



QCon 全球软件开发大会
INTERNATIONAL SOFTWARE
DEVELOPMENT CONFERENCE

BEIJING 2017

FreeWheel大数据处理实践

SPEAKER / 姜冰 bjiang@freewheel.tv

孙炎 ysun@freewheel.tv



背景

大数据驱动的FreeWheel



FreeWheel简介



FreeWheel视频广告大数据的特点

数据规模

- 十亿记录/天
- 大于3T日志/天
- 保留18个月数据
- 跨DC

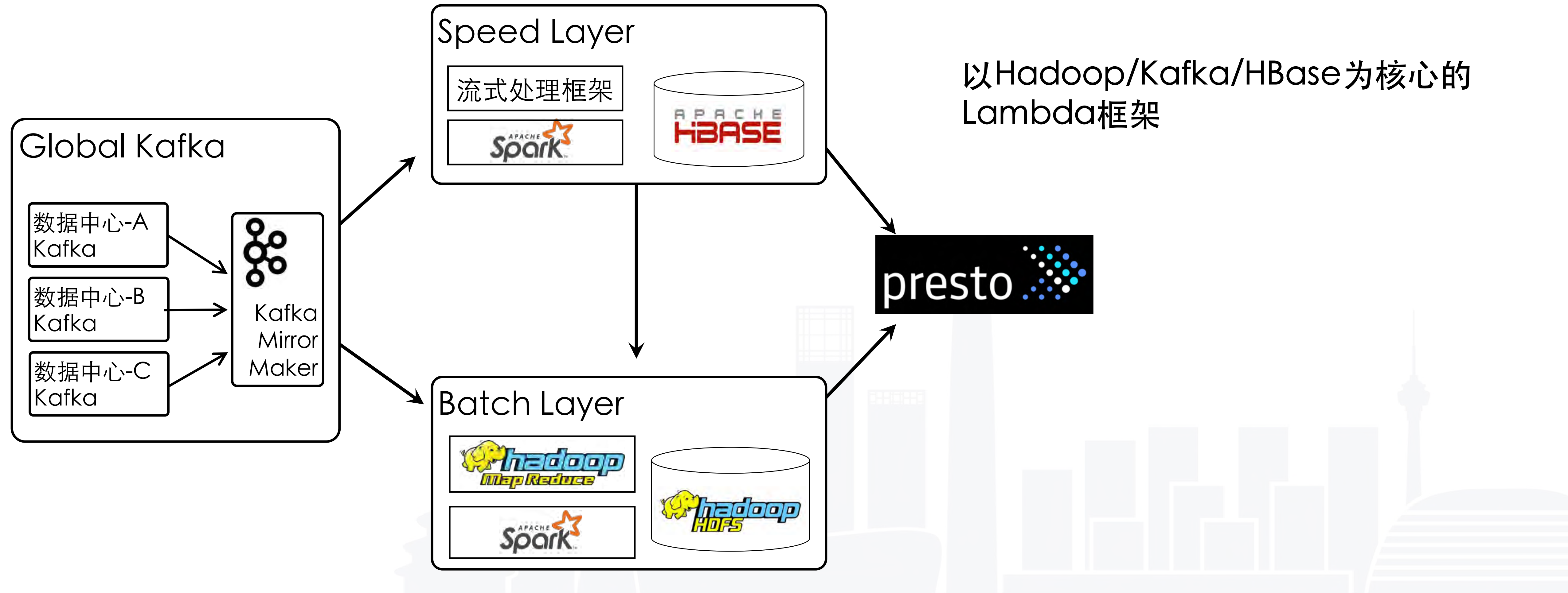
数据类型

- 会话性数据
- 嵌套结构日志数据
- 结构化属性数据

数据应用类型

- 报表统计
- 业务流量追踪
- 实时分析查询
- 预测反馈

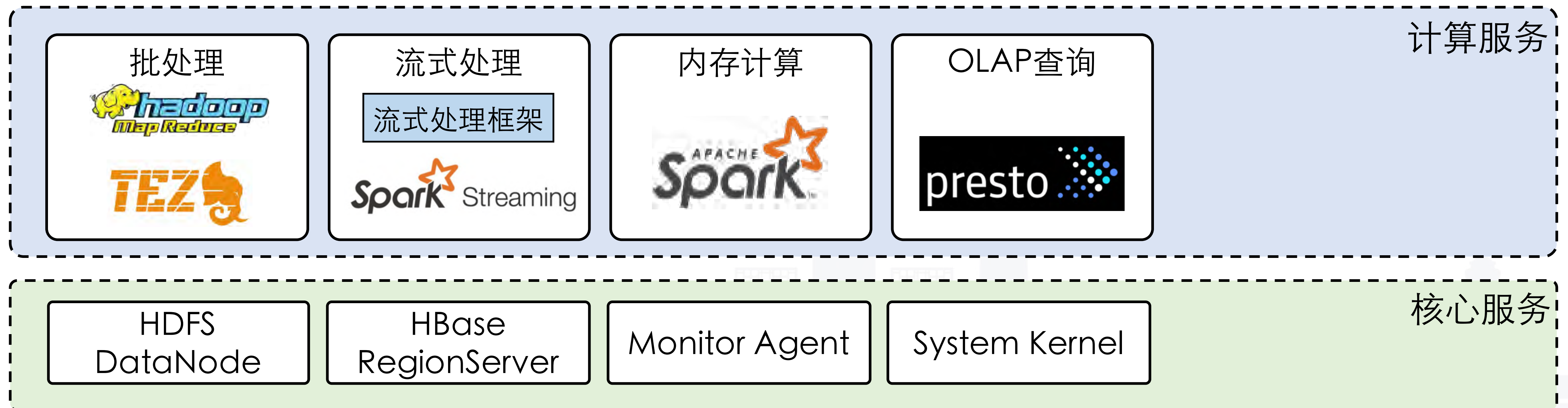
大数据处理平台架构



分布式资源管理平台

资源利用率

- 核心服务与计算服务
- 多种类型作业、服务共享资源



资源管理

多种类型作业共享资源

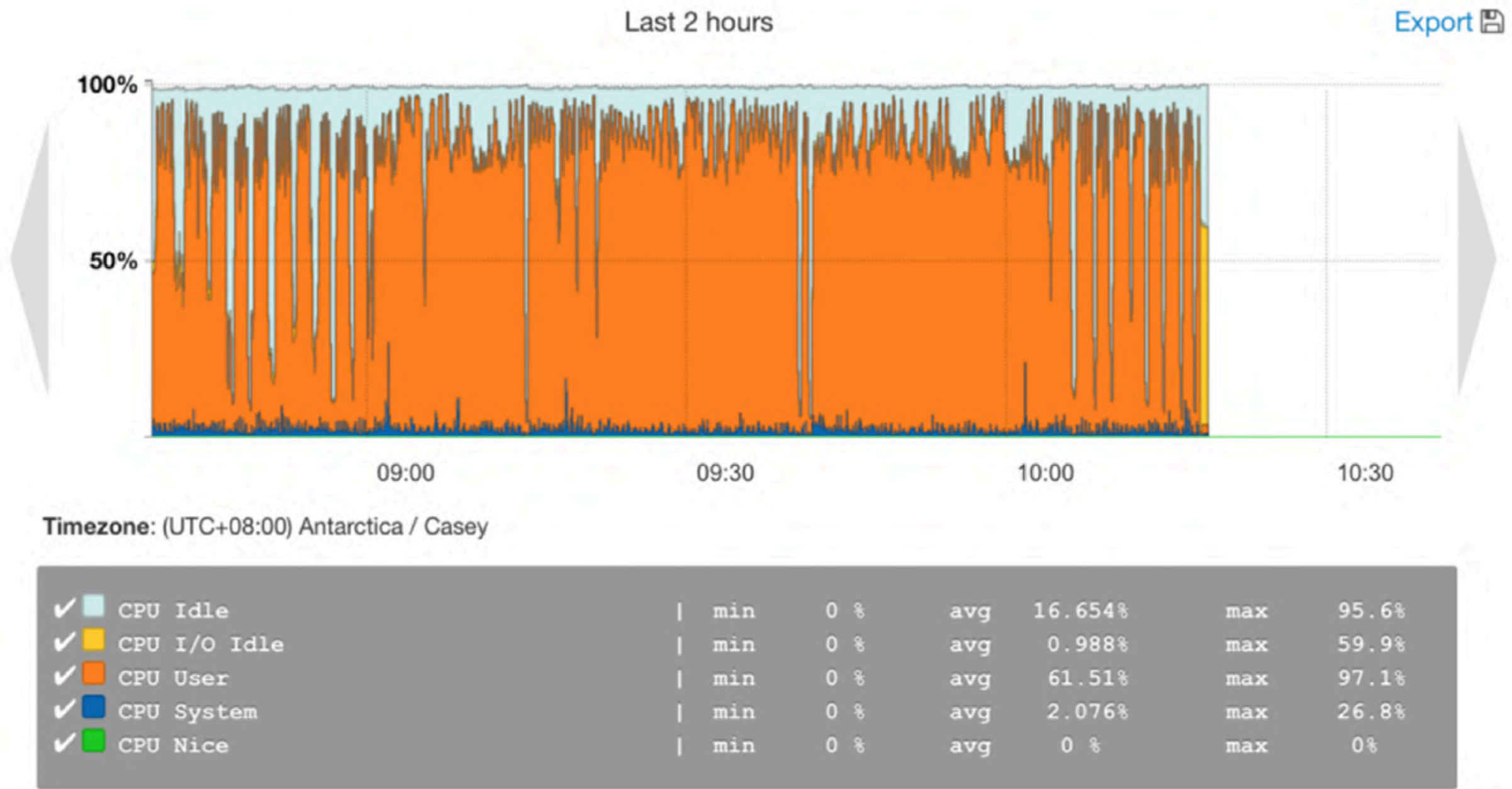


问题 - 计算服务与核心服务竞争

- 计算服务与核心服务相互影响
- 没有做CPU限制的计算任务可以导致机器宕机!

%Cpu(s): 90.4 us, 2.5 sy, 0.0 ni, 5.2 id, 1.5 wa, 0.0 hi, 0.4 si, 0.0 st											
KiB Mem : 13173728+total, 525268 free, 12420185+used, 7010152 buff/cache											
KiB Swap: 8388604 total, 5035196 free, 3353408 used. 7024724 avail Mem											
PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
27292	df	20	0	43.399g	0.026t	6424	S	2321	21.1	60644:15	presto-server
7631	data	20	0	3485280	1.390g	11172	S	105.2	1.1	5:30.43	java
7623	data	20	0	3569812	1.417g	11168	S	98.4	1.1	5:33.23	java
7663	data	20	0	3490792	1.399g	11164	S	96.1	1.1	5:28.66	java
7669	data	20	0	3559296	1.421g	11172	S	90.6	1.1	5:35.08	java
12606	data	20	0	3498328	1.409g	11496	S	89.3	1.1	3:44.39	java
25901	yarn	20	0	3269628	749056	6688	S	30.9	0.6	1750:41	java
22973	fr	20	0	17.718g	0.010t	8528	S	29.0	8.1	31:43.69	java
19974	hdfs	20	0	3310740	1.396g	4364	S	22.5	1.1	2199:42	java
8038	df	20	0	5516984	1.858g	3284	S	11.1	1.5	2461:58	java

