BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016

华为大数据多租户技术探索与实践

孙桂林/guilin.sun@gmail.com

自我介绍

2年的HWer
10年大规模分布式系统从业者
华为大数据系统架构

分布式系统
大规模分布式存储
海量数据处理
大数据云服务
……

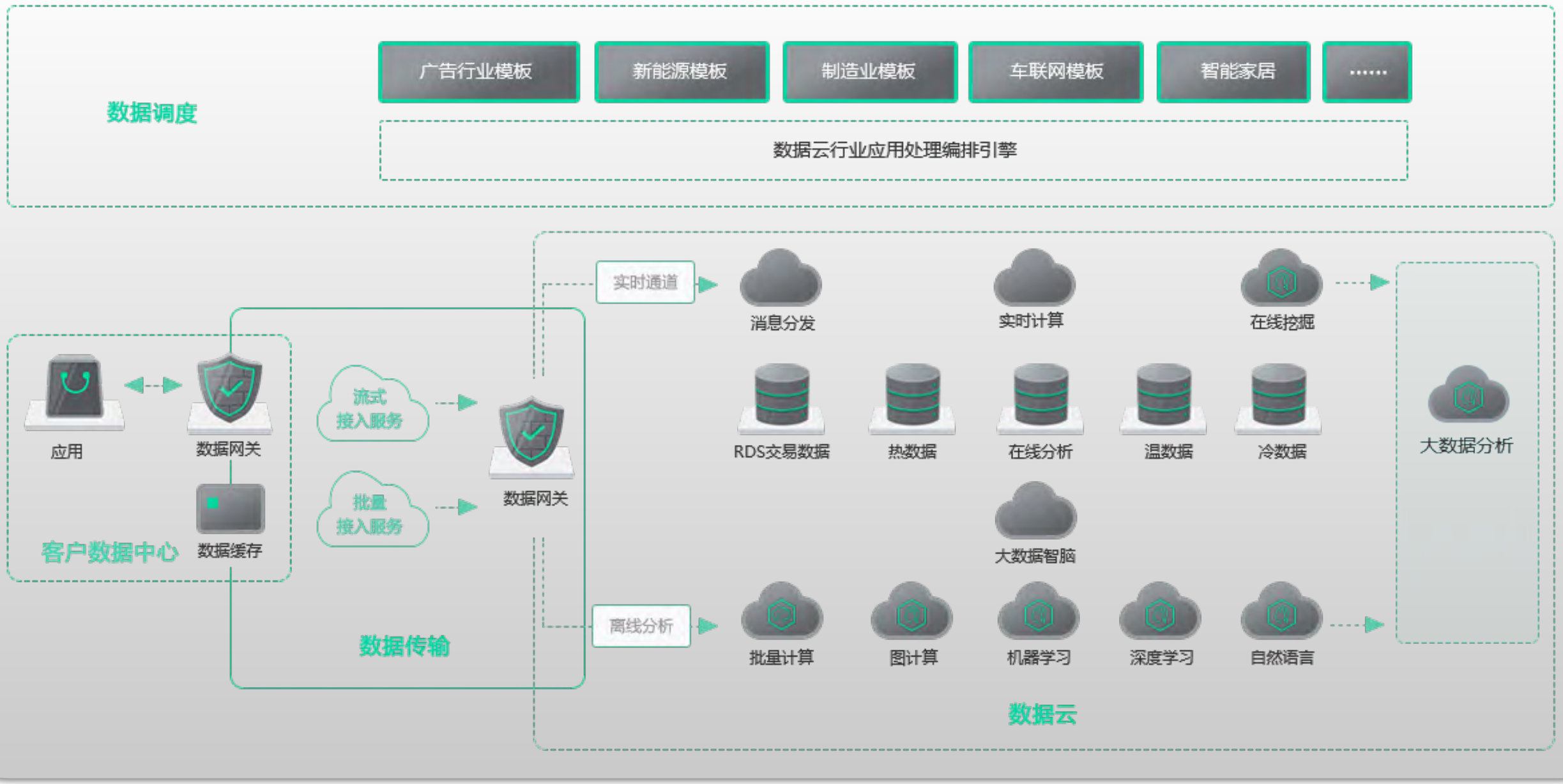
华为大数据业务与产品

电信、企业、消费者……

大数据业务举例

电信	企业	消费者
SmartCare SEQ Analyst	**银行	EMUI
智能网络规划优化 离网分析 个性化套餐包推荐 投诉处理 用户体验管理	精准营销 历史交易明细查询 实时事件营销 实时征信 异常交易预警	智能应用商店 智能帮助

