



数据技术嘉年华

Data Technology Carnival

云·数据·智能 - 数聚价值智胜未来

关注公众号回复help,
可获取更多经典学习
资料和文档, 电子书



我们是如何构建 金融级数据库云

上海富麦信息科技有限公司



第七屆



数据技术嘉年华

Data Technology Carnival

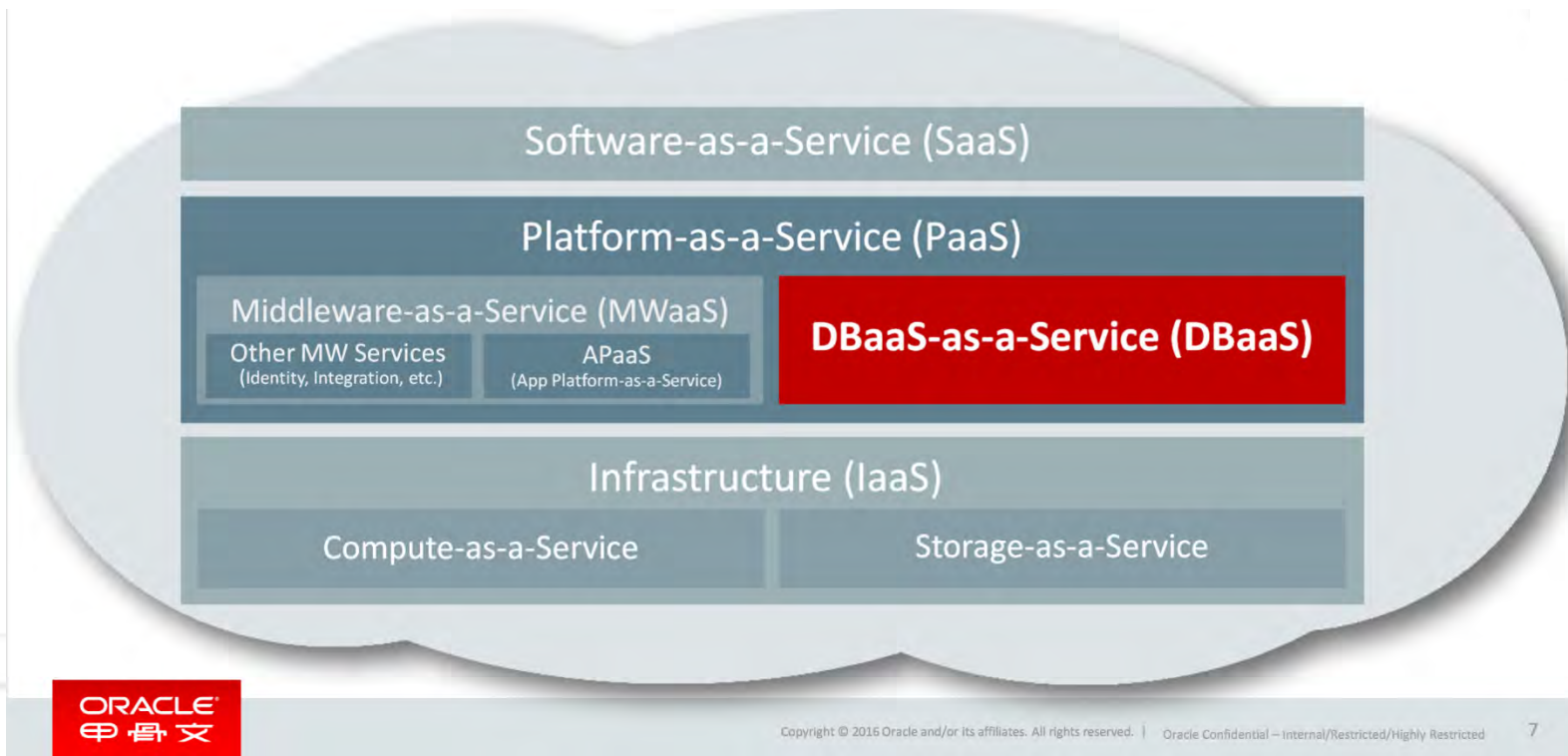


About ME

- 鲍琳
- 架构师，中国银联DBaaS项目
- 首席架构师，金融级数据库云(DBscale)
- R&D负责人，上海富麦信息科技
- 研究和使用的开源技术，并将数据库云(DBscale)在金融行业客户生成环境落地运行



