

网站存储经验谈

林钰

阿里巴巴-运维部



- 存储**不仅仅**是存储

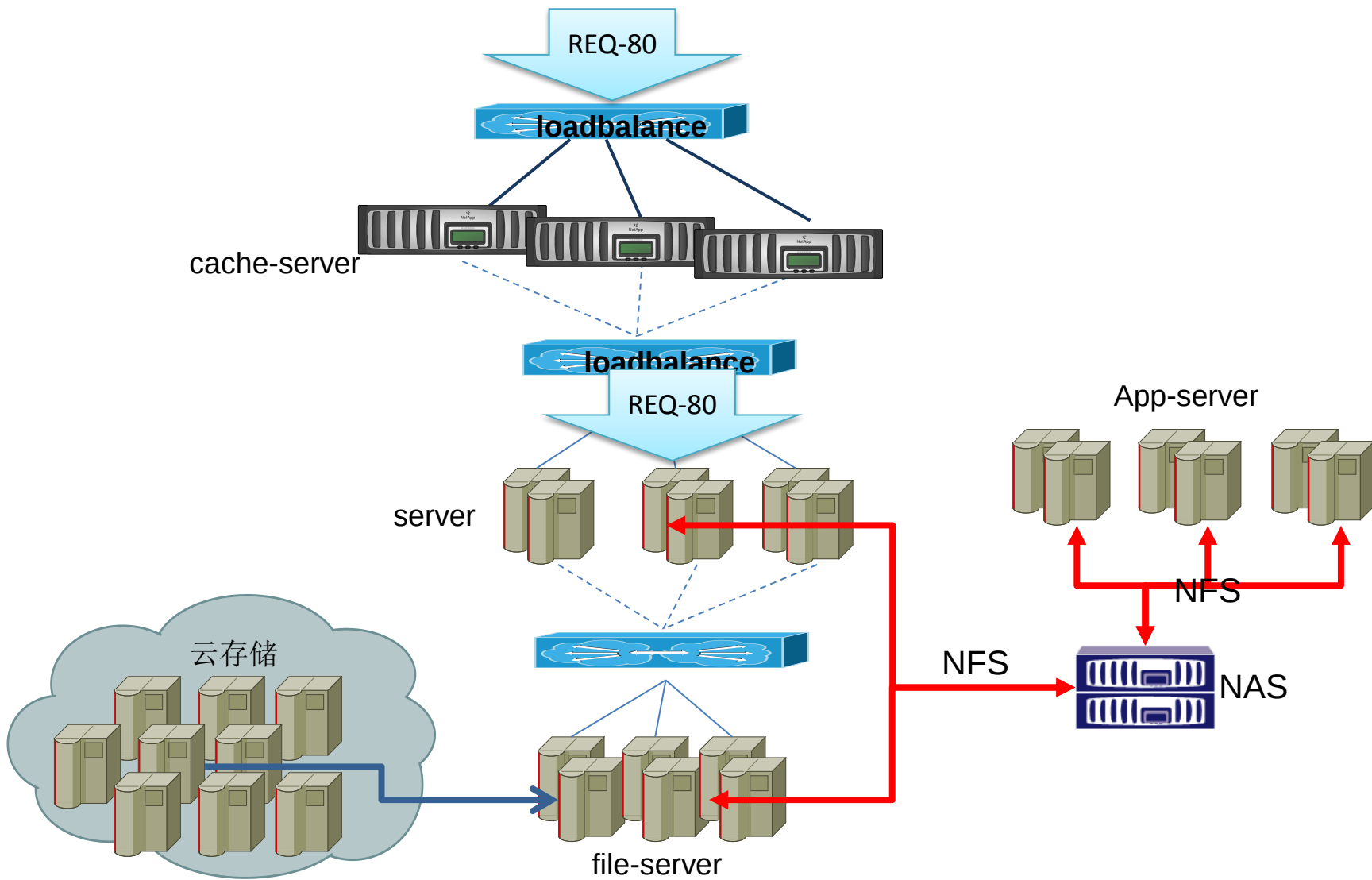
- 评估规划

- 维护优化

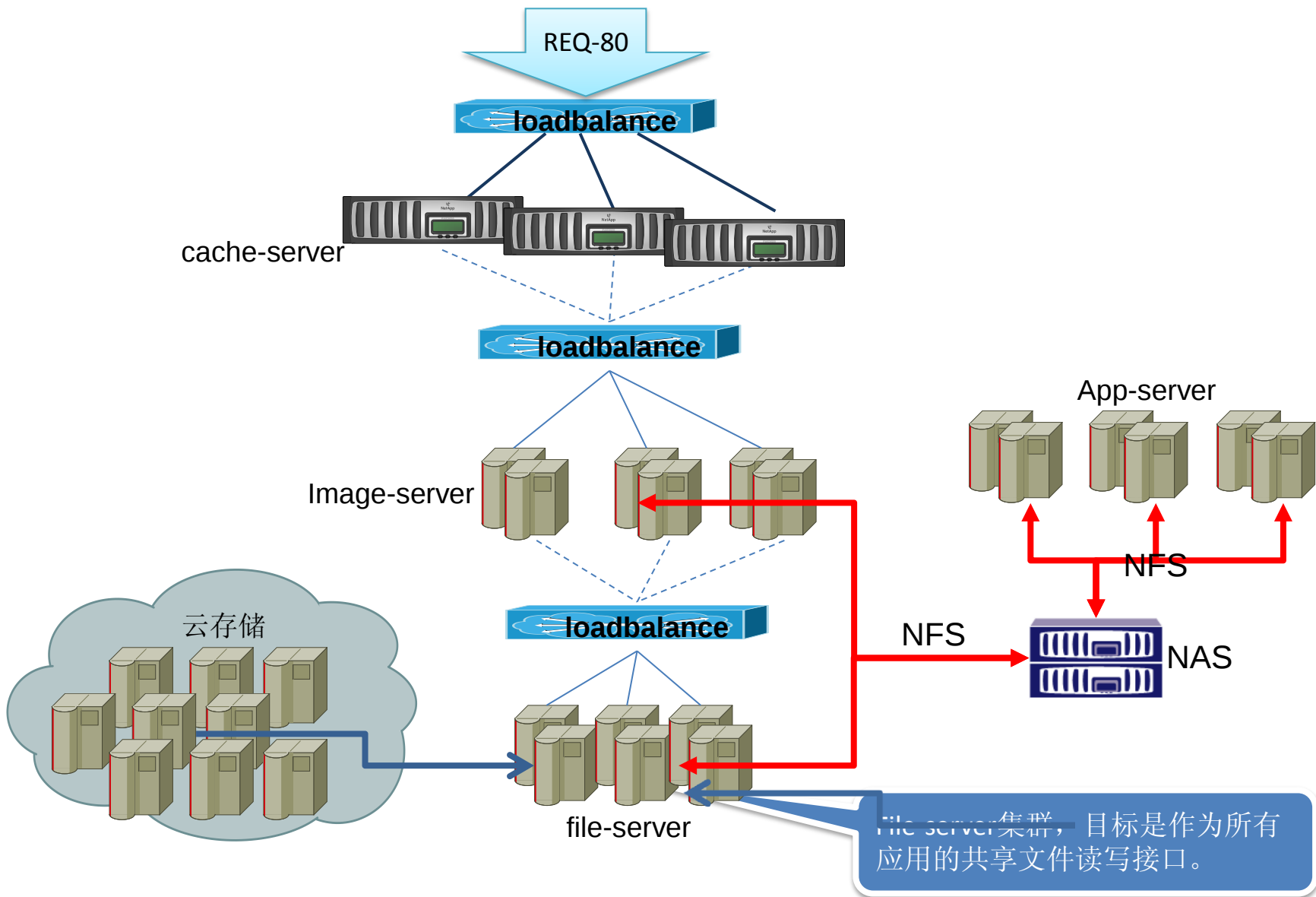
- 未来趋势



web共享系统发展



典型网站web共享系统



存在的问题

- 无规划
 - 不同应用的数据混合在一个目录中
 - 660000—— 单个目录文件数无控制
 - 100000—— 子目录数量过多
 - 随意的软链接
- 数据增长导致共享空间达上限（16TB）
- 访问增长导致存储性能成瓶颈



评估规划的经验

- 共享文件系统的组织
 - 不要等到积重难返：目录数量、层次、切分规则、文件数量
- 不同类型IO的相互影响
- 硬件降级的影响
 - RTO 个
 - 方案评估时鲜有涉及但非常重要



