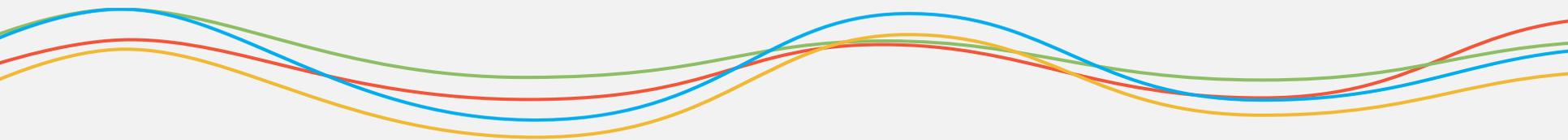


# 从传统银行到互联网金融

- Oracle数据库架构设计与性能优化实践



# Who am I

## □ 盖国强 云和恩墨信息技术有限公司 创始人

□ 国内第一个Oracle ACE及ACE总监；

□ 致力于技术分享与传播

- 技术论坛ITPUB的主要倡导者之一；
- 已经出版了12本技术书籍；
- 和张乐奕共创 Oracle用户组 - ACOUG, 开展持续的公益活动；

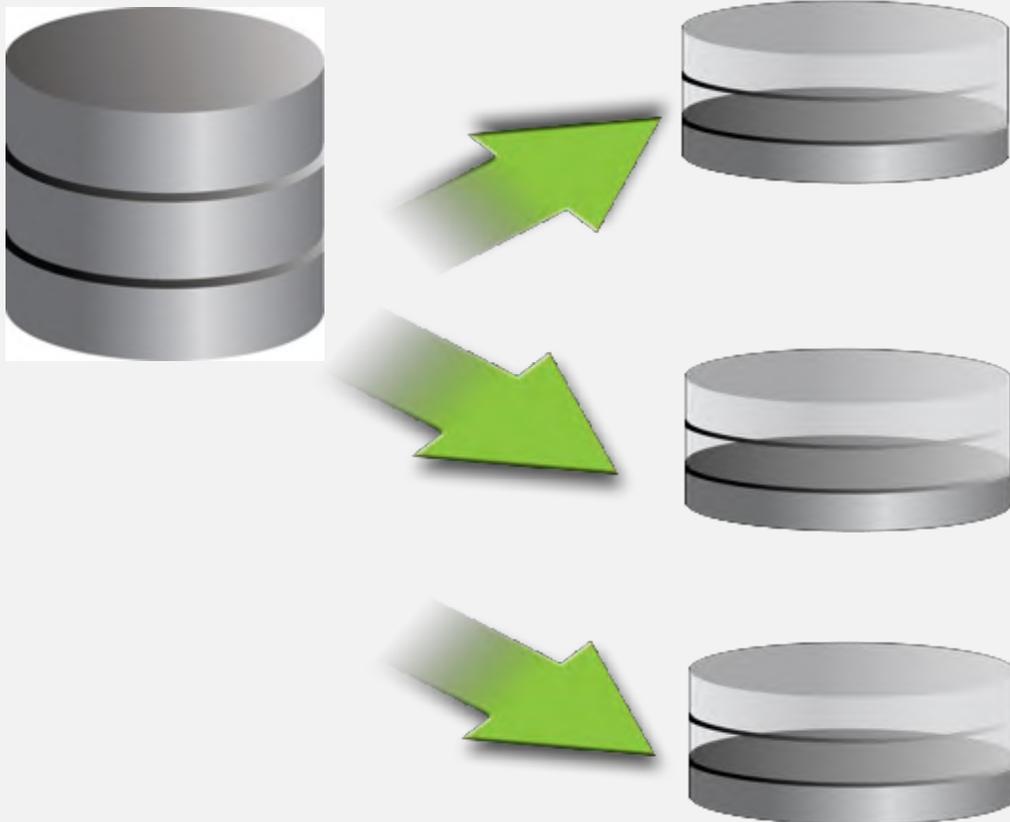


## □ 云和恩墨 国内综合数据服务领导者

□ 汇聚 Oracle ACE 总监(6/10), Oracle ACE, SQL大赛冠军, 以及数十位OCM专家, 同时具备MySQL和DB2专家；

□ 为包括电信、金融、保险、电商、能源等行业300多家客户提供服务和解决方案；





## 企业的经历:

- 数据累积 性能衰减
- 拆分数据表
- 分割数据库
- 分布式数据库
- 异构与迁移
- 业务驱动的数据库分拆

## 企业的目标:

- 提升性能
- 提高稳定性
- 保障数据安全
- 降低TCO

## 现实与挑战

- 客户在IT 建设中累积了大量的数据系统，分散割裂的部署导致了成本提升、运维复杂；
- 迫切希望通过数据整合和集中，降低软硬件成本，节省Oracle 数据库License；
- 改善IT运维，降低运维复杂度；
- 随着硬件能力的逐步提升，使得分散系统的整合归并成为可能；

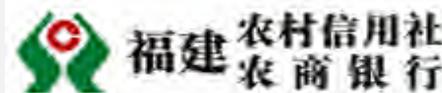


序号	名称	字符集	操作系统	需要整合 SCHEMA	数据量 (GB)	数据库版本
1	监控用户 电网理赔	ZHS16GBK	HP-UX	MONITOR	5.45	10.2.0.2
				Yaws	20	
2	统计/生产	ZHS16GBK	Linux	Stat	428	10.2.0.4
3	销售系统	ZHS16GBK	WIN2008			

序号	名称	字符集	操作系统	需要整合 SCHEMA	数据量 (GB)	数据库版本	应用类型 (OLTP/OLAP/混合)	数据库	数据库版本	字符集	数据量(GB)	SGA(GB)	高峰 DB TIME centi-second/soc	
4	精友库/生产	ZHS16GBK	WIN2008				数据仓库	SPTDB	OLAP	11.1.0.7	ZHS16GBK	2328.60	45.48	1006.14489
								MEDATA	OLTP	11.1.0.7	ZHS16GBK	8.64	4.11	141.265051
5	六个能力	ZHS16GBK	WIN2003				标准化监管	QIDB	OLTP	11.1.0.7	AL32UTF8	2.53	9.41	230.193504