CPU Usage



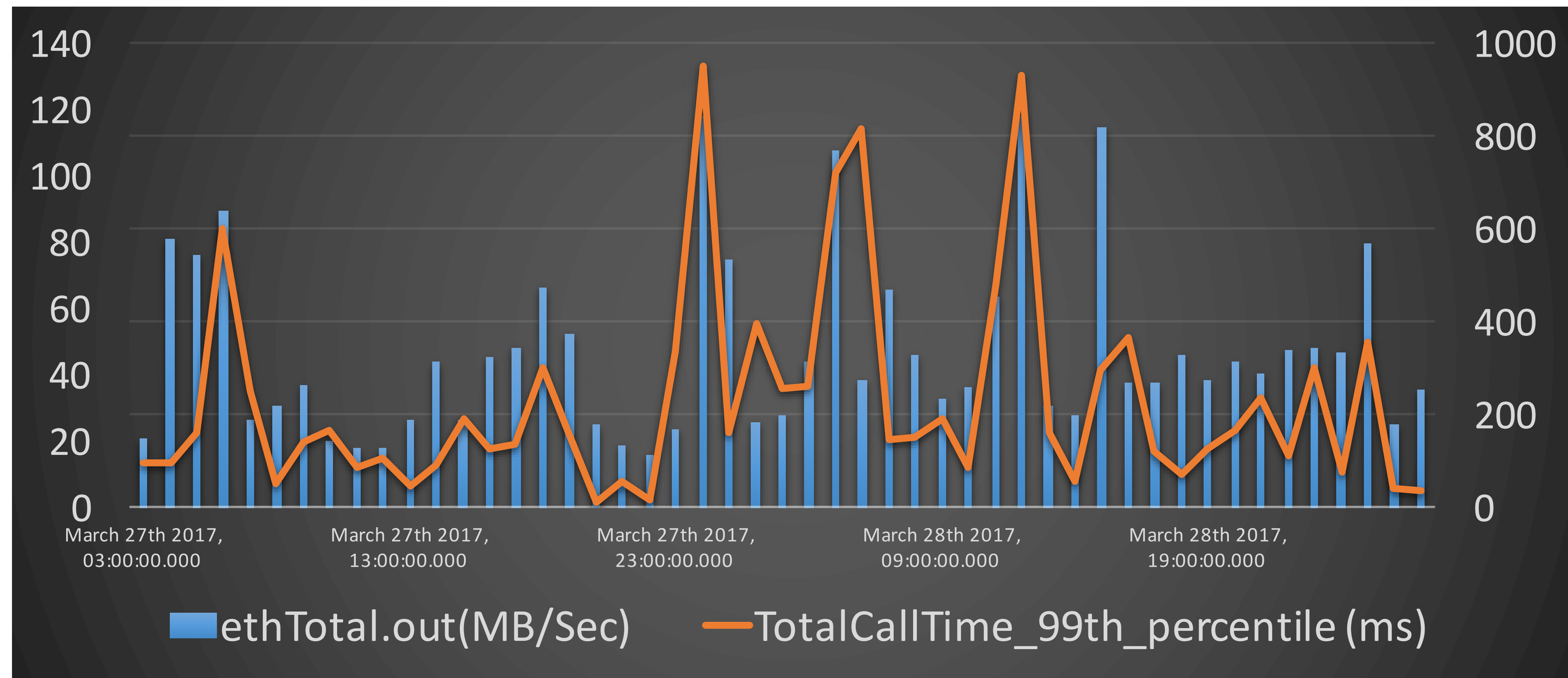
问题 - 网络IO竞争

- 计算服务占用大量IO

ATOP - hadoop179														2017/03/28 12:55:01		-----		60s elapsed	
PRC	sys	20.61s	user	2m30s	#proc	457	#trun	2	#tslpi	1639	#tslpu	0	#zombie	0	clones	973	#exit	0	
CPU	sys	26%	user	249%	irq	6%	idle	2914%	wait	5%	steal	0%	guest	0%	curf	2.60GHz	curscal	100%	
CPL	avg1	4.37	avg5	10.02	avg15	12.39			csw	1730201			intr	2107180			numcpu	32	
MEM	tot	125.6G	free	5.3G	cache	61.5G	buff	1.6G	slab	3.7G	shmem	27.1M	vmbal	0.0M	hptot	0.0M	hpuse	0.0M	
SWP	tot	8.0G	free	6.7G											vmcom	58.4G	vmlim	70.8G	
PAG	scan	0	steal	0	stall	0									swin	43	swout	0	
DSK		sdb	busy	4%	read	710	write	775	KiB/r	365	KiB/w	431	MBr/s	4.22	MBw/s	5.45	avio	1.74 ms	
DSK		sdd	busy	3%	read	749	write	548	KiB/r	329	KiB/w	412	MBr/s	4.02	MBw/s	3.68	avio	1.36 ms	
DSK		sde	busy	3%	read	598	write	728	KiB/r	408	KiB/w	438	MBr/s	3.98	MBw/s	5.20	avio	1.22 ms	
NET	transport		tcpi	1877270	tcpo	5269557	udpi	60	udpo	60	tcpao	850	tcppo	384	tcprs	14	udpie	0	
NET	network		ipi	1877367	ipo	779383	ipfrw	19	deliv	1877e3					icmpi	20	icmpo	0	
NET	eth0	98%	pcki	2642830	pcko	5248755	si	286 Mbps	so	985 Mbps	erri	0	erro	0	drpi	0	drpo	0	
NET	lo	----	pcki	33041	pcko	33041	si	58 Mbps	so	58 Mbps	erri	0	erro	0	drpi	0	drpo	0	
NET	docker0	----	pcki	11	pcko	12	si	1 Kbps	so	0 Kbps	erri	0	erro	0	drpi	0	drpo	0	
PID	TID	TCPRCV	TCPRASZ	TCPSND	TCPSASZ	UDPRCV	UDPRASZ	UDPSND	UDPSASZ	BANDWI	BANDWO	NET	CMD	1/58					
29420	-	697046	72	195171	20566	0	0	0	0	6750 Kbps	535 Mbps	83%	presto-server						
18039	-	156475	2658	92394	3145	0	0	0	0	55 Mbps	38 Mbps	14%	java						
11752	-	10460	70	1987	31560	0	0	0	0	97 Kbps	8361 Kbps	1%	java						

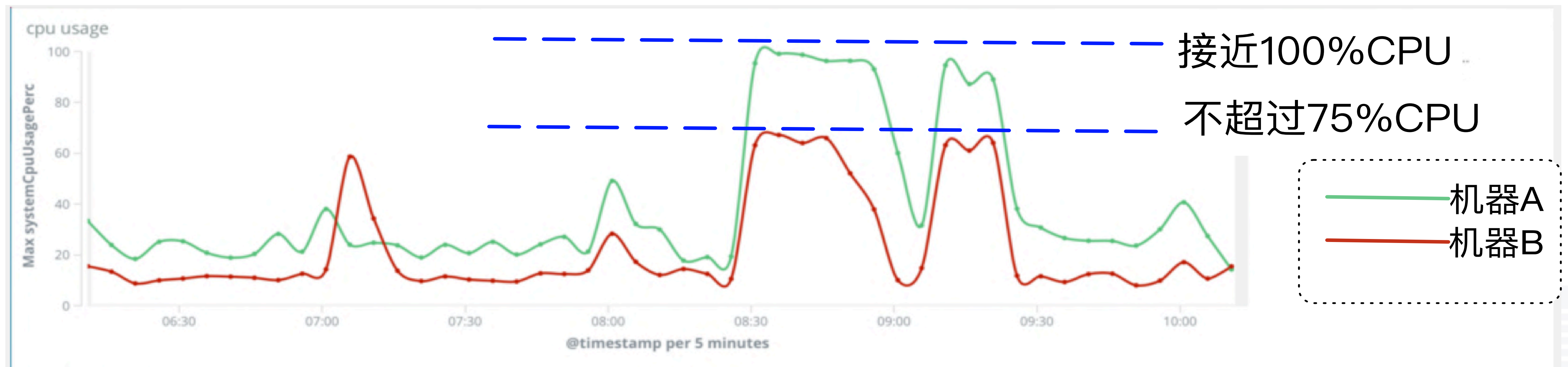
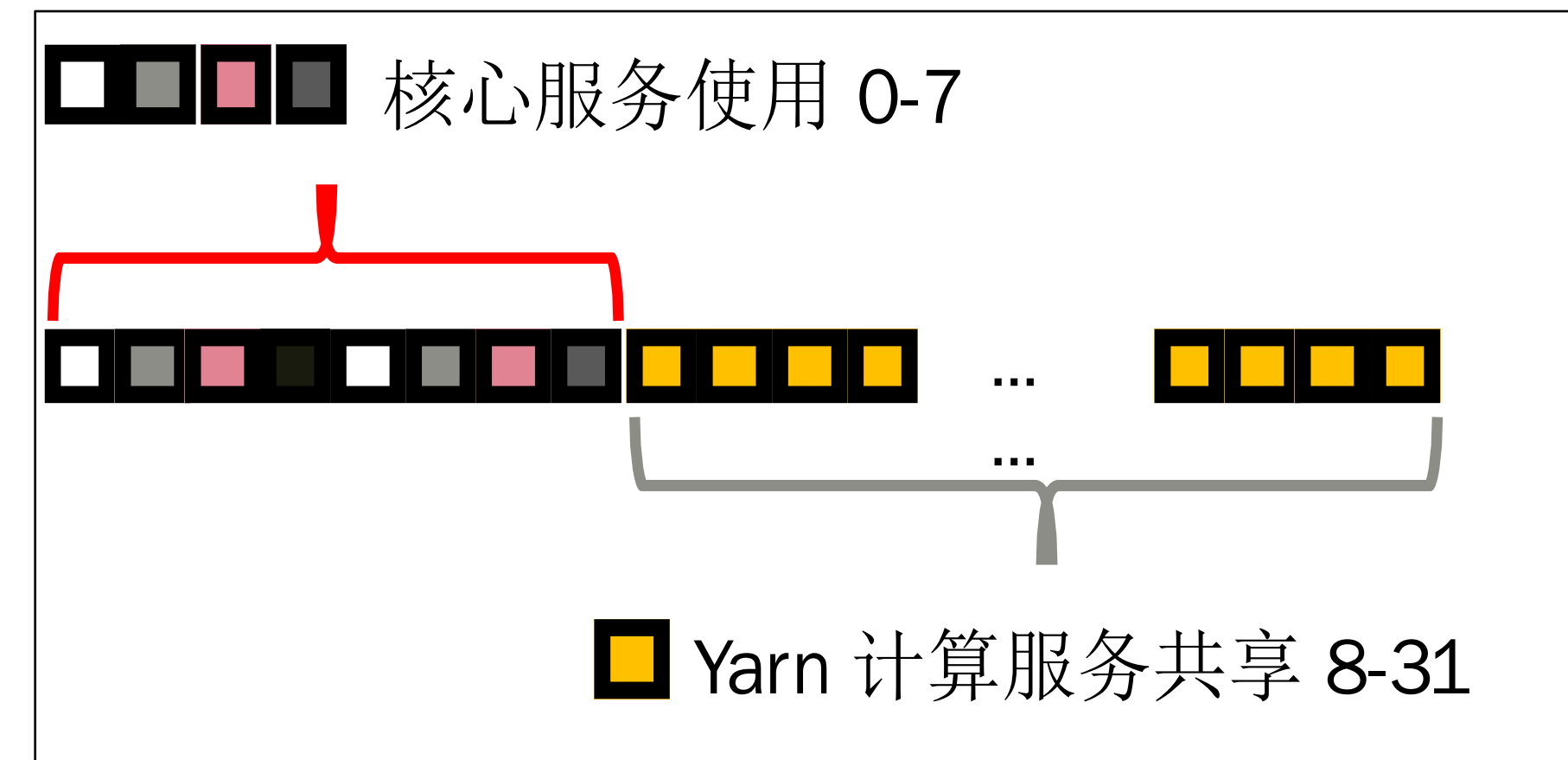
问题 - 网络IO竞争

