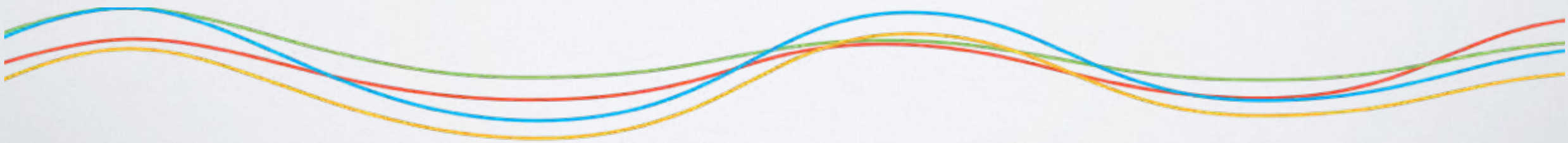
数据库优化经验谈

肖鹏



目录

- 介绍
- 优化思路
- 优化实践
- 总结

介绍

个人介绍

- 肖鹏
- 新浪微博研发中心数据库技术经理
- 10年数据库从业经验，专注于数据库领域的高可用，高可靠和架构优化方面，对于大规模集群运维管理有一定的心得

服务类型

消息队列 MCQ Kafka

缓存层 MC Redis

持久存储 MySQL Redis HBase

新浪微博数据库概括

MySQL

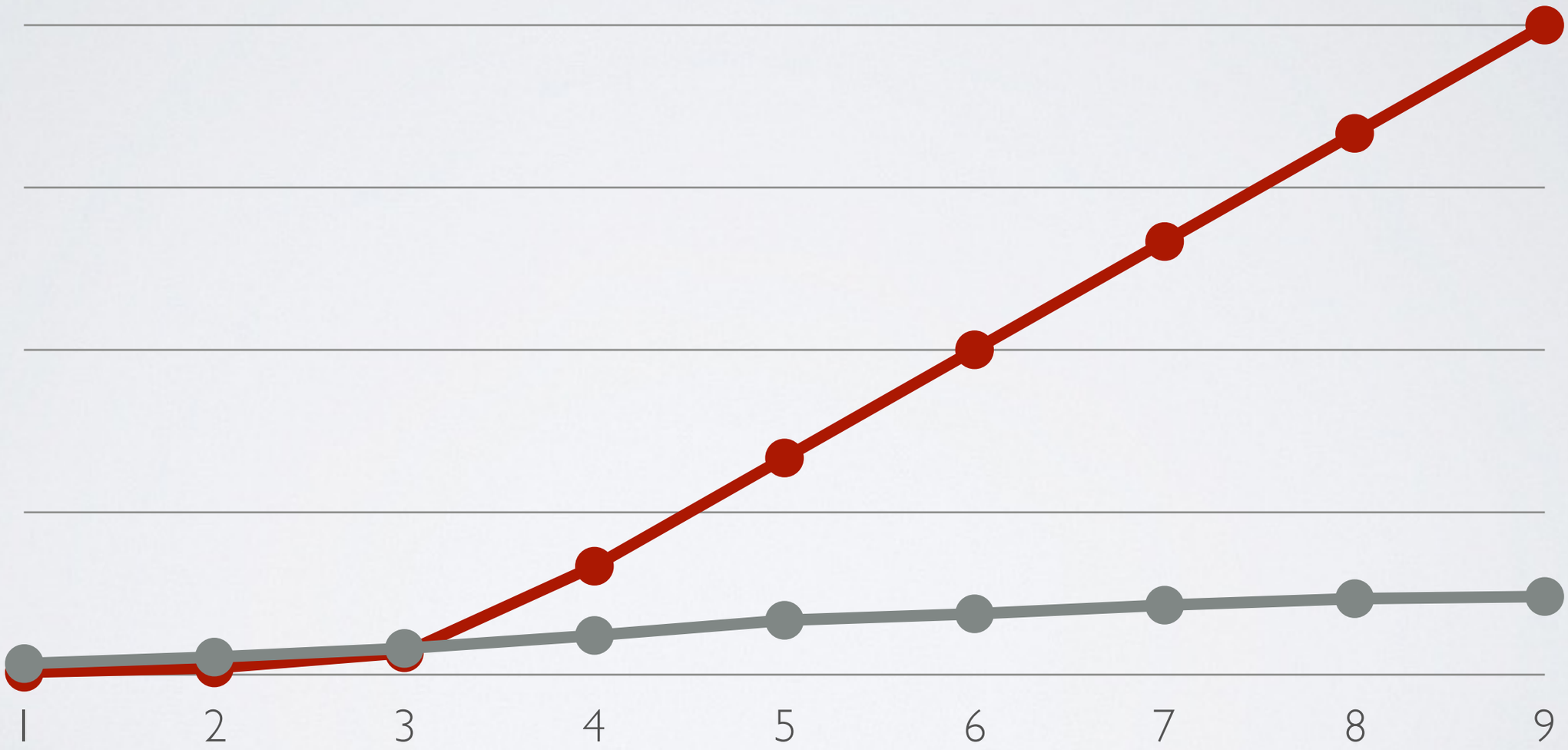
2k＋服务器，300T，4k的实例，百亿

NoSQL

2k＋服务器，70T，1.6w的实例，万亿

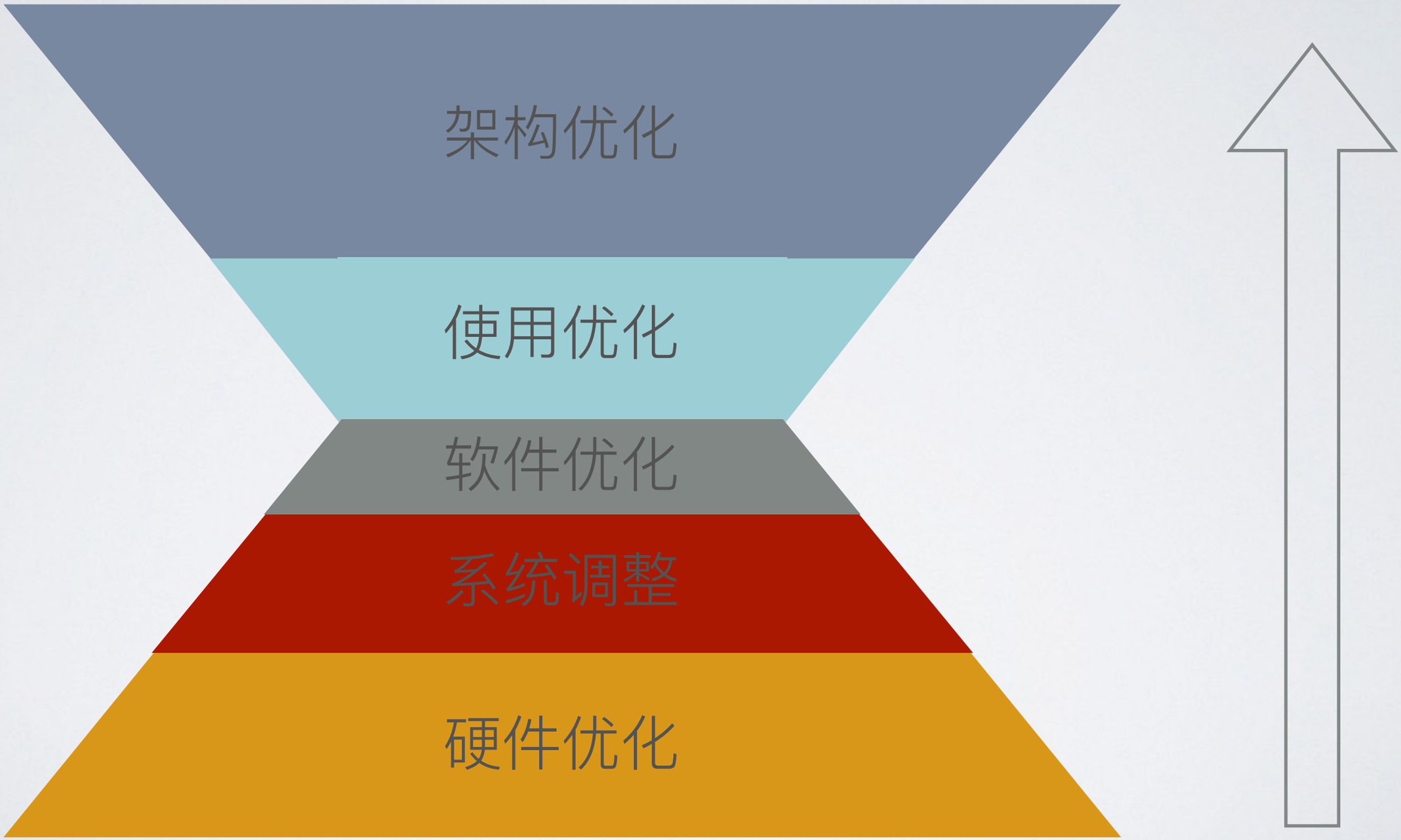
优化思路

业务发展趋势



A / B＝单机可支撑量

投入产出比



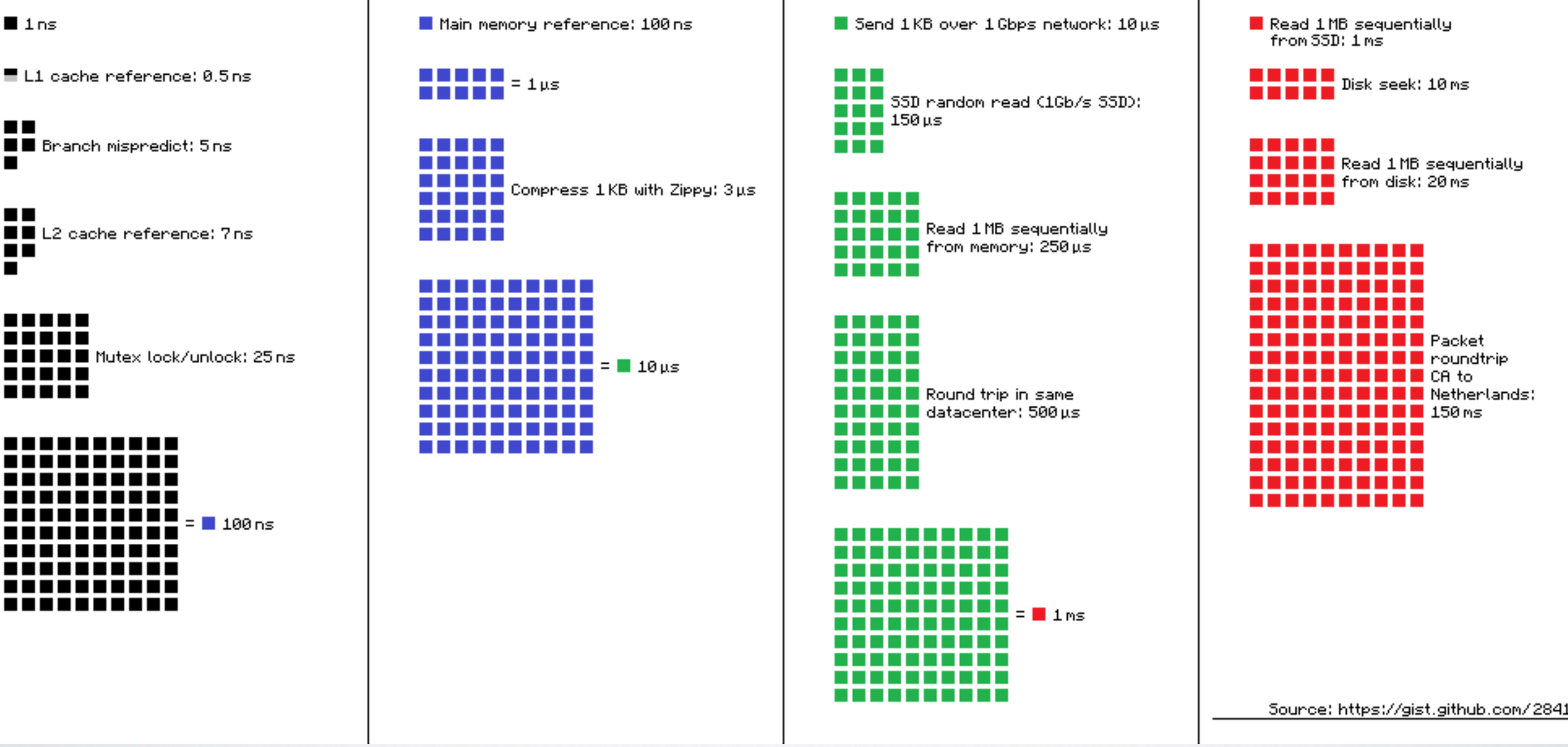
三条朴素的优化原则

- less is more
- Keep It Simple and Stupid(KISS)
- Performance is all about code path