								HPTDB	OLAP	11.2.0.3	ZHS16GBK	2332.20	17.47	2092.86676
								HCI0B	OLTP	11.2.0.3	ZHS16GBK	17.83	5.16	158.998505
								RPTPL	OLTP	11.2.0.3	ZHS16GBK	309.39	24.67	449.315235
6	影像系统	ZHS16GBK	Linux				财管	PPPL	OLTP	11.2.0.3	ZHS16GBK	27.31	17.47	138.78581
								DVDB	OLTP	10.2.0.4	ZHS16GBK	1877.08	41.30	2625.66546
7	门户数据库	AL32UTF8	WIN2003				信管	CMIS	混合	10.2.0.4	ZHS16GBK	597.81	34.78	183.031178
								NCD	OLTP	10.2.0.5	ZHS16GBK	566.20	4.13	169.415402
								ID08W	OLTP	11.1.0.7	ZHS16GBK	86.86	13.17	331.011067
								SPTDB2	OLAP	11.1.0.7	ZHS16GBK	29704.19	287.79	4383.11067
8	COGNOS 服务器(5.22)	AL32UTF8	WIN2003				财管EIL	JCBW	OLTP	11.1.0.7	AL32UTF8	3.72	21.58	88.3034074
	COGNOS 服务器(10.57)							ITIL	OLTP	11.2.0.3	ZHS16GBK	9.63	12.91	580.751594
	COGNOS 服务器(10.58)							门户网站	ORCL	OLTP	11.1.0.7	AL32UTF8	4.34	30.84
9	统计预审	ZHS16GBK	LINUX				计算机辅助审计	WPDB	OLAP	11.2.0.4	ZHS16GBK	1670.86	51.39	2463.97477
	统计预审							ICPTDB	OLTP	11.2.0.4	ZHS16GBK	109.65	25.31	7912.72836
							小微信息管理	ICC1DB	OLTP	11.2.0.4	ZHS16GBK	8.60	12.69	216.960056

## 现实与挑战

- 客户在IT 建设中累积了大量的数据系统，分散割裂的部署导致了成本提升、运维复杂；
- 迫切希望通过数据整合和集中，降低软硬件成本，节省Oracle 数据库License；
- 改善IT运维，降低运维复杂度；
- 随着硬件能力的逐步提升，使得分散系统的整合归并成为可能；



