# 统一资源调度平台建设实践



许令波/2017.10

# 内容简介

- 为什么需要资源调度
  - 什么是调度
  - 提升稳定性和效率
  - 建设数据中心
- 如何抽象资源
  - 资源的收集和管理。
  - 虚拟化
  - 数据持久化
- 如何调度资源
  - 技术选型
  - 兼容考虑
- 遇到的难题

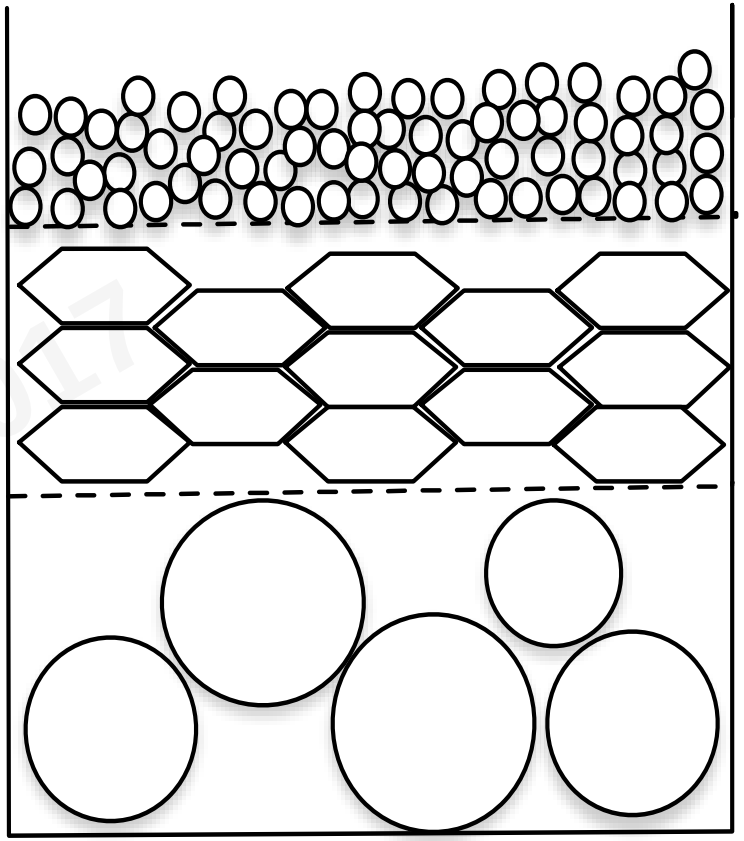
# 关于我

- 联系我：xulingbo0201@163.com
- 目前负责滴滴的资源调度平台建设，主要关注资源调度和容器化相关技术
- 之前有在阿里7年的性能优化经验
- 著有《深入分析Java Web技术内幕》

# 什么是调度

所谓资源调度一般分为两个阶段：

- 虚拟化（即资源的抽象）：虚拟机或容器技术来隔离资源
- 编排：更细粒度在时间和空间上优化资源的使用



# 一些数据

- 从物理机迁入oceanbank,一台GIFT-PROXY可省：22核，120G内存。一台GIFT-FS可省9核CPU，40G内存
- 节约成本统计：50多台GIFT-PROXY和60多台GIFT-FS共节约500多核，2T内存。折算成物理机共省下约：节省了约50% M机型物理机。

# 提升稳定性和效率

- 提升运维效率
  - 标准化运维，只要一个镜像就可以部署起来
  - 自动化，应用依赖和部署变自动化减少了人为的干预
- 提升稳定性：
  - 应用和运行的机器解耦了
  - 相应的就可以做弹性伸缩和调度

