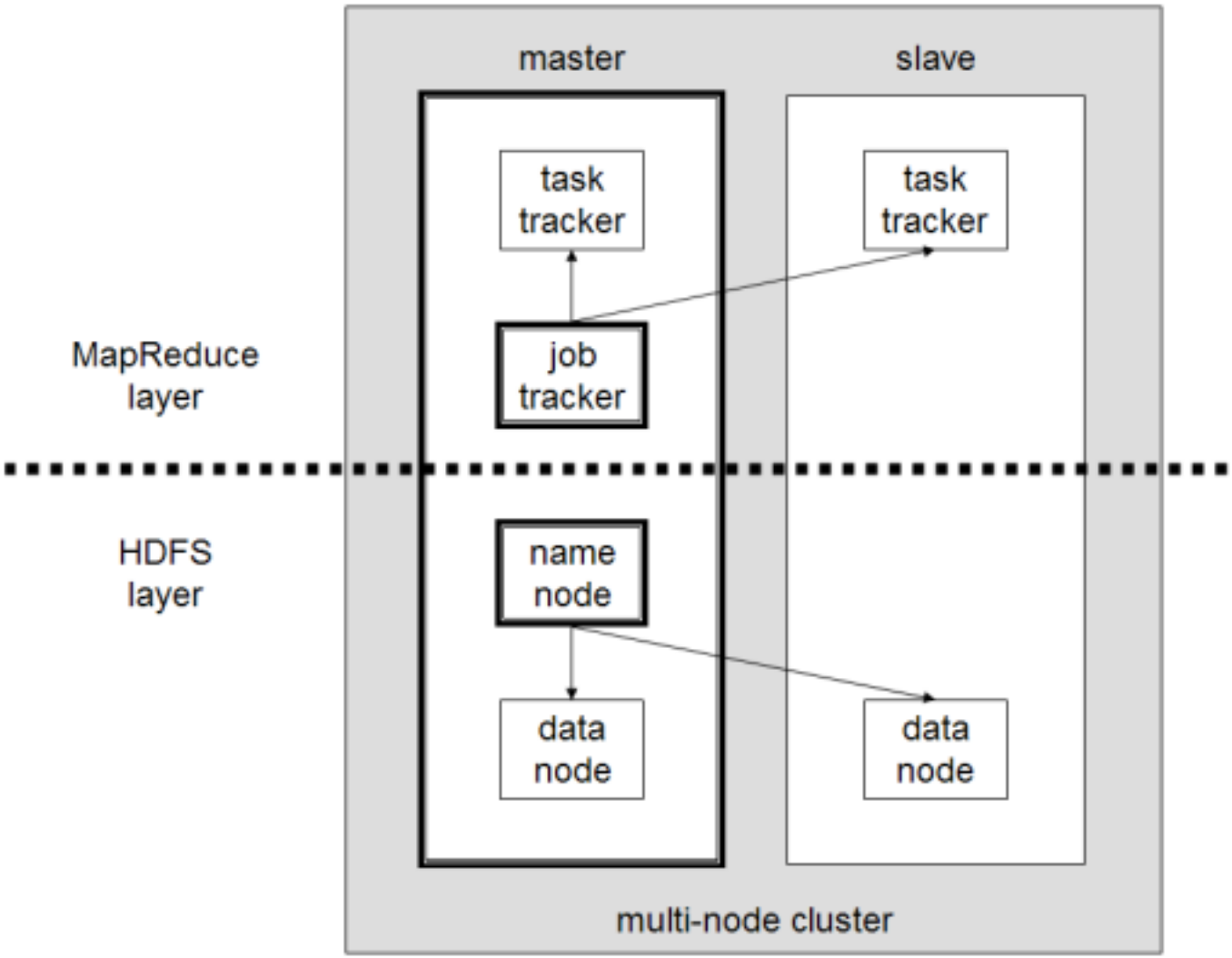
Tencent 腾讯

- Introduction To Typhoon
- Scheduling Hadoop Jobs to Typhoon
- Mapping HDFS to XFS

Why Support Hadoop Jobs

- Allow smooth transition.
  - A lot of existing hadoop jobs:
    - webpage analysis, query analysis, data mining, …
  - A lot of small hadoop clusters.
    - Owned by different teams.
- Take advantage of advances in Hadoop development.
  - Shared experiment platform
  - Advanced Hadoop features