- HBase延迟受到Network IO的影响



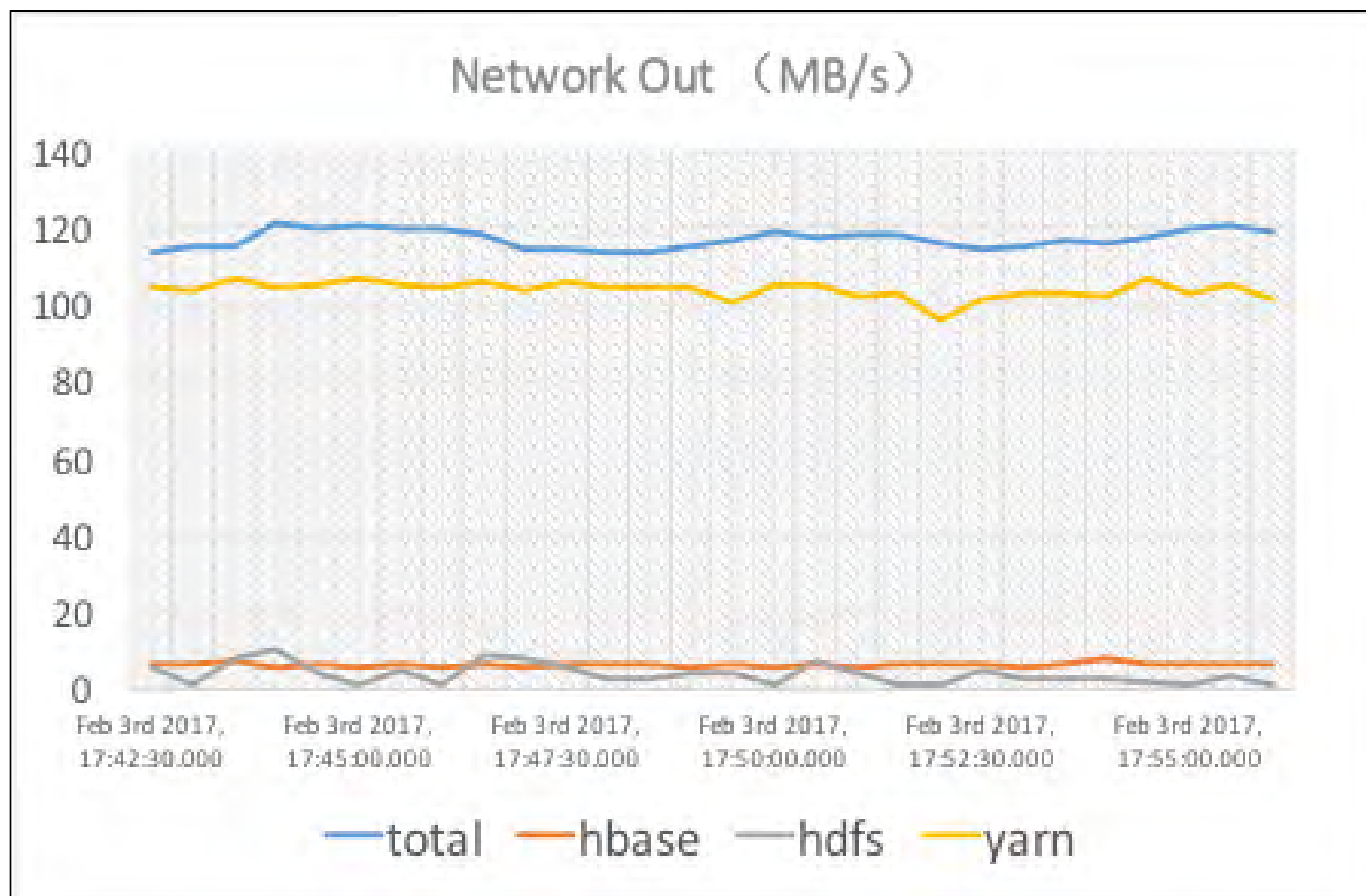
Linux CGroup 隔离 — CPU资源

- 计算服务和核心服务cpuset隔离
- 每一层内设置cpu.shares竞争共享
- 粗粒度cpuset隔离
- 细粒度cpu.shares竞争



Linux CGroup 隔离 — Network IO资源

- 对于HDFS, HBase, YARN计算服务, 使用tc(htb)+net_cls cgroup控制
- 粗粒度网络IO隔离
- 设置安全边界!



#设置tc htb类别和流量控制规则

```
$ tc class add dev eth0 parent 1: classid 1:1 htb rate 1000mbit
$ tc class add dev eth0 parent 1:1 classid 1:10 htb rate 480mbit ceil 1000mbit
$ tc class add dev eth0 parent 1:1 classid 1:11 htb rate 260mbit ceil 1000mbit
$ tc class add dev eth0 parent 1:1 classid 1:12 htb rate 260mbit ceil 1000mbit
```

#创建cgroup net_cls分组 (yarn-service, hbase-service, hdfs-service)

```
$ cgcreate -g net_cls:/yarn-service
$ cgset -r net_cls.classid=0x10010 yarn-service
$ cgcreate -g net_cls:/hbase-service
$ cgset -r net_cls.classid=0x10011 hbase-service
$ cgcreate -g net_cls:/hdfs-service
$ cgset -r net_cls.classid=0x10012 hdfs-service
```

#设置tc filter按照cgroup进行过滤

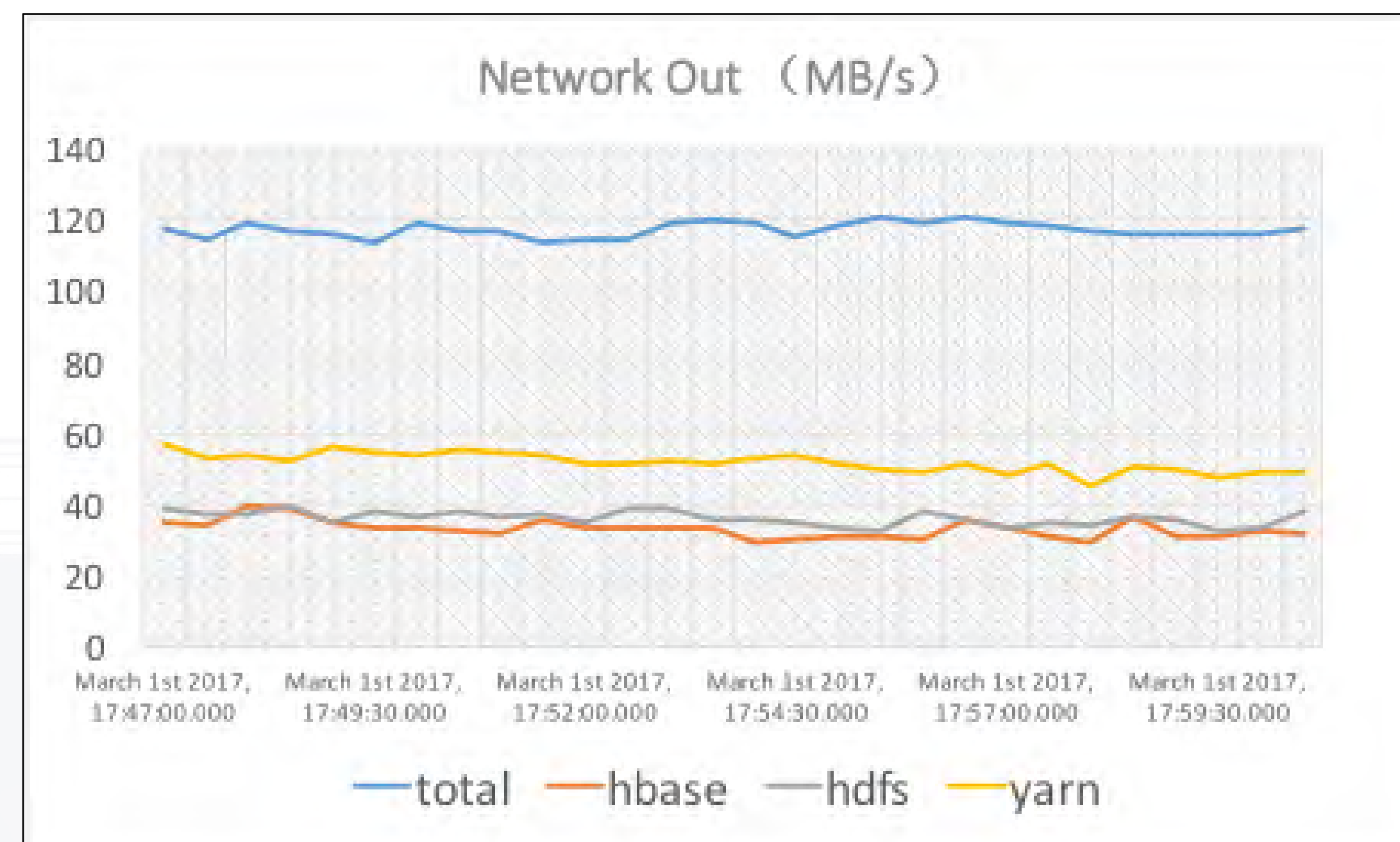
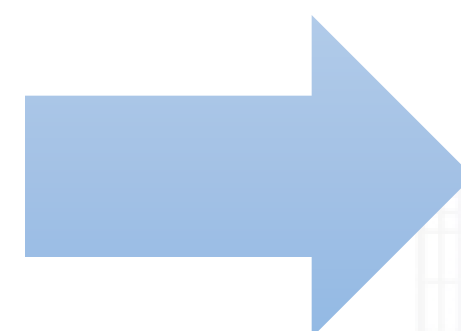
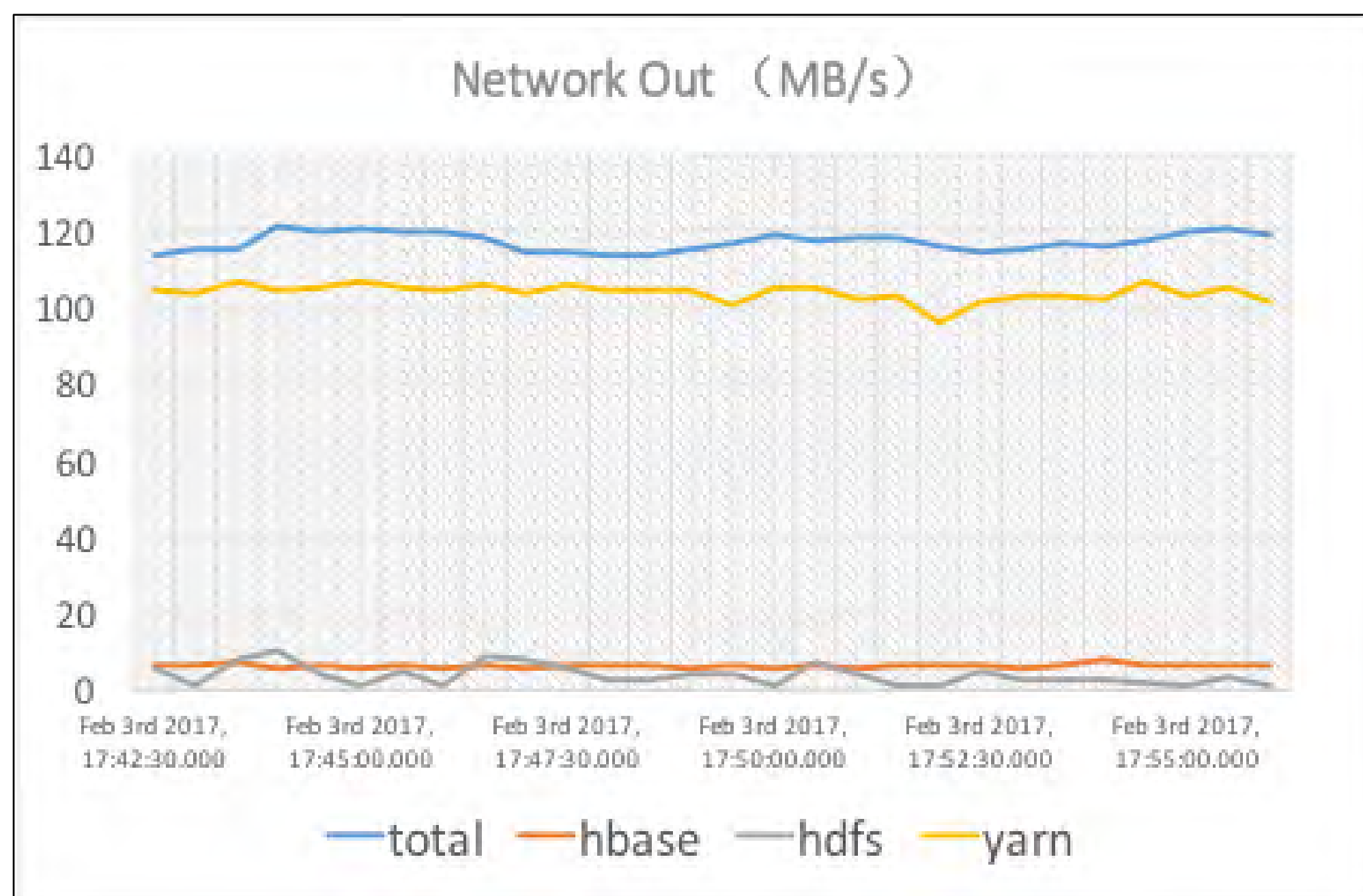
```
$ tc filter add dev eth1 parent 1: protocol ip handle 1: cgroup
```