# 建设数据中心

资源的统一抽象对上层应用屏蔽时间（不会down机）和空间（机器、机房、区域等）上的差异

# 如何抽象资源

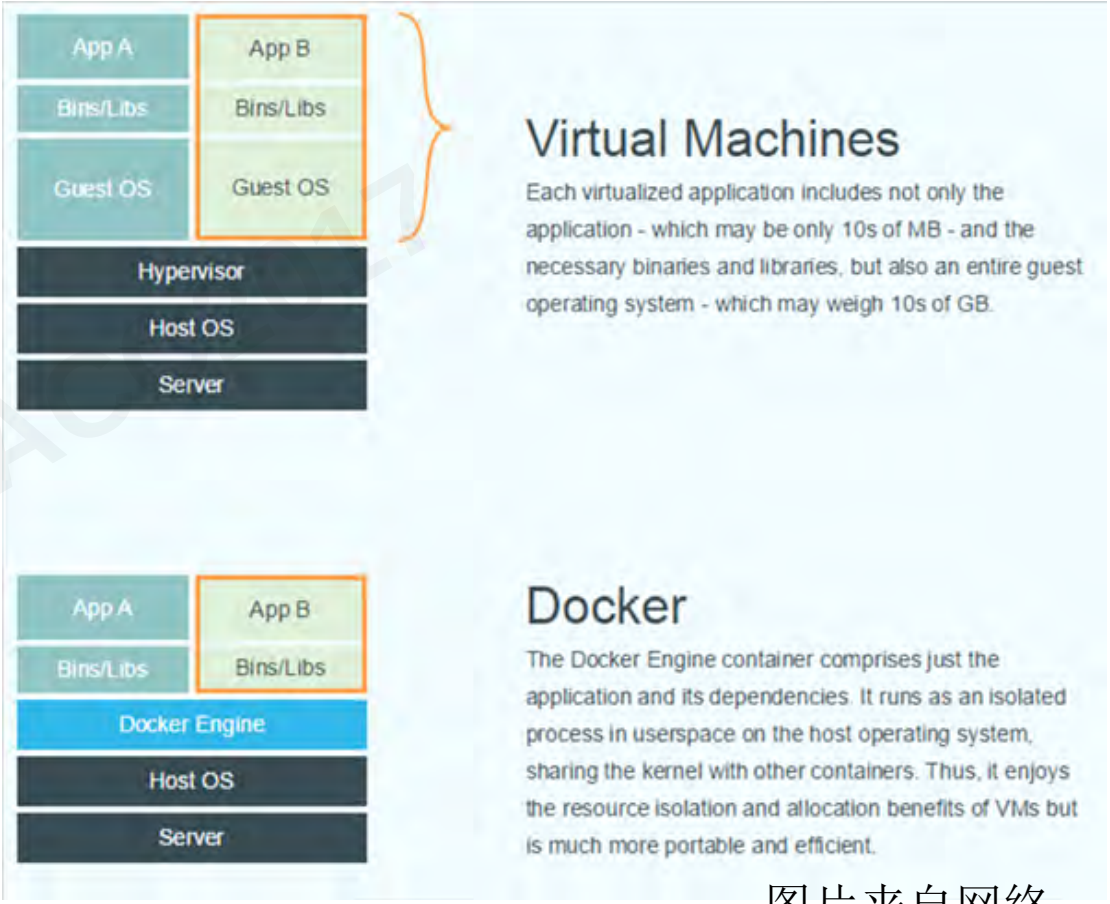
- 对具体物理资源的收集和管理。
- 虚拟化。将抽象的硬件资源属性进行重新封装，封装成上层可以使用的一个实体，可以是容器也可以是虚拟机或者一个资源集合。
- 数据持久化。业务少不了会有数据存储在本机，这样就会把机器变成有状态，不利于对资源的全局调度，所以有三种种场景需要解决：一是不需要永久本地存储但是会实时写到本地如应用的日志；二是需要永久存储如DB。三是分布式存储场景，做到存储与计算分离

# 物理资源的收集和管理

- 资源的信息管理。有多少，用了多少，还有多少。
- 大量物理机器的集群管理。除了通常几十万台的机器管理功能外，还有一分部的任务管理，如负责接收master的创建容器的task等
- 资源的合理分配策略和算法。上层的资源请求最终会在每台的物理机上进行分配以及如何分配，这里面有很多优化空间，如何保证每台物理机资源的合理利用需要合理的分配策略和算法支撑。