Hadoop Structure

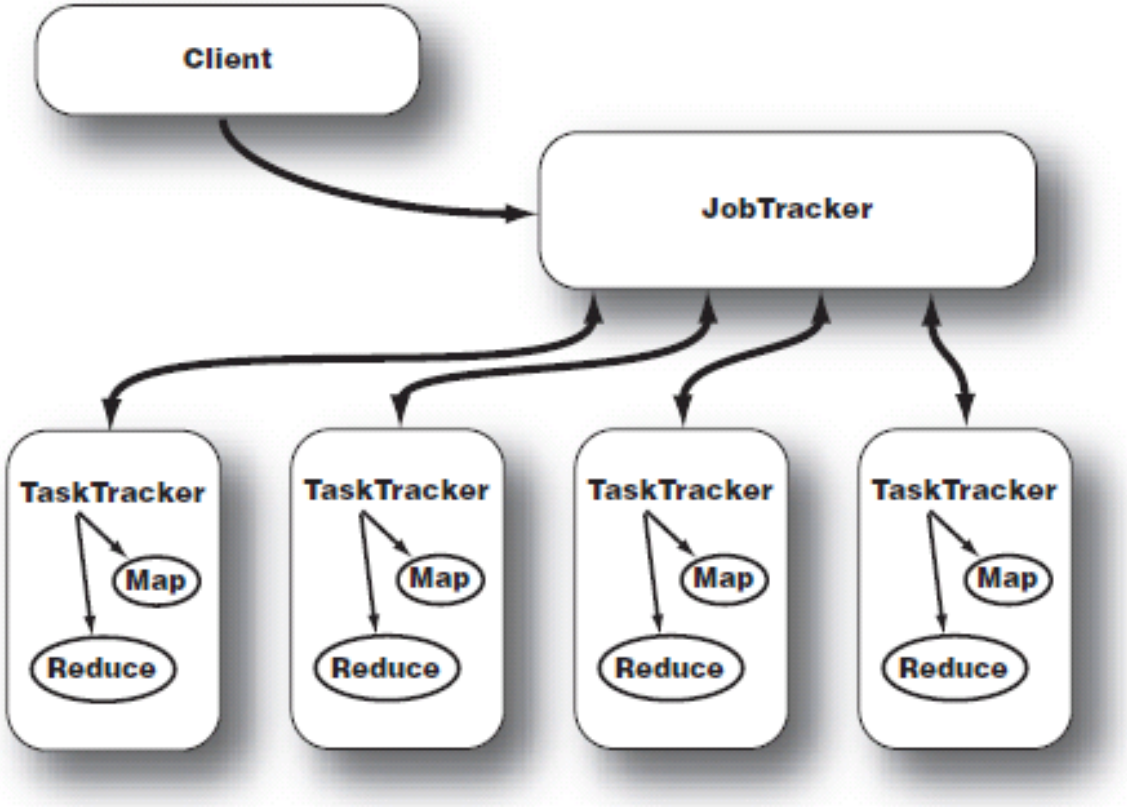


Support Hadoop Jobs

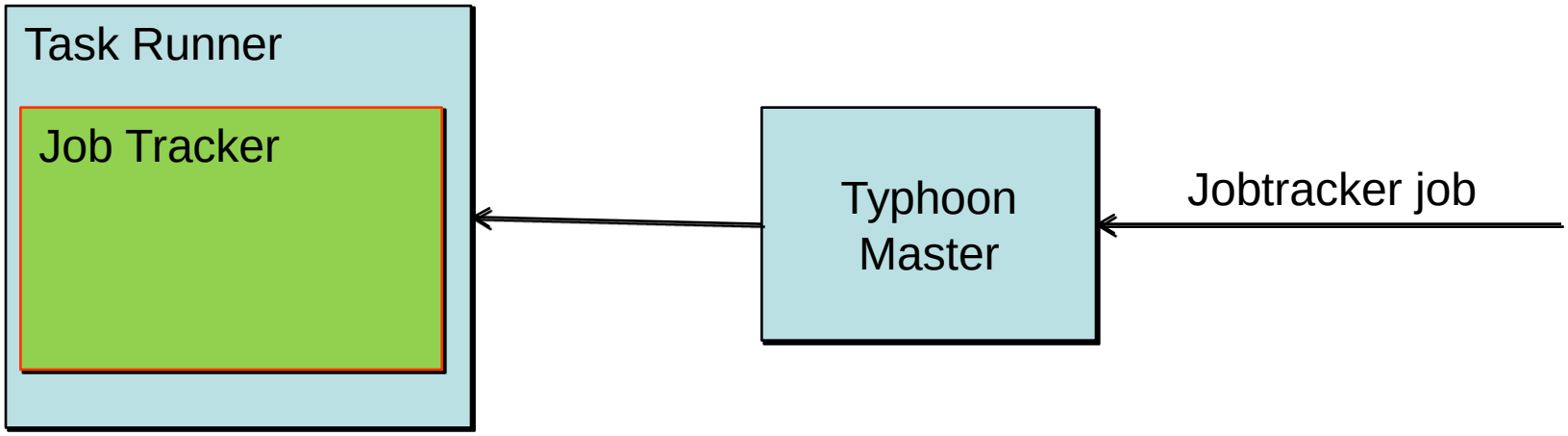
Tencent 腾讯

- Run Hadoop Jobs in Typhoon w/o modification.
- Challenges:
  - Mostly written in Java
  - Job scheduling API different from Typhoon API.
  - Mostly use HDFS as storage.