#启动服务进程并加入对应的cgroup

```
$ cgexec -g net_cls:${group_name} start-service-command
```


Linux CGroup 隔离 — Network IO资源

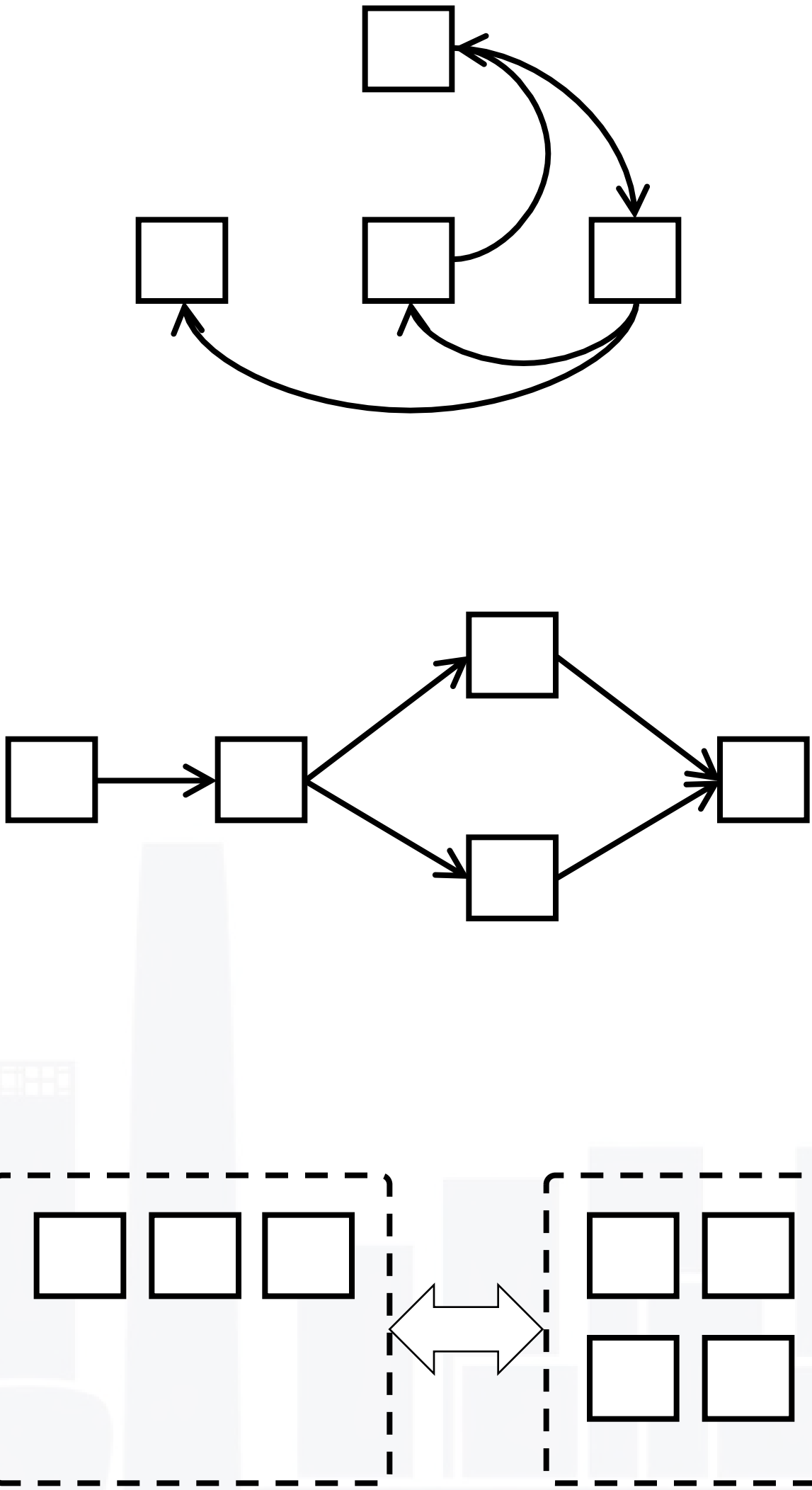
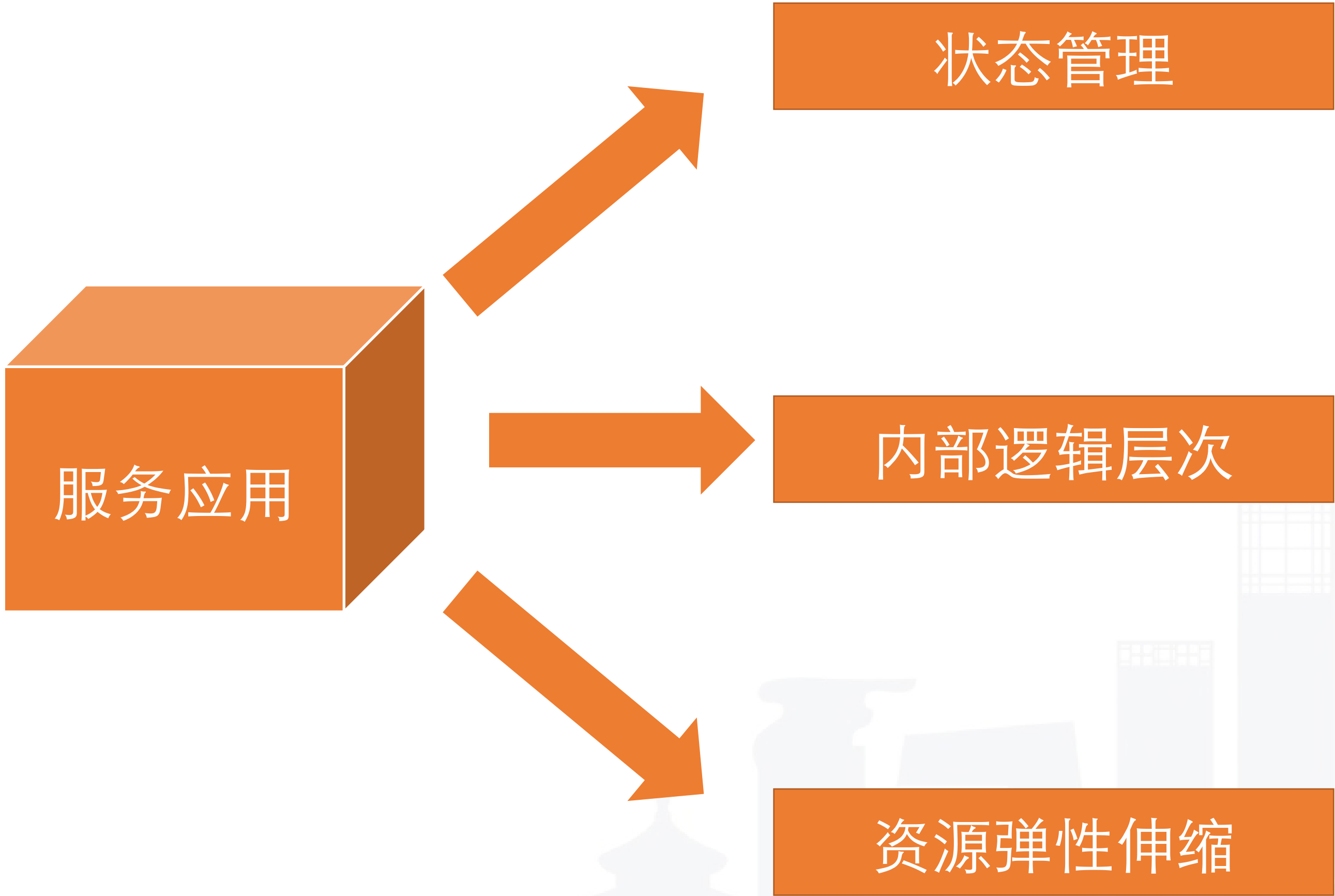
使用tc(htb)+net_cls cgroup调整前后



应用编排和CD

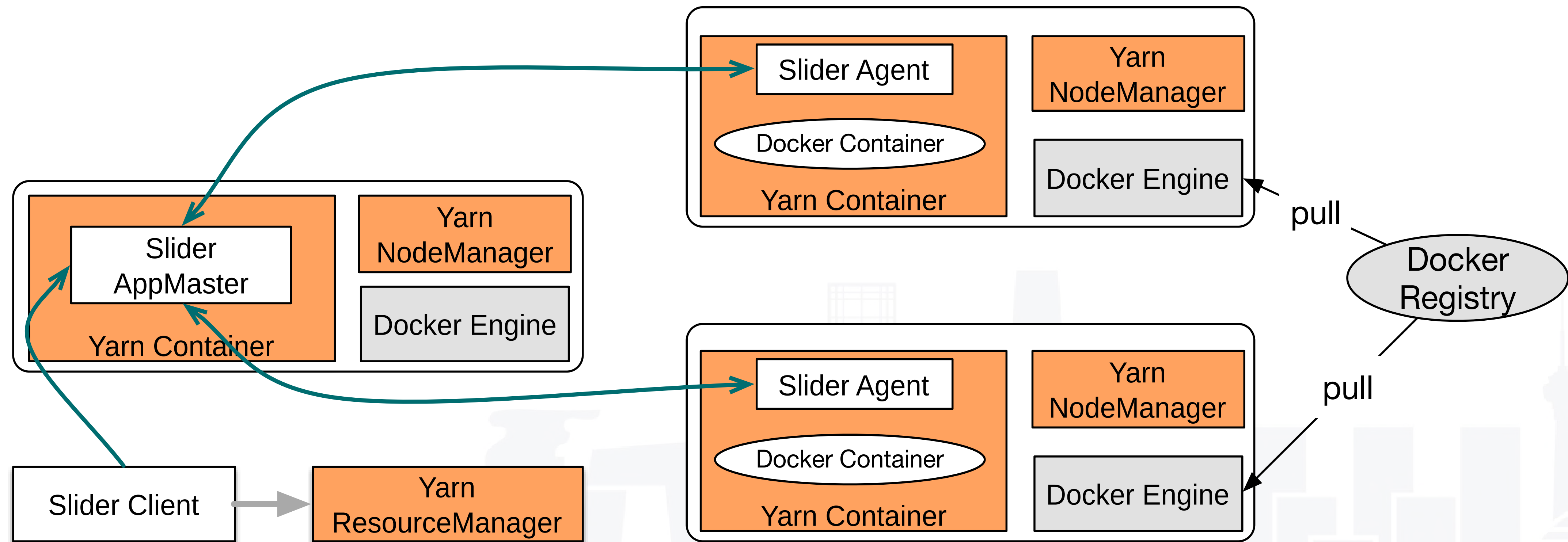
基于Slider+Docker实现弹性部署策略

服务类型应用



Slider + Docker

- Slider管理“服务”的状态和逻辑
- Docker负责“一体化”组装



Slider + Docker 弹性伸缩

改造：

- 扩展Slider动态伸缩功能
- 保证Docker Container和Slider Agent状态一致
<https://issues.apache.org/jira/browse/SLIDER-1213>
- 动态更新升级Docker Image功能