# 虚拟化

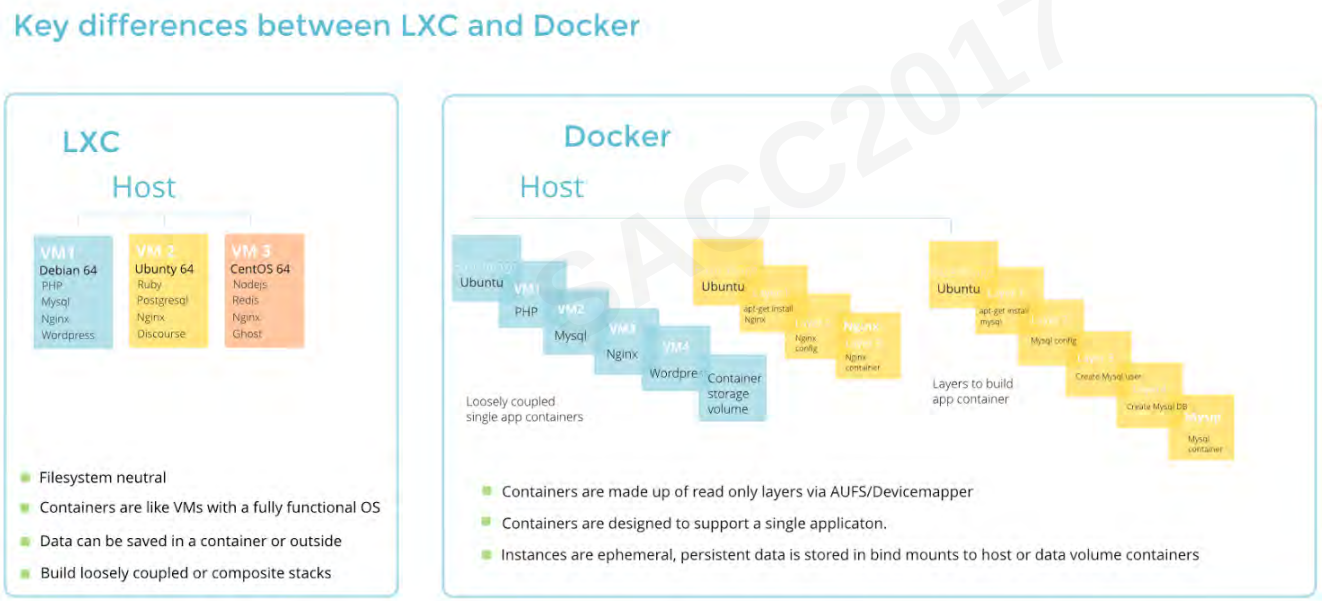
vm、docker和lxc的选择问题:




图片来自网络

# Docker vs LXC

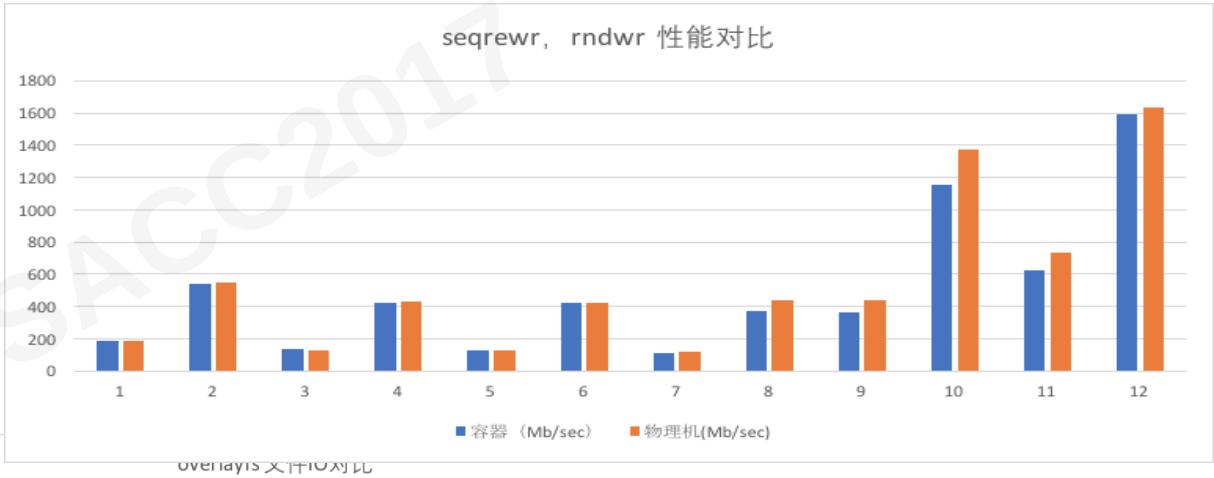
- copy on write技术可以节省磁盘
- 非volume挂载磁盘会有性能损耗