什么是数据库云



第七届



数据技术嘉年华

Data Technology Carnival



我们如何构建数据库云



第七屆



数据技术嘉年华

Data Technology Carnival



我们如何构建数据库云

- 如何构建资源池模型
- 如何实现高效的调度算法
- 如何构造网络模型
- 如何构建存储模型
- 如何获得高效运维能力



第七届

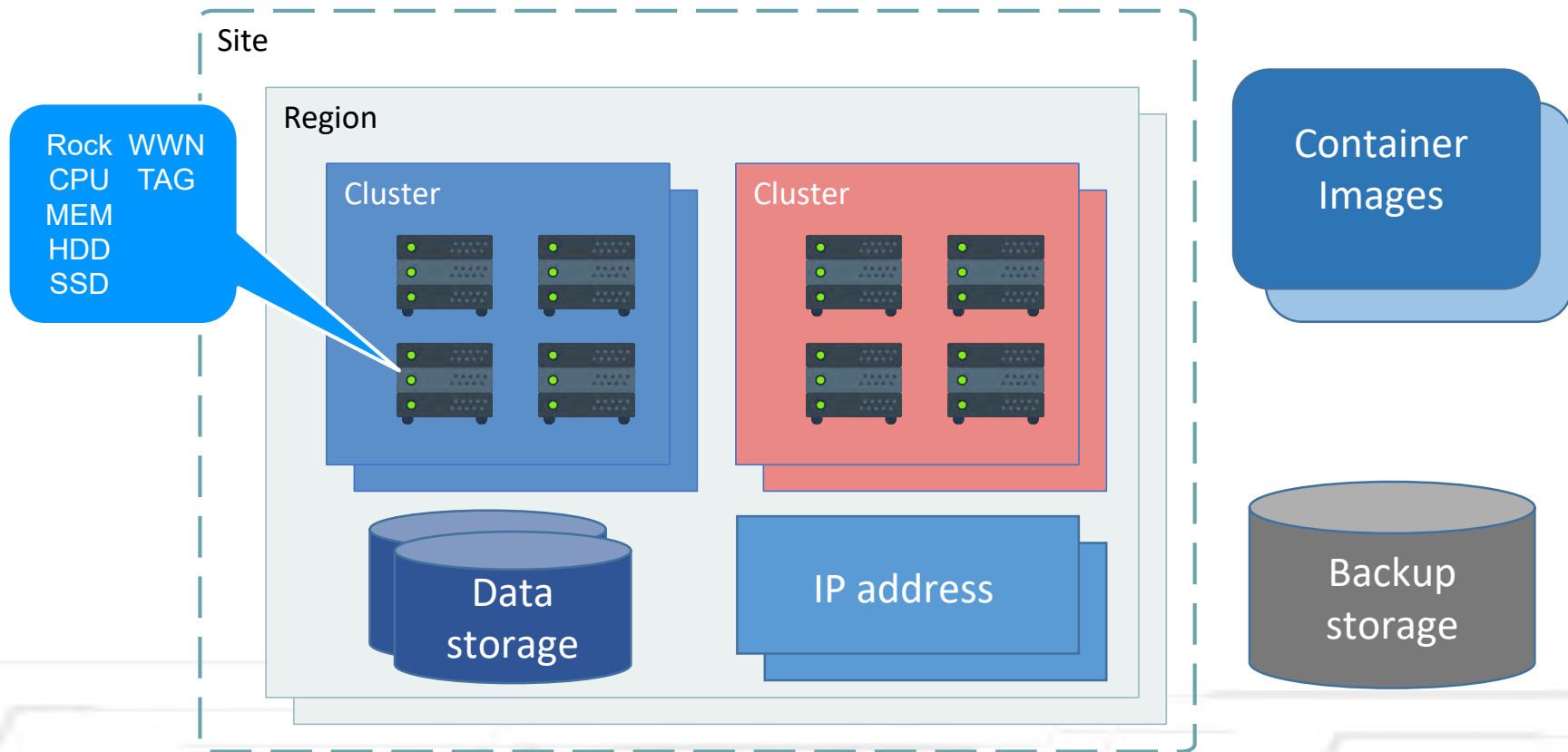


数据技术嘉年华

Data Technology Carnival



如何构建资源池模型



第七届



数据技术嘉年华

Data Technology Carnival



如何实现高效的调度算法

求解之路的探索

- 现有方案是否解决了我们的问题？
- 我们的研究和探索



第七届



数据技术嘉年华

Data Technology Carnival



Algorithm 1 DRF pseudo-code

$R = \langle r_1, \dots, r_m \rangle$ ▷ total resource capacities
 $C = \langle c_1, \dots, c_m \rangle$ ▷ consumed resources, initially 0
 s_i ($i = 1..n$) ▷ user i 's dominant shares, initially 0
 $U_i = \langle u_{i,1}, \dots, u_{i,m} \rangle$ ($i = 1..n$) ▷ resources given to user i , initially 0

pick user i with lowest dominant share s_i

$D_i \leftarrow$ demand of user i 's next task

if $C + D_i \leq R$ **then**

$C = C + D_i$ ▷ update consumed vector

$U_i = U_i + D_i$ ▷ update i 's allocation vector

$s_i = \max_{j=1}^m \{u_{i,j}/r_j\}$

else

return ▷ the cluster is full

end if

- ① Mesos 采用了DRF(Dominant Resource Fairness) 调度机制。
- ② Mesos中的DRF调度算法过分的追求公平，没有考虑到实际的应用需求。在实际生产线上，往往需要类似于Hadoop中Capacity Scheduler的调度机制，将所有资源分成若干个queue，每个queue分配一定量的资源，每个user有一定的资源使用上限。
- ③ Mesos采用了Resource Offer机制，这种调度机制面临着资源碎片问题，即：每个节点上的资源不可能全部被分配完，剩下的一点可能不足以让任何任务运行，这样，便产生了类似于操作系统中的内存碎片问题。