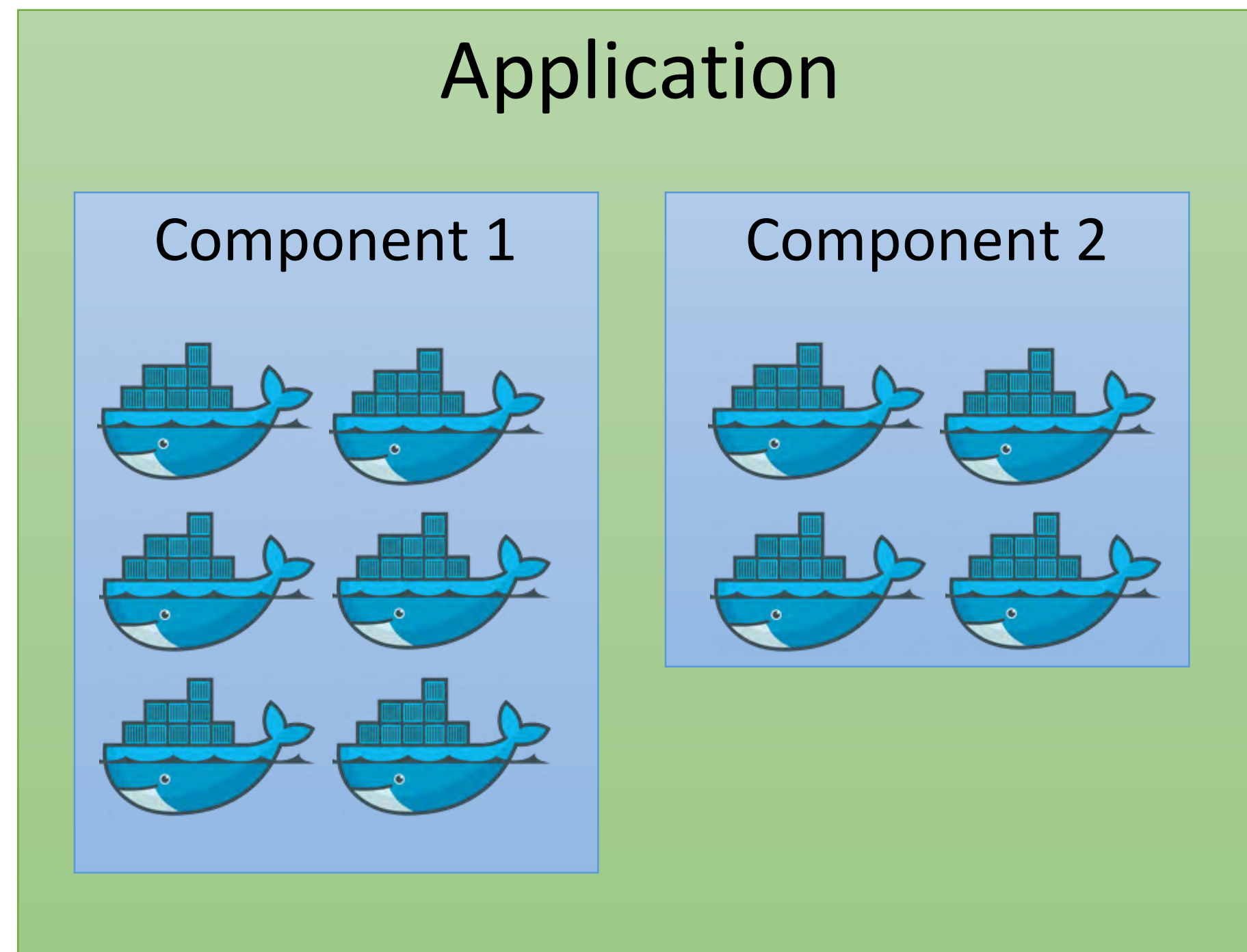
```
#在hadoop101上增加一个matcher实例：
```

```
$ slider flex dip -component matcher +1 -host hadoop101
```

```
#关闭在hadoop105上的flusher实例：
```

```
$ slider flex dip -component flusher -1 -host hadoop105
```

Slider + Docker — CD

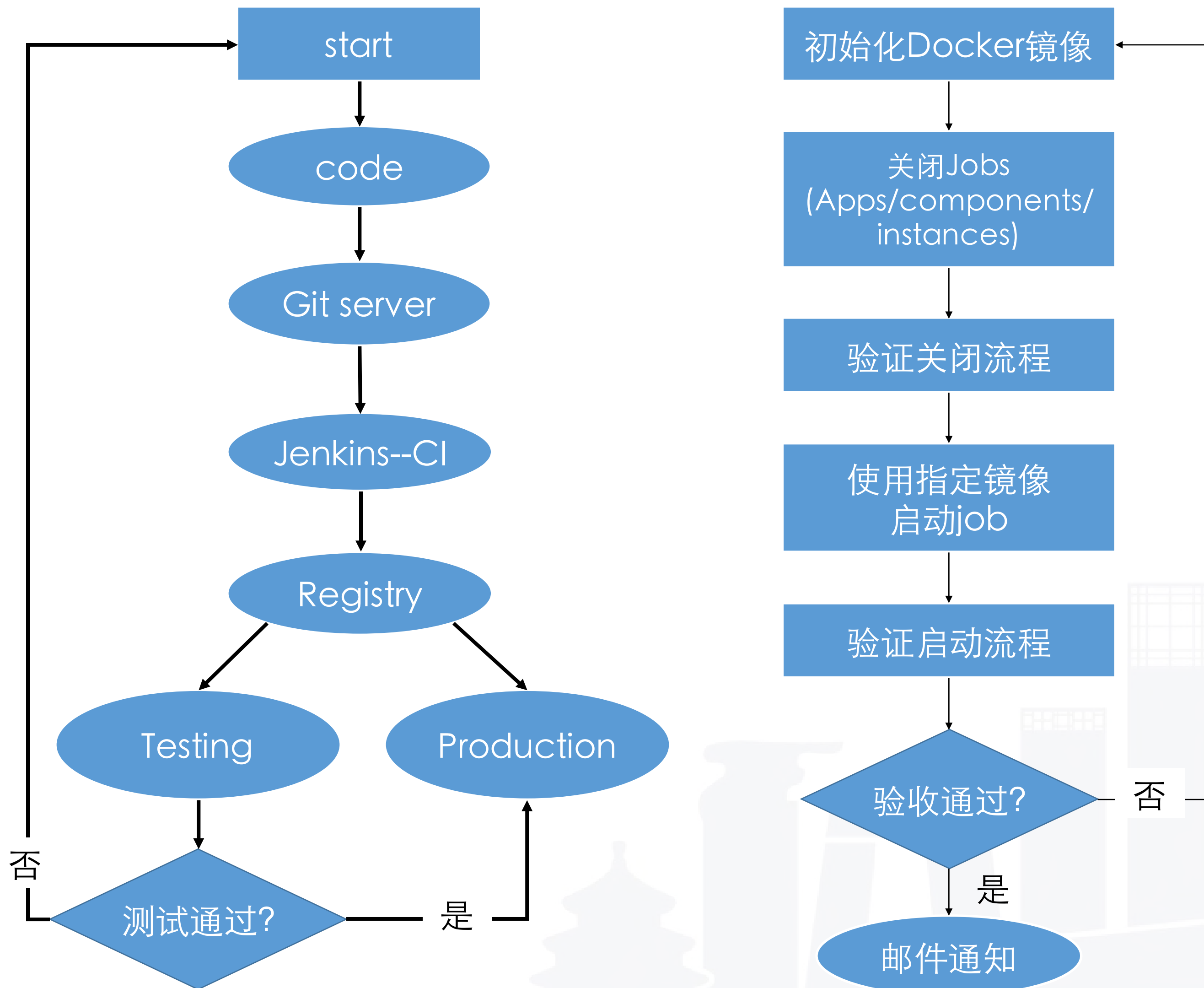


Docker提供一致的部署环境和测试环境。
基于Slider灵活自动化部署

- Application
- Component
- Instance

整体升级和灰度发布

Slider + Docker — CD



- 选择合适的发布策略
- 基于模块和instances设计验证流程
- 改造Slider，动态升级docker镜像
- slider flex灰度发布

AWS部署实践

一套代码，多套环境



公有云

公有云的优点：

- 资源按需实时供应
- 资源配置的灵活性
- 资源隔离-多租户

混合云部署是当下的趋势

- 76%公司采用混合云(Gartner)
- 82%公司采用多云的策略(RightScale)
- 50%公司因公有云开销的原因重新回归私有云(451 Research)

迁移公有云考虑的因素

应用可移植性

公有云开销

功能

数据容量

Gartner, Market Trends: Cloud Adoption Trends Favor Public Cloud With a Hybrid Twist 2015

RightScale 2016 State of the Cloud Report

451 Research: AWS Lambda: new and exciting, old and rehashed, more vendor lock-in (or all the above)?, November, 2016

AWS方案选型

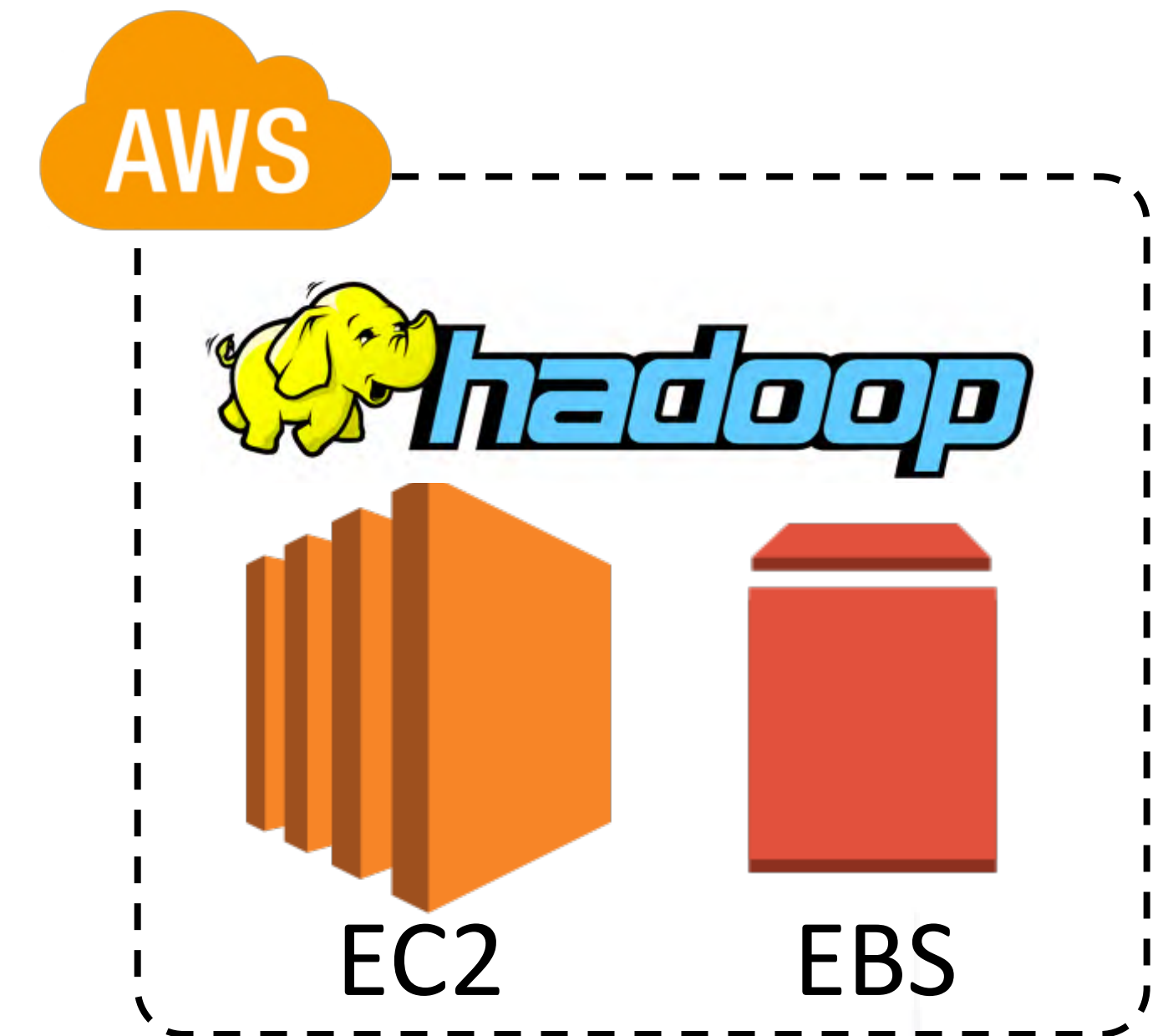
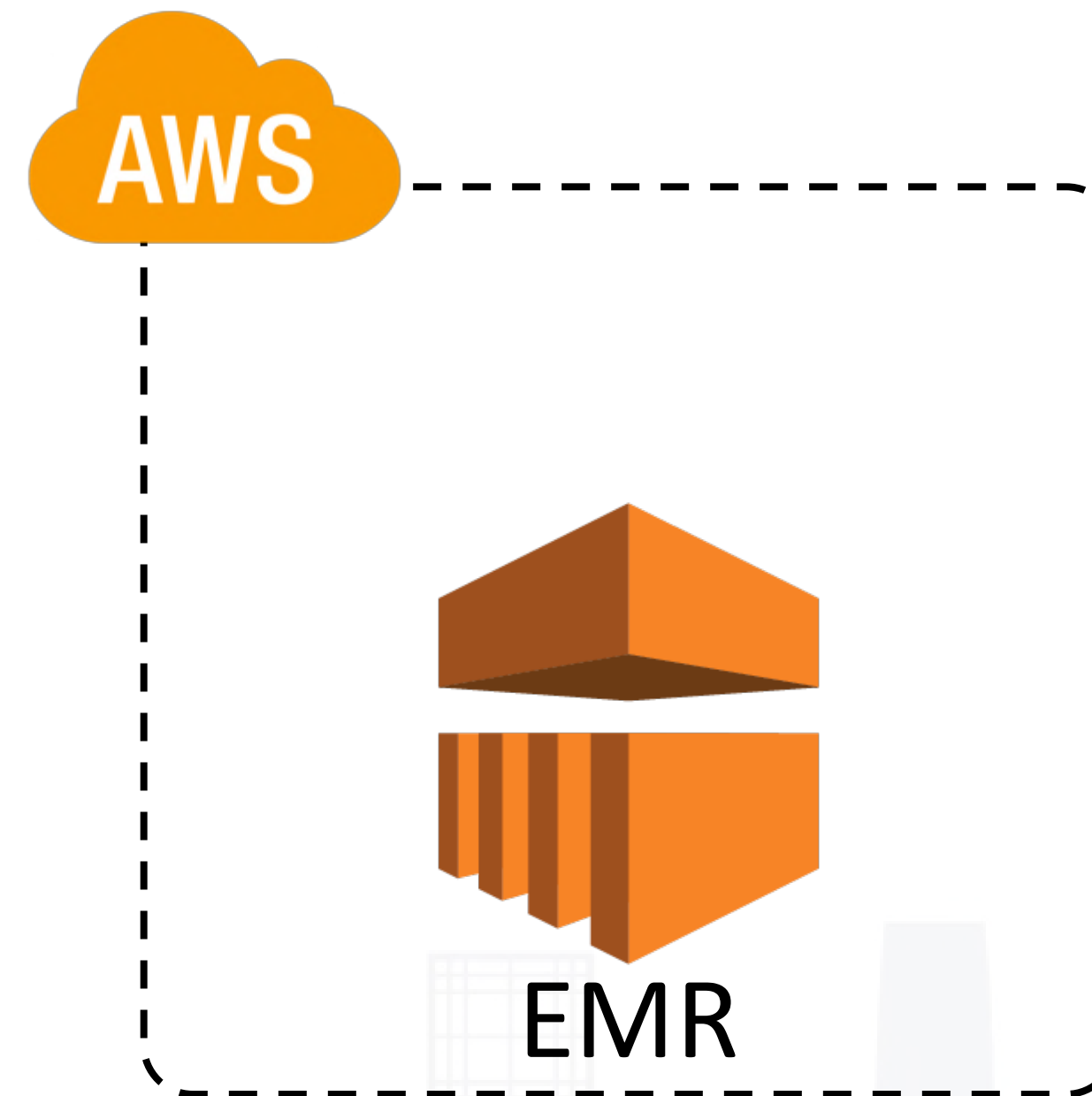
两种方案：