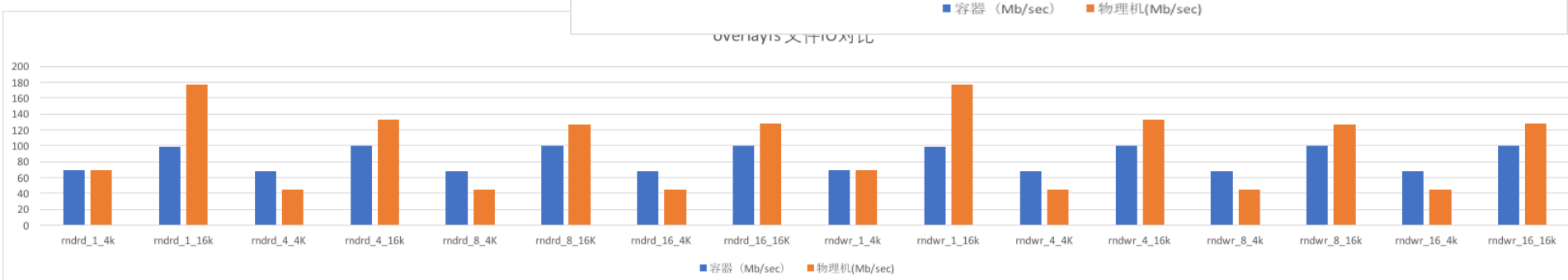



图片来自网络

# 容器存储选型

- Devicemapper性能对比
- OverlayFS性能对比

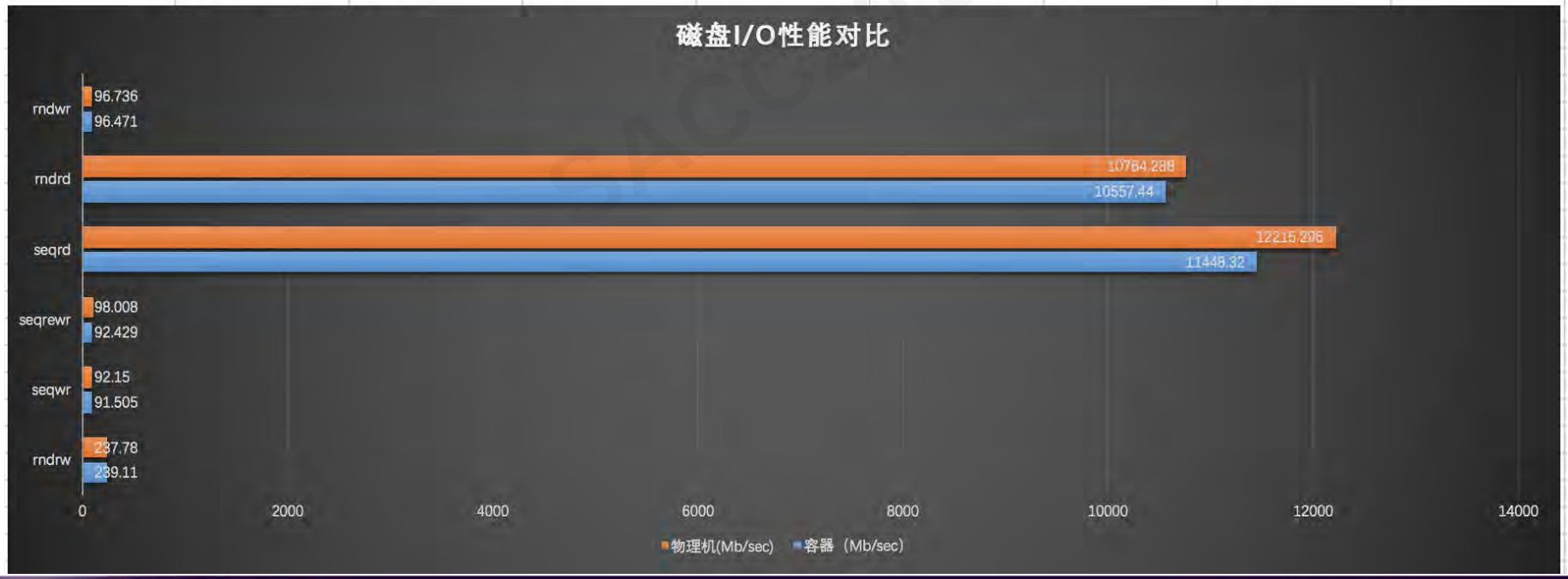


# Lxc的磁盘性能

## ■LXC与宿主机磁盘I/O性能测试对比

| 序号 | 磁盘读写方式 | 容器（Mb/sec） | 物理机(Mb/sec) | 容器(Requests/sec) | 物理机(Requests/sec) | 并发 | 块大小（字节） | 文件大小（G） |
|---|---|---|---|---|---|---|---|---|
| 1 | rndrw | 239.11 | 237.78 | 15292.89 | 15217.78 | 16 | 16384 | 3 |
| 2 | seqwr | 91.505 | 92.15 | 5856.34 | 5897.61 | 16 | 16384 | 3 |
| 3 | seqrewr | 92.429 | 98.008 | 5915.48 | 6272.49 | 16 | 16384 | 3 |
| 4 | seqrd | 11448.32 | 12215.296 | 732709.18 | 781810.57 | 16 | 16384 | 3 |
| 5 | rndrd | 10557.44 | 10764.288 | 675648.2 | 688908.61 | 16 | 16384 | 3 |
| 6 | rndwr | 96.471 | 96.736 | 6174.13 | 6191.1 | 16 | 16384 | 3 |