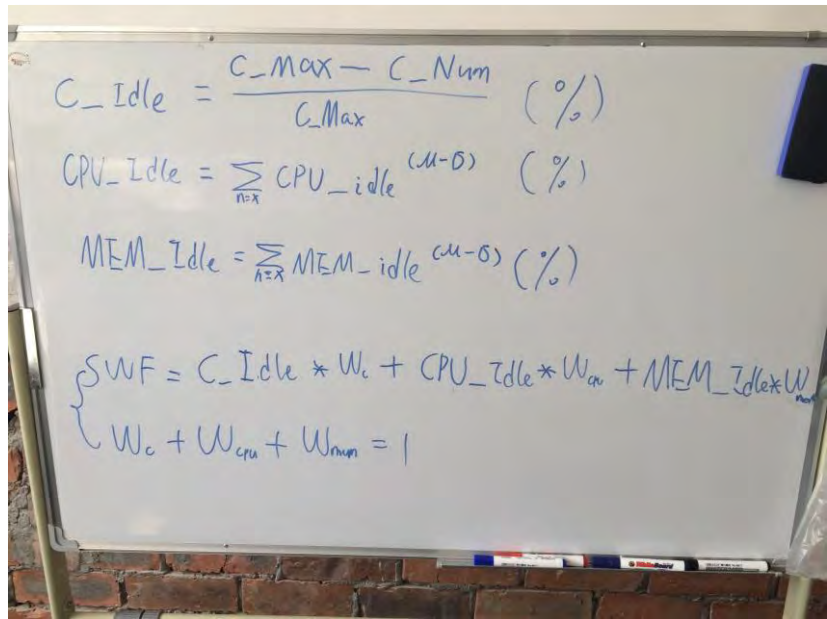


kubernetes

- ① Kubernetes 仅仅是实现了一个极其简单的调度器。鼓励开发者编写自己的调度器注册进框架
- ② 调度策略分为两大类：Predicates和Priorities，其中Predicates判断是否将pod调度到特定 minion(host)上运行，而Priorities则是在Predicates的计算基础上，通过积分Score方式，决定调度量。
- ③ Predicates包括：PodFitsPorts、PodFitsResources、NoDiskConflict、MatchNodeSelector和 HostName，即一个minion能够被选中的前提是需要经历前面提到的这5个Predicates的检验，而 Priorities又包括：LeastRequestedPriority、ServiceSpreadingPriority和EqualPriority，分别为通过Predicates检验的minion计算优先级（score），score是一个范围是0-10的整数，0代表最低优先级，10代表最高优先级。
- ④ 调度机制还是过于平均，Predicates本质上作为一个过滤器。



基于场景加权调度算法



Handwritten formulas on a whiteboard:

$$C_Idle = \frac{C_Max - C_Num}{C_Max} (\%)$$
$$CPU_Idle = \sum_{n=x} CPU_idle^{(u-0)} (\%)$$
$$MEM_Idle = \sum_{n=x} MEM_idle^{(u-0)} (\%)$$
$$\begin{cases} SWF = C_Idle * W_c + CPU_Idle * W_{cpu} + MEM_Idle * W_{mem} \\ W_c + W_{cpu} + W_{mem} = 1 \end{cases}$$

- ① 基于不同应用的场景数据做资源的实时计算。
- ② 场景数据的短期切片和中长期切片可以适应资源池投产的不同阶段。
- ③ 实现了(人工)可干预的分配机制(阈值)。
- ④ 通过权重比对利用率优先，容量优先和可用性优先进行调控。
- ⑤ 具体实现采用较为独立的模块方式，方便将来开源后被第三方使用，定制和集成。
- ⑥ 面向金融行业应用场景，进行持续的演进和调整。



第七届



数据技术嘉年华

Data Technology Carnival



物理机剩余CPU资源百分率

$$= \frac{\text{物理机未分配的CPU核数}}{\text{物理机CPU核数}} \times 100\%$$

物理机剩余内存资源百分率

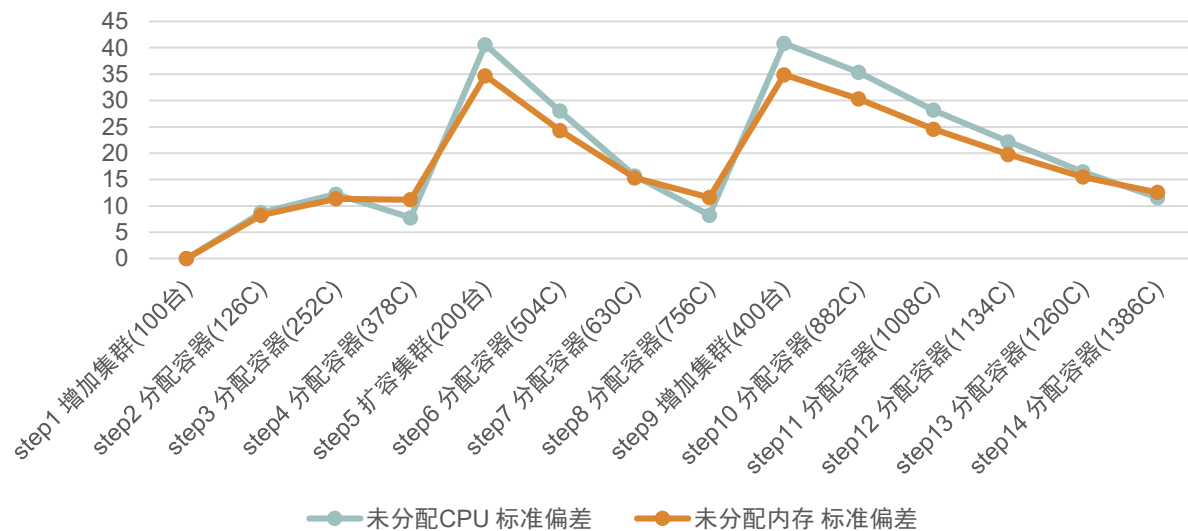
$$= \frac{\text{物理机未分配的内存大小}}{\text{物理机内存大小}} \times 100\%$$