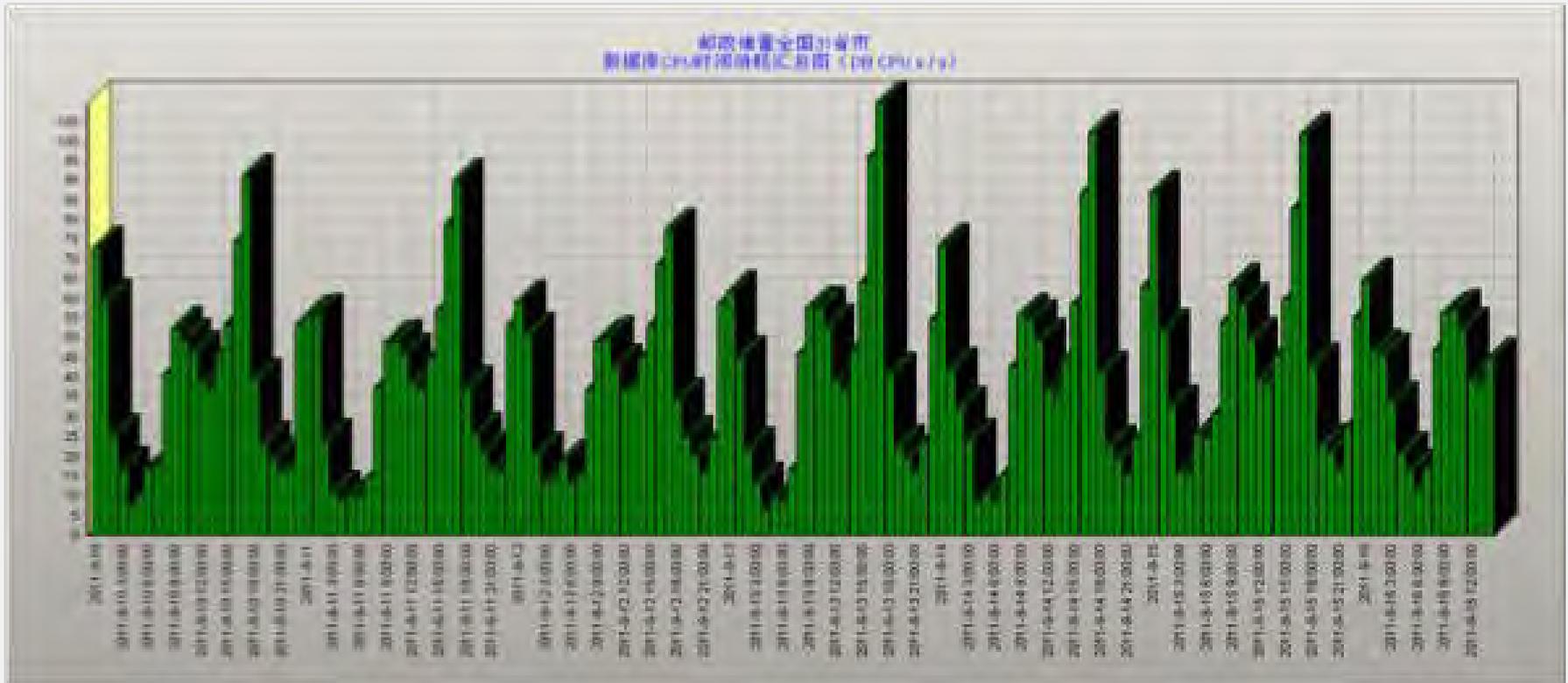
序号	名称	字符集	操作系统	需要整合 SCHEMA	数据量 (GB)	数据库版本
1	监控用户 电网理赔	ZHS16GBK	HP-UX	MONITOR	5.45	10.2.0.2
				Yaws	20	
2	统计/生产	ZHS16GBK	Linux	Stat	428	10.2.0.4
3	销售系统	ZHS16GBK	WIN2008			
4	精友库/生产	ZHS16GBK	WIN2008			
5	六个能力	ZHS16GBK	WIN2003			
6	影像系统	ZHS16GBK	Linux			
7	门户数据库	AL32UTF8	WIN2003			
8	COGNOS 服务器(5.22)	AL32UTF8	WIN2003			
	COGNOS 服务器(10.57)					
	COGNOS 服务器(10.58)					
	统计预审					
9	统计预审	ZHS16GBK	LINUX			