Hadoop Scheduling Overview

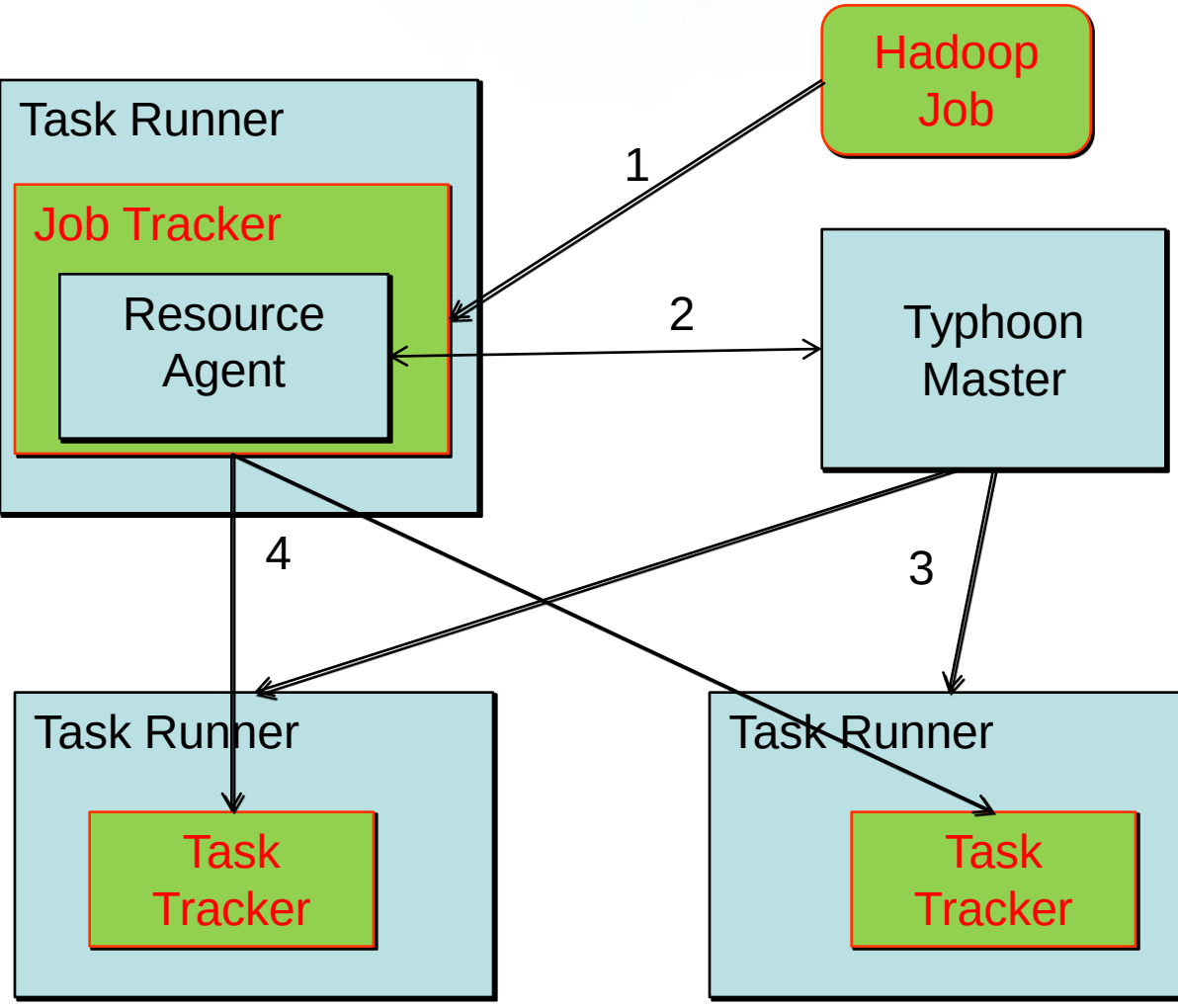


Hadoop Jobs on Typhoon



Run Jobtracker as a Typhoon job

Hadoop Jobs on Typhoon



1. Submit hadoop to Jobtracker
2. RA asks master to add more task trackers
3. Master adds more.
4. Jobtracker asks tasktrackers to run new tasks.

Other Issues

Tencent 腾讯

- When to recycle task trackers to make room for non-hadoop jobs?
- How to preempt running hadoop jobs?

Talk Overview

Tencent 腾讯

- Introduction To Typhoon
- Scheduling Hadoop Jobs to Typhoon
- Mapping HDFS to XFS (HDFS on XFS)

What and Why?

Tencent 腾讯