$$\text{百分率平均值} \mu = \frac{1}{N} \sum_{i=1}^N \text{物理机}i\text{资源百分率}$$

集群剩余CPU标准差 =

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (\text{物理机}i\text{未分配资源百分率} - \mu)^2}$$

集群剩余资源百分率标准差折线图 示例图



第七届



数据技术嘉年华

Data Technology Carnival



如何构造网络模型

求解之路的探索

- 现有方案是否解决了我们的问题？
- 我们的研究和探索



第七届

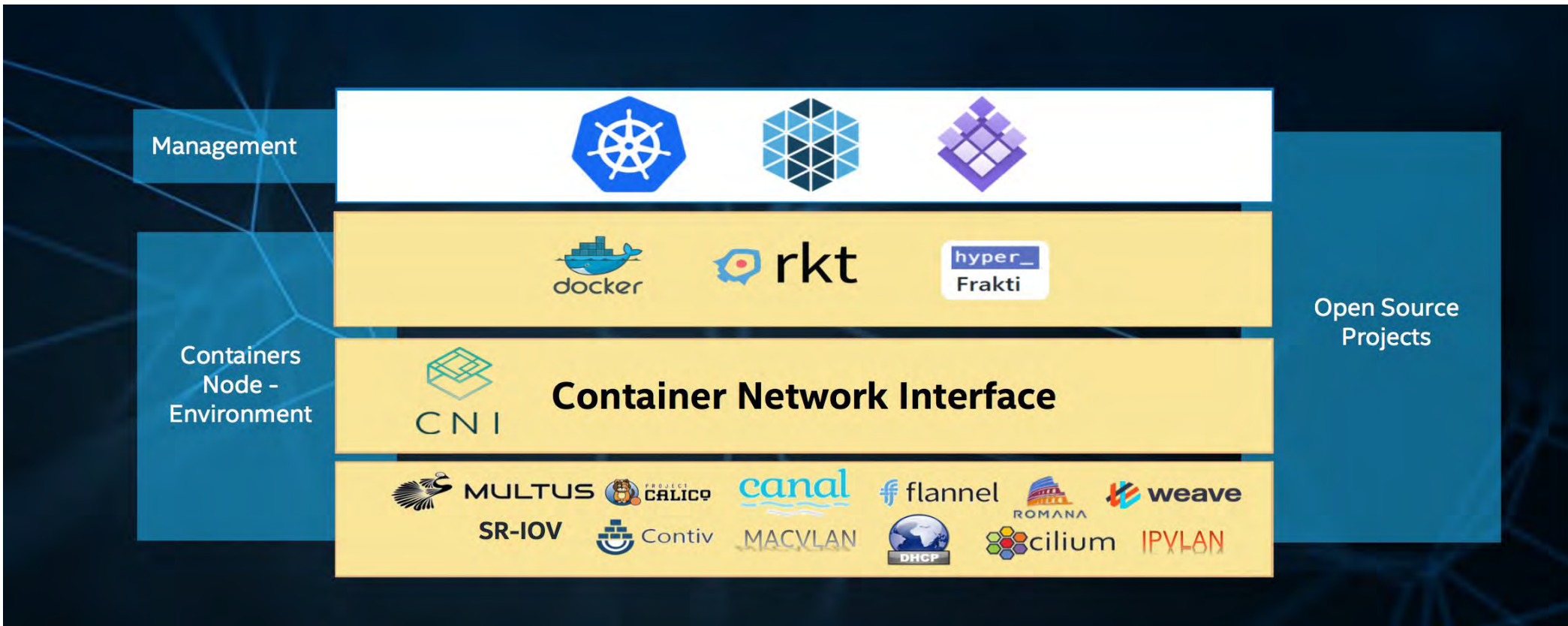


数据技术嘉年华

Data Technology Carnival



容器网络现状



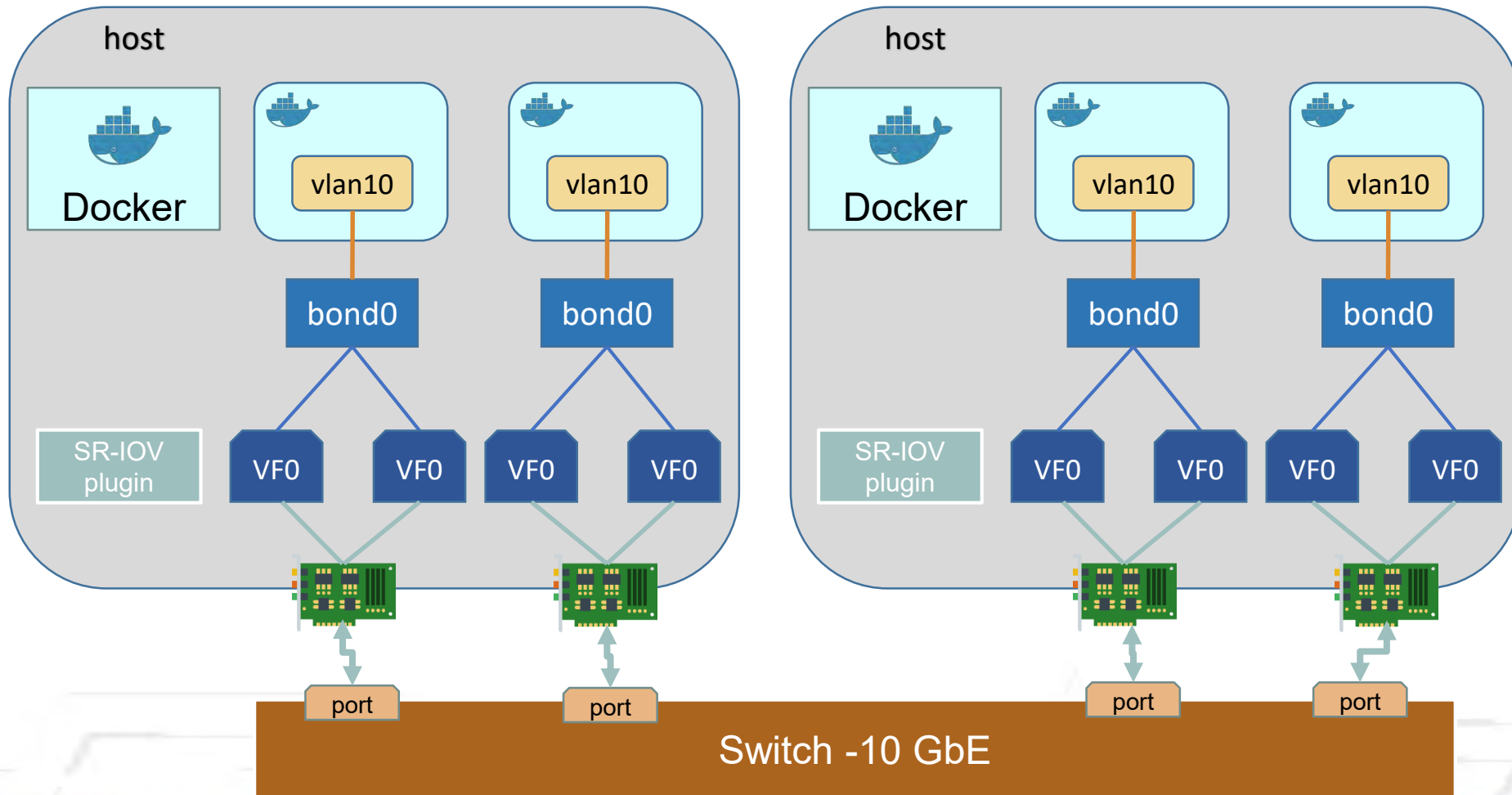
第七届



数据技术嘉年华

Data Technology Carnival



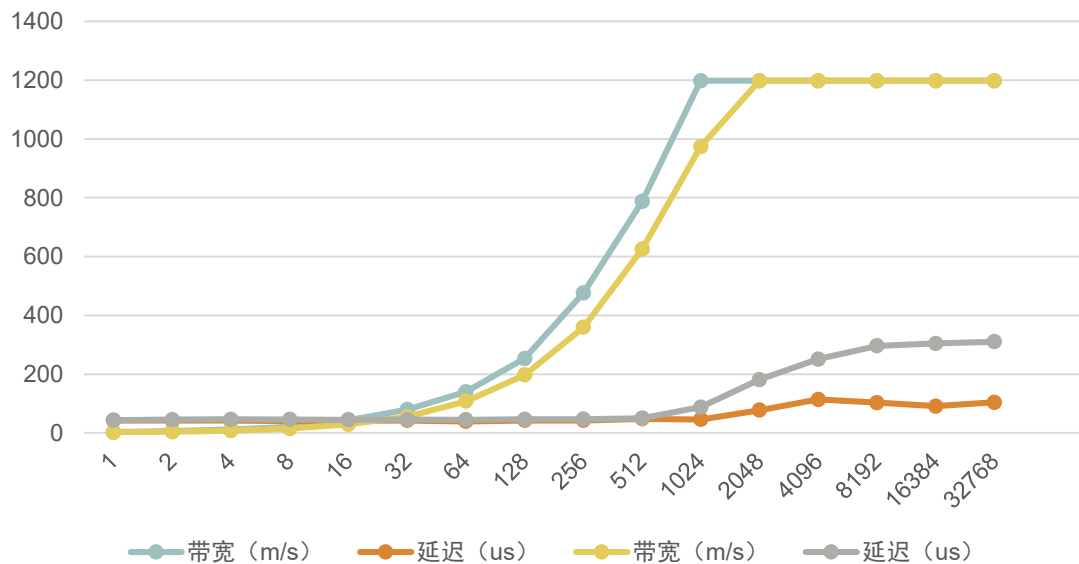


第七届



数据技术嘉年华
Data Technology Carnival





- 支持VLAN网络的支持
- 支持网卡bond高可用结构
- 每个容器使用独立的网卡，物理级隔离流量
- 支持Tx流控



第七届



数据技术嘉年华

Data Technology Carnival



如何构建存储模型

■我们的研究和探索