AWS EMR

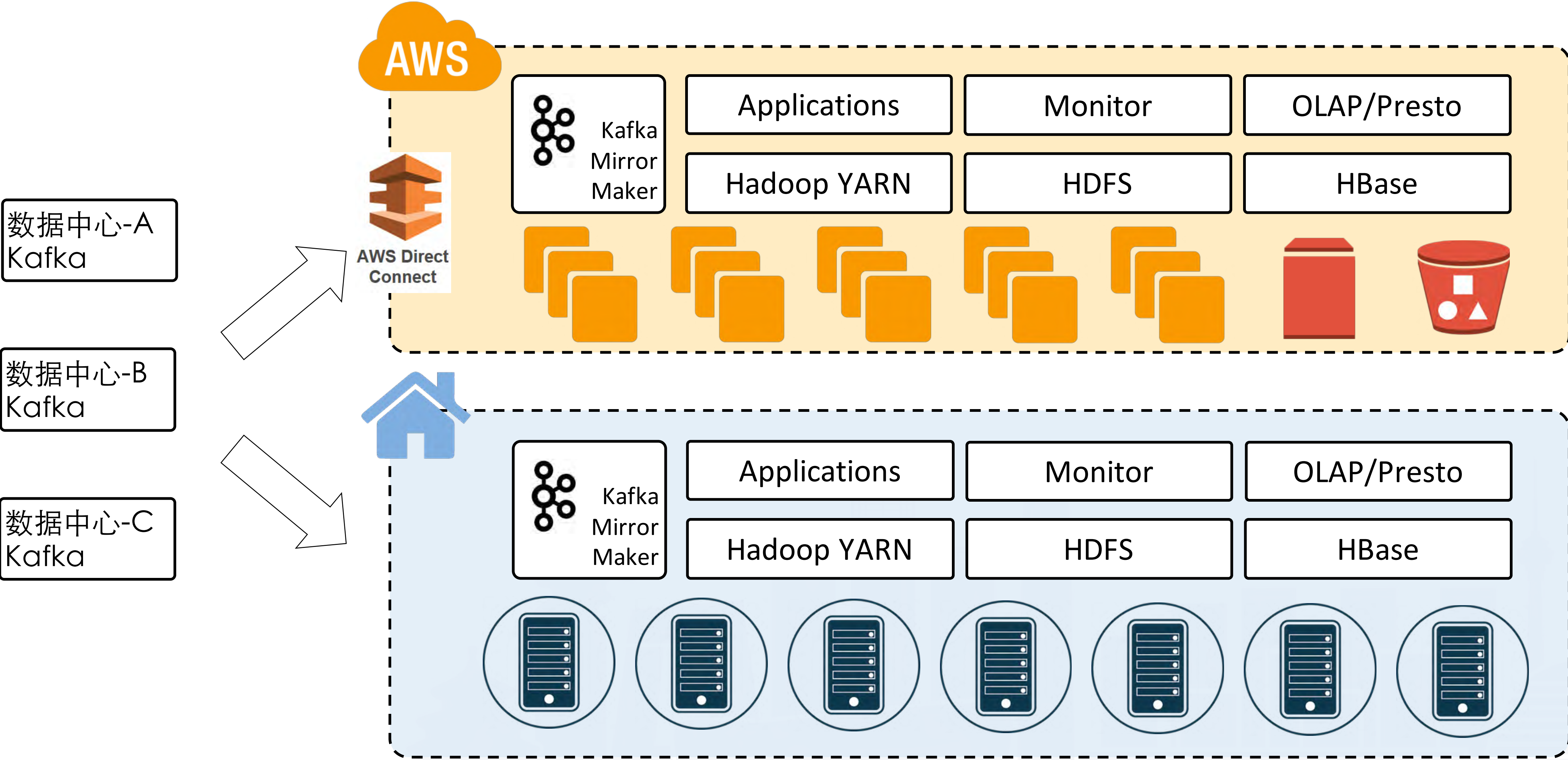
- 单点故障
- HBase on S3 不稳定

AWS EC2 + EBS

- 存储与计算分离
- Hadoop HA
- 实时数据与历史数据分离



On Premise + AWS Cloud



OLAP查询演进

基于Presto的数据查询服务



数据特点

- 超过十亿条日志/天
- 保留至少18个月
- 近200T数据，持续增加
- 将近600列，持续增加
- 列式存储  Parquet



TREASURE DATA

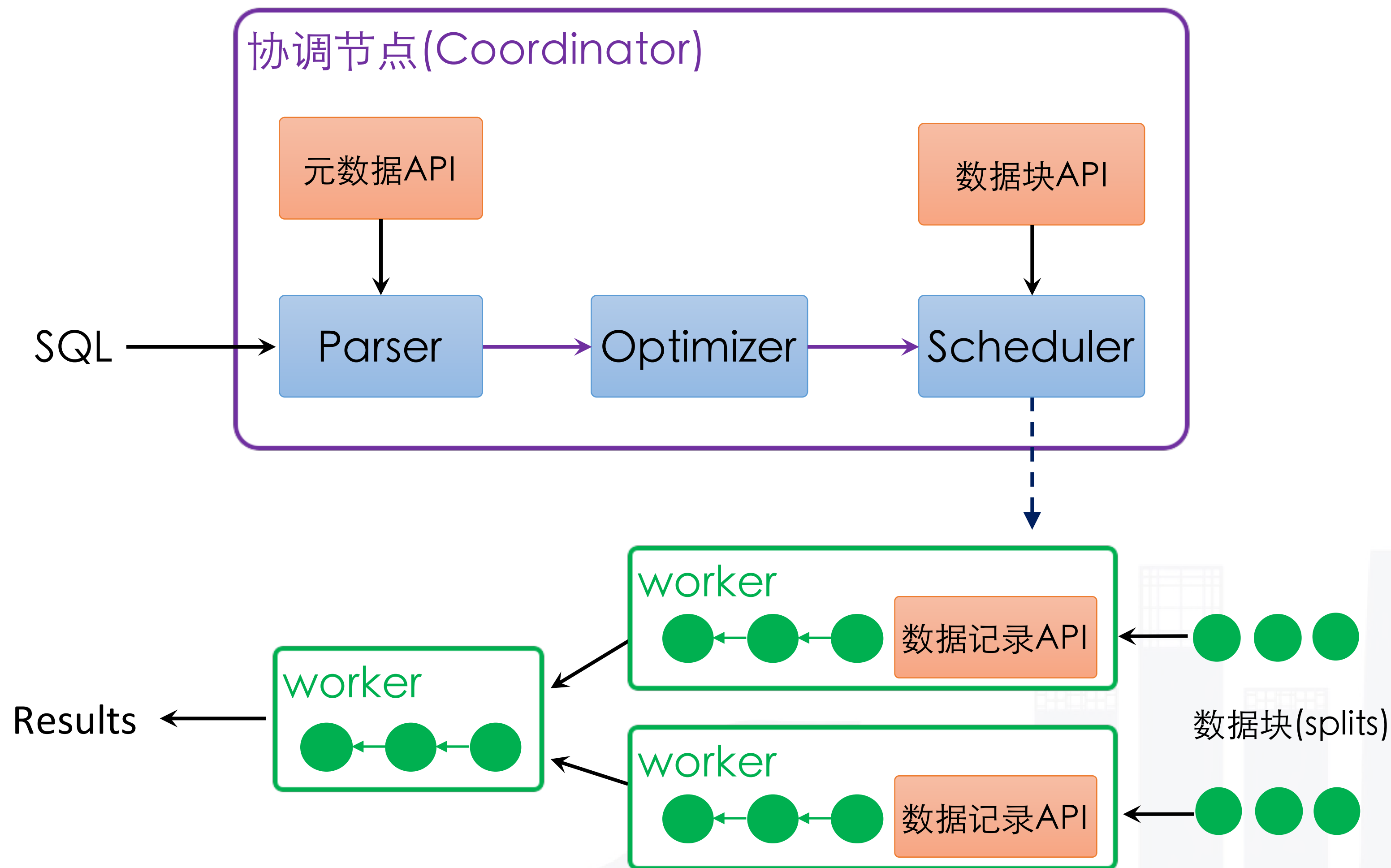
facebook



NETFLIX

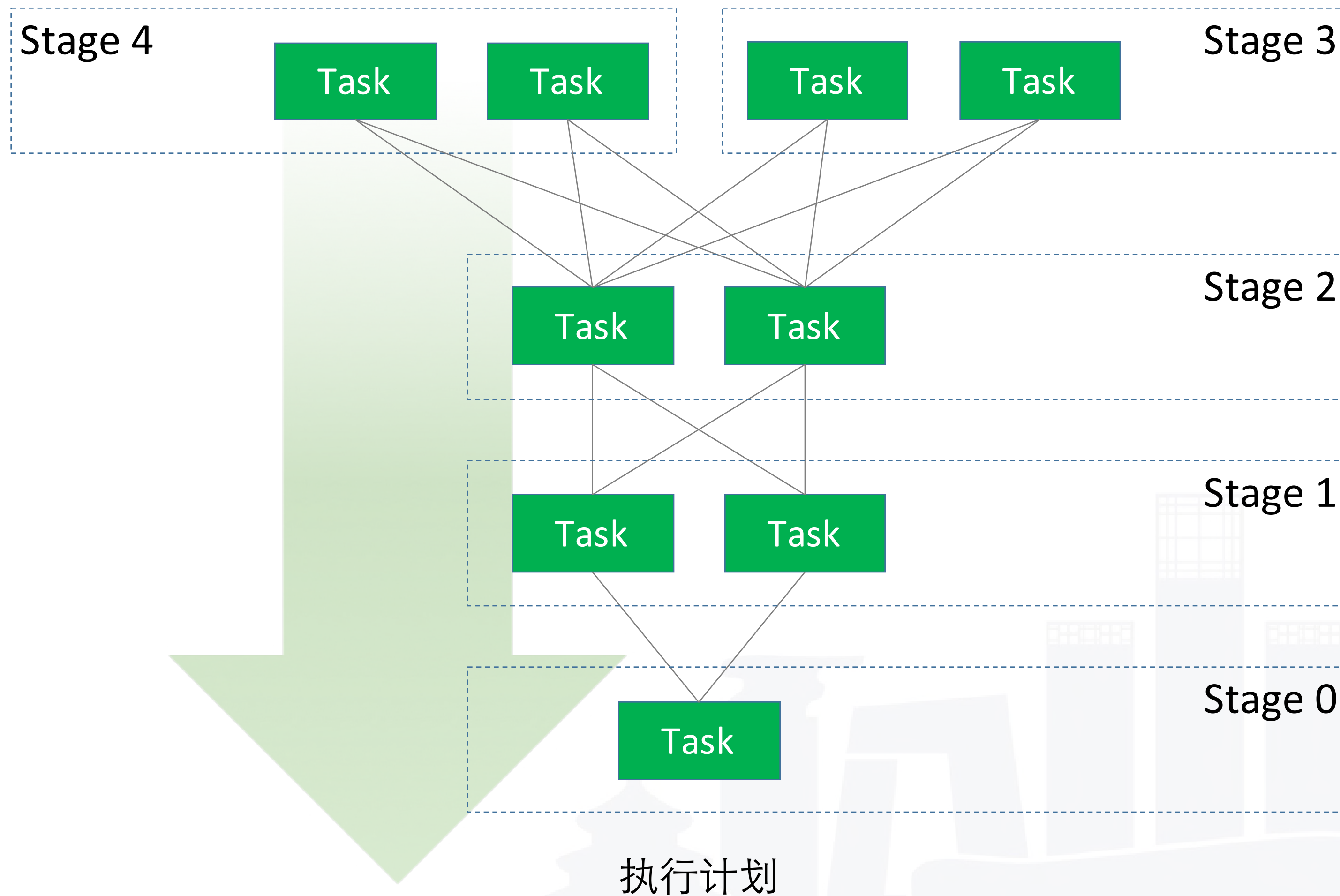
TERADATA

Presto架构及优势