- XFS: Tencent file system,
  - similar to GFS and HDFS.
  - better scalability: metadata distributed to multiple machines.
- Why HDFS on XFS?
  - Store data used by HDFS apps on XFS
  - Run hadoop apps without little or no code modification.

Feasibility

Tencent 腾讯

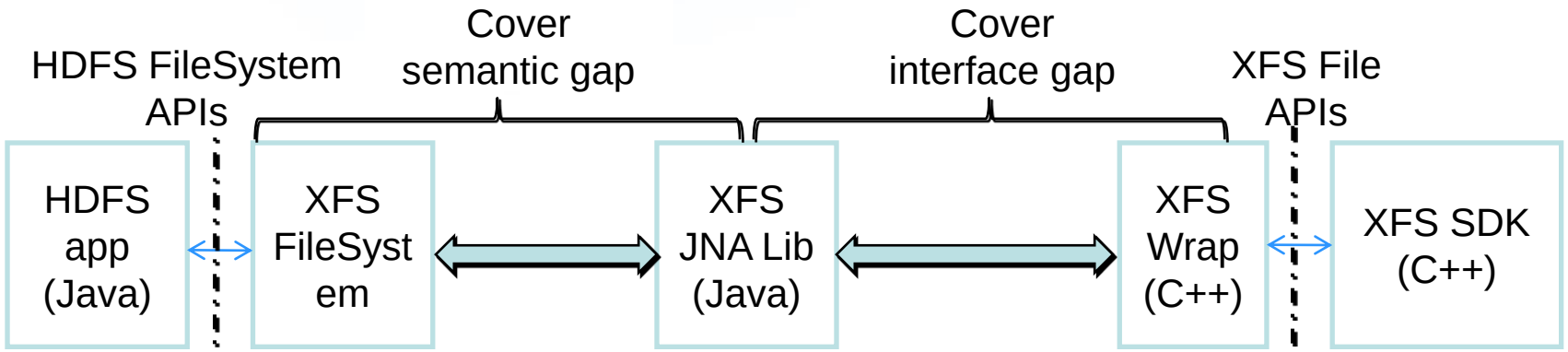
- HDFS “FileSystem” interface
  - Support extension for new File Systems
  - Java
- XFS “File” interface
  - Factory pattern. Support registering new File Systems
  - C++
- Solution: Implements an XFSFileSystem class in Java that wraps XFS “File” interface as HDFS “FileSystem”.