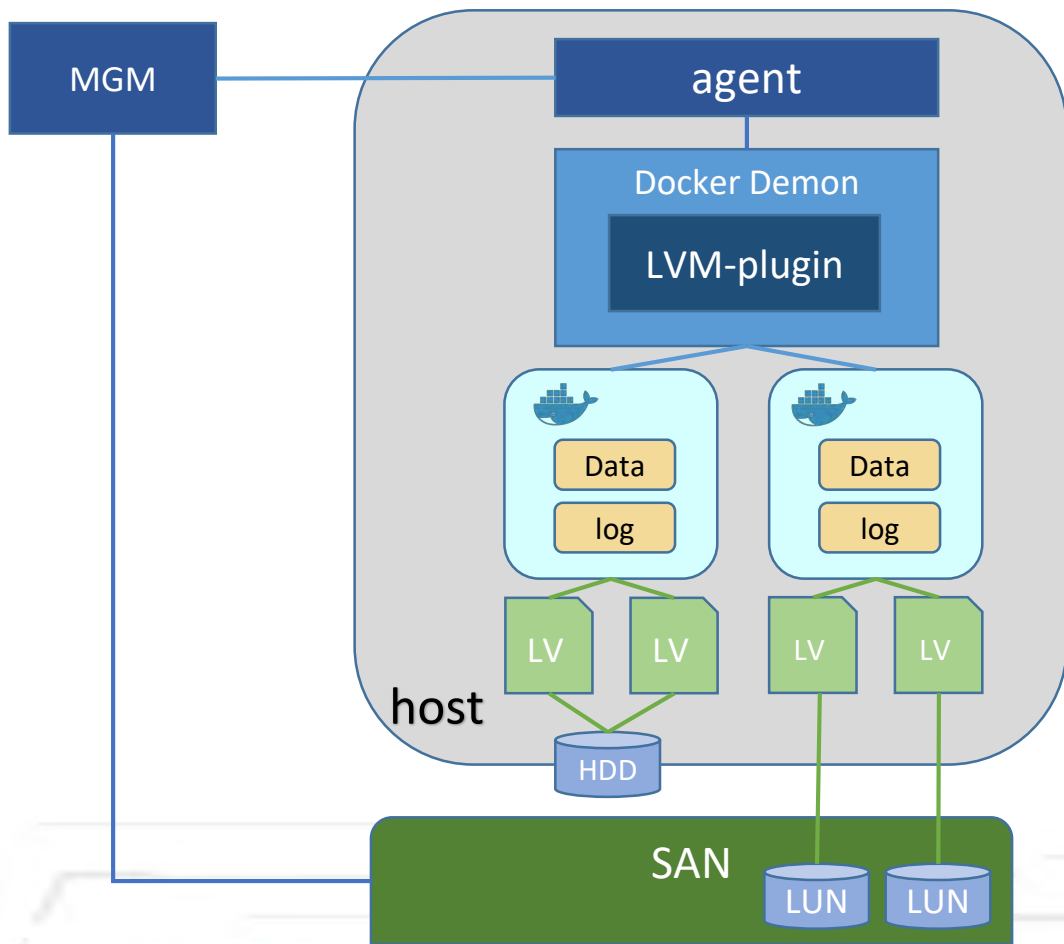


第七届



数据技术嘉年华
Data Technology Carnival

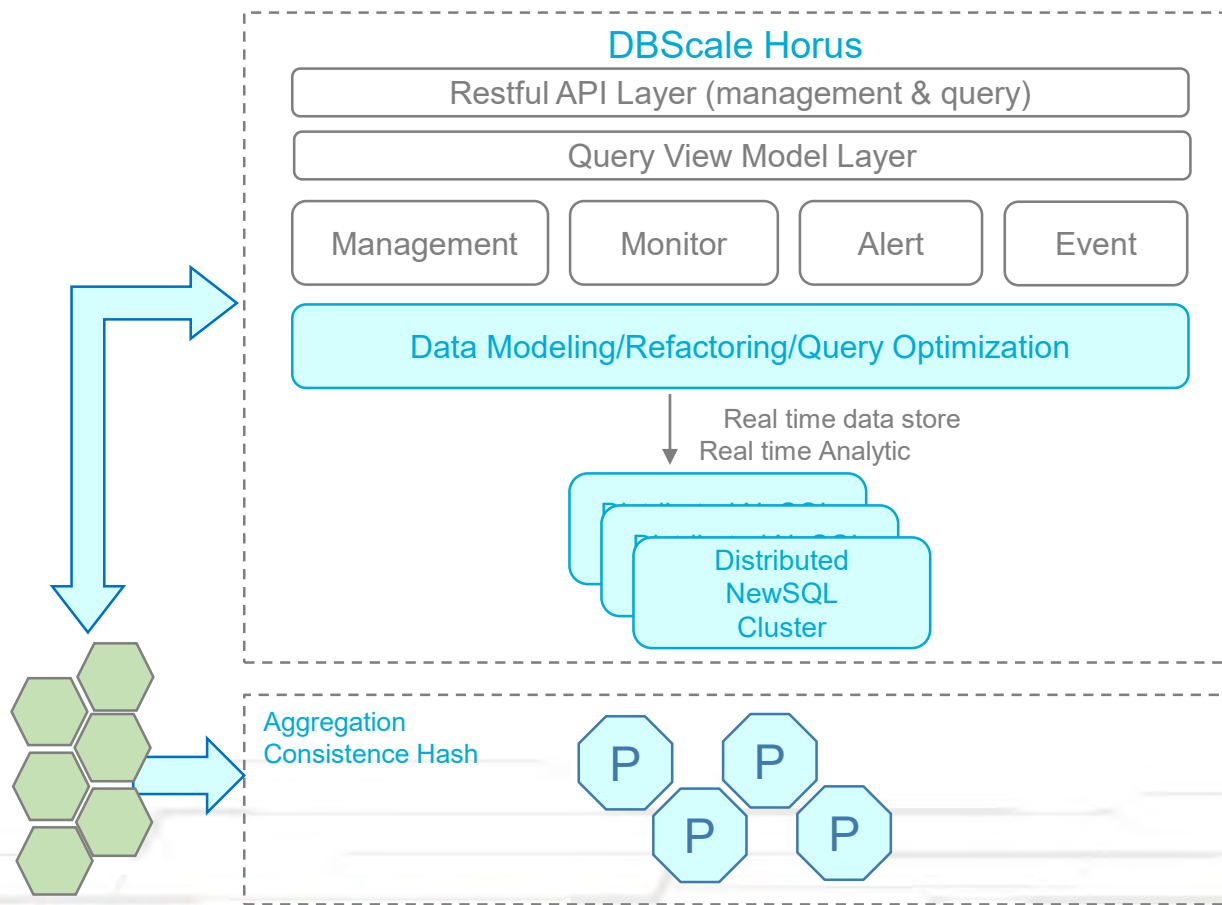




- 混合模式存储架构
- 支持本地磁盘和SAN
- 数据日志分离存储提高性能
- 使用LVM技术在线扩容



如何获得高效运维能力



- 高度可扩展性
- 整体分布式设计
- 实时告警
- 监控项动态配置
- 灵活的数据模型





平台管理资源的分配概览

CPU
分配率

内存
分配率

存储
分配率

内置
盘分配率

IP资源
分配率

端口
资源
分配率

集群
实例
分配数



第七届



数据技术嘉年华
Data Technology Carnival

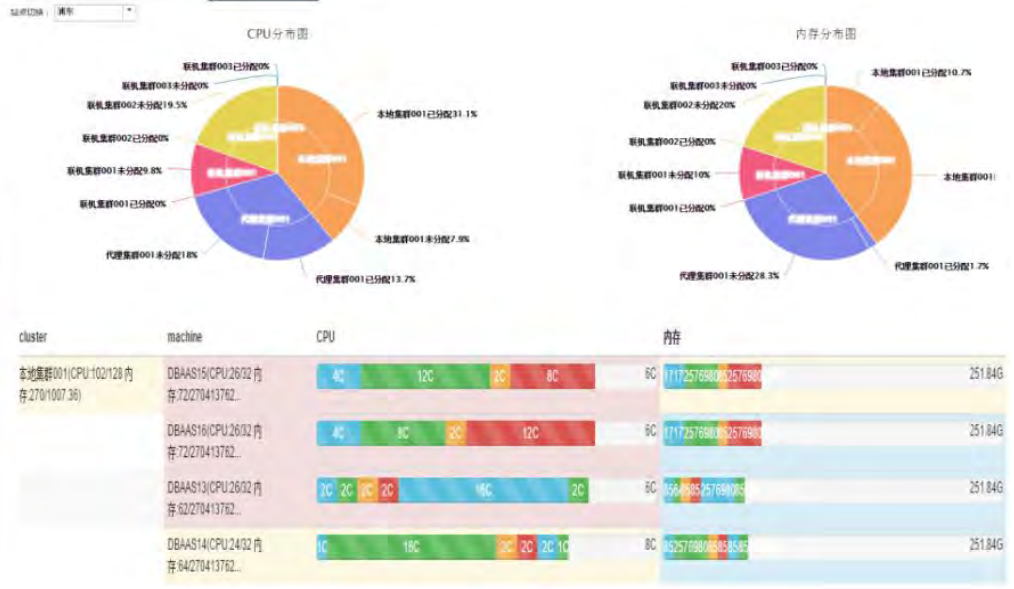


首页 站点资源监控 主机状态监控 实例状态监控 主机资源监控

所属站点: 浦东 所属集群: 代理集群001 使用状态: 可用 查询 重置

当前页码: 80 使用默认排序

主机名	所属站点	所属集群	网络地址	CPU使用率		内存使用率		HDD分配率		SDD使用率		状态	监控
				状态	分配值-总值	状态	分配值-总值	状态	分配值-总值	状态	分配值-总值		
DBAAS01	浦东	代理集群001	145.4.245.11	可用	0/40	0/252	0/2513	0/0	0/0	0/0	停用	🔍	