优化实践－硬件篇

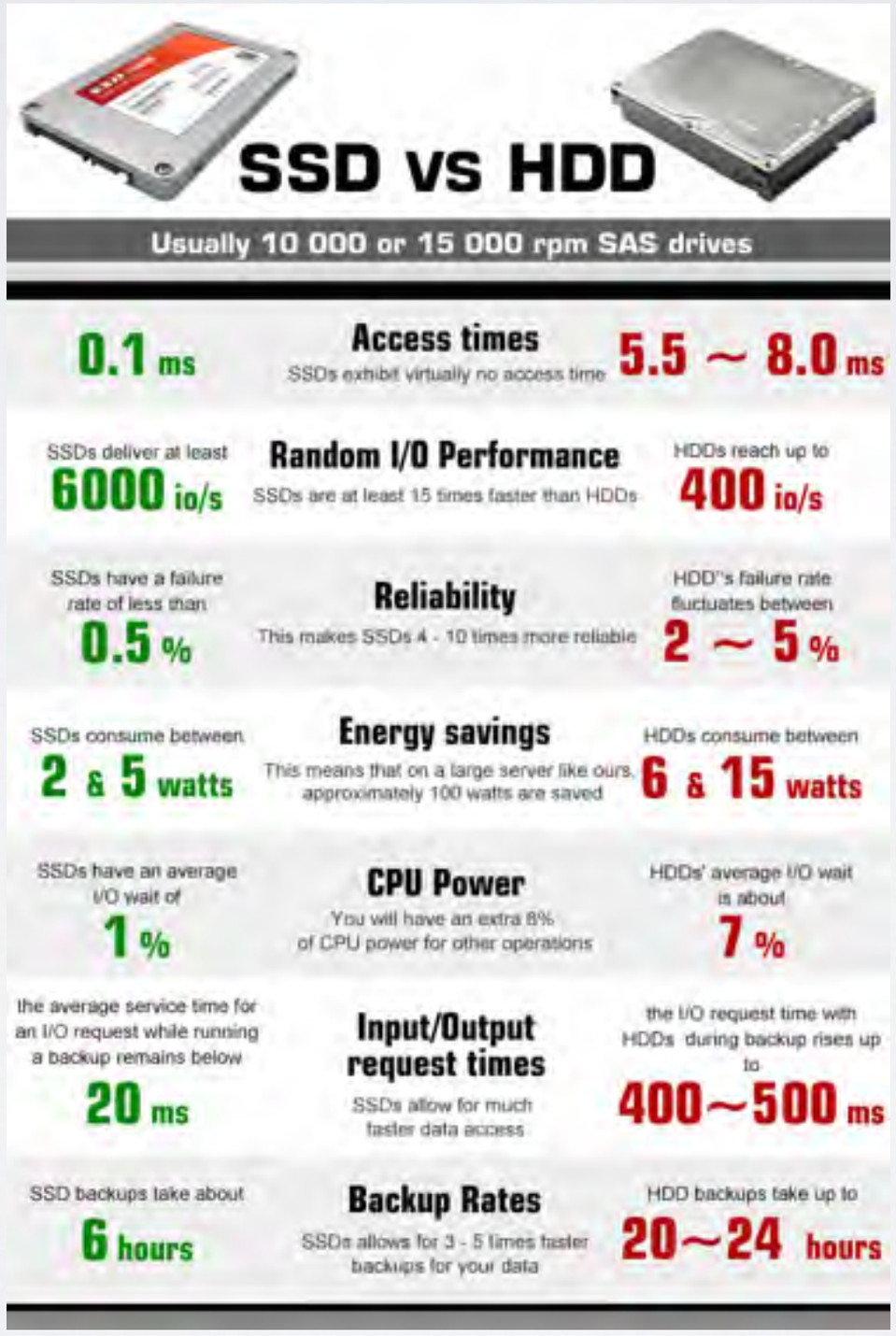
响应时间

Latency Numbers Every Programmer Should Know



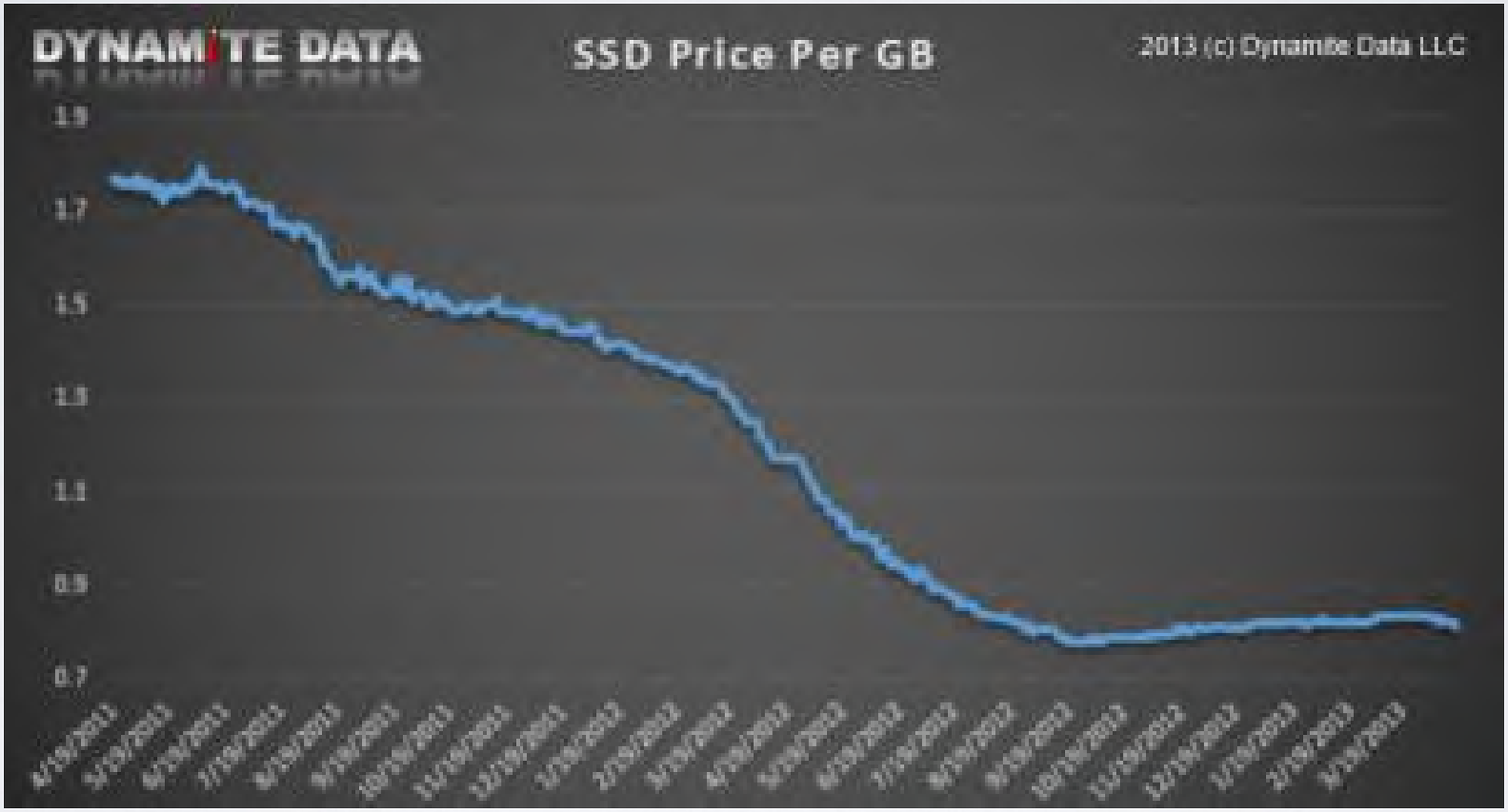
Source: https://gist.github.com/2841832

SSD VS HDD



图片来自网络

价格


DYNAMITE DATA
SSD Price Per GB
2013 (c) Dynamite Data LLC
1.9
1.7
1.5
1.3
1.1
0.9
0.7