调研项目	数据库	高崎 DB
数据仓库	SPTDB	LME centi-
	MEDATA	second/soc
	QIDB	006.14489
标准化监管	HPTDB	11.265057
	HCI08	30.193504
	RPTPL	792.86676
	PPFL	58.998505
财管	DVDB	49.315235
信管	CMIS	58.78581
财管EIL	NCD	525.66546
身份核查	IDDB#	53.031478
数据仓库	SPTDB2	54.415402
纪检检查	JCB#	91.011067
ITIL	ITILD	383.11067
门户网站	ORCL	5.3034074
计算机辅助审计	WPDB	60.751594
小微信息管理	ICPTDB	59.556172
	ICCIDB	463.97477
		912.72836
		10.960656

依赖数据库的详细统计数据建模 Oracle 12.1.0.2 Stat 1178 Event 1650

Oracle 11.2.0.4 Stat 679 Event 1367

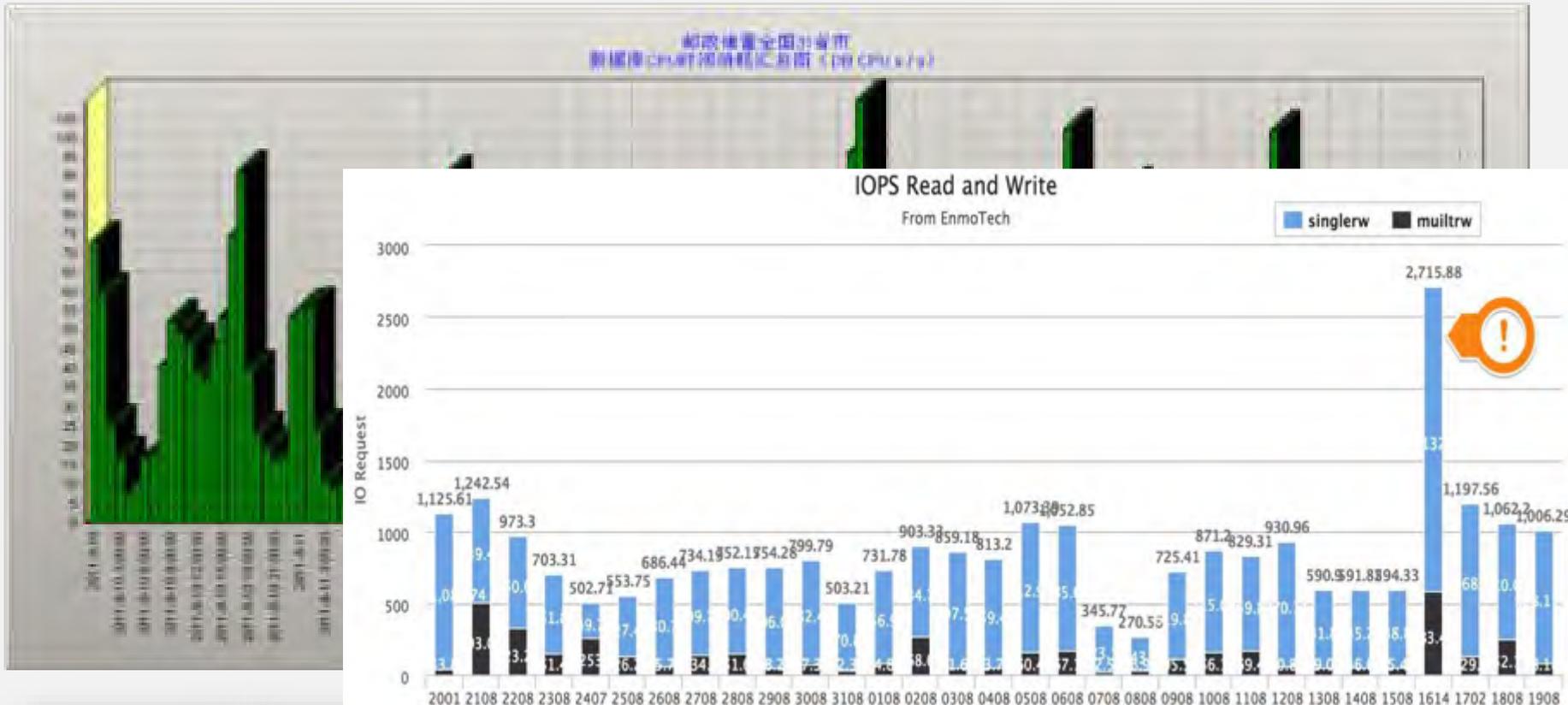