DBAAS11	浦东	代理集群001	145.4.245.21	可用	30/32	28/251	219/2511	0/0	0/0	0/0	停用	🔍	
DBAAS12	浦东	代理集群001	145.4.245.22	可用	15/32	14/251	109/2511	0/0	0/0	0/0	自用	🔍	
DBAAS13	浦东	本地集群001	145.4.245.23	可用	26/32	62/251	1380/2511	0/0	0/0	0/0	自用	🔍	
DBAAS14	浦东	本地集群001	145.4.245.24	可用	24/32	64/251	1380/2511	0/0	0/0	0/0	自用	🔍	
DBAAS15	浦东	本地集群001	145.4.245.25	可用	26/32	72/251	920/2511	0/0	0/0	0/0	自用	🔍	
DBAAS16	浦东	本地集群001	145.4.245.26	可用	26/32	72/251	920/2511	0/0	0/0	0/0	自用	🔍	
DBAAS17	浦东	联机集群001	145.4.245.27	可用	0/32	0/251	0/2511	0/0	0/0	0/0	自用	🔍	
DBAAS18	浦东	联机集群001	145.4.245.28	可用	0/32	0/251	0/2511	0/0	0/0	0/0	自用	🔍	
DBAAS19	浦东	联机集群002	145.4.245.29	可用	0/32	0/251	0/2511	0/0	0/0	0/0	自用	🔍	
DBAAS20	浦东	联机集群002	145.4.245.30	可用	0/32	0/251	0/2511	0/0	0/0	0/0	自用	🔍	



