数据库性能量化

叶正盛
阿里巴巴-运维部

About me

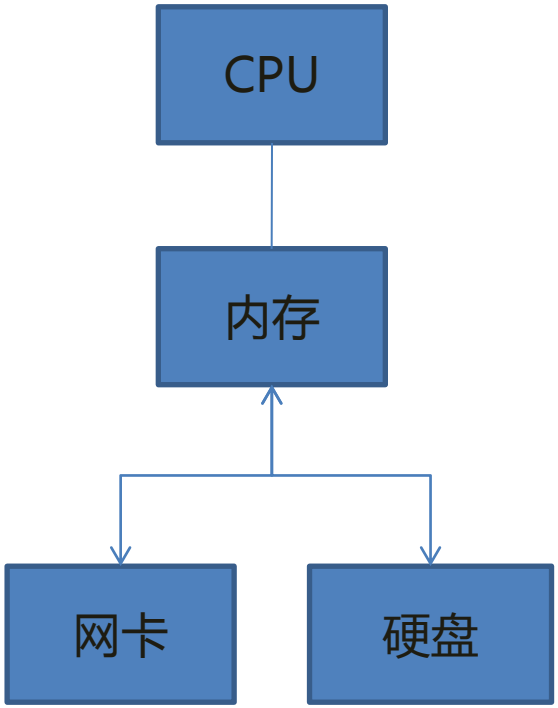
- 姓名：叶正盛
- 阿里巴巴数据库技术专家
- 国家认证系统分析师、高级项目经理
- 10余年软件开发及管理经验
- 从事过微机监控、外贸、进销存、ERP系统设计开发
- 从事过省级电力信息化建设
- 我的博客：http://blog.csdn.net/yzsind
- 新浪微博：http://weibo.com/yzsind

Agenda

- 硬件与数据库相关性能指标介绍
- 业务指标转变为数据库技术指标实例
- 什么时候做数据库拆分？
- SSD给数据库带来什么变化？

硬件与数据库相关性能指标

- 磁盘
  1秒钟可以从磁盘随机访问多少次？
- 网络
  网络延时与网络带宽
- 内存
  访问内存一个数据要多少时间？
- CPU
  对数据库CPU最重要的是什么？



存储磁盘性能量化

	10K 3.5寸 SAS	15K 3.5寸 SAS	10K 2.5寸 SAS	15K 2.5寸 SAS
延时 (等待时间)	3ms	2ms	3ms	2ms
延时 (寻道时间)	3.5ms	3.5ms	3ms	3ms
IOPS-8KB	153(333)	181(500)	150(333)	200(500)
内部平均带宽	130MB/s	160MB/s	130MB/s	160MB/s