案例：共享目录规划

- /app/[0-9]{2}/[0-9]{2}/[0-9]{2}/[0-9]{2}/[0-9]/
- 总空间接近NAS设备单卷上限



- /app/2010/[0-9]{3}/[0-9]{3}/



- 解决单卷容量受限问题
- NAS的IO负荷下降明显



改造前后对比

- 路径形式

<http://i02.c.aliimg.com/img/offer/53/03/11/32/5/530311325.summ.jpg>

Server nfs U3: (27256945 calls)

null	getattr	setattr	lookup	access	readlink	read
0 0%	4888025 18% 0 0%		9478499 35% 11953746 44% 0 0%		925427 3%	
write	create	mkdir	symlink	mknod	remove	rmdir
0 0%	0 0%	0 0%	0 0%	0 0%	8424 0%	2809 0%
rename	link	readdir	readdir+	fsstat	fsinfo	pathconf
0 0%	0 0%	0 0%	13 0%	2 0%	0 0%	0 0%
commit						
0 0%						

Read request stats (version 3)

0-511	512-1023	1K-2047	2K-4095	4K-8191	8K-16383	16K-32767	32K-65535	64K-131071	> 131071
2528	2673	64874	351746	140002	72852	100718	190043	0	0

Write request stats (version 3)

0-511	512-1023	1K-2047	2K-4095	4K-8191	8K-16383	16K-32767	32K-65535	64K-131071	> 131071
0	0	0	0	0	0	0	0	0	0



改造前后对比 cont.

• 路径形式

http://i02.c.aliimg.com/img/offer2/2011/028/411/114820_656187364.gif

```
Server nfs U3: (637200 calls)
null      getattr  setattr  lookup   access   readlink  read
0 0%      44913 7%   10106 2%   239320 38% 160924 25% 0 0%      36863 6%
write     create   mkdir    symlink  mknod    remove    rmdir
57324 9%   48769 8%   28469 4%   0 0%     0 0%     10168 2% 0 0%
rename    link     readdir  readdir+ fsstat    fsinfo    pathconf
0 0%      0 0%     7 0%     335 0%    2 0%     0 0%     0 0%
commit
0 0%
```



```
Read request stats (version 3)
0-511     512-1023 1K-2047   2K-4095   4K-8191   8K-16383 16K-32767 32K-65535 64K-131071 > 131071
166       213       419       2622      4892      6156      8572      13823      0           0
```



```
Write request stats (version 3)
0-511     512-1023 1K-2047   2K-4095   4K-8191   8K-16383 16K-32767 32K-65535 64K-131071 > 131071
282       353       627       4966      9791      9948      12695     18662      0           0
```