华为企业云



数据接入服务 | 多维交互式分析服务 | MapReduce服务 | 机器学习服务 | 数据调度服务

大数据、云、多租户

云上的大数据集群 or 大数据集群的云

大数据多租户的核心挑战

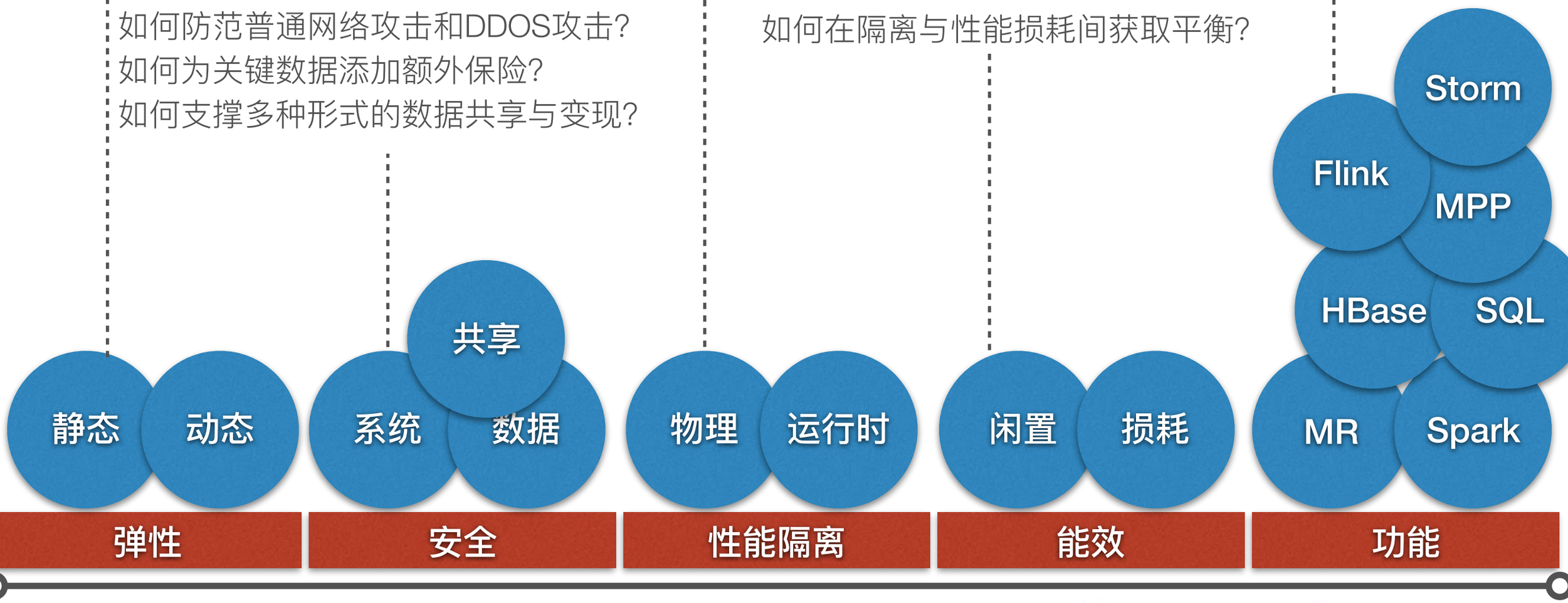
能否快速地申请、释放预留资源?
能否运行时根据资源用量快速扩容、缩容?

如何防范和预警本地提权漏洞攻击?
如何防范普通网络攻击和DDOS攻击?
如何为关键数据添加额外保险?
如何支撑多种形式的数据共享与变现?

能否支持重要租户、应用的强性能隔离?
如何控制租户对于共享资源的竞争?

如何利用租户的闲置资源?
如何在隔离与性能损耗间获取平衡?

如何对接现有大数据生态?



两种路线

存储密集，计算稀疏

IaaS能力超强

性能追求

零运维

业务增长快，业务需求不稳定

重计算

数据变现

体量小

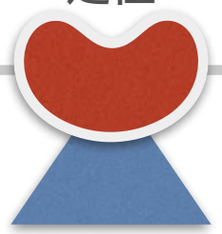
体量足够大

Adhoc访问

重I/O

体验新的大数据分析软件

大数据软件云端部署、托管能力	方向	大数据技术内部的租户隔离能力
虚拟机上的真实大数据集群	形态	物理机上的虚拟大数据集群
运行在云服务上的大数据集群	定位	大数据集群提供的云服务



弹性	安全	性能隔离	能效	功能

两种路线

存储密集，计算稀疏

IaaS能力超强

性能追求

零运维

业务增长快，业务需求不稳定

重计算

数据变现

体量小

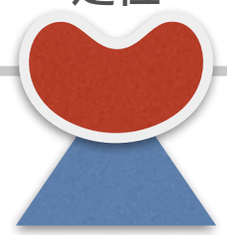
体量足够大

体验新的大数据分析软件

Adhoc访问

重I/O

大数据软件云端部署、托管能力	方向	大数据技术自身的租户隔离能力
虚拟机上的真实大数据集群	形态	物理机上的虚拟大数据集群
运行在云服务上的大数据集群	定位	大数据集群提供的云服务