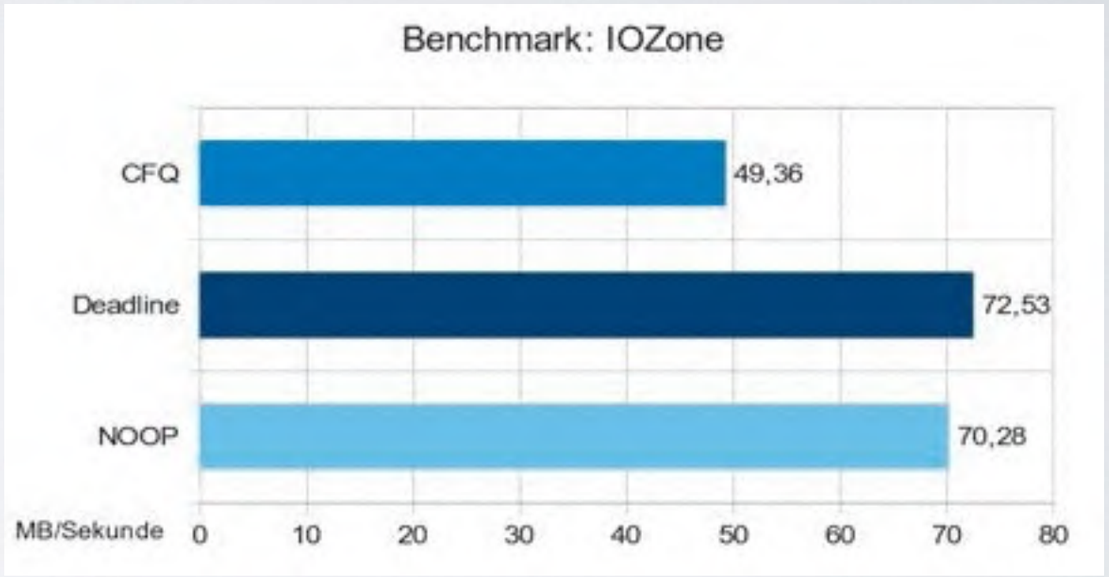
经验

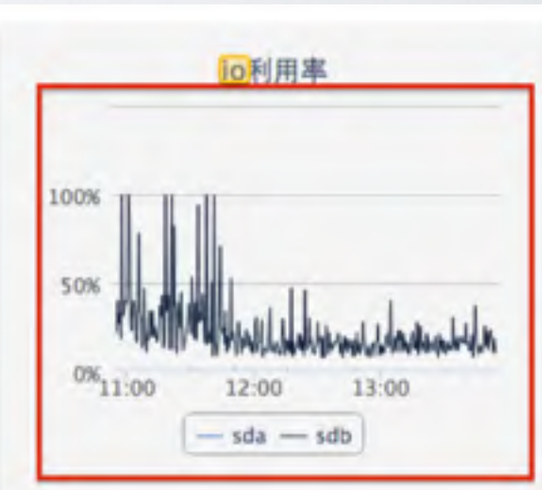
- 时间机会成本，稍纵即逝
- 人力成本远高于服务器成本
- 减少不必要的拆分

优化实践－系统篇

IO调度算法

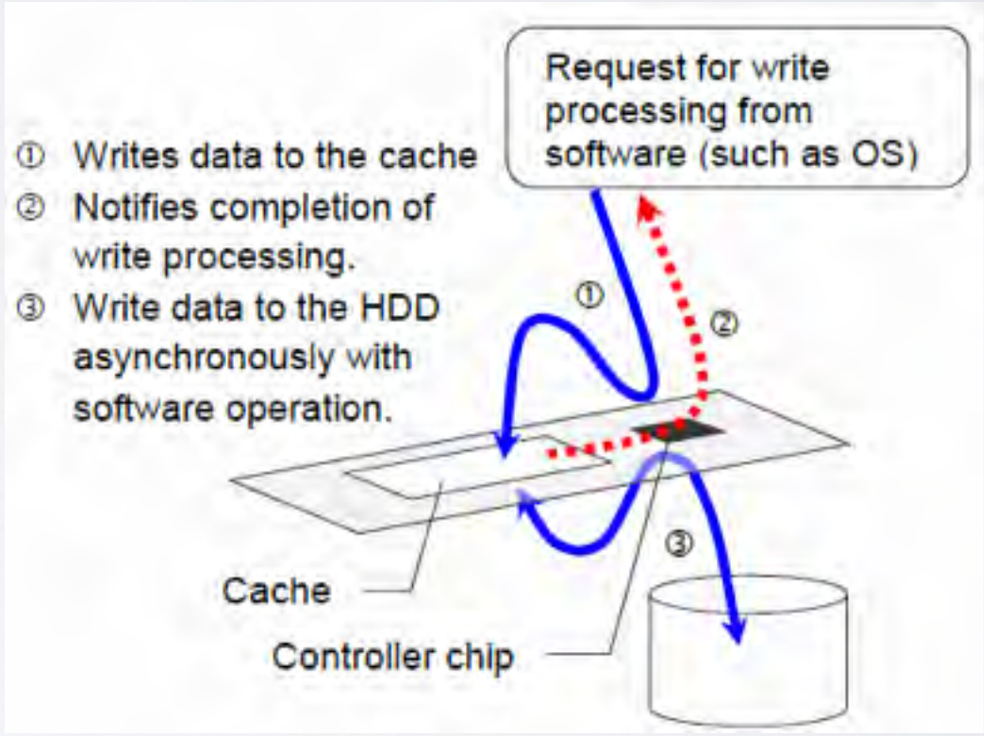
- CFQ
- DEADLIN
- NOOP





RAID策略

- write back
- write through
- BBU



http://m.osslab.com.tw/@api/deki/files/2318/=wb.gif

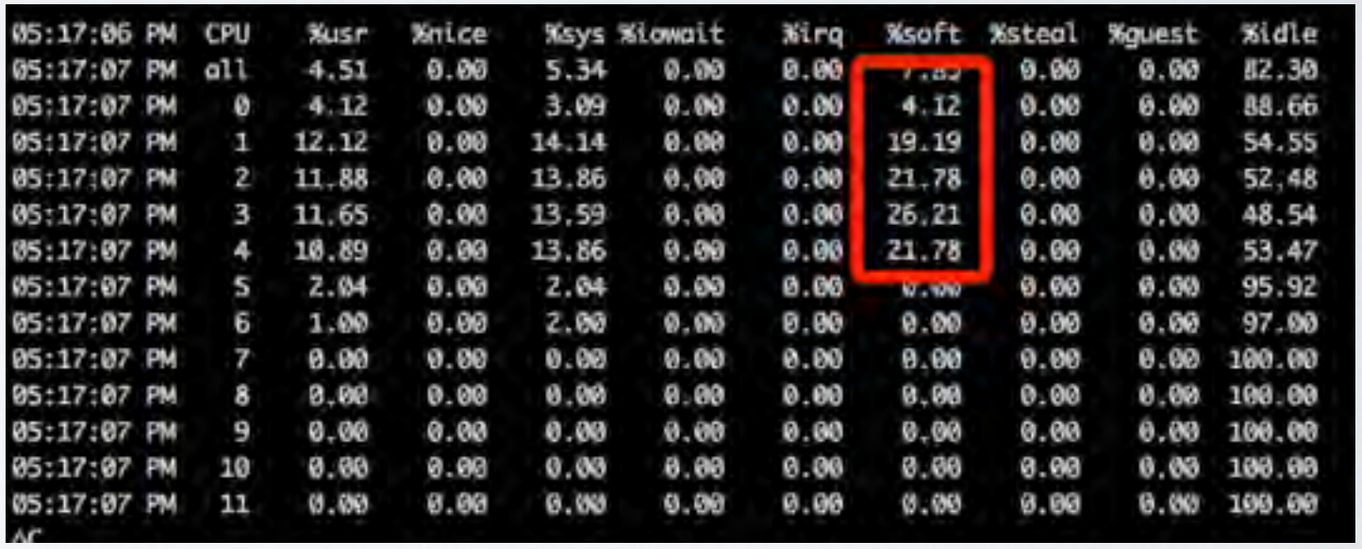
网卡多队列

- cpu soft irq

- lspci -vvv

- /proc/interrupts

- irqbalance



05:17:06 PM	CPU	%usr	%nice	%sys	%iowait	%irq	%soft	%steal	%guest	%idle
05:17:07 PM	all	4.51	0.00	5.34	0.00	0.00	7.85	0.00	0.00	82.30
05:17:07 PM	0	4.12	0.00	3.09	0.00	0.00	4.12	0.00	0.00	88.66
05:17:07 PM	1	12.12	0.00	14.14	0.00	0.00	19.19	0.00	0.00	54.55
05:17:07 PM	2	11.88	0.00	13.86	0.00	0.00	21.78	0.00	0.00	52.48
05:17:07 PM	3	11.65	0.00	13.59	0.00	0.00	26.21	0.00	0.00	48.54
05:17:07 PM	4	10.89	0.00	13.86	0.00	0.00	21.78	0.00	0.00	53.47
05:17:07 PM	5	2.04	0.00	2.04	0.00	0.00	0.00	0.00	0.00	95.92
05:17:07 PM	6	1.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	97.00
05:17:07 PM	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
05:17:07 PM	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
05:17:07 PM	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
05:17:07 PM	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
05:17:07 PM	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

^C

TIPS

- swapness
- XFS or EXT4
- ulimit
- NUMA

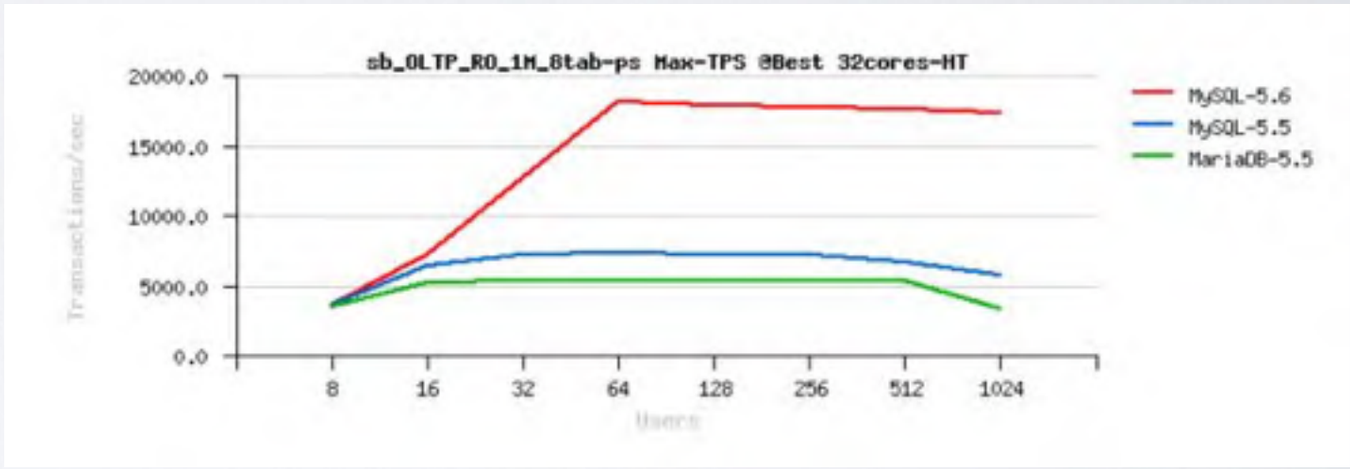
优化实践 — 软件篇

MYSQL

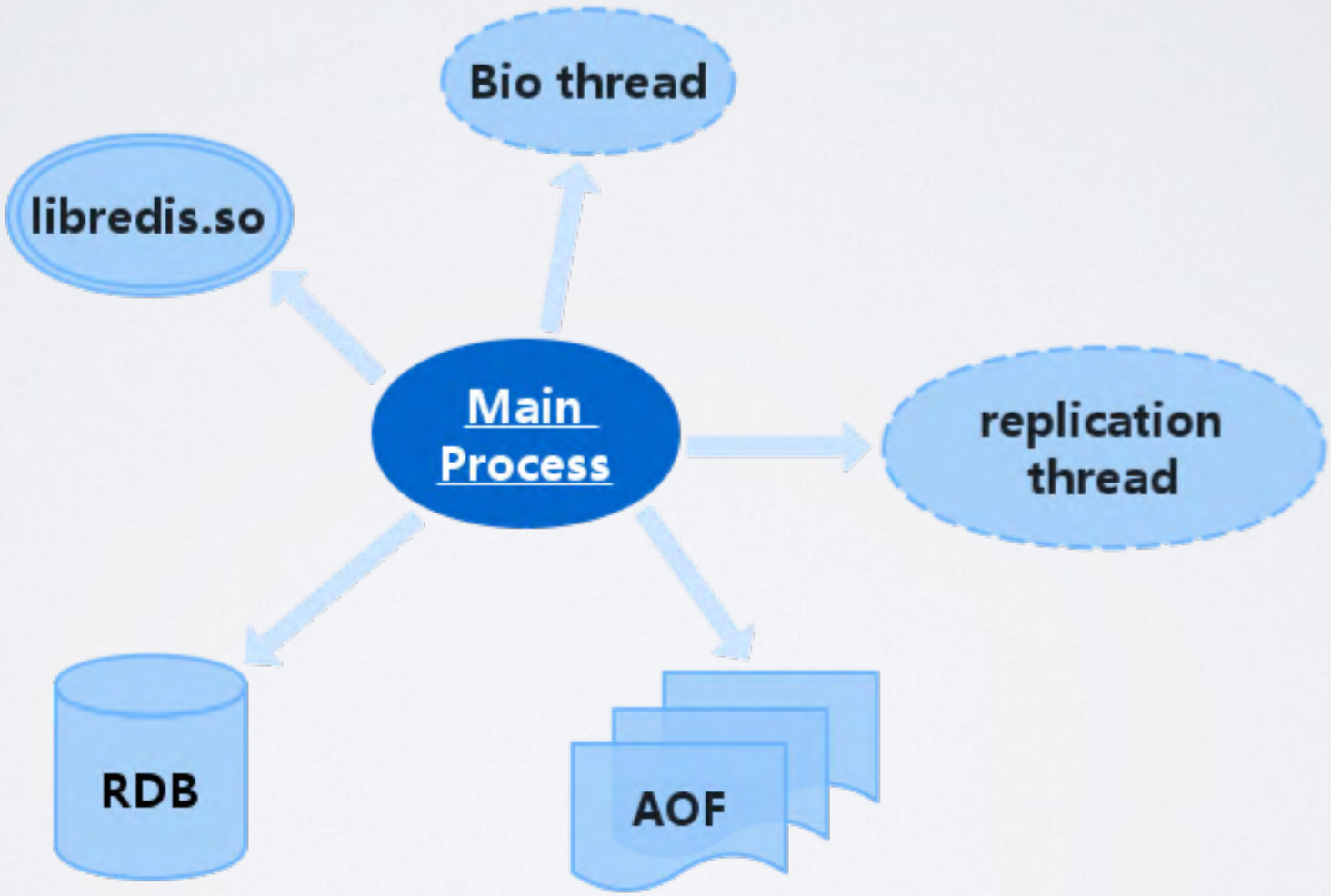
- innodb_buffer_pool_size
- innodb_log_file_size
- innodb_flush_log_at_trx_commit
- sync_binlog
- query_cache
- innodb_io_capacity
- innodb_buffer_pool_instances