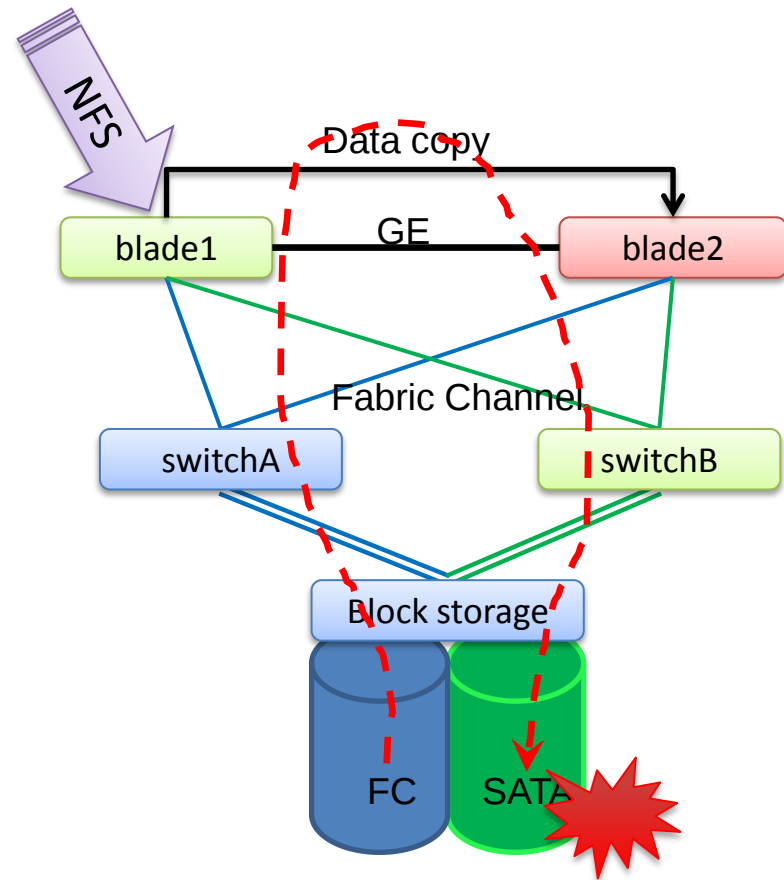
案例：硬件降级影响分析

- 正常的主备复制方案

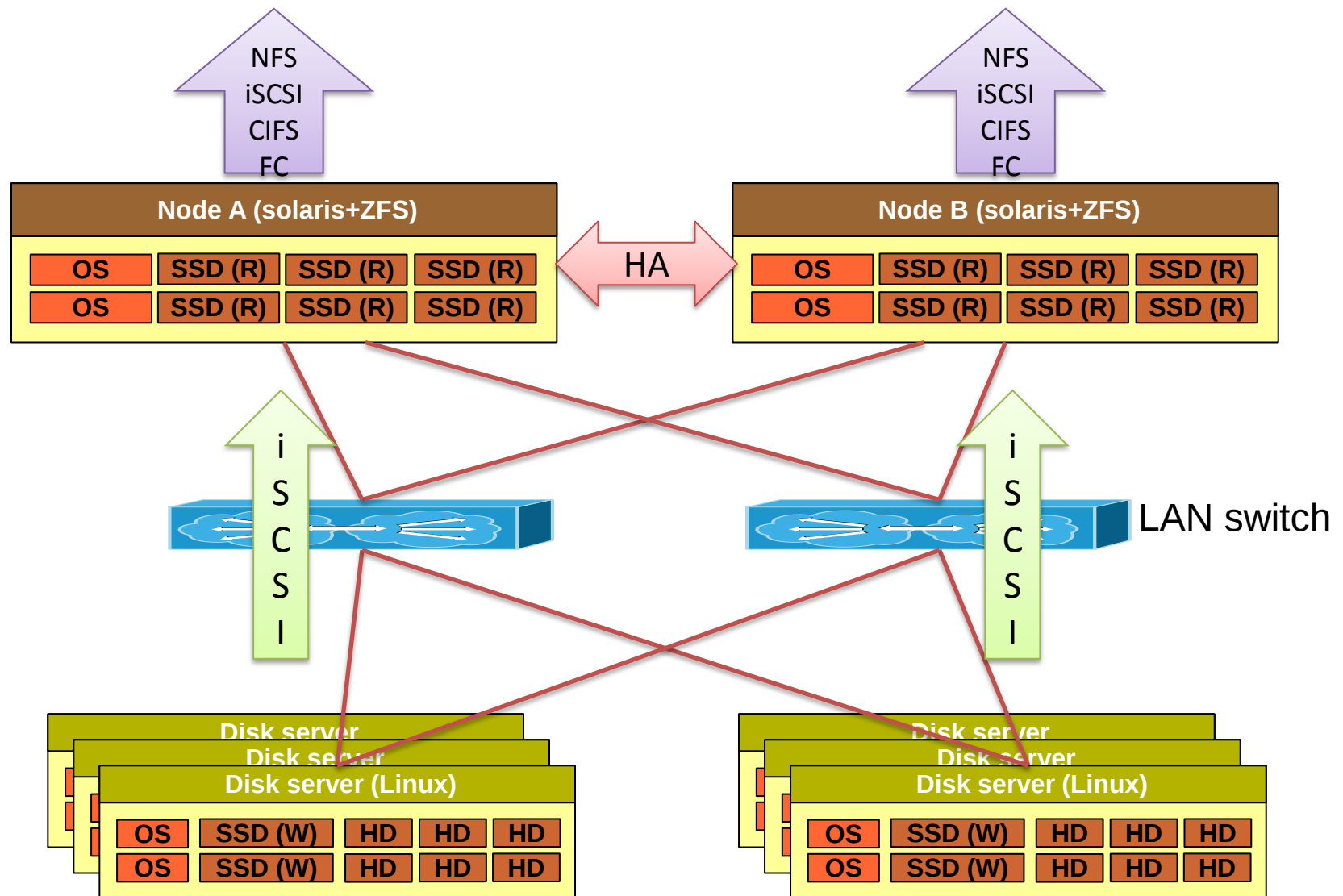
- 性能 ✓
- 成本 ✓
- 可靠性 ✓

- 硬件降级之后？

- RTO≈20h
- Cache occupied
- NFS slowly



案例：FS3存储的尝试





水漫帝都

北京市城市規劃局

sina 新闻中心 新浪首页 > 新闻中心 > 国内新闻 > 正文

三峡大坝固若金汤，可以抵挡万年一遇洪水

<http://www.sina.com.cn> 2003年06月01日08:54 金羊网-新快报

泄洪坝段曾出现的多条裂缝已全部处理万无一失

新华社记者 施勇峰

sina 新闻中心

新闻中心 > 国内新闻 > 正文