# 如何调度资源

- 调度框架选型
- 调度策略

# 调度框架选型

■ K8s mesos swarn等比较

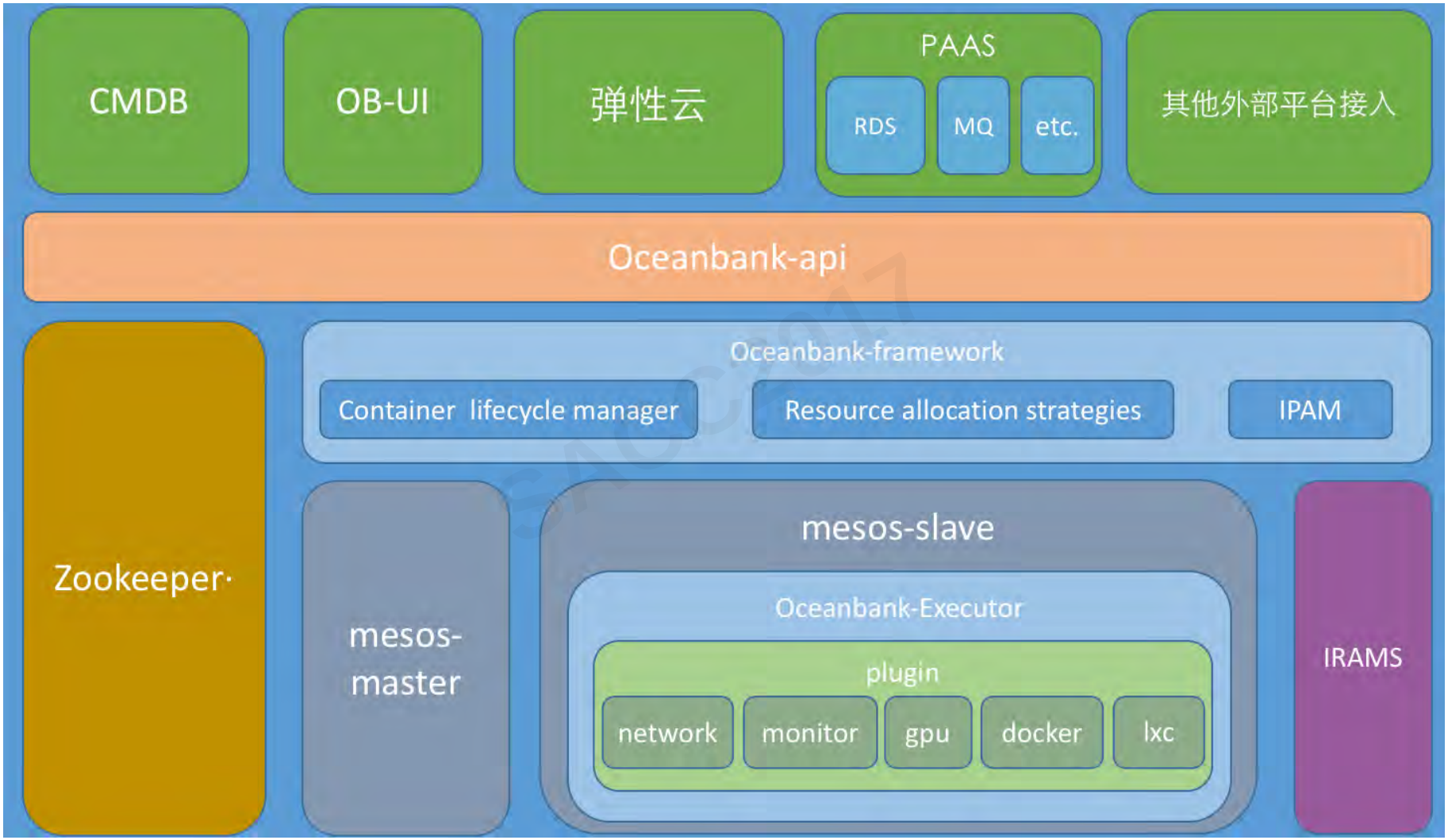
| | Framework | Architecture | Resource granularity | Multi-scheduler | Pluggable logic | Priority preemption | Re-scheduling | Oversubscription | Resource estimation | Avoid interference |
|---|---|---|---|---|---|---|---|---|---|---|
| OPEN | Kubernetes | monolithic | multi-dimensional | N[v1.2, DD, Issue] | Y[DD] | N[Issue] | N[Issue] | Y[DD] | N | N |
| | Swarm | monolithic | multi-dimensional | N | N | N[Issue] | N | N | N | N |
| | YARN | monolithic/ two-level | RAM/CPU slots | Y | N[app-lvl. only] | N[IIRA] | N | N[IIRA] | N | N |
| | Mesos | two-level | multi-dimensional | Y | Y[framework-lvl.] | N[IIRA] | N | Y[v0.23, Doc] | N | N |
| | Nomad | shared-state | multi-dimensional | Y | Y | N[Issue] | N[Issue] | N[Issue] | N | N |
| | Sparrow | fully-distributed | fixed slots | Y | N | N | N | N | N | N |
| CLOSED | Borg | monolithic[2] | multi-dimensional | N[2] | N[2] | Y | Y | Y | Y | N |
| | Omega | shared-state | multi-dimensional | Y | Y | Y | Y | Y | Y | N |
| | Apollo | shared-state | multi-dimensional | Y | Y | Y | Y | N | N | N |

