高可用架构设计与实践

讲师：孙玄@58

法律声明

【声明】

本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关于我

- 58集团技术委员主席
- 58同城高级系统架构师
- 即时通讯、转转、C2C技术负责人
- 前百度高级工程师
- 代表58同城对外嘉宾分享
  - QCon
  - SDCC
  - DTCC
  - Top100
  - 程序员
  - UPYUN
  - TINGYUN
  - ……

代表58对外交流

- Qcon(全球软件开发大会)
- SDCC(中国开发者大会)
- Top100(全球案例研究峰会)
- DTCC(中国数据库技术大会)
- 《程序员》撰稿2次
- 58技术发展这10年[计划中]

炼数成金课程

课程

- 《MongoDB实战》
  - 已开课
  - 欢迎大家报名学习
- 《大规模高性能分布式存储系统设计与实现》
  - 已开课
  - 欢迎大家报名学习

上次课回顾

- 为什么需要缓存?
- 缓存适用的场合
- 高可用架构使用缓存类型有哪些？各自作用是什么?
  - local、进程、分布式
- 高可用架构使用分布式缓存类型？如何选择?
- 高可用架构缓存冗余如何设计?
- 高可用架构缓存一致性如何保证?
- 高可用架构缓存命中率如何保证?
- 高可用架构缓存设计的最佳实践是什么?
- 我们的实践案例;

OutLine

- 性能评估目的
- 性能相关环节
- 性能评估工具
- 性能评估方法
- 扩容考虑因素

性能评估目的

- **找出系统性能瓶颈**
  - 硬件瓶颈
    - CPU、内存、磁盘I/O、网络I/O
  - 软件瓶颈
    - 应用程序设计缺陷（数据库连接释放时机）、数据库连接数滥用、数据库查询的滥用等等
- **提供性能优化方案**
  - 硬件问题？升级硬件
    - Scale UP或者Scale Out
  - 软件问题？
    - 改进应用程序架构、设计
- **达到硬件和软件的合理配置，使系统资源使用平衡**
  - CPU型
  - 存储型
  - ……

性能相关环节

- 硬件
  - CPU
  - 内存
  - 磁盘I/O
  - 网络I/O
- 应用软件
  - 软件瓶颈

性能评估工具(Linux)

- 硬件
  - CPU性能
    - vmstat
  - 内存性能
    - free
  - 磁盘I/O性能
    - iostat、sar
  - 网络I/O性能
    - ifstat
  - 系统整体性能
    - top
- 软件
  - 压力测试

硬件性能评估

CPU性能评估

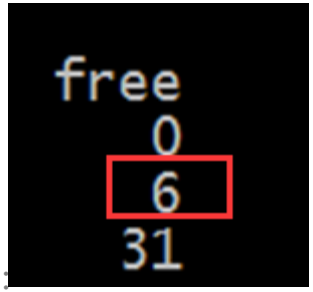
- vmstat工具
  - vmstat 1

```
[im_work@im20 ~]$ vmstat 1
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 4  0      0 213892  12524 7049076    0    0     6    65    0    0  2  1 97  0  0
 1  0      0 208072  12524 7050968    0    0     0   556 31192 41392  4  1 94  0  0
 2  0      0 204312  12532 7053008    0    0     0   252 34326 50376  5  2 93  0  0
 0  0      0 195672  12540 7055372    0    0     0   104 33586 47914  6  2 92  0  0
 3  0      0 192604  12540 7058540    0    0     0     0 39280 77646  4  2 94  0  0
 1  0      0 190352  12540 7061228    0    0     0     0 34599 54351  5  2 94  0  0
 0  0      0 187120  12540 7063924    0    0     0 10312 34161 50275  5  1 94  0  0
 2  0      0 185444  12540 7066816    0    0     0     0 32462 45793  4  1 94  0  0
 6  0      0 177496  12548 7069380    0    0     0   184 33140 52328  4  2 94  0  0
 4  0      0 202492  12540 7041264    0    0     0     0 42355 91337  6  3 91  0  0
 4  0      0 199524  12540 7044392    0    0     0     0 32830 50397  4  1 95  0  0
 2  0      0 198040  12540 7047392    0    0     0   440 33855 53143  5  1 94  0  0
 1  0      0 198352  12540 7049808    0    0     0     0 34321 55552  5  2 94  0  0
 3  0      0 197116  12548 7052508    0    0     0   180 33786 51308  5  2 94  0  0
 1  0      0 192628  12548 7055792    0    0     0     0 40719 90678  5  3 92  0  0
 2  0      0 191080  12548 7059340    0    0     0     0 33282 55541  4  2 94  0  0
 0  0      0 188712  12548 7061784    0    0     0 12136 35160 55368  5  2 93  0  0
 3  0      0 186648  12548 7064400    0    0     0     0 33265 51725  5  1 94  0  0
 1  0      0 184192  12556 7067064    0    0     0   156 34288 50773  5  1 94  0  0
 2  0      0 179004  12556 7071820    0    0     0     0 46595 108615  7  3 90  0  0
 3  0      0 202308  12532 7048324    0    0     0     0 35622 58495  5  2 93  0  0
 1  0      0 198588  12540 7051252    0    0     8 62972 34592 56494  5  2 93  0  0
```