东方网-文汇报

三峡大坝 今年起可防千年一遇洪水

<http://www.sina.com.cn> 2007年05月08日08:57 东方网-文汇报

据新华社宜昌5月7日电今年三峡工程防汛腾库工作已经启动，至7日8时，三峡大坝上游坝前水位已降至147.51米。中国长江三峡工程开发总公司负责人表示，三峡大坝今年可以满足千年一遇洪水的防洪度汛标准。

搜狐新闻 | 新闻中心 > 国内新闻 > 纪念98抗洪胜利：十年生聚 十年教训 > 纪念98抗洪胜利十周年消息

三峡大坝可抵御百年一遇特大洪水(图)

2008年10月21日08:49

【我说说西包】[字号：大 中 小]

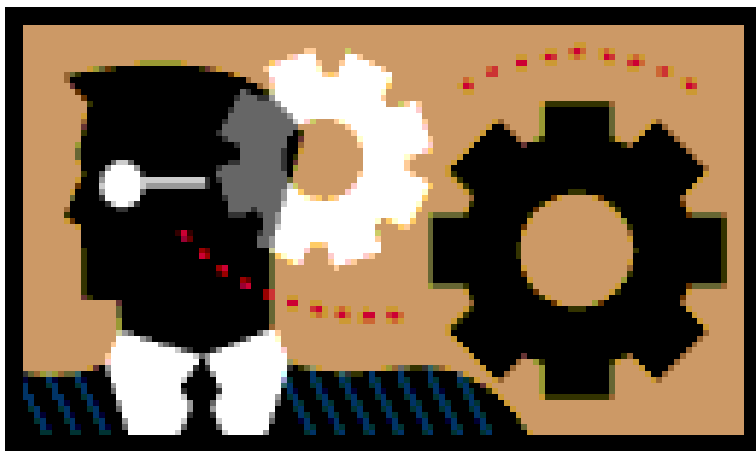
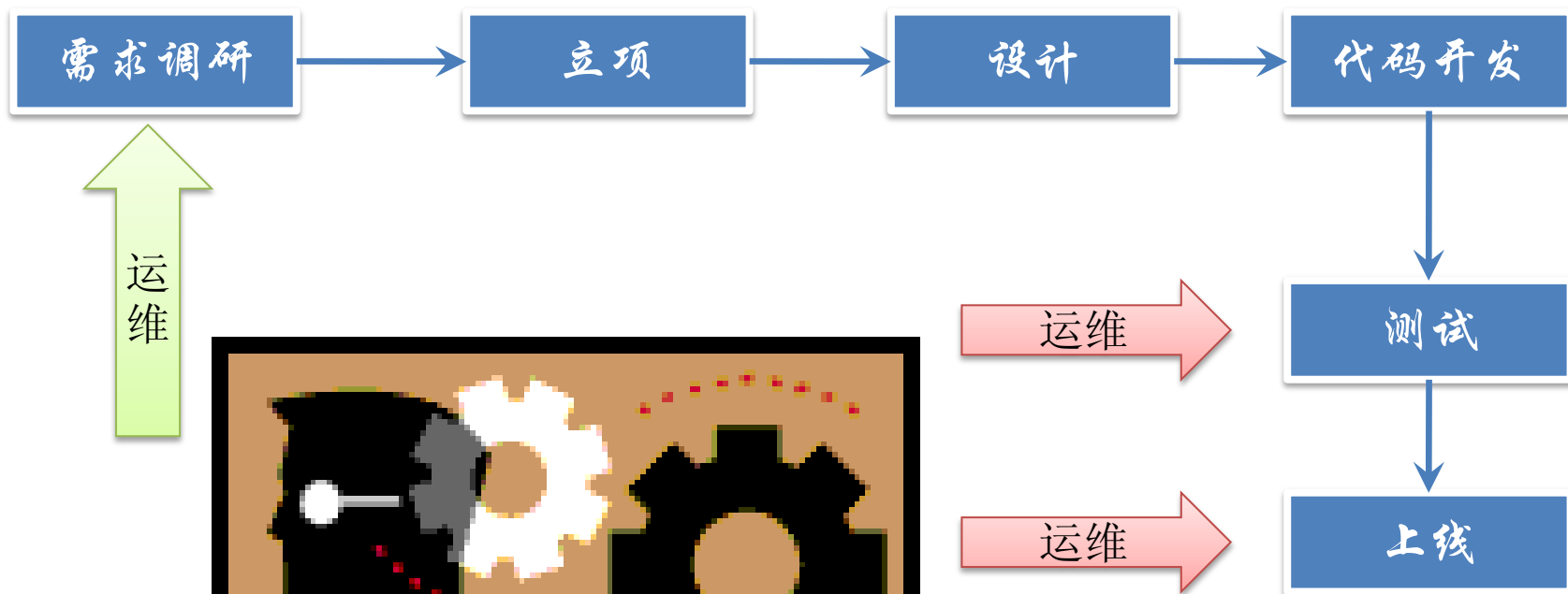