版本选择

- group replication
- MRR
- ICP
- innodb_purge_threads



REDIS BGSAVE



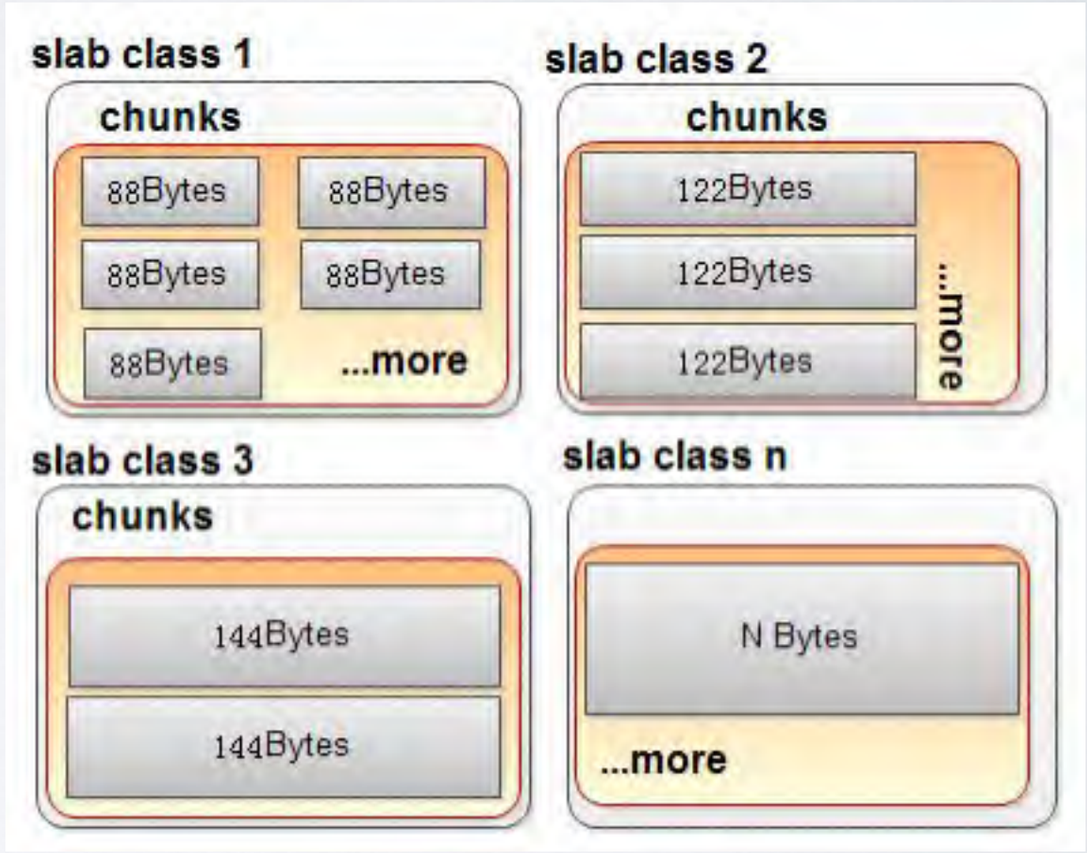
感受

- 没有内核能力在软件层能做的优化非常有限

优化实践－使用篇

MC之SLAB

- memcache-itool.py



REDIS之KEY分析

- Redis-rdb-Tools
- Redis-sampler
- Redis-audit

```
Example Output
==============

Sampling localhost:6379 DB:4 with 1000000 RANDOMKEYS

TYPES
=====
 zset: 873268 (87.33%)    string: 124995 (12.50%)  set: 1022 (0.10%)
 hash: 576 (0.06%)        list: 139 (0.01%)

STRINGS, SIZE OF VALUES
=======================
 6: 61222 (48.98%)        7: 17056 (13.65%)        13: 8274 (6.62%)
 15: 7991 (6.39%)         5: 4629 (3.70%)          31: 3263 (2.61%)
 20: 2670 (2.14%)         2: 2518 (2.01%)          27: 1675 (1.34%)
 42: 1270 (1.02%)         159: 893 (0.71%)         1: 705 (0.56%)
 47: 641 (0.51%)          34: 594 (0.48%)          41: 521 (0.42%)
 38: 493 (0.39%)          28: 413 (0.33%)          22: 406 (0.32%)
 139: 351 (0.28%)         29: 343 (0.27%)          83: 337 (0.27%)
(suppressed 172 items with perc < 0.5% for a total of 6.98%)
 Average: 15.97 Standard Deviation: 26.52
 Min: 0 Max: 1123

Powers of two distribution: (NOTE <= p means: p/2 < x <= p)
 <= 8: 82913 (66.33%)     <= 16: 16789 (13.43%)    <= 32: 9571 (7.66%)
 <= 64: 6232 (4.99%)      <= 128: 3333 (2.67%)     <= 256: 2682 (2.15%)
 <= 2: 2518 (2.01%)       <= 1: 740 (0.59%)        <= 4: 199 (0.16%)
 <= 512: 14 (0.01%)       <= 1024: 3 (0.00%)       <= 2048: 1 (0.00%)
```

MYSQL之SLOWLOG

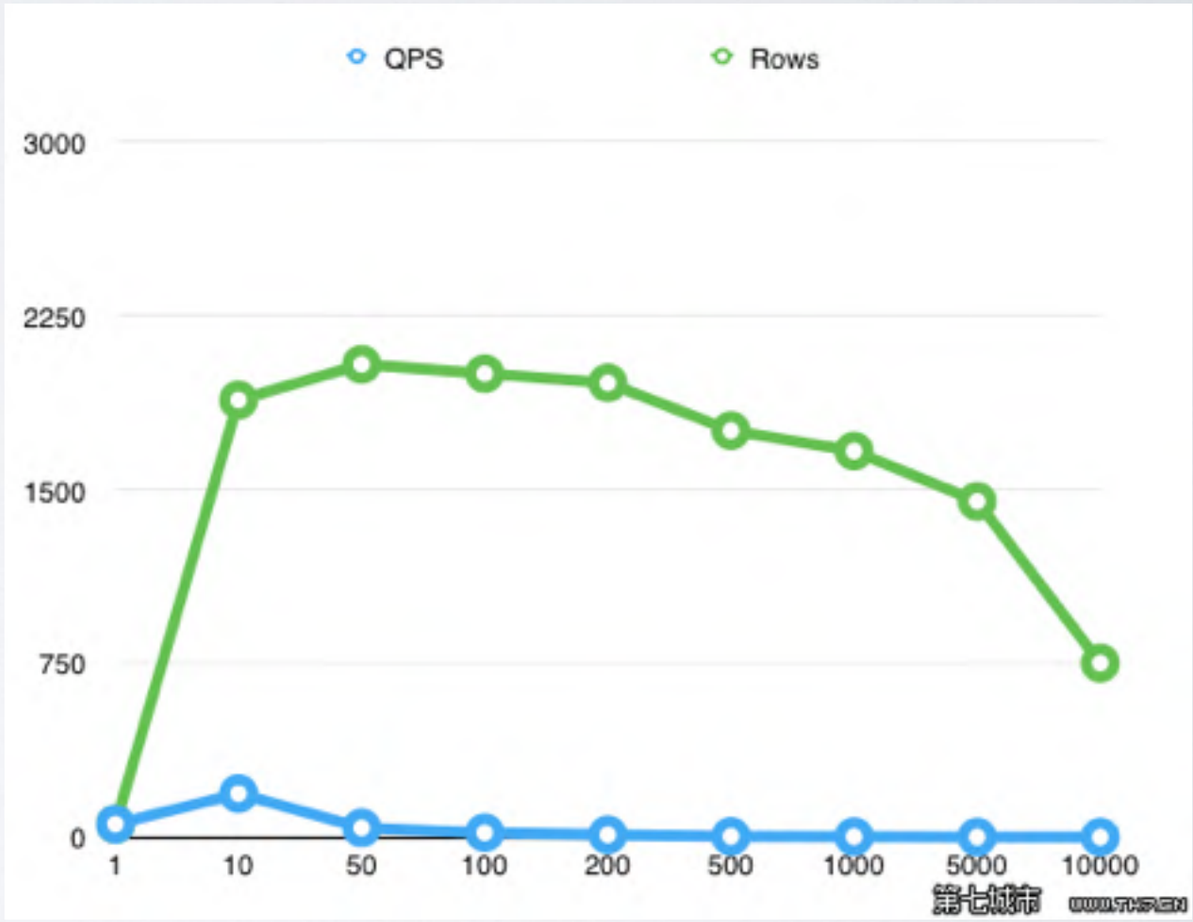
- pt-query-digest
- box anemometer
- all because index

1.32s vs 0.03s

MYSQL之BATCH INSERT

- 500个insert into tb values ()

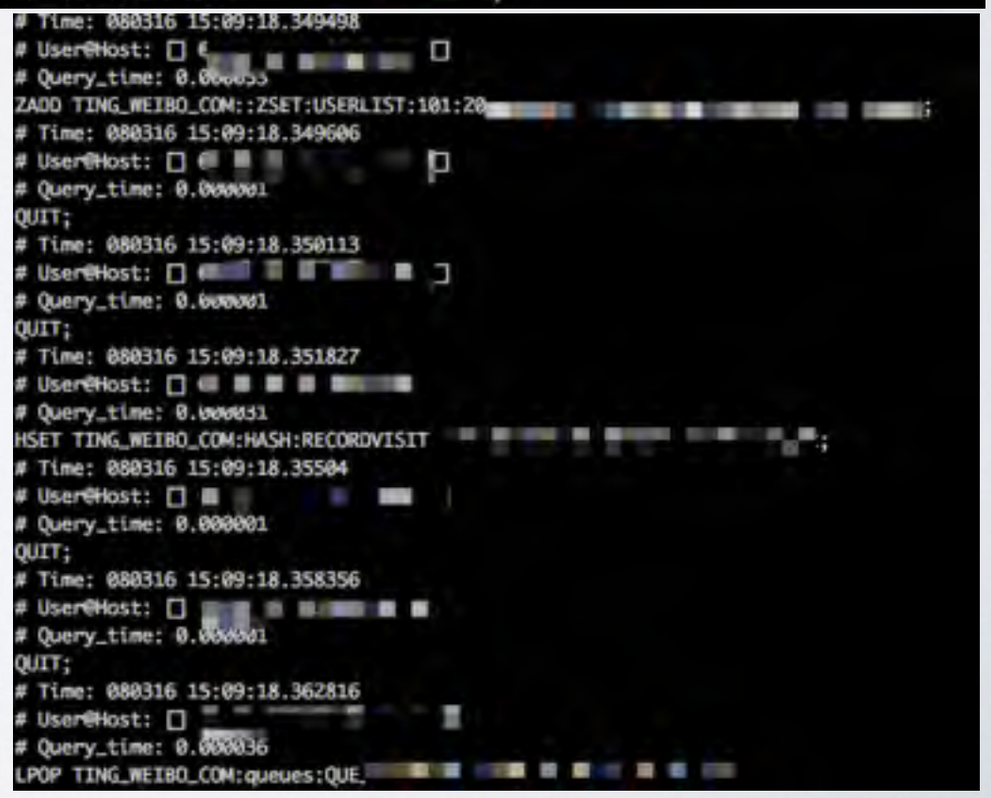
- 10个insert into tb values () ()



