HadoopYarn@toutiao

徐鹏

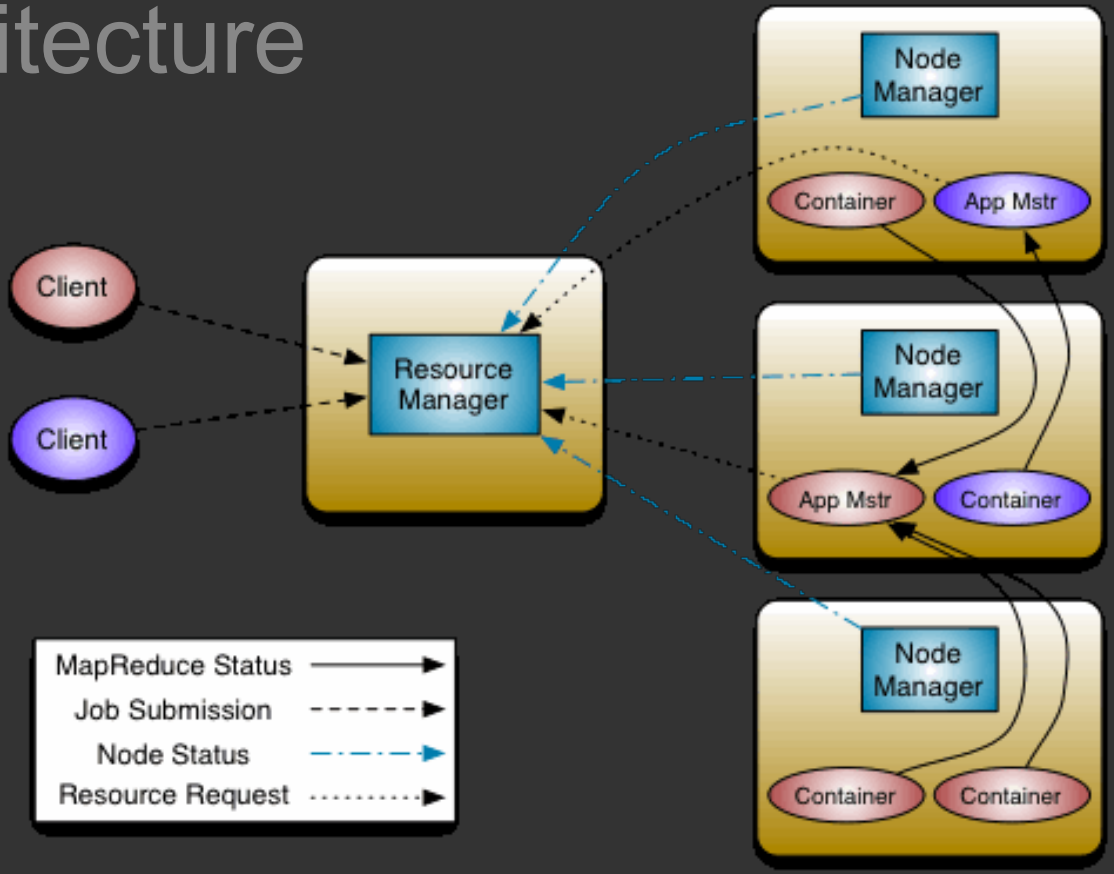
今日头条基础架构工程师

Apache HDFS contributor

Apache Yarn contributor

01 Intro

Architecture



Overview

nodes	-	1w+
vcore	-	40w+
memory	-	2PB
cluster	-	5
region	-	3

Overview

jobs - 10w

mapreduce/spark - 1: 2

long_running - 1000

queues - 70+

labels - 10+

Performance

throughput(container, sls)	-	4w
online	-	1 ~ 3k

SchedulePolicy

FairScheduler + DRF Policy

HumanResource

Dev + Ops + Test - 2

02 Challenge

Challenge

Container throughput

Challenge

Resource fragment brought by DRF schedule policy

Challenge

Resource utilization

Challenge

NodeManager > 5000

ResourceManager internal event avalanche

Challenge

Easy management

multi region + multi cluster + multi label

Challenge

Isolation

03 What we done

Scheduler

MultiThread version of FairScheduler

1 - r/w lock against synchronized
2 - multi thread against single thread
3 - 100x

Scheduler

CPU/Memory fragment

1 - delay scheduling for node resource balance
2 - cpu utilization increase 8%
3 - memory fragment ratio decrease 62.5%

Scheduler

Preemption refactor

Preemption for Yarn!!! Not for mapreduce

Scheduler

LabelScheduling

10+ label for cluster isolation(env/resource)

Scalability

Event avalanche

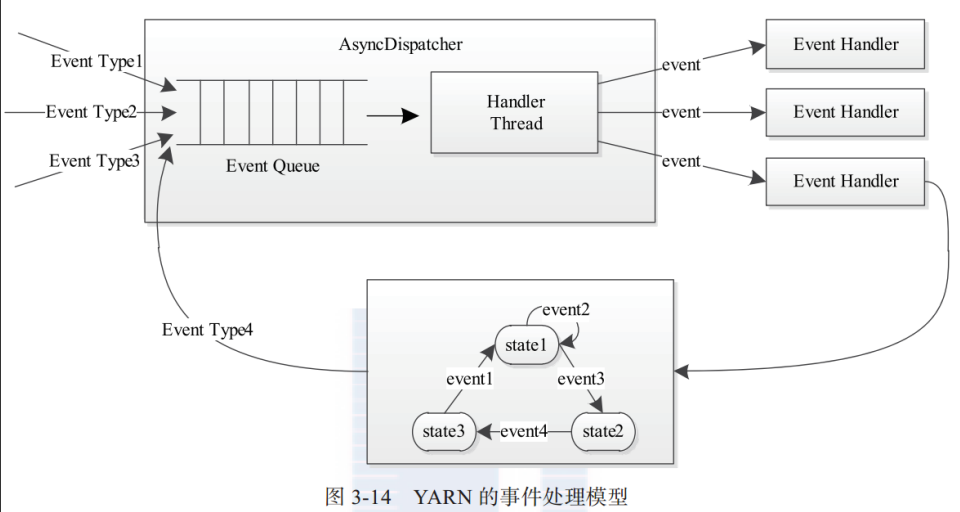


图 3-14　YARN 的事件处理模型

Scalability

Event avalanche

1 yarn safemode
2 heartbeat interval auto-adjustment
3 ignore unnecessary event
4 800w event -> less than 1w
5 30min -> less than 1min

Scalability

NodeManager

1 Single event processor
2 Single thread for uploading logs to s3/hdfs
3 Resource localize not considering io util

Scalability

Yarn Proxy

1 - substitution of yarn federation
2 - easy to migrate for yarn user
3 - route ApplicationProtocol request by yarn queue

Isolation

1 - cgroup with memory support

2 - cgroup cpu-set for important streaming job

Management

ClusterManager

- centralized node resource & label manager

Management

QueueManager

- manage queue by team owner

Management

DtopViewer

- actual resource usage per application
- realtime : kafka -> sparkstreaming -> opentsdb -> portal
- offline analyze : kafka -> sparkstreaming -> hdfs -> hive

Management

HistoryAnalyzer - offline

- mapreduce history analyze
- spark history analyze

Management

History Analyzer - offline

- history files -> hive
- mapreduce & spark
- daily report

Management

History Analyzer - realtime

- log -> kafka -> es
- realtime analyze

xp

Xicheng District, Beijing

Scan the QR code to add me on WeChat

QA

Thanks