影响性能的主要因素：转速、盘片大小、磁存储密度

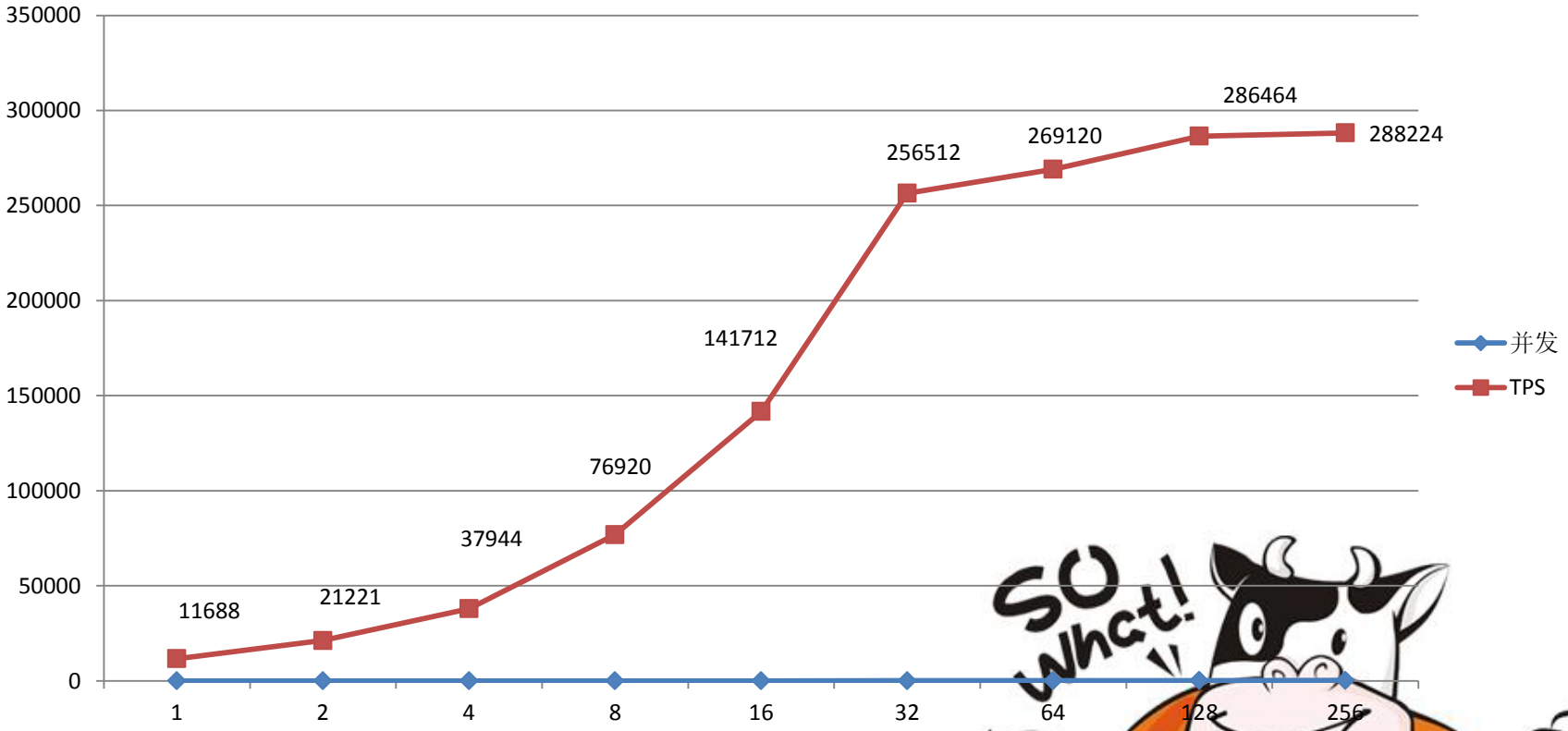
网络性能量化

- 100Mbps/1Gbps/10Gbps
- 带宽：10MB/s, 100MB/s ,1000MB/s
- 本地机房延时：50us-1ms

```
mking>ping 10.20.149.82
PING 10.20.149.82 (10.20.149.82) 56(84) bytes of data.
64 bytes from 10.20.149.82: icmp_seq=0 ttl=64 time=0.124 ms
64 bytes from 10.20.149.82: icmp_seq=1 ttl=64 time=0.109 ms
64 bytes from 10.20.149.82: icmp_seq=2 ttl=64 time=0.110 ms
64 bytes from 10.20.149.82: icmp_seq=3 ttl=64 time=0.109 ms
64 bytes from 10.20.149.82: icmp_seq=4 ttl=64 time=0.110 ms
```