小结

- 除了考虑高端中端、磁盘数量、raid规划之外.....
 - 数据布局影响存储的IO效率
 - IO类型小变化导致大影响
 - 硬件降级一直不受重视



XX项目流程



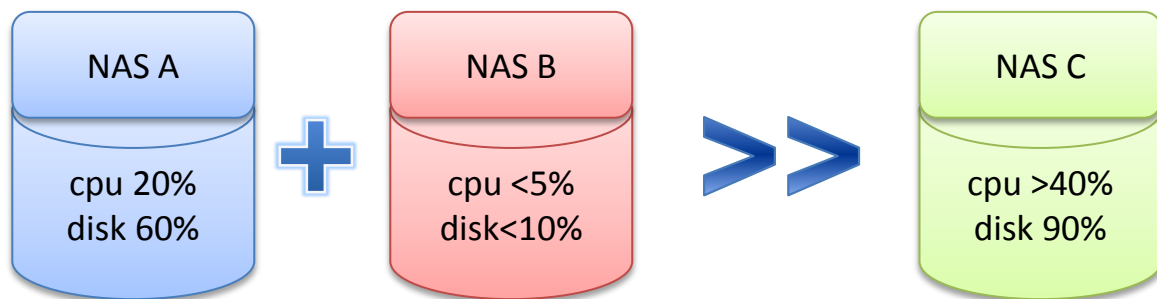
维护优化的经验

- 功夫在存储之外
 - 积累不合理、滥用的现象
 - 了解具体的业务
 - 天时、地利
- 逆向推动
 - 设备性能统计
 - 周边关联系统的表现



案例：为什么要了解业务

- 设备搬迁场景， $1+1<2$
 - A、B、C硬件性能完全一致
 - A、B数据迁移至异地C的不同磁盘组
- 发现迁移之后读缓慢
 - 唯一的遗漏点：内存
 - A应用特点是大量的元数据读



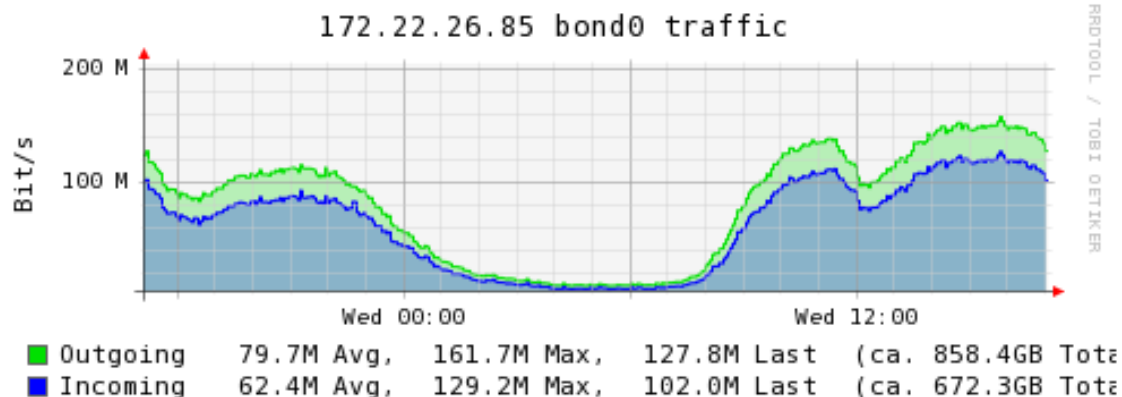
案例：关联系统影响

- 存储IO量、流量增加
 - 从存储本身无从查找增量来源
- Web-cache系统出现高频次MISS特征
 - http://img.china.alibaba.com/img/mysite/42/54/27/b3_hulingli0802_1247022222728.jpg?r=0.6438864282476
- 原因：为“所见即所得”，绕过Cache