DataNode Proxy

存储计算分离

集群动态伸缩

VM

VPC

Virtualization

Pool

精简内核

P2P

I/O直通

临时集群

适配更多的大数据组件

弹性 | 安全 | 性能隔离 | 能效 | 功能

两种路线

存储密集，计算稀疏

IaaS能力超强

性能追求

零运维

业务增长快，业务需求不稳定

重计算

数据变现

体量小

体量足够大

体验新的大数据分析软件

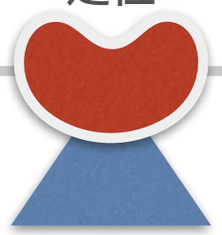
Adhoc访问

重I/O

大数据软件云端部署、托管能力 ------ 方向 ---------- 大数据技术自身的租户隔离能力

虚拟机上的真实大数据集群 ------ 形态 ------- 物理机上的虚拟大数据集群

运行在云服务上的大数据集群 ---------- 定位 --------- 大数据集群提供的云服务



弹性	安全	性能隔离	能效	功能
资源管理	加密	分区	资源管理	为更多的大数据组件增加多租户能力
调度算法	分区	份额	调度算法	
	沙箱			

安全

上古时代的Hadoop没有任何安全能力

无安全机制

鉴权

信任用户声称的身份，改善了『误越界』的问题

实现了『用户到服务』以及『服务到服务』的认证

Kerberos认证

内部使用

企业

数据表、视图、列

细粒度权限控制

数据行

文件、行、列

数据加密

公网

VM

VPC

EMR

提权
&
DoS

Partition

Quota

Sandbox

Guard

“Hadoop”-a-a-S

提权攻击应对思路及手段

缩小攻击面

延长攻击线

异常早可见

手段	作用	举例
Partition	降低提权的影响。	Label Partition
Sandbox	不能隔离的通过Sandbox防护。	VM、Container、JVM
Subtraction	只开放高层、安全的接口。	SQL-a-a-S
Guard	异常行为探测与处理。	Apache Eagle

实践中往往需要组合多种安全机制。

VM != Safe

VMware Workstation target at the PwnFest hacking competition

Posted on November 10, 2016 by Bo Fu

VMware Workstation is among the targets of the PwnFest hacking competition. At this event, which is organized along the Power of Community security conference in Seoul, security researchers are demonstrating their attack capabilities. The event is modeled after the well-known Pwn2Own competition.

Earlier today at the event, the 360 Marvel Team and security researcher Lokihardt (JungHoon Lee) used the same issue to demonstrate that they could execute code on the VMware Workstation host from the guest. We have received details on this issue directly from the researchers and we are now working on a solution. We have confirmed that the issue is limited to VMware Workstation and VMware Fusion and that ESXi is not affected.

We would like to thank the organizers of the event, the 360 Marvel Team, and Lokihardt for working with us to address the issue.

November 13 update

Today, we've published VMware Security Advisory VMSA-2016-0019 which documents the release of VMware Workstation 12.5.2 and VMware Fusion 8.5.2. These new Workstation and Fusion versions address the issue that was demonstrated at the PwnFest event. The issue has been assigned CVE identifier CVE-2016-7461.