1Gbps网络Netperf 测试结果

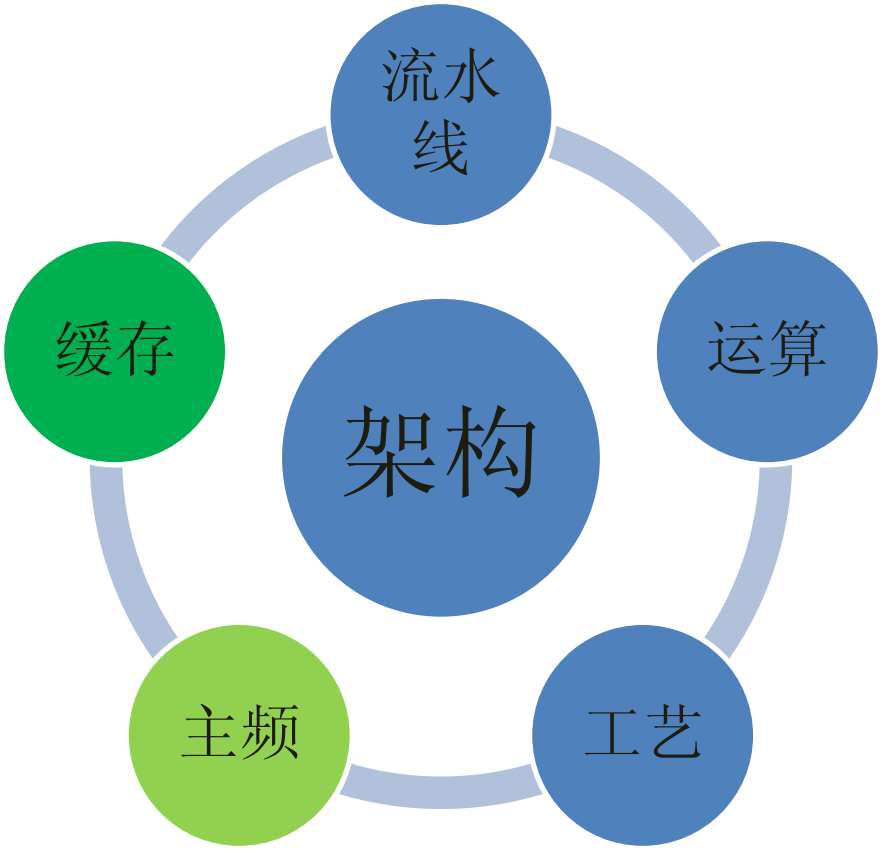
- 数据库TCP包请求表现



网络延时与网络带宽

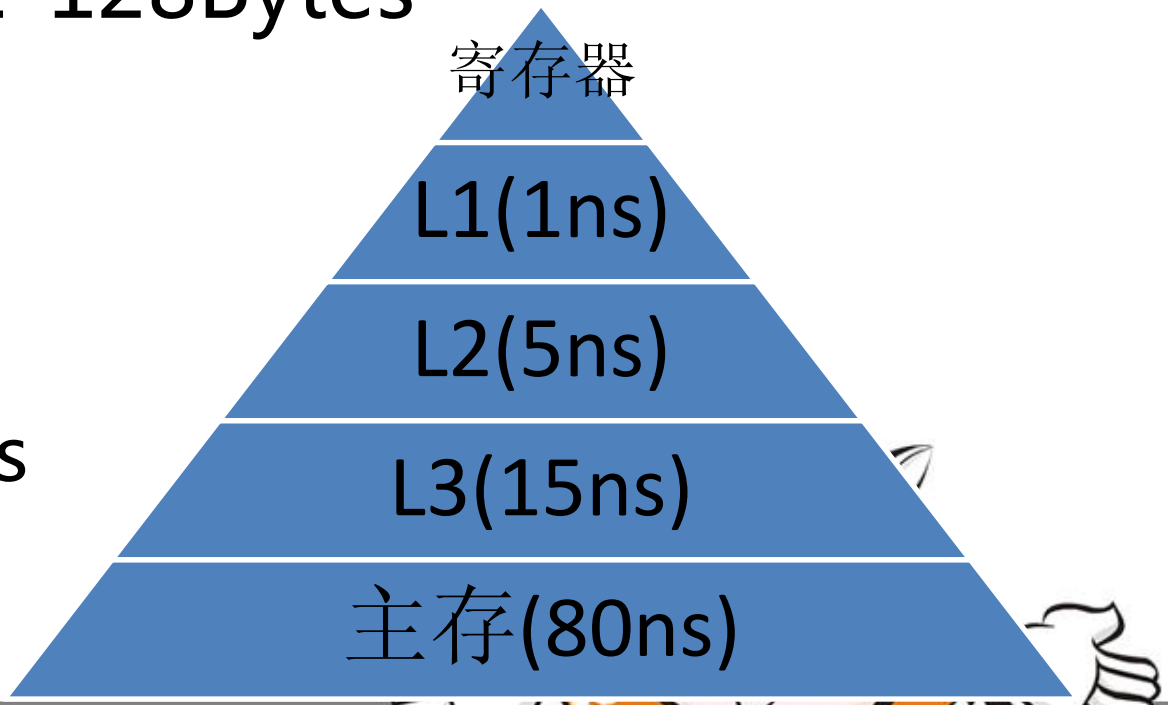
- 网络延时=处理时间+传输时间+传播时间
  - 处理时间=网络设备数据包处理时间（主机、交换机、路由器等等）
  - 传输时间=数据量/物理链路网络带宽
  - 传播时间=两地距离*2/200000
- Socket缓冲区大小(buffer_size)
- 远距离网络单线程带宽≈ buffer_size/2/latency
- 实例，A到B网络延时15ms，
  单线程测试结果：
- 缓冲区大小16K，传输带宽约600KB/s
- 缓冲区大小40K，传输带宽约1.6MB/s
- 缓冲区大小400K，传输带宽约15MB/s