- procs
  - r：运行和等待CPU时间片的进程数，如果长期大于系统CPU的个数，CPU遇到瓶颈，需要扩展CPU。
  - b：等待资源的进程数，比如正在等待磁盘I/O、网络I/O等。
- cpu
  - us：用户进程消耗CPU时间百分比，us值高，用户进程消耗CPU时间多，如果长期大于50%，优化程序；
  - sy：内核进程消耗的CPU时间百分比；
  - us + sy参考值为80%，如果us + sy大于80%，说明可能存在CPU不足。
- 问题
  - 系统CPU整体利用率不高，而应用响应缓慢？
    - 单线程？多个线程？
    - CPU闲置

硬件性能评估

内存性能评估



- free工具
  - free [-m/-g]
  - 应用程序可用内存数量：
- 经验值
  - 应用程序可用内存/系统物理内存＞70% 内存充足
  - 应用程序可用内存/系统物理内存<20% 内存不足，需要增加内存
  - 20%<应用程序可用内存/系统物理内存<70%内存基本够用
  - 下图：
    - 6/30=0.2

```
[im_work@im20 ~]$ free -g
             total       used       free     shared    buffers     cached
Mem:            31         31          0          0          0          6
-/+ buffers/cache:         24          6
Swap:           31          0         31
```

硬件性能评估

磁盘I/O性能评估

– iostat工具

```
[im_work@imtest ~]$ iostat -xdk 1
Linux 2.6.18-348.el5 (imtest)   2015年07月13日

Device:         rrqm/s   wrqm/s   r/s   w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
sda               4.74   262.75  4.54 28.13   289.82  1204.05    91.47     0.06    1.97   0.10   0.34
sda1              0.00     0.00  0.00  0.00     0.00     0.00    25.11     0.00    3.20   3.18   0.00
sda2              3.56     4.21  0.77  0.44    17.32    18.67    59.57     0.06   51.08   4.39   0.53
sda3              1.18   258.54  3.77 27.68   272.50  1185.39    92.69     0.00    2.46   0.03   0.08
sdb             111.33   409.28 155.66  4.49 18636.49  1655.11   253.40     0.07    0.44   0.23   3.69
sdc              25.68   396.86 44.56 51.33  4889.59  2046.69   144.68     0.07    0.76   0.66   6.31

Device:         rrqm/s   wrqm/s   r/s   w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00     0.00  2.00  2.00    12.00     8.00    10.00     0.03    8.00   5.75   2.30
sda1              0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda2              0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda3              0.00     0.00  2.00  2.00    12.00     8.00    10.00     0.03    8.00   5.75   2.30
sdb               0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdc               0.00     0.00  0.00  3.00     0.00    12.00     8.00     0.08   26.67  12.33   3.70

Device:         rrqm/s   wrqm/s   r/s   w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00  2816.00  0.00 86.00     0.00 11608.00   269.95    10.46  121.62   3.20  27.50
sda1              0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda2              0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda3              0.00  2816.00  0.00 86.00     0.00 11608.00   269.95    10.46  121.62   3.20  27.50
sdb               0.00     0.00  0.00  0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdc               0.00     0.00  0.00  1.00     0.00     4.00     8.00     0.01   11.00  11.00   1.10
```