– VMware Security Response Center and VMware Workstation Team

2016年11月10号的PwnFest擂台中，来自国内和韩国的两只团队分别在VMware上实现了虚拟机逃逸，可在宿主机上执行任意代码。

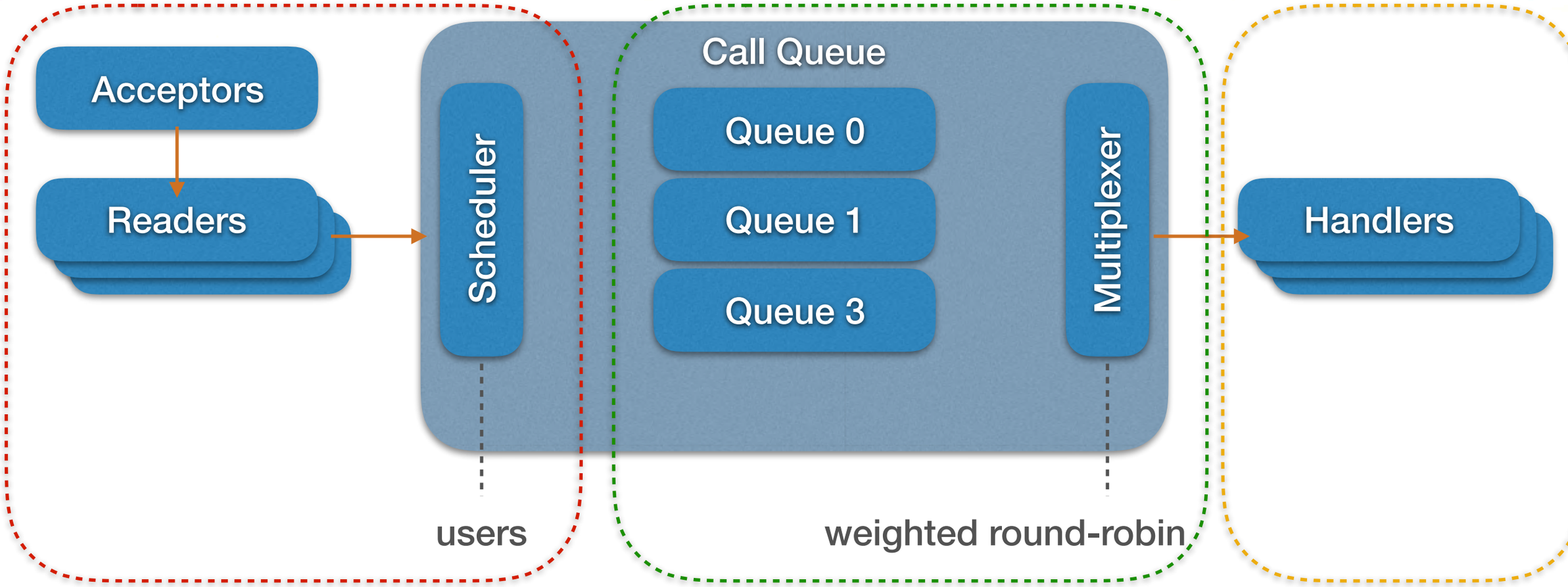
虽然非常困难构造，逃逸可能是对虚拟机最大的安全威胁

DoS攻击应对思路及手段

手段	作用	举例
RPC Fair Share	防止RPC的DDoS	Hadoop FairCallQueue
Language Sandbox	禁止敏感API的调用	JVM安全策略禁止访问网络
Federation	租户不共享瓶颈节点	HDFS/YARN Federation
Container/VM	租户不共享集群	EMR
Subtraction	只开放高层、安全的接口	SQL-a-a-S
Guard	异常行为探测与处理	网络流量清洗