CPU



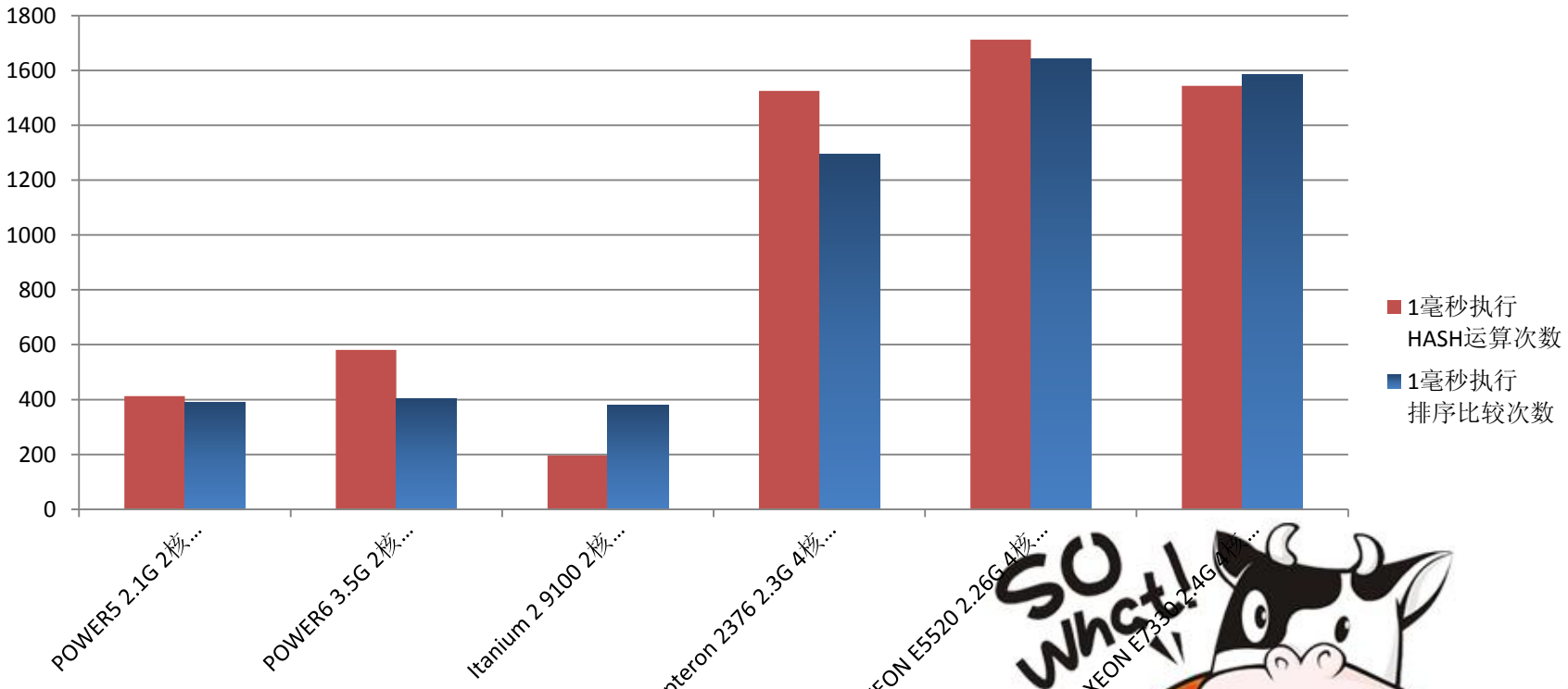
CPU缓存、内存

- Cache 延时 0.5-30ns
- Cache 带宽 10-100GB/s
- Cache Line 32-128Bytes
- 主存延时
  - 30-200ns
- 主存带宽
  - 2GB/s-12GB/s



CPU单核性能

- 执行Oracle数据库的hash及排序比较运算



Oracle数据库在1秒可以做什么

CPU：INTEL 2GHz，单核测试
以下数据与机器硬件性能、Oracle版本、参数关系密切，数据仅供数量级内的参考，仅用于快速评估

次数	动作
10	连接数据库
100	磁盘物理读，注：非SSD硬盘
1000	简单SQL硬解析，select * from t where pk=?
10000	简单SQL软解析
100000	逻辑读
1000000	Hash运算，10字节排序，取Sysdate
4000000	简单函数运算，如substr、lower之类的函数

实例分析

- 普通商品管理子系统
- 20万商家，5万活跃会员
- 2000万商品
- 平均每个商品信息基本信息300字节，详细信息8K
- 业务高峰期4小时