使用分析

- vc-MySQL-sniffer

- vc-Redis-sniffer



- vc-pg-sniffer

www.vividcortex.com

```
# 890ms user time, 10ms system time, 17.47M rss, 107.61M vsz
# Current date: Wed Aug  3 15:27:08 2016
# Hostname: hebe210
# Files: 123
# Overall: 2.61k total, 37 unique, 652.50 QPS, 0.01x concurrency ________
# Time range: 2008-03-16 15:24:33 to 15:24:37
# Attribute          total     min     max     avg     95%  stddev  median
# ============     ======= ======= ======= ======= ======= ======= =======
# Exec time           59ms      1us    108us     22us     44us     16us     25us
# Query size       156.84k         4     256   61.53   246.02   62.72   51.63

# Profile
# Rank Query ID           Response time Calls R/Call Apdx V/M   Item
# ==== ================== ============= ===== ====== ==== ===== ==========
#    1 0x13838FA0F8DCD689  0.0196 33.0%   758 0.0000 1.00  0.00
#    2 0x3C103B3D1DE813EA  0.0183 30.8%   510 0.0000 1.00  0.00
#    3 0xAA85C671683C0E93  0.0081 13.6%   231 0.0000 1.00  0.00
#    4 0xAE5B980DAA7278FF  0.0017  2.8%    35 0.0000 1.00  0.00
#    5 0x5ACE65C74881A338  0.0014  2.4%    53 0.0000 1.00  0.00
#    6 0x5F35D3DC79094F92  0.0014  2.4%    53 0.0000 1.00  0.00
#    7 0x89902ABA8334F631  0.0014  2.4%    29 0.0000 1.00  0.00
#    8 0x06872EB401E2791B  0.0008  1.4%    17 0.0000 1.00  0.00
#    9 0x8ED794C0D5413D5A  0.0008  1.4%   745 0.0000 1.00  0.00
#   10 0xA1F96C852D775771  0.0006  1.1%    17 0.0000 1.00  0.00 SET
#   11 0x532EBCB9E10F06F0  0.0006  1.0%    19 0.0000 1.00  0.00
#   12 0x703CC0FE9215D371  0.0006  0.9%    17 0.0000 1.00  0.00
#   13 0xA330AA9E035F3573  0.0005  0.8%    15 0.0000 1.00  0.00
#   14 0x3E4D03D07CDDC867  0.0005  0.8%    13 0.0000 1.00  0.00
#   15 0xD9C0AA6E015C82C2  0.0004  0.7%    14 0.0000 1.00  0.00
# MISC 0xMISC              0.0026  4.4%    84 0.0000   NS   0.0 <22 ITEMS>

# Query 1: 189.50 QPS, 0.00x concurrency, ID 0x13838FA0F8DCD689 at byte 228769
# This item is included in the report because it matches --limit.
# Scores: Apdex = 1.00 [1.0], V/M = 0.00
# Query_time sparkline: | ^_     |
# Time range: 2008-03-16 15:24:33 to 15:24:37
# Attribute    pct   total     min     max     avg     95%  stddev  median
# ============ === ======= ======= ======= ======= ======= ======= =======
# Count         29     758
# Exec time     33    20ms    14us   106us    25us    38us     8us    23us
# Query size    25  39.23k      53      53      53      53       0      53
# Query_time distribution
#   1us
#  10us  ################################################################
# 100us  #
#   1ms
#  10ms
# 100ms
#    1s
#  10s+
LPOP TING_WEIBO_COM:queues:async-action-queue-level-4\G
```