图片来自网络

# 我们最关心的几个因素

- 支持的规模、稳定性
  - Mesos支撑>30000nodes;250k containers
  - Twiter/Apple/Uber/ebay/airbnb等
- 兼容性、接入成本
  - 支持大数据集群的接入
  - 支持各种中间件的接入

# 统一调度的架构


CMDB
OB-UI
弹性云
PAAS
RDS
MQ
etc.
其他外部平台接入
Oceanbank-api
Zookeeper·
Oceanbank-framework
Container lifecycle manager
Resource allocation strategies
IPAM
mesos-master
mesos-slave
Oceanbank-Executor
plugin
network
monitor
gpu
docker
lxc
IRAMS

# 统一调度的架构

# 中间件接入

- 提供容器asVM，将容器做成VM一样的体验
- 利用LXCFS隔离CPU、men和Load信息

# 实现LXCFS的load隔离

- 计算公式：**load(t) = load(t-1) *exp(-5/60R) + n(t)* (1 – exp(5/60R))**
- 如果想在容器中获取正确Loadavg信息的那么就要具备以下几点：
  - 获取运行在容器中的所有进程（包括：线程）。
  - 获取运行在容器中的进程总数。
  - 获取运行在容器中的所有进程运行状态。
  - Loadavg计算公式。
- 难点是如何控制性能消耗
  - 控制对进程信息获取的系统调用

# 大数据接入

- 直接提供容器as vm给大数据使用，直接将yarn部署在容器里
- 在离线混部需要考虑磁盘的存储容量问题
- 资源动态调度支持

# 遇到的坑

1. Java线程夯住的问题排查过程
2. lxc容器cgroup被修改问题分析
3. 【源码分析】LXC存在配置文件解析的问题
4. Ob系统内存分配失败oom问题分析
5. LXCFS获取proc文件长度失败排查

# 示例1-Java线程夯住的问题排查过程

```
[arius@arius-compute-vir002 ~]$ pstack 568
Thread 1 (process 568):
#0  0x00007f15272399ed in read () from /lib64/libc.so.6
#1  0x00007f15271c99b0 in __GI__IO_file_underflow () from /lib64/libc.so.6
#2  0x00007f15271ca93e in __GI__IO_default_uflow () from /lib64/libc.so.6
#3  0x00007f15271c5bce in getc () from /lib64/libc.so.6
#4  0x00007f1445389cb8 in get_totalticks () from /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.65-3.b17.el7.x86_64/jre/lib/amd64/libmanagement.so
#5  0x00007f144538a12e in get_cpuload_internal () from /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.65-3.b17.el7.x86_64/jre/lib/amd64/libmanagement.so
#6  0x00007f144538a2e5 in get_cpu_load () from /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.65-3.b17.el7.x86_64/jre/lib/amd64/libmanagement.so
#7  0x00007f15117584ae in ?? ()
#8  0x000000046c3df050 in ?? ()
#9  0x00007f11bc000a30 in ?? ()
#10 0x00007f11c4002ac8 in ?? ()
#11 0x0000000000000000 in ?? ()
```

```
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, "", 4096)                    = 0
read(7392, ^CProcess 568 detached
```