业务指标->技术指标

活跃会员数：5万，业务高峰时段：4小时 (9:30-11:30,14:30-16:30)

业务功能	会员操作次数	总操作次数	返回记录数	总返回记录数	数据大小	总数据大小
登录	2	100000	1	100000	1000	100000000
商品列表	50	2500000	20	50000000	6000	15000000000
查看商品明细	200	10000000	1	10000000	8000	80000000000
新增商品	2	100000	1	100000	8000	800000000
修改商品	50	2500000	1	2500000	8000	20000000000
删除商品	1	50000	1	50000	500	25000000
总计		15250000		62750000		1.15925E+11
每秒指标		应用QPS 1059		存储IOPS 4358		网络带宽 8050347（8MB/s）

分表、分区

人员待办工单查询

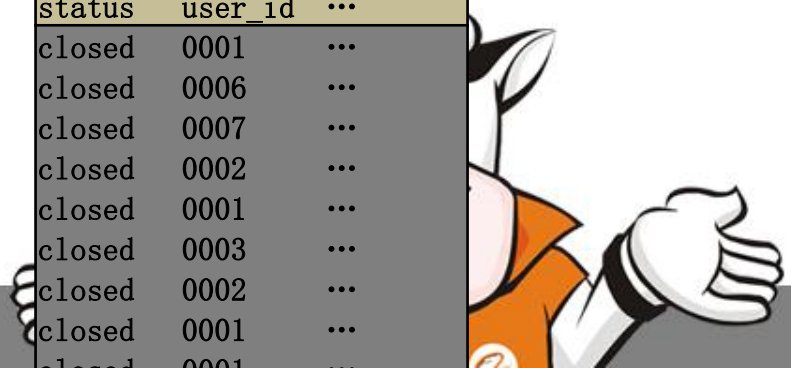
Select * from bpm_work where user_id = ‘0001’ and status= ‘new’