pt-query-digest

优化实践－架构篇

选型

Redis

memcache

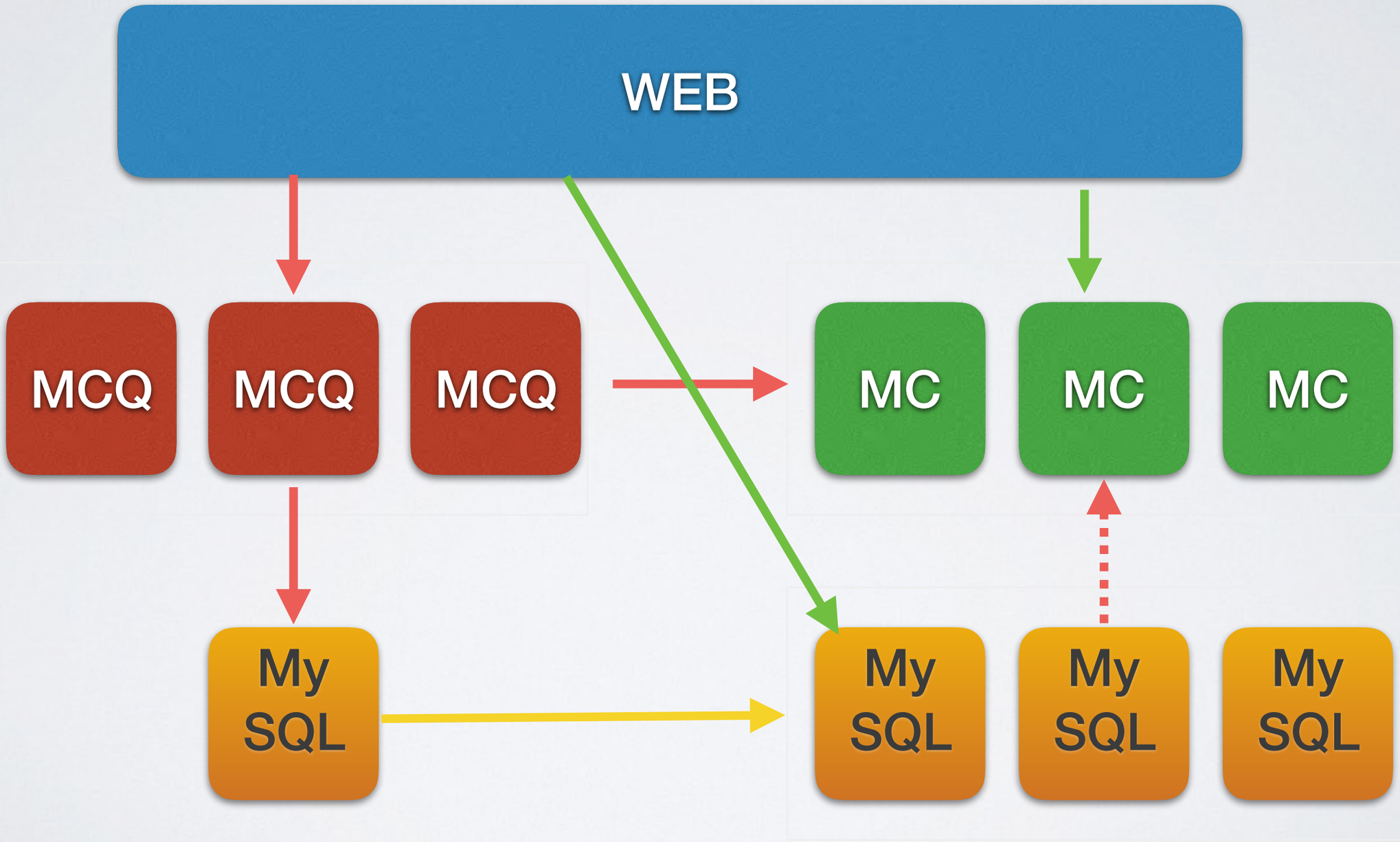
MySQL

Pika

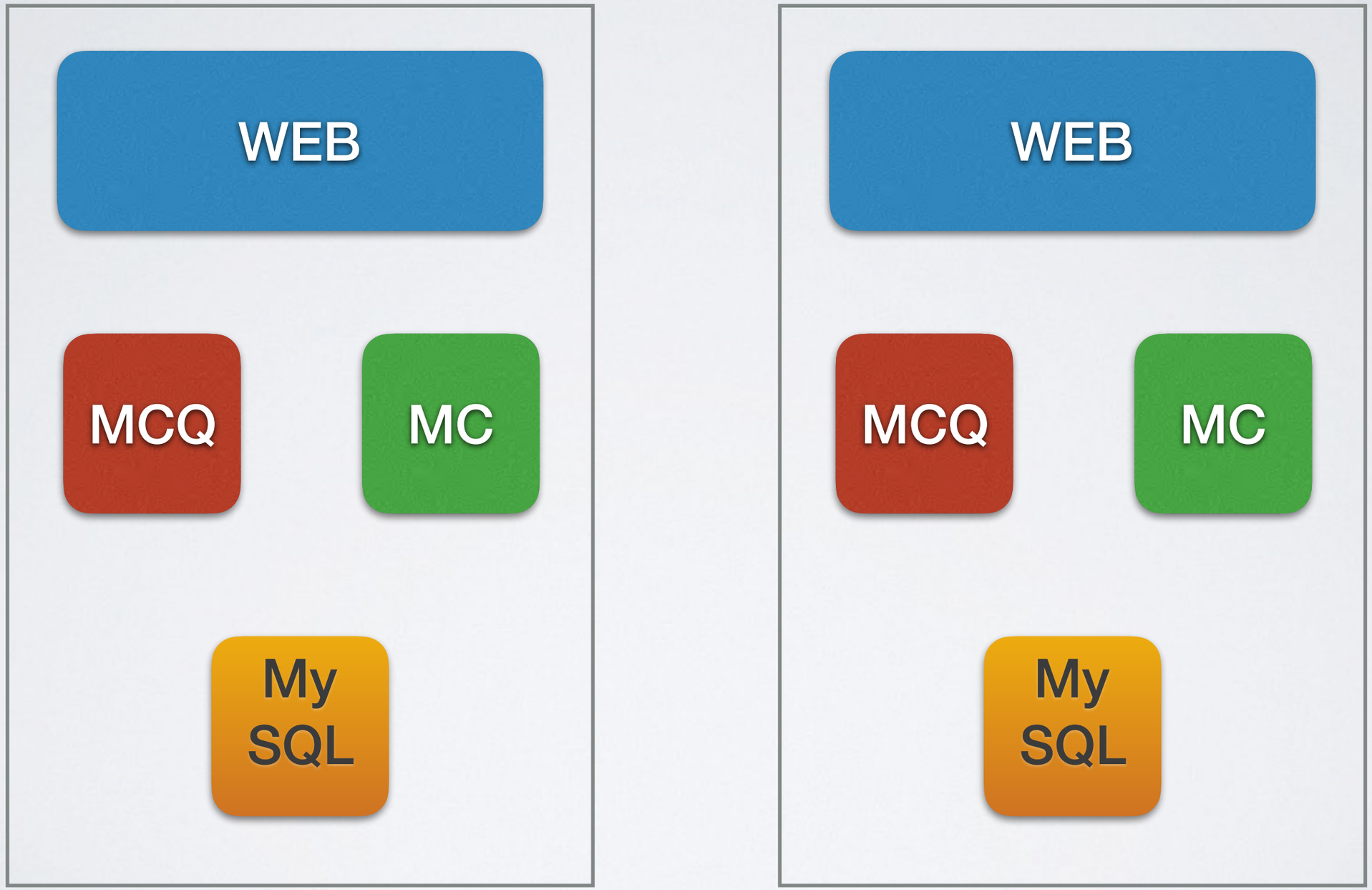
MongoDB

HBase

设计



本地化



分层

MC (L0) | MC (L0) | MC (L0) | MC (L0) | MC (L0) | MC (L0) | MC (L0) | MC (L0)