首页 站点资源监控 主机状态监控 实例状态监控

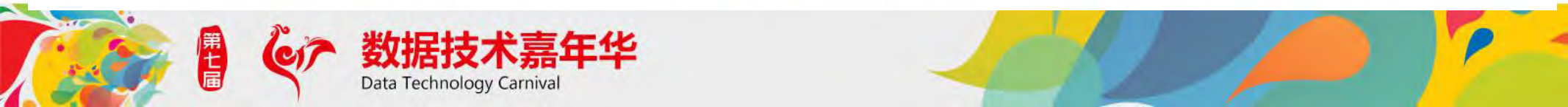
所属站点: 浦东 所属用户: 实例名称: 实例状态: 可用 查询 重置

所属用户	业务名称	实例信息			
		实例名称	实例状态	备份空间使用量	所属站点
admin	本地单机1.0	xn_bed03	可用	0.00/200.00G	浦东
		xn_bed02	可用	0.00/200.00G	浦东
		xn_bed05	可用	0.00/200.00G	浦东
		xn_bed04	可用	0.00/200.00G	浦东
		xn_bed01	可用	0.00/200.00G	浦东
	本地单机2.0	xn_bed04	可用	0.00/100.00G	浦东
		xn_bed05	可用	0.00/100.00G	浦东
		xn_bed03	可用	0.00/100.00G	浦东
		xn_bed01	可用	0.00/100.00G	浦东
		xn_bed02	可用	0.00/100.00G	浦东
本地集群1.0	xn_rep05	可用	0.00/20.00G	浦东	
	xn_rep04	可用	0.00/20.00G	浦东	
	xn_rep03	可用	0.00/20.00G	浦东	
	xn_rep01	可用	0.00/20.00G	浦东	
	xn_rep02	可用	0.00/20.00G	浦东	

首页 站点资源监控 主机状态监控 实例状态监控

所属站点: 浦东 所属集群: 可用 查询 重置