DoS攻击可能是应用层无意识产生的。

FairCallQueue

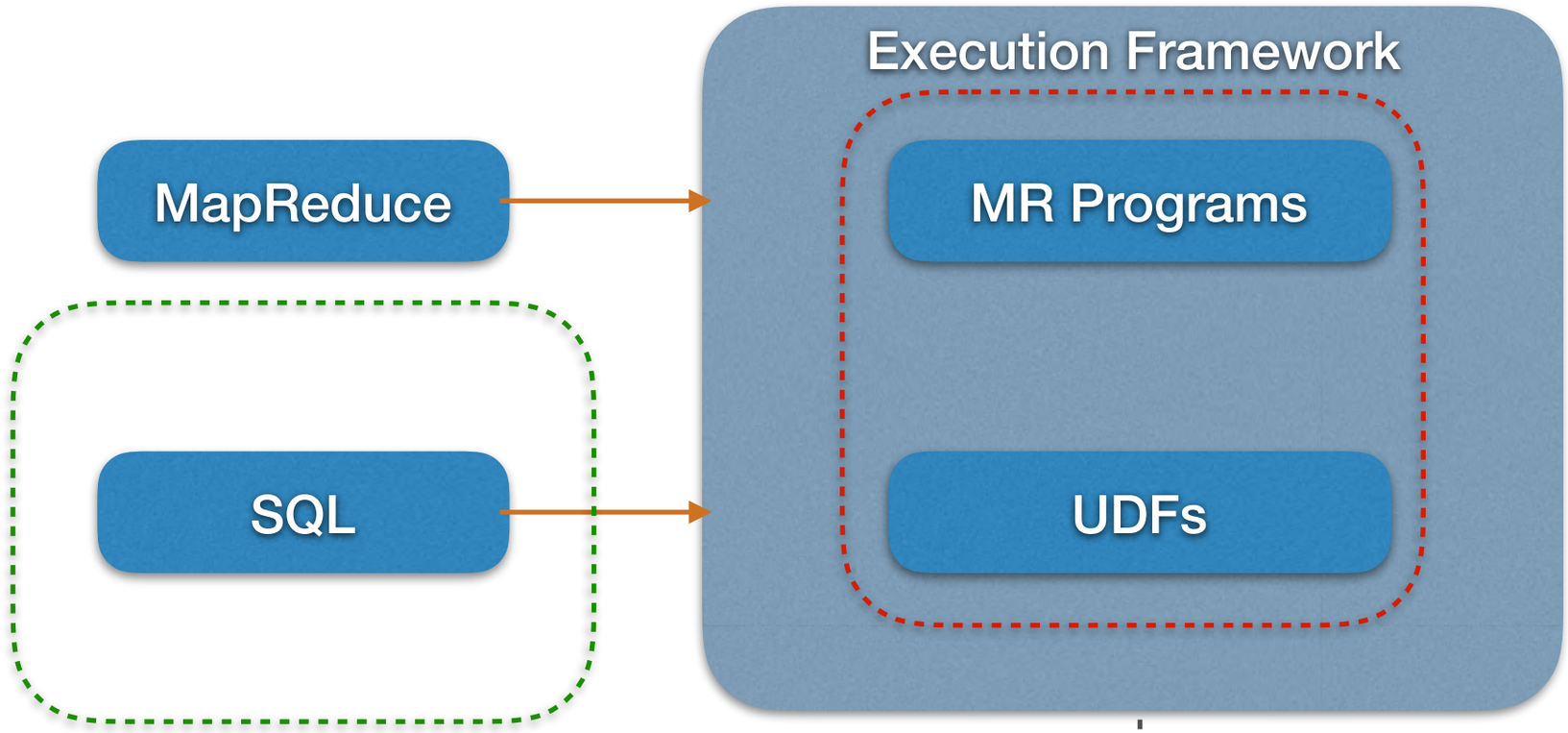


FairCallQueue依赖于获取请求的用户信息来做后面的调度，但获取用户信息前的连接处理、请求头处理、用户信息获取没有办法做到公平。

不同操作的代价可能差别很大，执行时间、并发能力上都有很大不同。

防范无意识的DDoS攻击，也提升RPC资源分配的公平性。

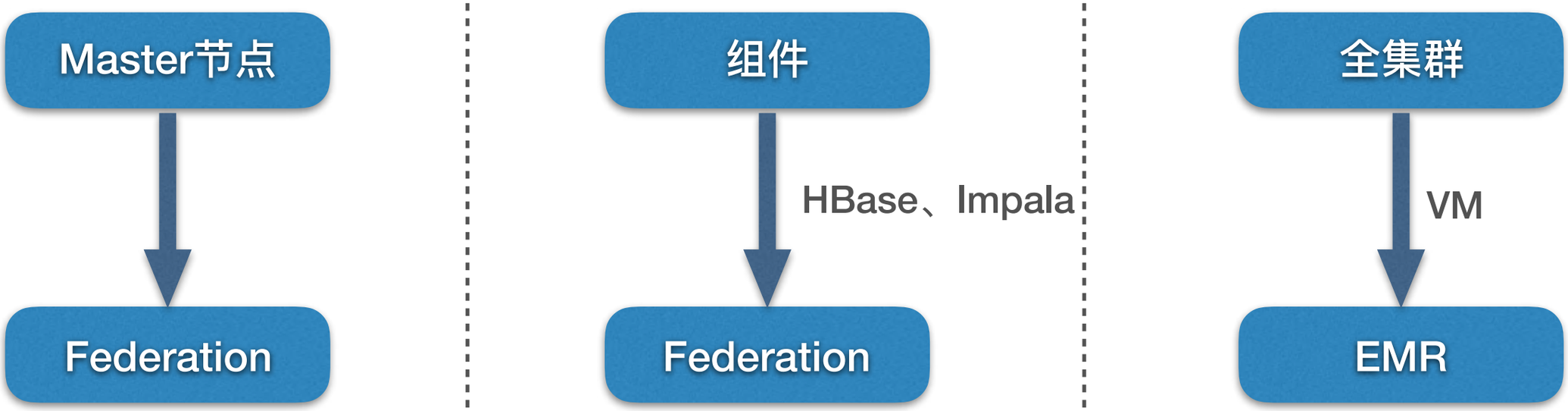
Sandbox



受限的SQL接口，平台产生的的代码更加容易控制和优化。目前国内在公网开放的单实例大数据服务也都是从SQL开始。

JVM的安全策略可以做到限制文件系统功能、网络等资源的访问，这样可以把用户代码框定在处理自己的内存数据上，与外部世界的交互由可信的框架代码来完成。

多实例



Federation是相对轻量的多实例方案，但为每一个租户Federation是不现实的。而且Worker节点上还是要做类似Fair Share的工作。

总有些组件自身不具备或者暂时不具备多租户能力。通常更多是为了性能隔离，结合Partition一起使用。

全集群多实例的方案往往会基于VM来做，这样对集群各组件的负载隔离都比较，但是即使从DoS的角度，VM也不是高枕无忧的。

VM DoS

VM间的共享资源也存在安全风险

MEMORY COMPONENTS	ATTACKER'S TECHNIQUE	CONTENTION TYPE	RUNTIME SLOWDOWN
Shared LLC	LLC cleansing	storage-based	1~5.5X
Buses	bus locking	scheduling-based	1~7.9X
IMC	memory flooding	scheduling-based	1~1.54X
DRAM		storage-based	