- iostat –xdk 1
- 磁盘块设备分布
- rkB/s每秒读取数据量kB;
- wkB/s每秒写入数据量kB;
- svctm I/O请求的平均服务时间，单位毫秒;
- await I/O请求的平均等待时间，单位毫秒；值越小，性能越好;
- util 一秒中有百分几的时间用于I/O操作。接近100%时，表示磁盘带宽跑满，需要优化程序或者增加磁盘;
- rkB/s、wkB/s根据系统应用不同会有不同的值，但有规律遵循：长期、超大数据读写，肯定不正常，需要优化程序读取。
- svctm的值与await的值很接近，表示几乎没有I/O等待，磁盘性能好，如果await的值远高于svctm的值，则表示I/O队列等待太长，需要优化程序或更换更快磁盘。

硬件性能评估

网络I/O性能评估

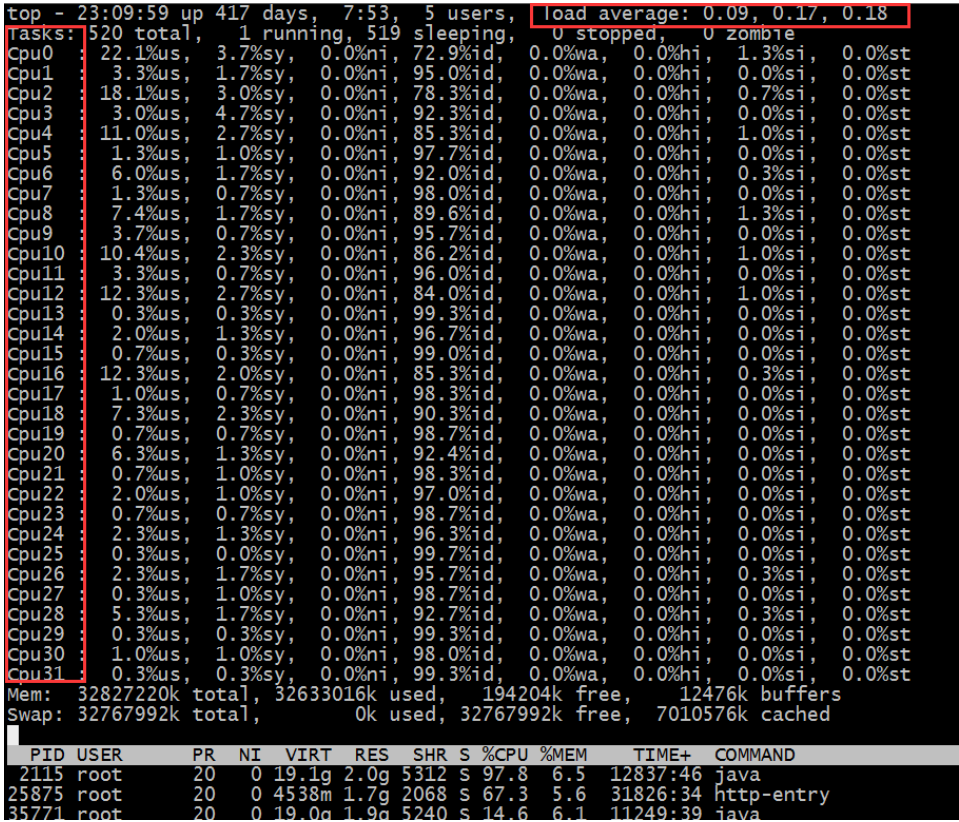
- ifstat工具
  - ifstat 1
  - 各个网卡的in、out
  - 观察网络负载情况
  - 程序网络读写是否正常
    - 程序网络I/O优化
    - 增加网络I/O带宽

```
[im_work@imtest ~]$ ifstat 1
       eth0                    eth1
 KB/s in  KB/s out   KB/s in  KB/s out
   54.47     13.49      0.07      0.00
   53.40      3.46      0.00      0.00
   49.79      5.00      0.07      0.00
   42.69      2.38      0.13      0.00
   46.98      3.16      0.00      0.00
   45.15      9.76      0.07      0.00
   40.76      4.86      0.13      0.00
   39.54      2.91      0.00      0.00
   40.50      9.37      0.00      0.00
```