- 活动数据与历史数据分离：(分表、分区、压缩）
  - 工作流（任务流、工单），按状态分表分区
  - 历年帐务记录，按年月分表分区

status	user_id	…
new	0002	…
closed	0001	…
new	0002	…
new	0003	…
closed	0003	…
new	0008	…
closed	0001	…
new	0001	…
new	0001	…
closed	0001	…
closed	0002	…
new	0002	…
closed	0003	…
closed	0006	…
new	0007	…

status	user_id	…
new	0001	…
new	0002	…
new	0001	…
new	0008	…
new	0003	…
new	0001	…

status	user_id	…
closed	0001	…
closed	0006	…
closed	0007	…
closed	0002	…
closed	0001	…
closed	0003	…
closed	0002	…
closed	0001	…

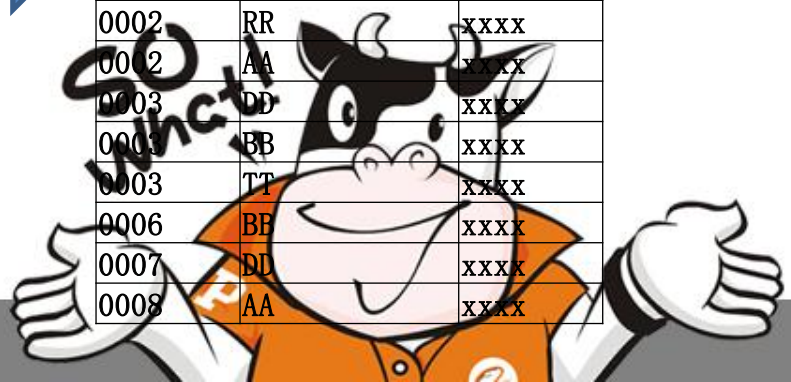


数据聚集

- 核心数据聚集（聚集索引、单表聚簇）
  - 一对多关系
  - 会员发布商品
  - 会员交易记录
  - 博客评论、反馈

blog_id	user_name	comment
0002	AA	xxxx
0001	AA	xxxx
0002	CC	xxxx
0003	DD	xxxx
0003	BB	xxxx
0008	AA	xxxx
0001	EE	xxxx
0001	DD	xxxx
0001	GG	xxxx
0001	BB	xxxx
0002	RR	xxxx
0002	AA	xxxx
0003	TT	xxxx
0006	BB	xxxx
0007	DD	xxxx

blog_id	user_name	comment
0001	EE	xxxx
0001	DD	xxxx
0001	GG	xxxx
0001	BB	xxxx
0001	AA	xxxx
0002	AA	xxxx
0002	CC	xxxx
0002	RR	xxxx
0002	AA	xxxx
0003	DD	xxxx
0003	BB	xxxx
0003	TT	xxxx
0006	BB	xxxx
0007	DD	xxxx
0008	AA	xxxx