主机名	所属站点	所属集群	运行状态	管理agent状态	监控agent状态	docker状态	docker plugin状态	实例单元信息
DBAAS01	浦东	代理集群001	可用	可用	可用	可用	可用	🔍
DBAAS11	浦东	代理集群001	可用	可用	可用	可用	可用	🔍
DBAAS12	浦东	代理集群001	可用	可用	可用	可用	可用	🔍
DBAAS13	浦东	本地集群001	可用	可用	可用	可用	可用	🔍
DBAAS14	浦东	本地集群001	可用	可用	可用	可用	可用	🔍
DBAAS15	浦东	本地集群001	可用	可用	可用	可用	可用	🔍
DBAAS16	浦东	本地集群001	可用	可用	可用	可用	可用	🔍
DBAAS17	浦东	联机集群001	可用	可用	可用	可用	可用	🔍
DBAAS18	浦东	联机集群001	可用	可用	可用	可用	可用	🔍
DBAAS19	浦东	联机集群002	可用	可用	可用	可用	可用	🔍
DBAAS20	浦东	联机集群002	可用	可用	可用	可用	可用	🔍



COME & **JOIN** *US*



第七屆



数据技术嘉年华
Data Technology Carnival



一个分享交流的地方



微信号: eyygle



Long Press QR Code To
Identify The Concern

长按二维码识别关注



扫一扫，加入我们，分享更多知识



第七届



数据技术嘉年华

Data Technology Carnival





THANKS