2016年3月，T. Zhang等人成功地在EC2上，利用VM间的共享资源进行了memory DoS攻击。

https://arxiv.org/pdf/1603.03404.pdf

性能隔离

多实例

静态隔离

动态隔离

上古时代的Hadoop没有任何性能隔离能力

FIFO

Capacity, Fair

可以控制不同租户和应用的计算资源份额

可以控制用户进程和平台组件的CPU与内存占用

cgroups

对象存储

虚拟机

虚拟网络

分区

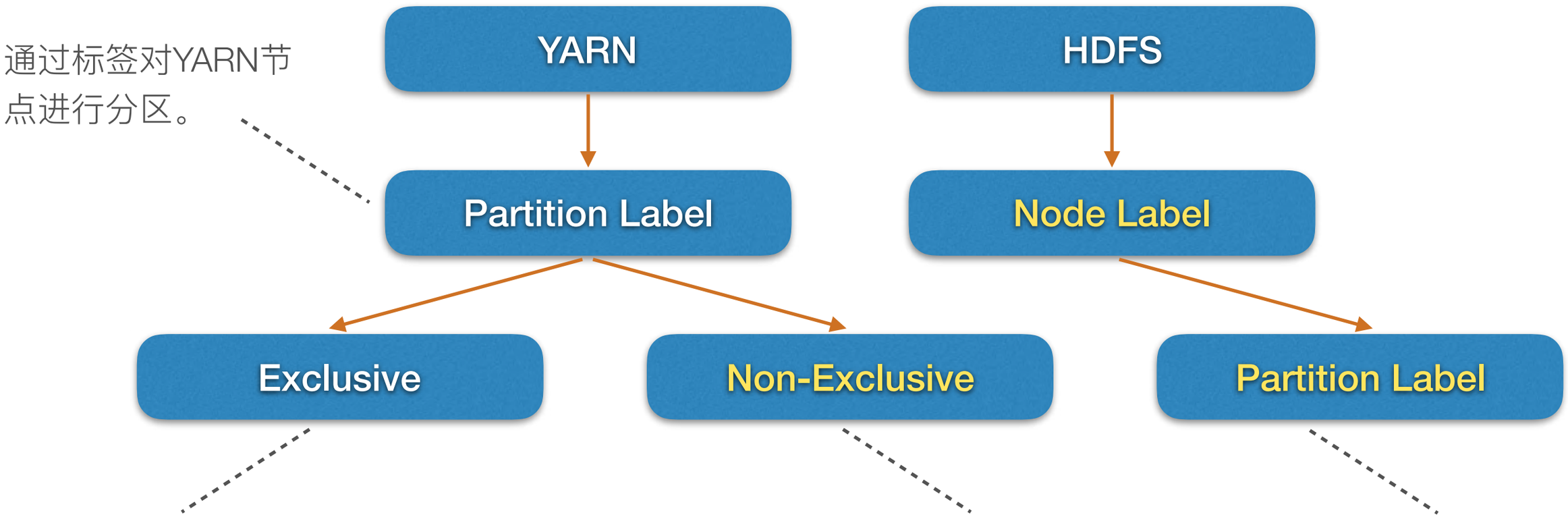
联邦

份额

优先级

与无意识的DDoS相比，性能隔离侧重于合理的资源使用与竞争。

分区



通过标签对YARN节点进行分区。

分区间不能共享计算资源。

分区间可以配置共享策略，在隔离的同时也兼顾资源利用率。

通过标签来对HDFS节点进行分区，并支持跨分区的数据块放置策略。

分区对弹性的影响?

份额控制

保留资源，空闲也不可以挪用

Reserved

Min

最小保障资源，有需要就得保证。

Weight

资源用满时，能占用的比例。

Max

最大资源用量。

FairCallQueue……

不同调度器支持不同程度的R-M-S-M语义。

?