# 示例2- Ob系统内存分配失败问题分析

```
[7180785.639634] odin-metrics invoked oom-killer: gfp_mask=0xd0, order=0, oom_score_adj=0
[7180785.640772] odin-metrics cpuset=e9605ef9cf58d248872a2f9c2992d26e0f8b299b923bb4705f5f4c7c7ef9b02c mems_allowed=0-1
[7180785.641817] CPU: 21 PID: 19149 Comm: odin-metrics Tainted: G        W  OE K------------   3.10.0-514.16.1.el7.x86_64
[7180785.642859] Hardware name: INSPUR SA5212M4/Shuyu, BIOS 4.0.7 03/29/2016
[7180785.642864]  ffff880bb2e89f60 000000005cb60e37 ffff880100aefcc0 ffffffff81686ac3
[7180785.642865]  ffff880100aefd50 ffffffff81681a6e ffff8804d5ff9d80 0000000000000001
[7180785.642866]  0000000000000000 0000000000000000 0000000000000046 ffffffff811847c6
[7180785.642867] Call Trace:
[7180785.642874]  [<ffffffff81686ac3>] dump_stack+0x19/0x1b
[7180785.642880]  [<ffffffff81681a6e>] dump_header+0x8e/0x225
[7180785.642886]  [<ffffffff811847c6>] ? find_lock_task_mm+0x56/0xc0
[7180785.642889]  [<ffffffff81184c7e>] oom_kill_process+0x24e/0x3c0
[7180785.642891]  [<ffffffff8118471d>] ? oom_unkillable_task+0xcd/0x120
[7180785.642896]  [<ffffffff81093c0e>] ? has_capability_noaudit+0x1e/0x30
[7180785.642899]  [<ffffffff811f3651>] mem_cgroup_oom_synchronize+0x551/0x580
[7180785.642901]  [<ffffffff811f2aa0>] ? mem_cgroup_charge_common+0xc0/0xc0
[7180785.642904]  [<ffffffff81185504>] pagefault_out_of_memory+0x14/0x90
[7180785.642905]  [<ffffffff8167f8da>] mm_fault_error+0x68/0x12b
[7180785.642909]  [<ffffffff81692885>] __do_page_fault+0x395/0x450
[7180785.642944]  [<ffffffffa0626c99>] ? xfs_file_aio_write_checks+0x99/0x1b0 [xfs]
[7180785.642946]  [<ffffffff81692975>] do_page_fault+0x35/0x90
[7180785.642948]  [<ffffffff8168eb88>] page_fault+0x28/0x30
[7180785.642951] Task in /docker/e9605
b4705f5f4c7c7ef9b02c
[7180785.642952] memory: usage 4194304
[7180785.642952] memory+swap: usage 41
[7180785.642953] kmem: usage 0kB, limi
[7180785.642969] Memory cgroup stats f
B inactive_anon:1487408KB active_anon:
[7180785.642969] [ pid ]   uid  tgid t
[7180785.643138] [ 5121]     0  5121
[7180785.643141] [19654]    81 19654
[7180785.643143] [19761]     0 19761
[7180785.643145] [20065]     0 20065
[7180785.643146] [22312]     0 22312
```