- 效率高 - 交互式查询
- 可扩展数据源
- 多数据源join
- 易于伸缩
- 支持ANSI SQL

Presto架构及优势



- 效率高 - 交互式查询
- 可扩展数据源
- 多数据源join
- 易于伸缩
- 支持ANSI SQL

Presto使用

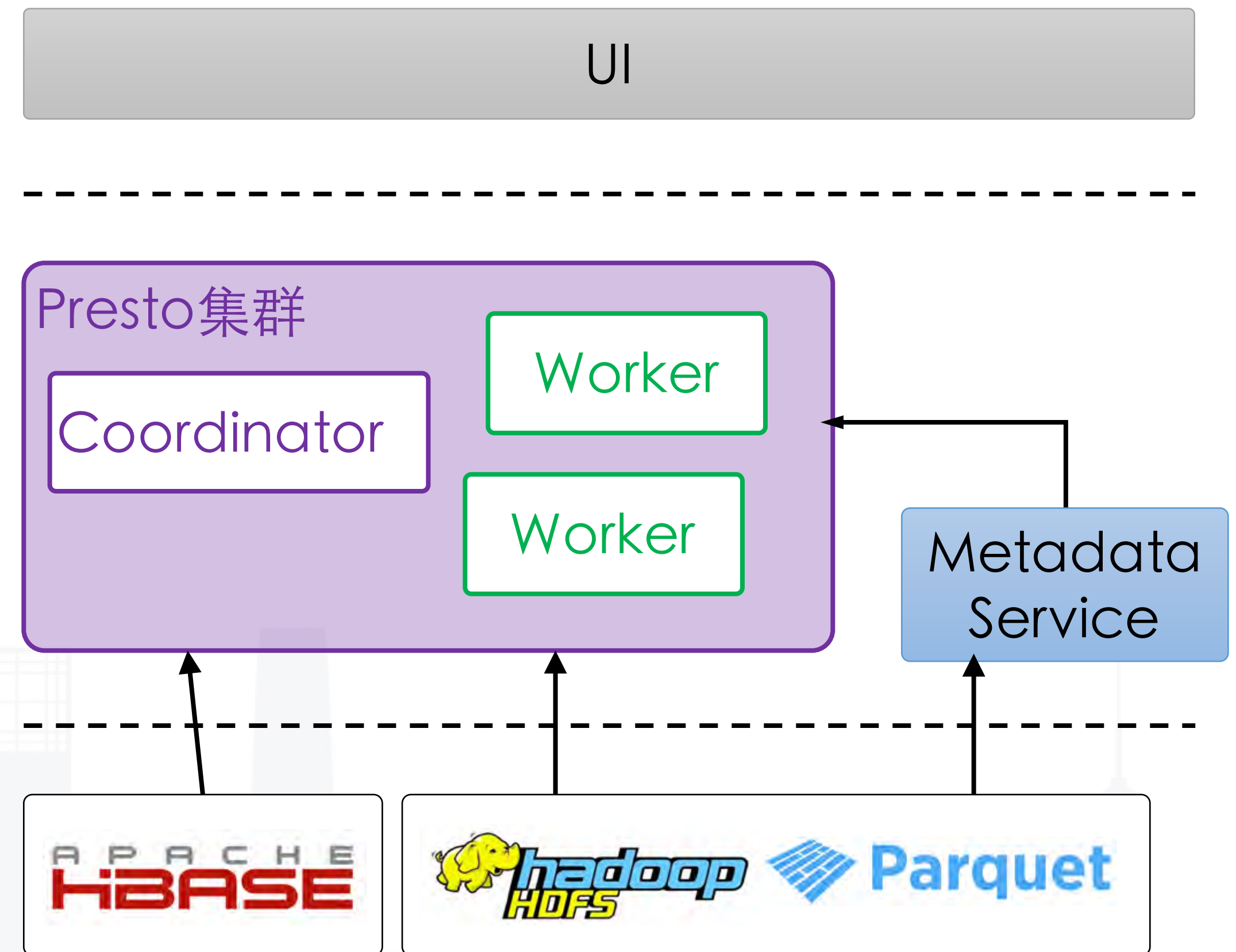
新增Connector

- 支持混合查询 HBase + HDFS
- 谓词下推
- UDF

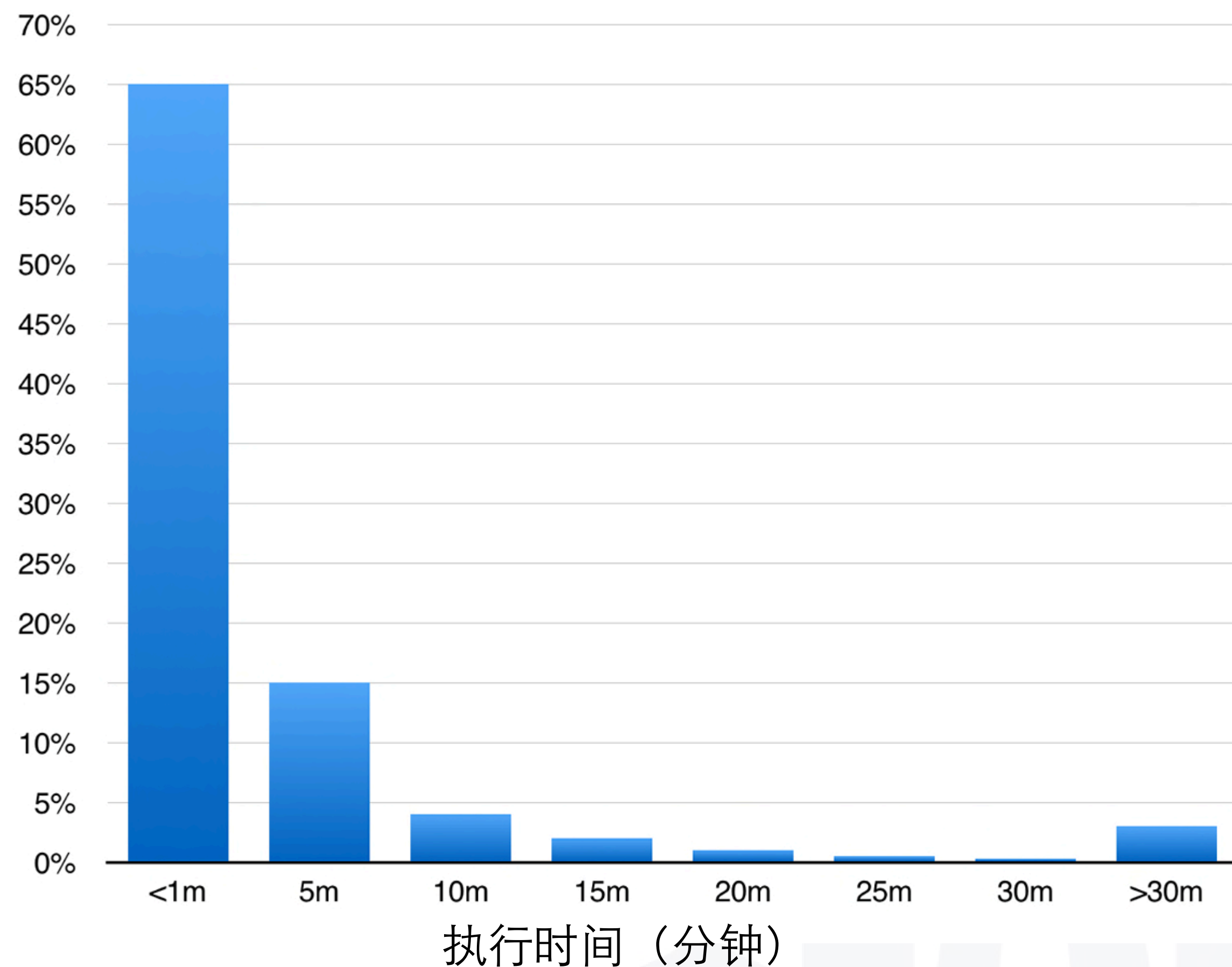
多数据源join: fact + dimension

元数据服务

查询UI



问题一 长查询



长查询

- group by 1, 2, 3, 4, 5, 6, 7, ...
- cross joins
- 复杂条件
- 扫描数据量大
- ...