Challenges

- Interface gap
  - Langurage: Java vs. C++. complex structures, memory mgmt.
  - Functionality: eg. List, getLocation, diff in replica factor, permission
- Semantic gap
  - Behavior: eg. Should mkdir be recursive?
  - Exception handling: eg. What if read non-exist file

Solution overview

– HDFS on XFS Layers



– Cover interface gap

- JNA, for simplicity than JNI and flexibility than SWIG

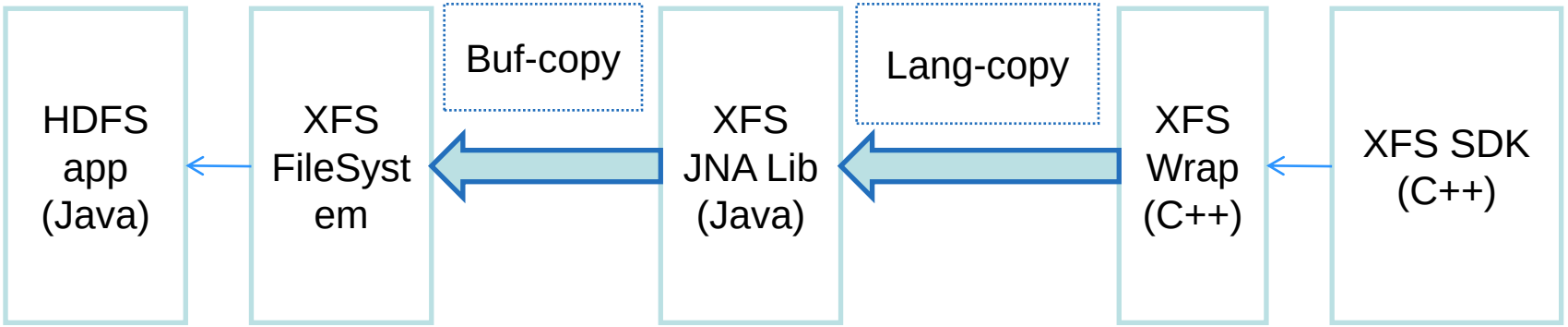
– Cover semantic gap

- Analyze both HDFS and XFS source code
- Perform exact HDFS behaviors by wrapper XFS
- Use exactly same exception handling policies

Improve Read/Write: why?

– Naïve HDFS on XFS: single client read/write performance: ~20% lower than XFS

– Why? Use read as example



– Two memory copy!

- Lang-copy: copy from C++ to Java due to no “pin” data
- Buf-copy: copy data to user provided buffer

Improve Read/Write: how?

- Java NIO + JNA
  - ByteBuffer with direct mode memory allocation in C++ heaps
  - JNA with native memory mgmt in C++ heaps by malloc/free
- HDFS on XFS: two approaches
  - HDFSonXFS-E :
    - No lang-copy by JNA, native malloc/free, retain buf-copy
  - HDFSonXFS-D : a new API,
    - No lang-copy or buf-copy
    - using JNA native malloc/free to get buffer and pass in call stack

Improve Namespace OPs

- Naïve HDFS on XFS: slow list, recursive rm, du
  - Cost for scalability: list and dus in XFS must check master and meta servers
  - Cost for stable serving: rmr and dus in XFS runs with limited speed to avoid user requests starving
- Towards fast and stable OPs with scalability
  - Iterable recursive dus and rmr in XFS Master
  - Paging operation in XFS SDK