cgroups

DRM Queue

RPC

I/O

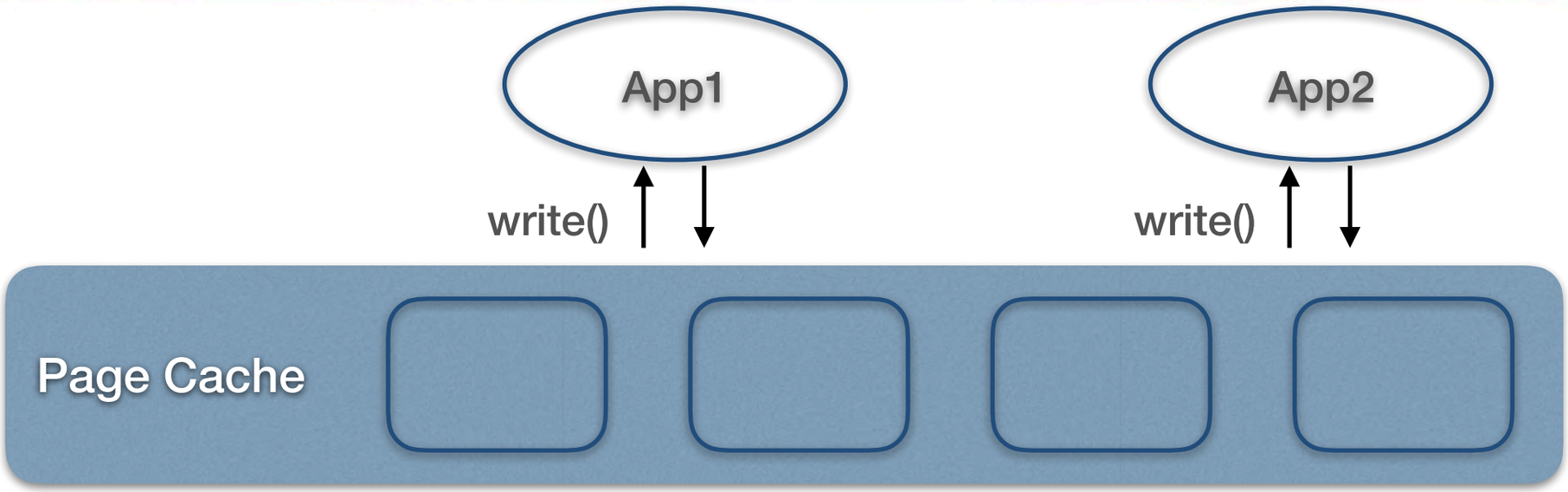
Memory

CPU

Cluster

Local

cfq & blkio



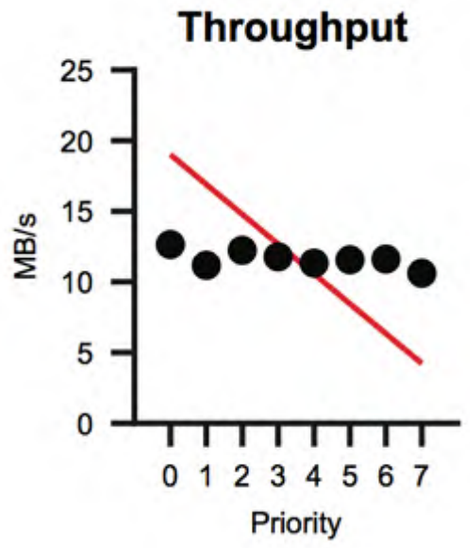
Loss of cause information!

不同的内核版本实现略有区别，某些较老的版本在page cache阻塞时会由应用进程触发刷盘。

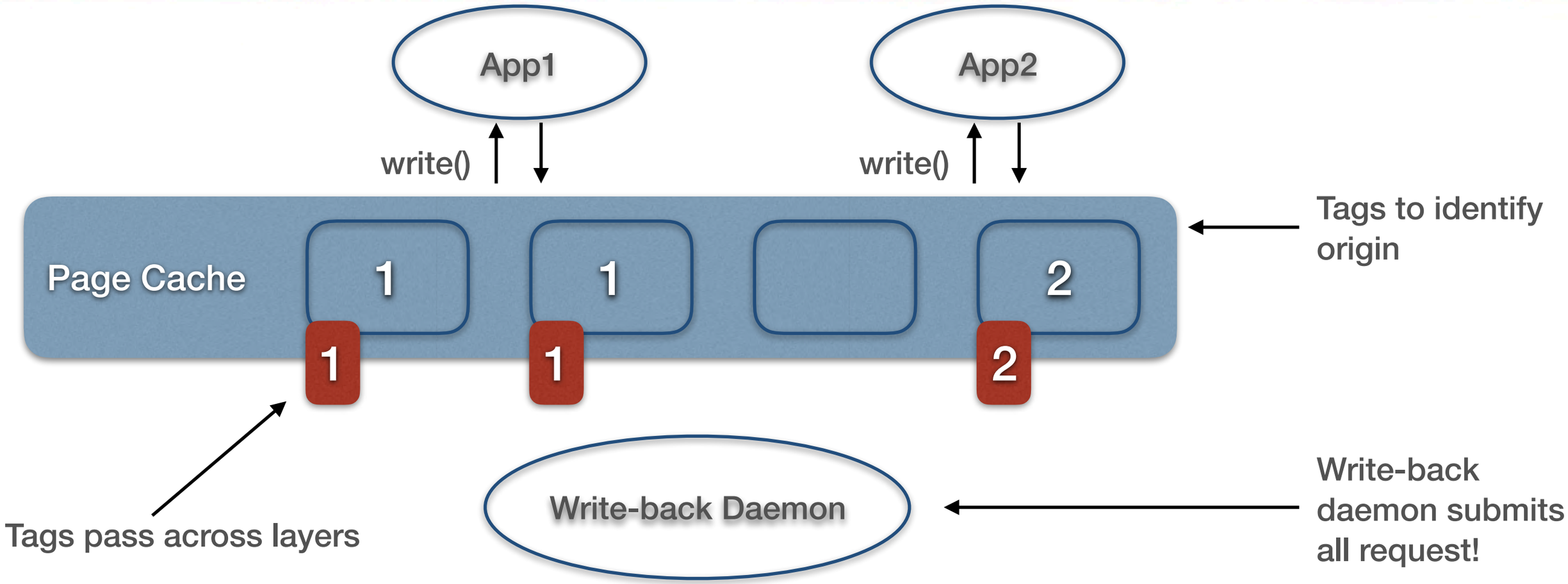
Write-back daemon submits all request!