影响

- 降低整体响应速度，增加失败概率
- Presto不容错，增加失败成本

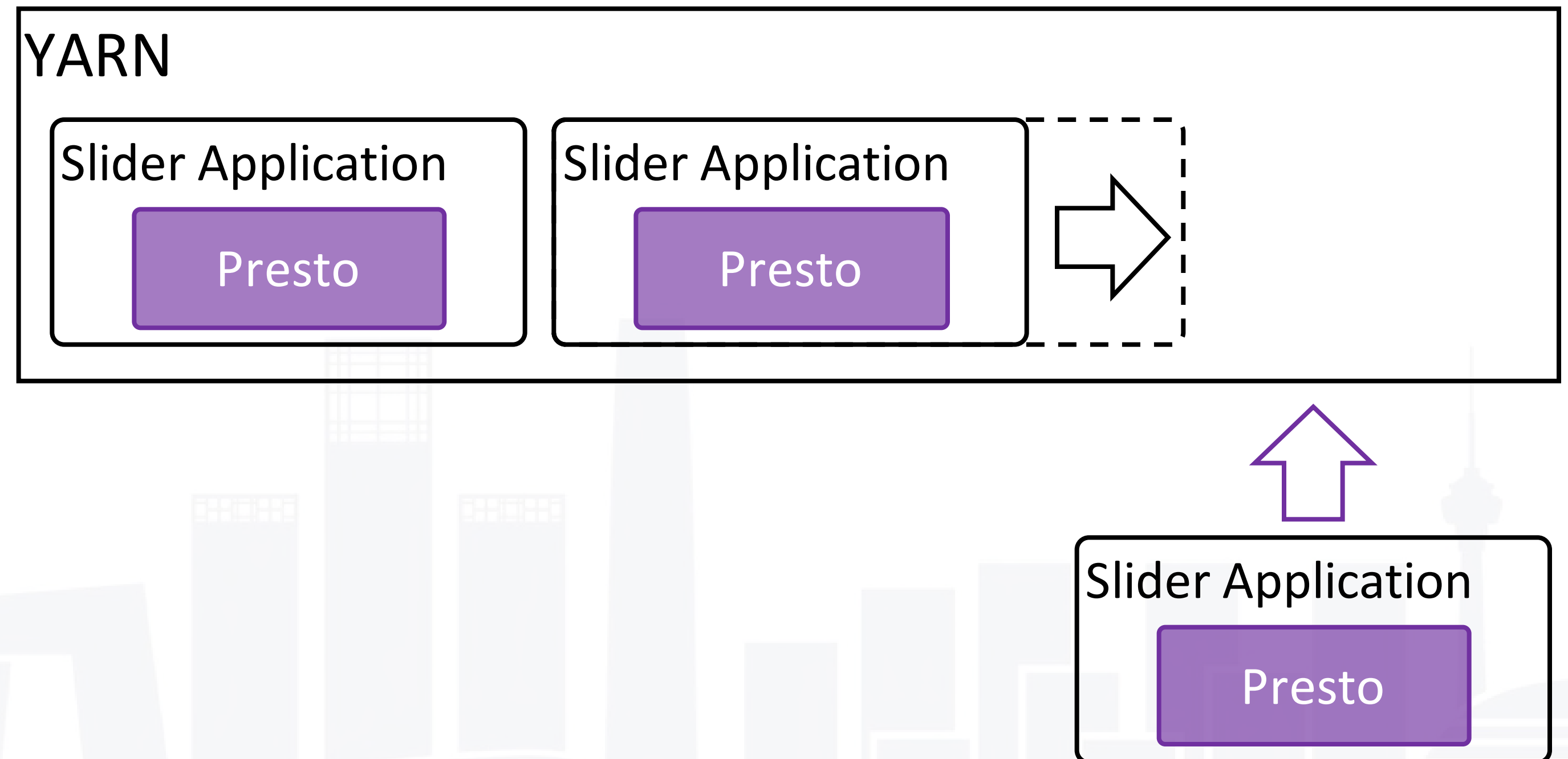
Presto in Slider

Presto + Slider + YARN

- Presto启动脚本
- Presto参数配置
- 资源配置
- 组件间依赖

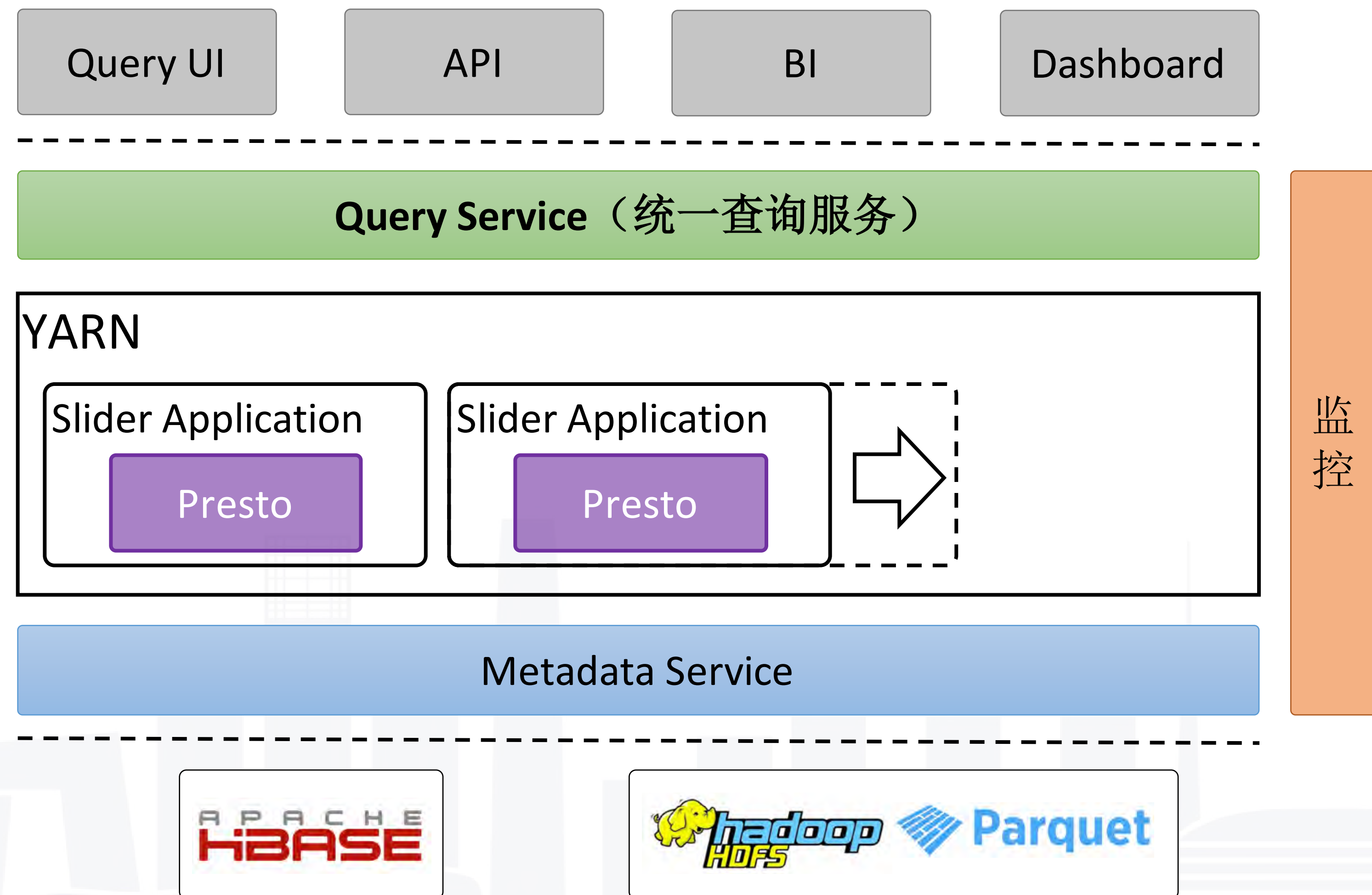
优势

- All in one, 运维友好
- 部署/伸缩方便
- 资源隔离

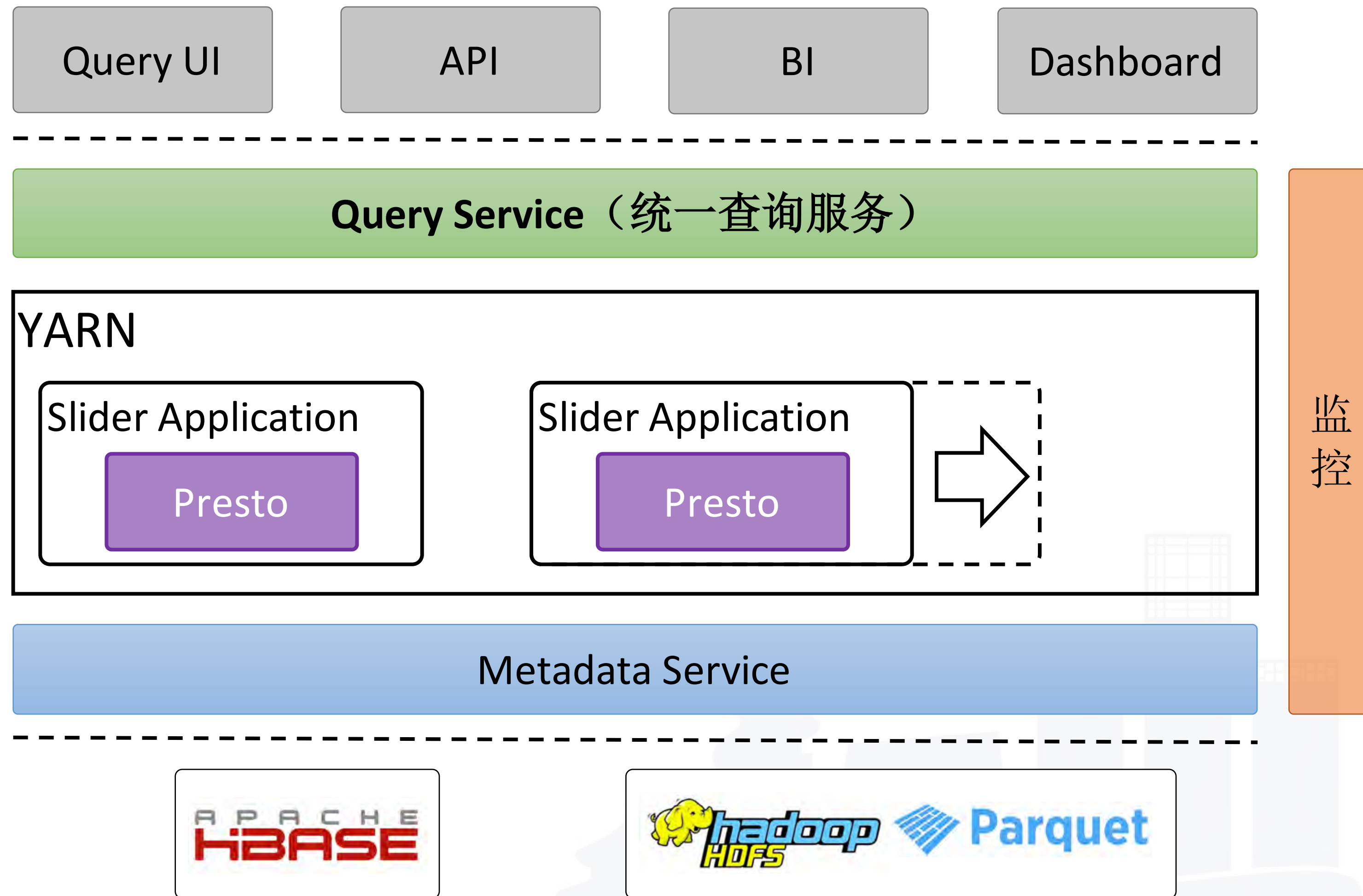


统一查询服务

- 查询限制与预测
- 负载均衡
- 容错
- 权限控制
- 灰度发布
- 审计



应用场景



数据科学家

- 广告投放策略咨询
- 网络视频货币化报告(VMR)

服务支持团队

- 解决客户问题
- 数据面板

工程师

- 数据验证
- 问题排查

客户

- 定制化报表
- BI



FreeWheel



- 姜冰
- Lead Software Engineer
- bjiang@freewheel.tv
- 孙炎
- Senior Software Engineer
- ysun@freewheel.tv



关注QCon微信公众号，
获得更多干货！

Thanks!



主办方 **Geekbang**  **InfoQ**
极客邦科技