单机性能瓶颈

- 拆分
  - 水平拆分
  - 垂直拆分
  - 读写分离
  - 异地容灾

- 过早拆分增加系统的复杂度及维护成本，过晚拆分影响业务发展。

- 设计师一定要心中有数，而不是人云亦云

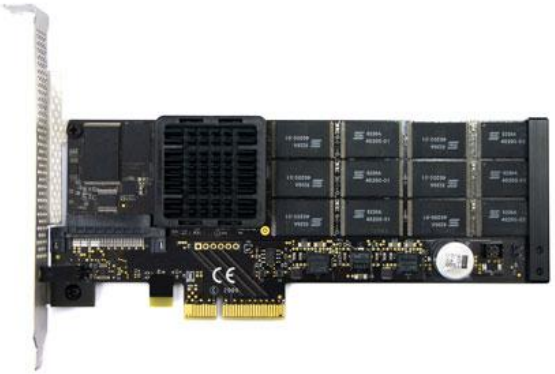
阿里巴巴 Alibaba.com

数据库拆分指标界限

- QPS？ 40000/s
- TPS？ 2000/s
- 日志数据写入量？ 20MB/s
- 数据容量？ 一天可以通过网络备份全部数据
- IOPS达到多少？ 没关系

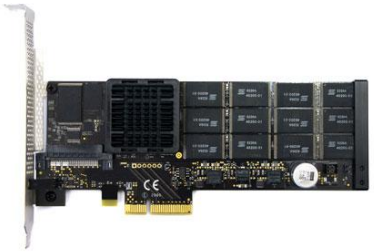
SSD

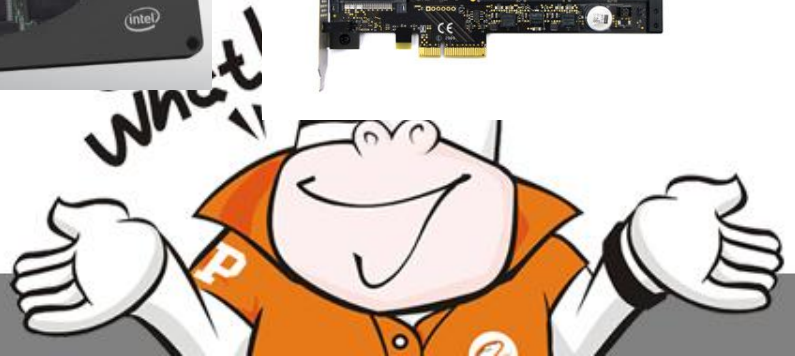
- 固态硬盘(Solid State Disk)
- 接口： USB、 eSATA、 SATA、 SAS、 FC、 PCI-E



SSD VS 磁盘

指标	15K SAS磁盘	普通企业应用SSD	PCI-E SSD
延时	5ms	100us	30us
带宽	150MB/s	250MB/s	700MB/s
IOPS(8KB)	200	15000	60000
价格	GB/5元	GB/20元	GB/100元
工作功耗	15W	5W	25W
空闲功耗	10W	0.1W	12W