Block-Level Scheduler - cfq/deadline/noop

失去了正确的cause信息，I/O调度器的优先级也就无从谈起，因此对于经过pagecache的所有I/O，cgroups blkio的weight配比也完全没办法保证。



Cross-Layer Tag

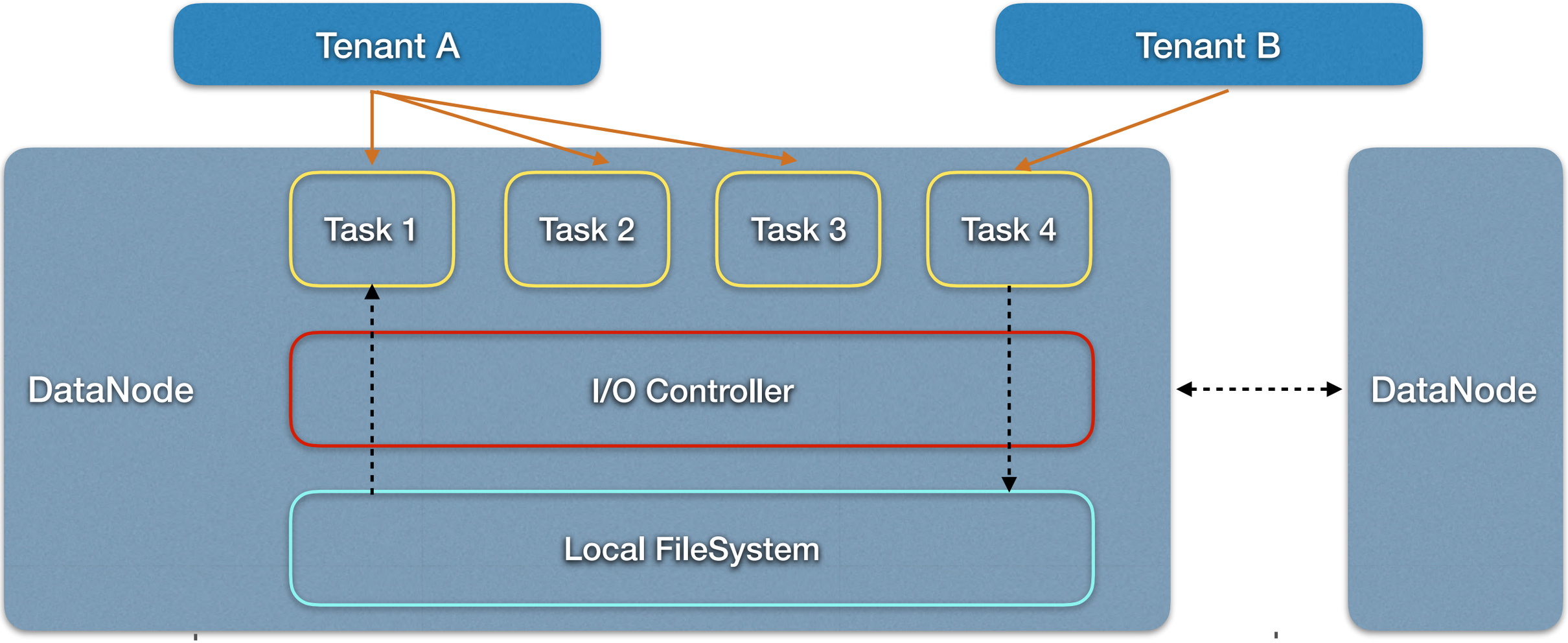


Block-Level Scheduler - cfq/deadline/noop

I/O来源信息的跨层传递是后续工作的基础。

http://sigops.org/sosp/sosp15/current/2015-Monterey/printable/168-yang.pdf

Global I/O Scheduling



DFS上的I/O其租户来源信息对本地文件系统也是不可见的。

Global范围内也要做到公平与高效。

https://www.usenix.org/system/files/fastpw13-paper19_0.pdf

回顾

API Restriction - SQL

I/O Weight

Language Sandbox

Multiple Instance

Partition

VM/LXCs

FairCallQueue

Guard

DRM Weight

弹性 | 安全 | 性能隔离 | 能效 | 功能

谢谢

Q&A

BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016

HUAWEI