| filename | size | total_pages | min_cached page | cached_pages | cached_size | cached_perc |
|---|---|---|---|---|---|---|
| system.journal | 240 | 1 | 0 | 1 | 4,096 | 100.00 |
| system@000555a54c4a84db-bbfee6cc48d586fb.journal~ | 67,108,864 | 16,384 | 0 | 14,579 | 59,715,584 | 88.98 |
| system@0000555a568e7ec9a-5d3332180b6192fd.journal~ | 50,331,648 | 12,288 | 0 | 9,131 | 37,400,576 | 74.31 |
| system@000555a57e6a45ae-af32585fb9da12f3.journal~ | 33,554,432 | 8,192 | 0 | 6,567 | 26,898,432 | 80.16 |
| system@0000555a5a5b8e743-576cc325e4e0a08c.journal~ | 25,165,824 | 6,144 | 0 | 4,391 | 17,985,536 | 71.47 |
| system@000555ab11589ab2-78c257f546752ea5.journal~ | 25,165,824 | 6,144 | 0 | 4,894 | 20,045,824 | 79.65 |
| system@0000555bc556b66bc-38ab27392a6a03a0.journal~ | 50,331,648 | 12,288 | 0 | 9,859 | 40,382,464 | 80.23 |
| system@000555bf5da3517a-0fb5086ee22ac6a7.journal~ | 16,777,216 | 4,096 | 0 | 2,851 | 11,677,696 | 69.60 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000000001-000554959a2b6a9a.journal | 75,497,472 | 18,432 | 0 | 16,442 | 67,346,432 | 89.20 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000013e27-000554b81b767e5a.journal | 83,886,080 | 20,480 | 0 | 17,035 | 69,775,360 | 83.18 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000027ad9-000554d85ffa6f97.journal | 75,497,472 | 18,432 | 0 | 16,520 | 67,665,920 | 89.63 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000003b803-000554fc69ced546.journal | 75,497,472 | 18,432 | 0 | 16,528 | 67,698,688 | 89.67 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000004f56f-0005552058daac89.journal | 75,497,472 | 18,432 | 0 | 16,479 | 67,497,984 | 89.40 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000063210-000555449933fd42.journal | 75,497,472 | 18,432 | 0 | 16,241 | 66,523,136 | 88.11 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000076 7fc-0005556a4daef4d3.journal | 75,497,472 | 18,432 | 0 | 16,020 | 65,617,920 | 86.91 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000008 97f7-000555591429ce345.journal | 92,274,688 | 22,528 | 0 | 20,977 | 85,921,792 | 93.12 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000009e097-0005559bc8a7a1e1.journal | 125,829,120 | 30,720 | 0 | 23,221 | 95,113,216 | 75.59 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000b409a-0005559c09ef1a7e.journal | 125,829,120 | 30,720 | 0 | 23,207 | 95,055,872 | 75.54 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000ca0fb-0005559c4ebd0065.journal | 125,829,120 | 30,720 | 0 | 23,271 | 95,318,016 | 75.75 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000e0297-0005559c96acd8d2.journal | 125,829,120 | 30,720 | 0 | 23,209 | 95,064,064 | 75.55 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000f622d-0005559ce0bc79f8.journal | 125,829,120 | 30,720 | 0 | 23,219 | 95,105,024 | 75.58 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000010c2dc-0005559d24b823c1.journal | 125,829,120 | 30,720 | 0 | 23,218 | 95,100,928 | 75.58 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000012333b-0005559d69b32972.journal | 125,829,120 | 30,720 | 0 | 23,260 | 95,272,960 | 75.72 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000013844d7-0005559dafdfe3f8.journal | 125,829,120 | 30,720 | 0 | 23,266 | 95,297,536 | 75.74 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000014e6f0-0005559dfd4fa694.journal | 125,829,120 | 30,720 | 0 | 23,237 | 95,178,752 | 75.64 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000164836-0005559e4d8f736a.journal | 125,829,120 | 30,720 | 0 | 23,274 | 95,330,304 | 75.76 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000017aae2-0005559e9cb0bd10.journal | 125,829,120 | 30,720 | 0 | 23,199 | 95,023,104 | 75.52 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000190b21-0005559eeb5b3f31.journal | 125,829,120 | 30,720 | 0 | 23,258 | 95,264,768 | 75.71 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000001a6d79-0005559f3ead46de.journal | 125,829,120 | 30,720 | 0 | 23,173 | 94,916,608 | 75.43 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000001bcd4a-0005559f99845e47.journal | 125,829,120 | 30,720 | 0 | 23,202 | 95,035,392 | 75.53 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000001d2e60-000555a000e36503.journal | 125,829,120 | 30,720 | 0 | 23,203 | 95,039,488 | 75.53 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000001e8f5d-000555a06173116b.journal | 125,829,120 | 30,720 | 0 | 23,190 | 94,986,240 | 75.49 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000001fefd6-000555a0c510cbe4.journal | 125,829,120 | 30,720 | 0 | 23,209 | 95,064,064 | 75.55 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000002150f7-000555a12f2444b5.journal | 125,829,120 | 30,720 | 0 | 23,151 | 94,826,496 | 75.36 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000022b0dd-000555a19e7f05e3.journal | 125,829,120 | 30,720 | 0 | 23,189 | 94,982,144 | 75.49 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000024114b-000555a20e075413.journal | 125,829,120 | 30,720 | 0 | 23,220 | 95,109,120 | 75.59 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000257256-000555a26ecdd448.journal | 125,829,120 | 30,720 | 0 | 23,161 | 94,867,456 | 75.39 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-000000000026d22d-000555a2d447521e.journal | 125,829,120 | 30,720 | 0 | 23,217 | 95,096,832 | 75.58 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000283376-000555a3480246cc.journal | 125,829,120 | 30,720 | 0 | 23,177 | 94,932,992 | 75.45 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-0000000000299456-000555a3d5fceda5.journal | 125,829,120 | 30,720 | 0 | 23,198 | 95,019,008 | 75.51 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000002af510-000555a43b836fd3.journal | 125,829,120 | 30,720 | 0 | 23,194 | 95,002,624 | 75.50 |
| system@aaee2d0cc3634b1ab3df1b79c81ff554-00000000002c5548-000555a4a9985542.journal | 125,829,120 | 30,720 | 0 | 23,216 | 95,092,736 | 75.57 |

total_cached size: 3,244,253,184


THANKS