MC L1 | MC L1 | MC L1 | MC L1

MC L1 | MC L1 | MC L1 | MC L1

MC（Master）

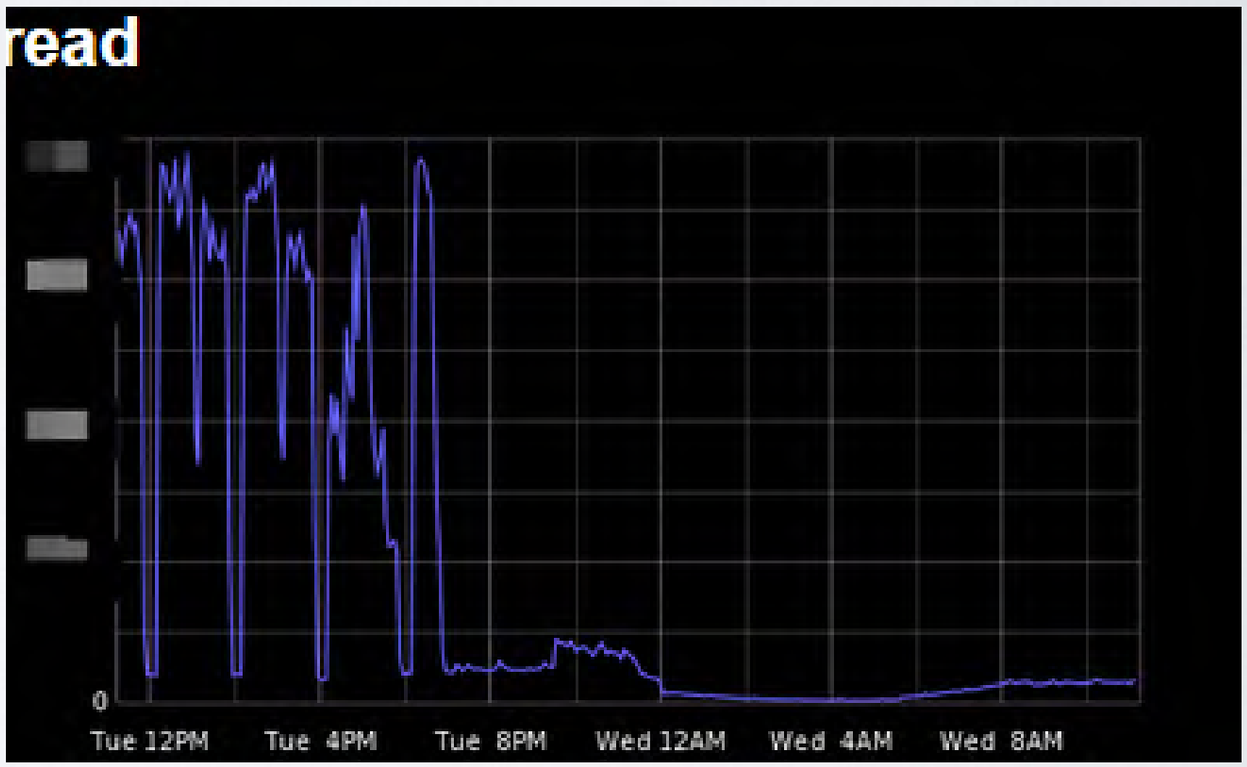
MySQL

拆分

- 垂直拆分
- 水平拆分
- 时间维度归档

产品逻辑

- Top 1 is don’t do

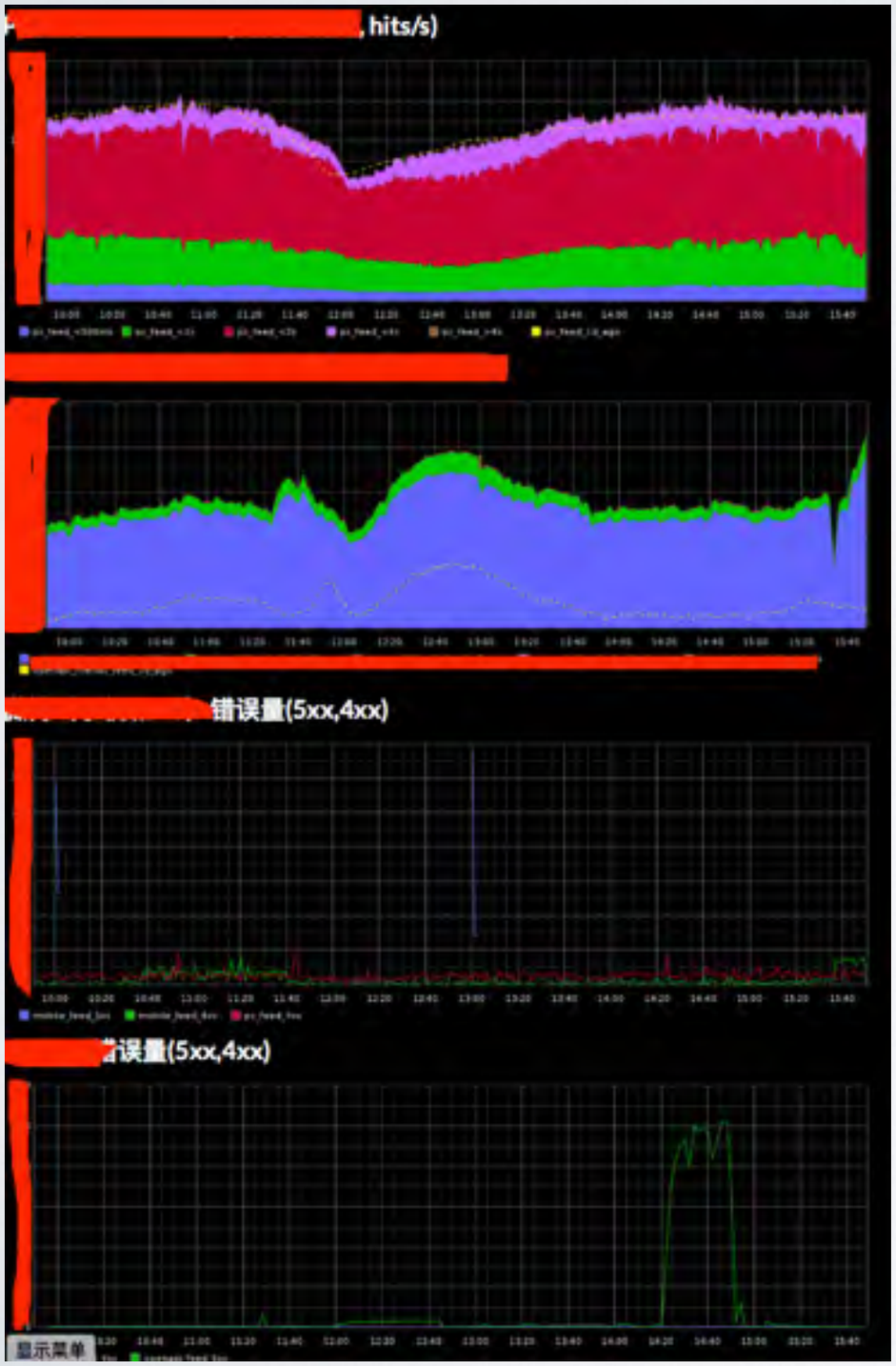


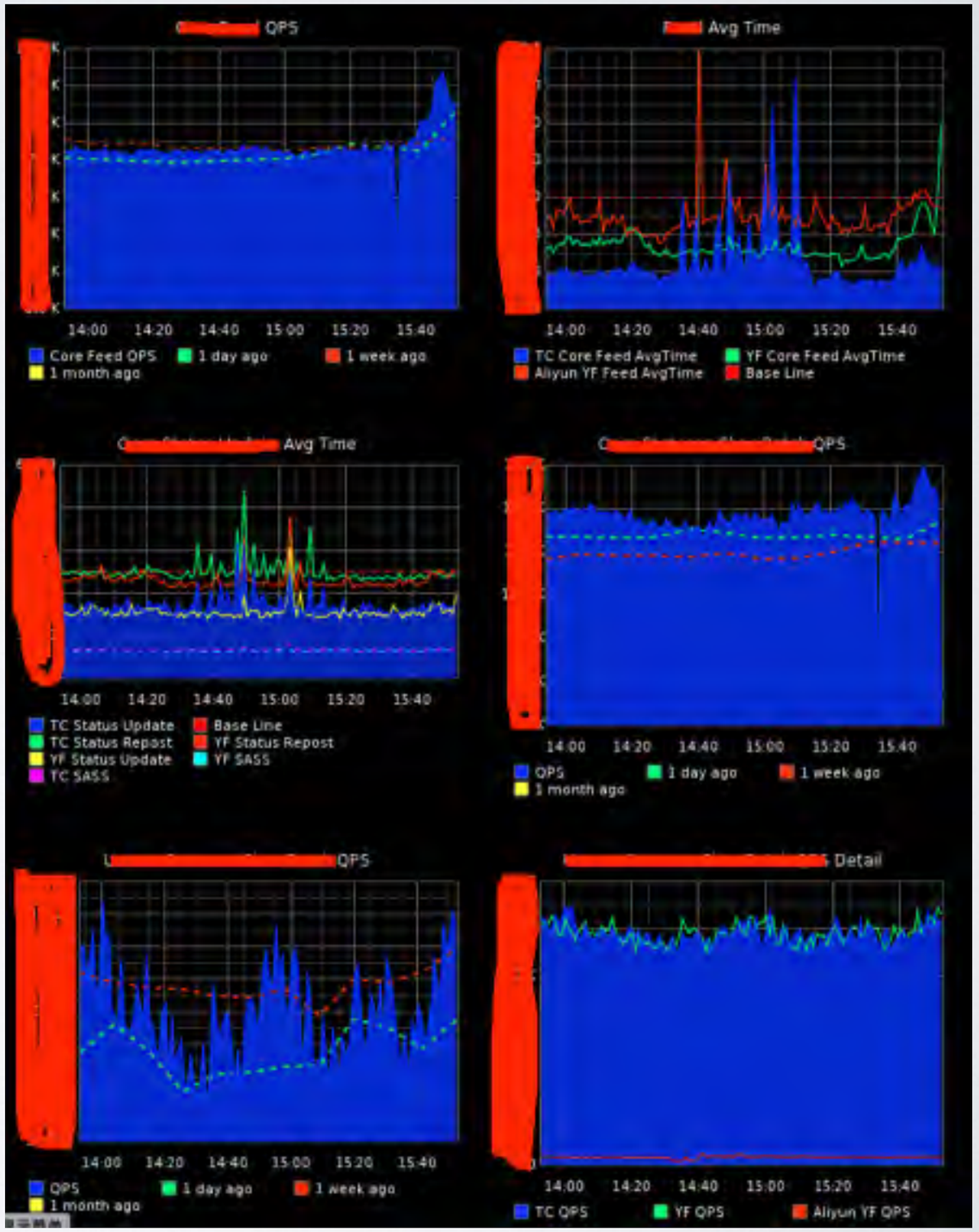
总结

"What gets measured gets managed."

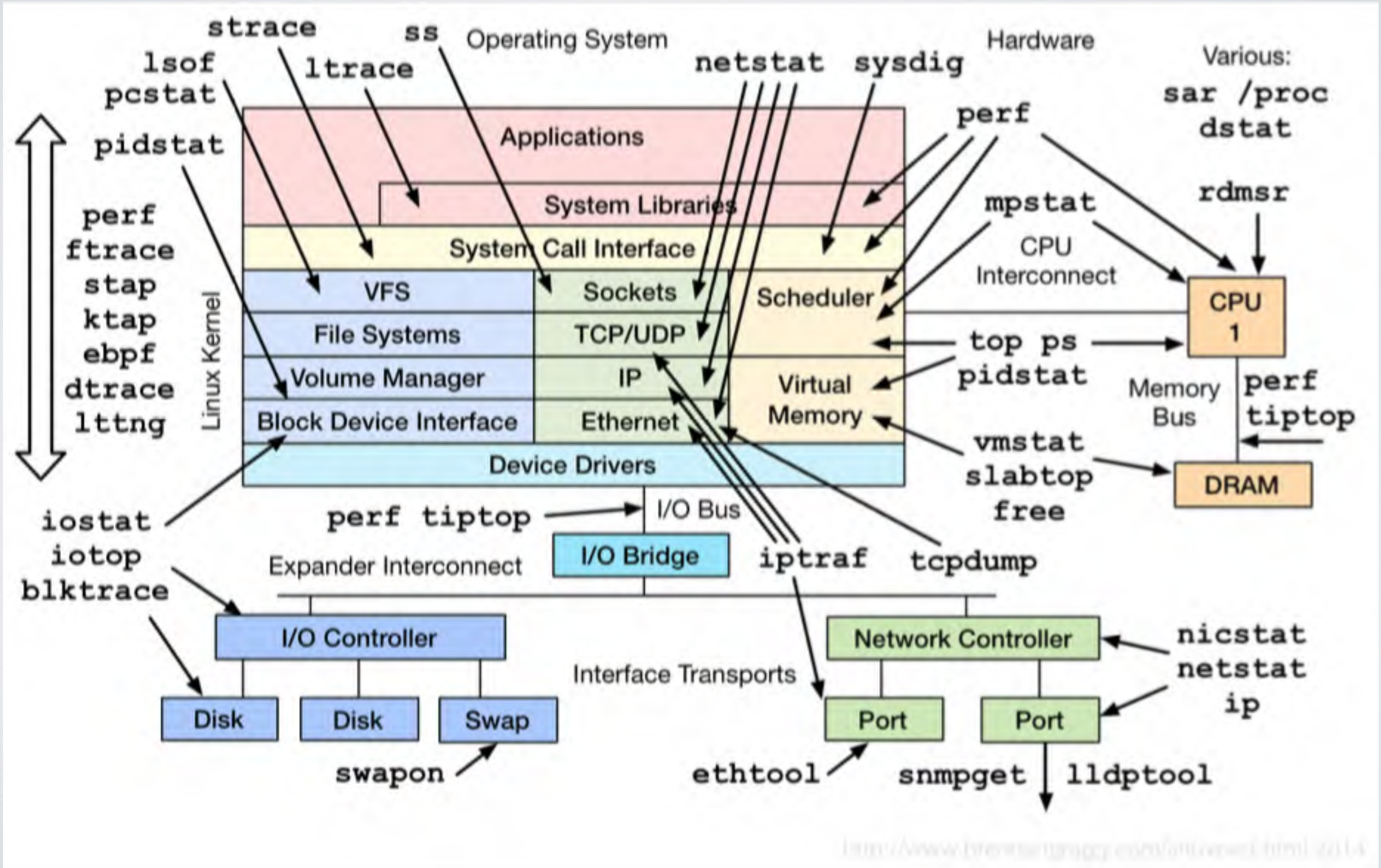
–Peter Drucker

数据驱动


, hits/s)
错误量(5xx,4xx)
错误量(5xx,4xx)
显示菜单


QPS
Avg Time
14:00 14:20 14:40 15:00 15:20 15:40
Core Feed QPS
1 day ago
1 week ago
1 month ago
TC Core Feed AvgTime
YF Core Feed AvgTime
Aliyun YF Feed AvgTime
Base Line
Avg Time
TC Status Update
TC Status Repost
YF Status Update
TC SASS
Base Line
YF Status Repost
YF SASS
QPS
QPS
1 day ago
1 week ago
1 month ago
QPS
Detail
TC QPS
YF QPS
Aliyun YF QPS

工欲善其事必先利其器


strace
ss
Operating System
Hardware
Various:
lsof
pcstat
ltrace
netstat
sysdig
sar /proc
dstat
perf
pidstat
Applications
System Libraries
System Call Interface
perf
ftrace
stap
ktap
ebpf
dtrace
lttng
Linux Kernel
VFS
Sockets
Scheduler
File Systems
TCP/UDP
Volume Manager
IP
Block Device Interface
Ethernet
Device Drivers
Virtual Memory
mpstat
CPU Interconnect
rdmsr
CPU 1
top ps
pidstat
Memory Bus
perf
tiptop
vmstat
slabtop
free
DRAM
iostat
iotop
blktrace
perf tiptop
I/O Bus
I/O Bridge
Expander Interconnect
iptraf
tcpdump
I/O Controller
Network Controller
Interface Transports
nicstat
netstat
ip
Disk
Disk
Swap
Port
Port
swapon
ethtool
snmpget
lldptool

经验

- case by case，不要经验主义
- 推测需要事实来证明
- 因果演绎有时候不如排除法
- 一次一个因素避免干扰


架构优化
使用优化
软件优化
系统调整
硬件优化
时间分布

THANK YOU