SSD方向

- 带宽接近内存（3年）
- 容量超过磁盘（2年）
- 价格GB/5元（3年）
- 新的硬盘外置接口，比SAS、SATA性能更好（5年）

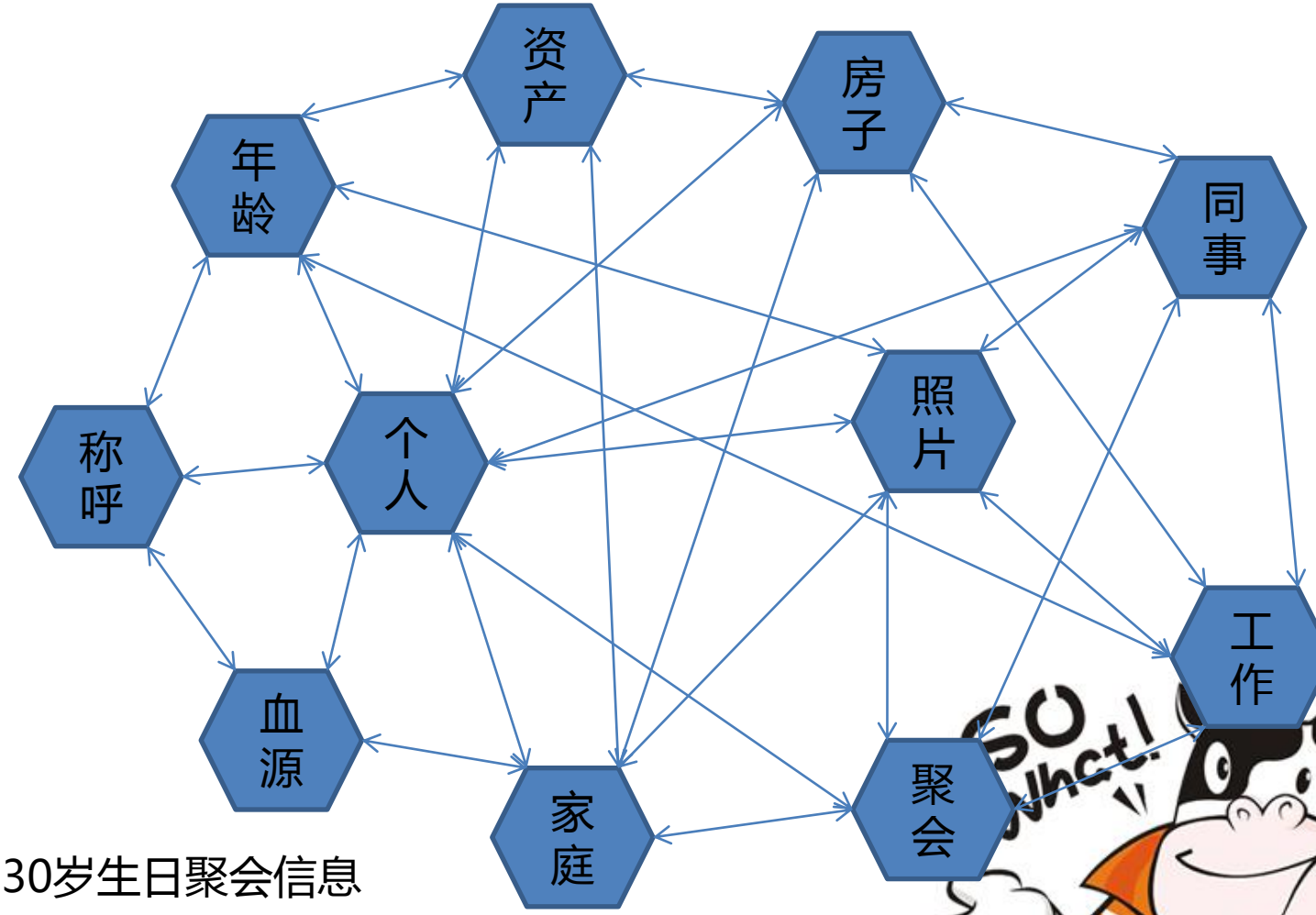
SSD对数据库性能的影响

- IOPS提高了100倍，按ID条件类型的查询性能大幅提升，memcached类上级缓存的提升性能不明显，缓存失效也不会产生雪崩效应；
- 索引的聚簇因子作用变小，聚集索引、簇表、索引组织表的性能提升不明显；
- SSD顺序写性能与磁盘没有优势，所以日志文件，归档文件放在SSD上性价比较低。

SSD对数据库发展的影响

- 采用SSD后，IOPS存在大量富余资源，传统关系型数据库已经不能满足硬件发展的需要；
- 关系型数据库更多从SQL技术性能方面考虑，适合于表格关系，但是人类思维及现实信息更像是网状关系，SSD可能会让网状关系数据库有新的崛起。

SSD与网状数据库



30岁生日聚会信息

KV vs RDBMS on SSD

KV数据库与传统数据库对SSD是同等起步，但SSD会让传统数据库满足更多性能需求场景，KV数据库在性能方向优势变小，所以需要在功能、易用性、可维护性方面突破，MongoDB就有它的亮点。

阿里巴巴 Alibaba.com

谢谢！