Evaluations

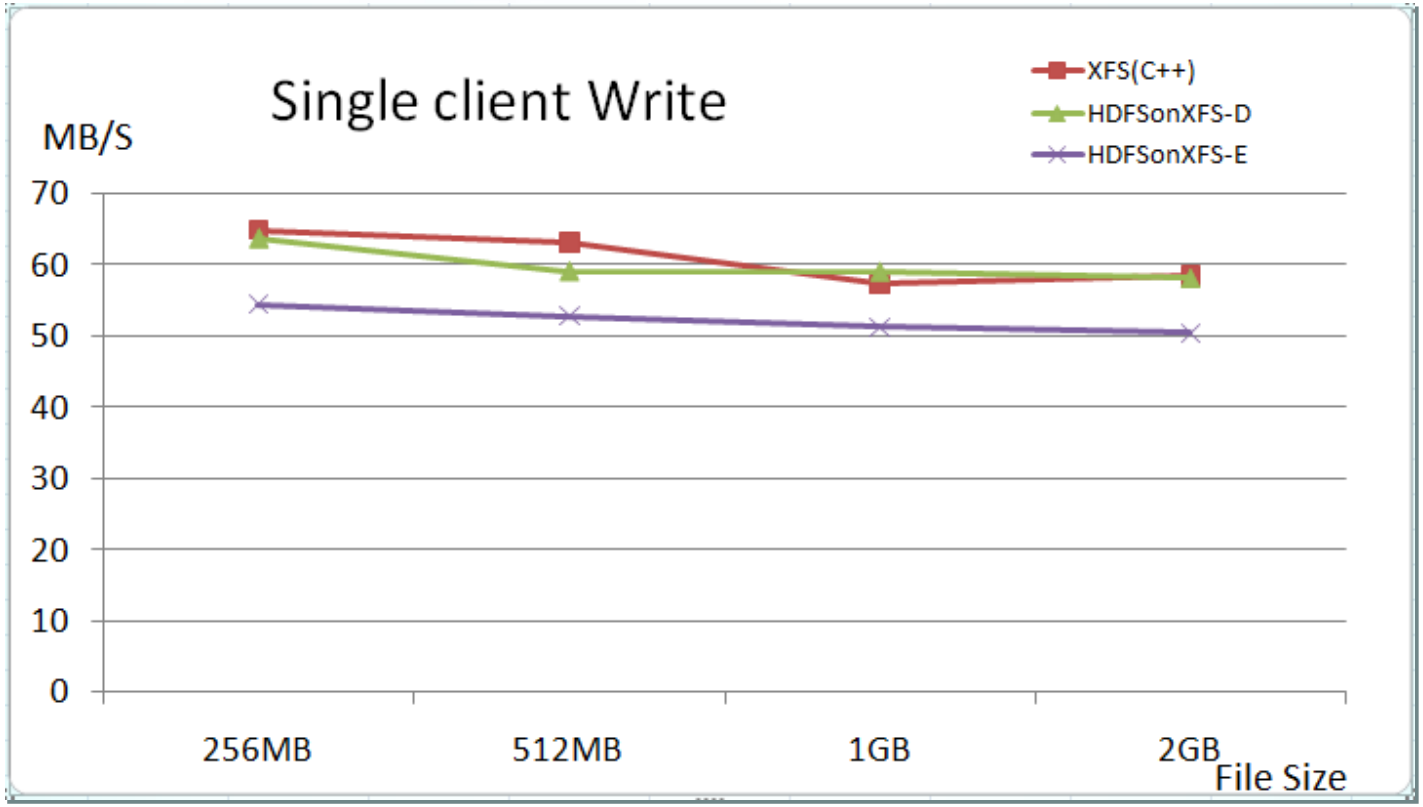
Tencent 腾讯

- Apps without code modication to verify correctness
  - Verified by Hadoop examples
- Performance by Micro-benchmarks
  - Single-client Read/Write
  - Multi-client Read/Write
  - Namespace operations
- Real world running jobs

Single client Read

– HDFSonXFS-D, 2GB file: 0.4% overhead of XFS

Single client Write



– HDFSonXFS-D: 2GB file: 0.6% overhead of XFS, but needs apps modifications

Multi-client Read

Tencent 腾讯

– Each node 5 threads, sequentially read 4MB blocks

– HDFSonXFS-E 9 clients: 2.6% overhead of XFS

Impl Summary

Tencent 腾讯

- Introduce JNA Lib and XFSFileSystem layers to cover interface and semantic gap
- JNA native memory mgmt to reduce read/write cost to about 0.4% and 0.6%
- XFS server side paging-recursive operations for fast, stable, and scalable namespace OPs
- HDFS on XFS is integrated to Hadoop cluster to submit jobs in Tencent

Conclusions

Tencent 腾讯

- Typhoon is our cloud computing system with clearly defined layers
- Typhoon supports running native Hadoop jobs by:
  - Running jobtrack and tasktrackers as jobs
  - Mapping HDFS to XFS
- Questions?