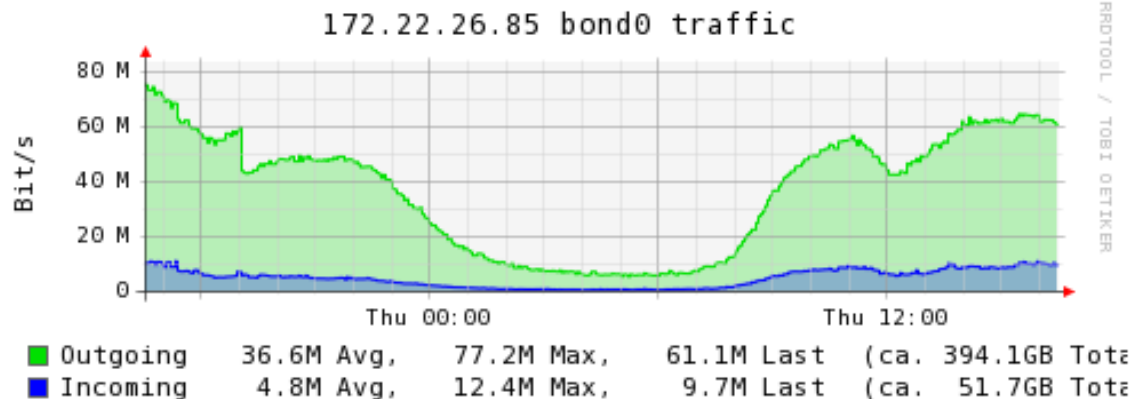


url优化前后对比

Daily

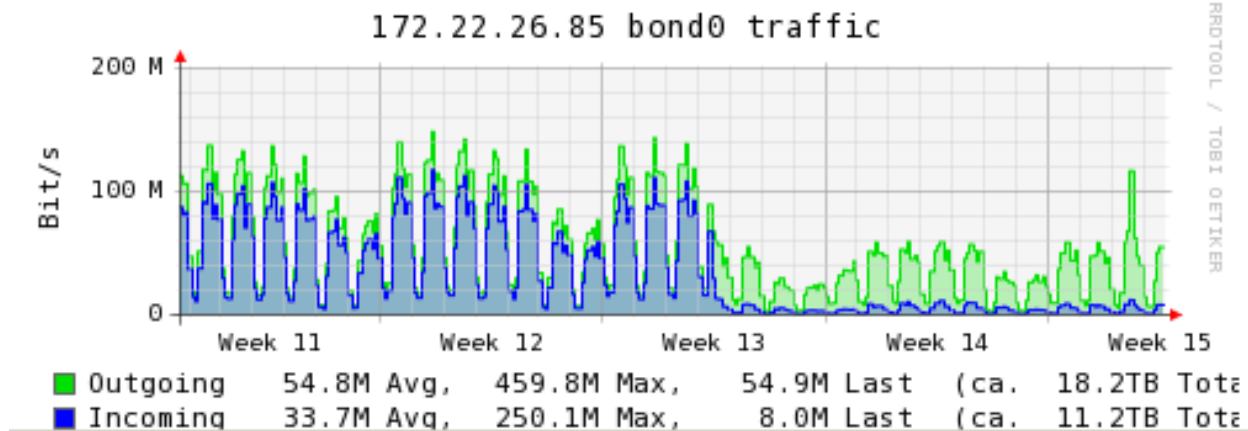


Daily



优化结果

Monthly



- 流量下降160Mbps+, 相应NAS存储总吞吐量大约300多Mbps, 相当于存储性能提升一倍
- 带宽费用效益 > 10000/月
- 用户体验改善



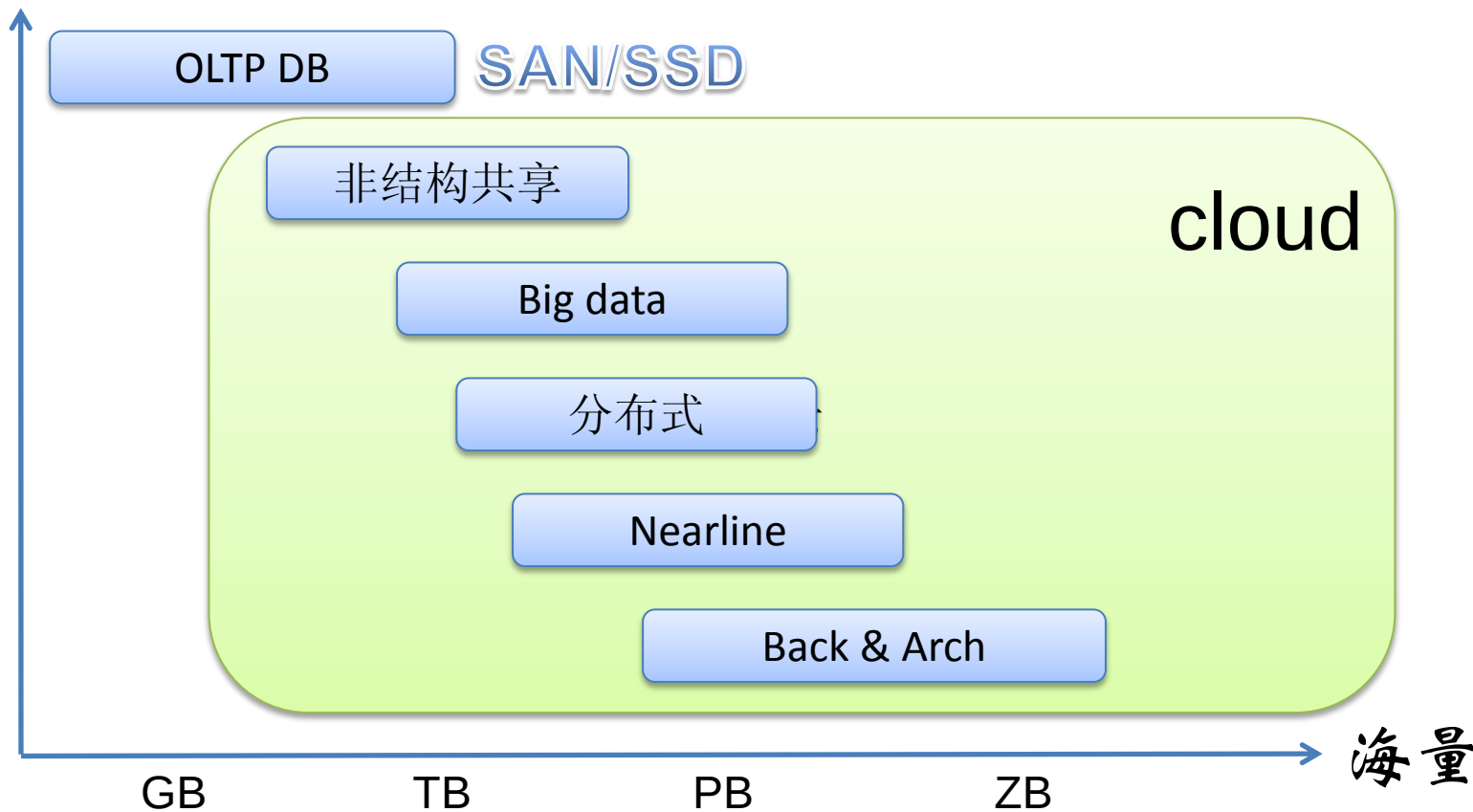
小结

- 除了“磁盘数不够了、Cache要扩容了”之外.....
 - 了解业务积累素材
 - 抓住机会搭顺风车
 - 周边关联系统的异常



存储场景

高性能



未来存储趋势

- SSD带来的变革
- 存储服务器流行
- 云存储、云计算



云存储时代的思考

- 云看起来很美 $\neq$ 撒手不管
 - 可靠性
 - 网络、电源、运维水平
 - 集群设计结构
 - 性能：并非所有场景适用
 - 可维护性
 - 用户可视可控
 - 可量化衡量
 - 数据可安全导出



回顾

- 评估规划阶段
 - 文件系统的组织
 - 注意不同IO类型的相互影响
 - 考虑硬件降级后的表现
- 维护、优化措施
 - 注意积累
 - 了解业务
 - 周边系统
- 存储趋势
 - PC+SSD: 运维成本、可靠性
 - 云存储: 运维人员仍有可为, 对OS本身的研究



Thanks. Q&A

- 欢迎关注运维领域的同学们共同探讨