硬件性能评估

系统整体性能评估

- top工具
  - top
- Load average
  - 任务队列的平均长度
    - 1分钟、5分钟、15分钟前到现在平均值
  - 这三个值的大小一般不能
  - 大于系统CPU的核数，如果
  - 三个值长期大于CPU的核数
  - 说明CPU很繁忙，负载很高
  - 影响机器整体系统；
  - 相反如果这三个值小于CPU
  - 个数，表示CPU比较空闲。
  - 图中：
    - 32个CPU，load average<32。



应用程序性能评估

应用程序性能评估

- 压力测试
  - 确认模块的最大吞吐量
    - QPS、TPS
  - 模块单机压力极限衡量标准
    - CPU极限
      - 线程32
      - CPU3200%
    - 内存极限
      - 内存32G
      - 应用占完了，排除内存泄露的可能
    - 磁盘I/O极限
      - util 100%
    - 网络I/O极限
      - 100Mb
      - in、out带宽跑满了
  - 日志
    - 根据日志统计出最大的QPS、TPS

应用程序性能评估

应用程序性能评估

- 压力测试
  - 压力工具
    - http
      - AB
    - TCP
      - 自己写
      - 开源
  - 线下机器与线上机器差异
    - 测试机器虚拟机或者配置低
    - 实际线上模块服务能力会更好
      - 1倍或者更多的提升
      - 根据机器配置差异以及程序本身
      - 或者拿线上机器同样配置做压力测试

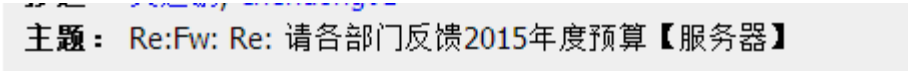
扩容考虑因素

系统使用和优化原则

- **始终保留一定量的空闲资源**
  - 多少合适？根据应用的特点，比如是否有突发性使用增长？
  - 一般情况下，保留至少50%的系统资源，以应付流量增长(流量翻一倍)；
  - 一般情况下，资源使用率达到80%时，需要考虑扩容的事儿。
- **系统硬件达到合理的配置，资源消耗均衡为目标**
  - 系统性能的木桶理论
- **应用程序对资源的使用要均衡（理想目标）**
  - 怎么样就算是均衡了？这个问题貌似比较难？！
  - 理想状况为：CPU消耗到50%的时候，磁盘I/O也到50%，网络I/O也到50%，内存使用也到50%；

扩容考虑因素

- 运维预计2015年机器
  - 如何评估机器，扩容?
    - 模块考虑2个因素
      - 目前机器的性能扩容
        - » CPU、内存、磁盘I/O、网络I/O、整体系统性能
      - 业务增长需要扩容机器
        - » 和PM同学目前业务目标
          - 帮帮装机XXW，日活XXW~XXW
        - » 目前业务目标到机器的转换关系
          - 一台机器最大承受的QPS（压力测试）
          - 帮帮日活到机器的折算方法
          - 得到需要扩容的机器
    - DB机器
      - 业务流量增长，DB带来的压力增长
      - DB每个库、表服务能力
      - 确定DB机器增加情况
    - 实例



本课总结

- 性能评估目的
- 性能相关环节
- 性能评估工具
- 性能评估方法
- 扩容考虑因素

THANK YOU