- 进行容量规划和预测
- 以量化数据指导整合迁移的资源配置



依赖数据库的详细统计数据建模 Oracle 12.1.0.2 Stat 1178 Event 1650

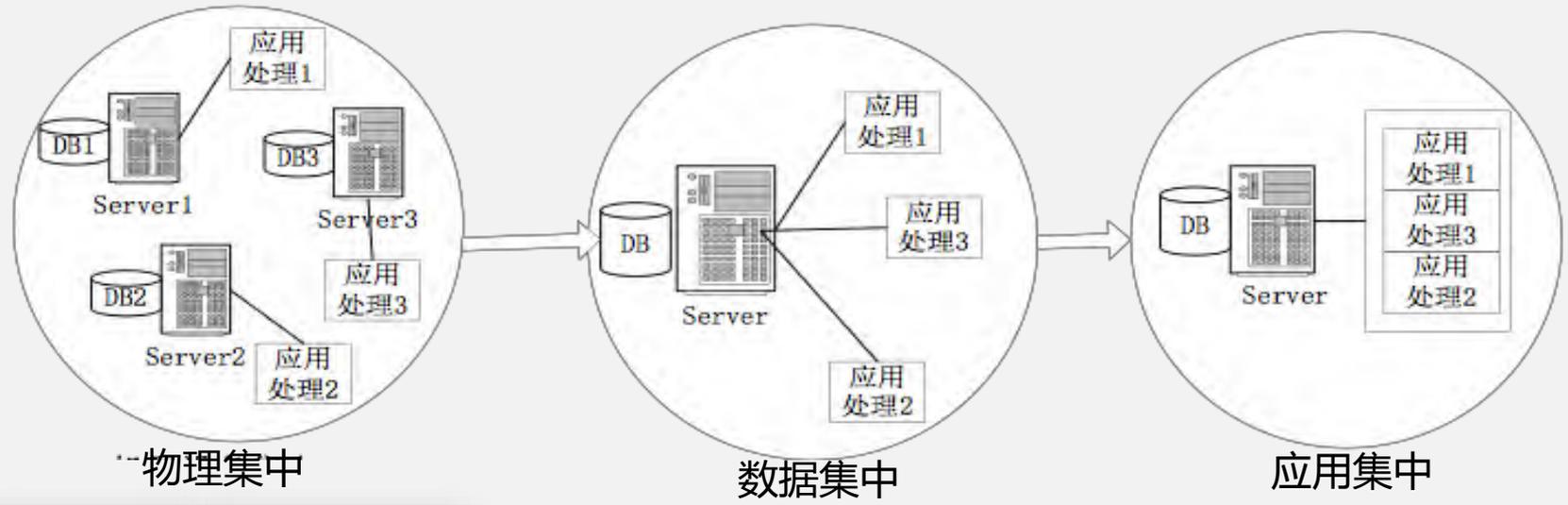
Oracle 11.2.0.4 Stat 679 Event 1367

- 进行容量规划和预测
- 以量化数据指导整合迁移的资源配置



## 金融行业率先经历了三个阶段的整合集中

- 从设备分散到集中 - 降低运行维护成本；
- 数据从分散到集中 - 提高整合能力、节约软硬件成本；
- 应用进行集中重构和服务化 - 支持业务创新与运营；
- 是企业业务发展发展到一定阶段的必然驱动；



## 邮储银行逻辑大集中

行长吕家进说，将以本次储蓄逻辑集中系统全面上线为契机，全面实施新一轮 IT 规划，致力于科技创新与业务革新的深度融合。



# 分析诊断：整合高并发的大量系统优化

Platform	CPUs	Cores	Sockets	Memory (GB)
HP-UX IA (64-bit)	62	62	16	1023.21

← 高配置

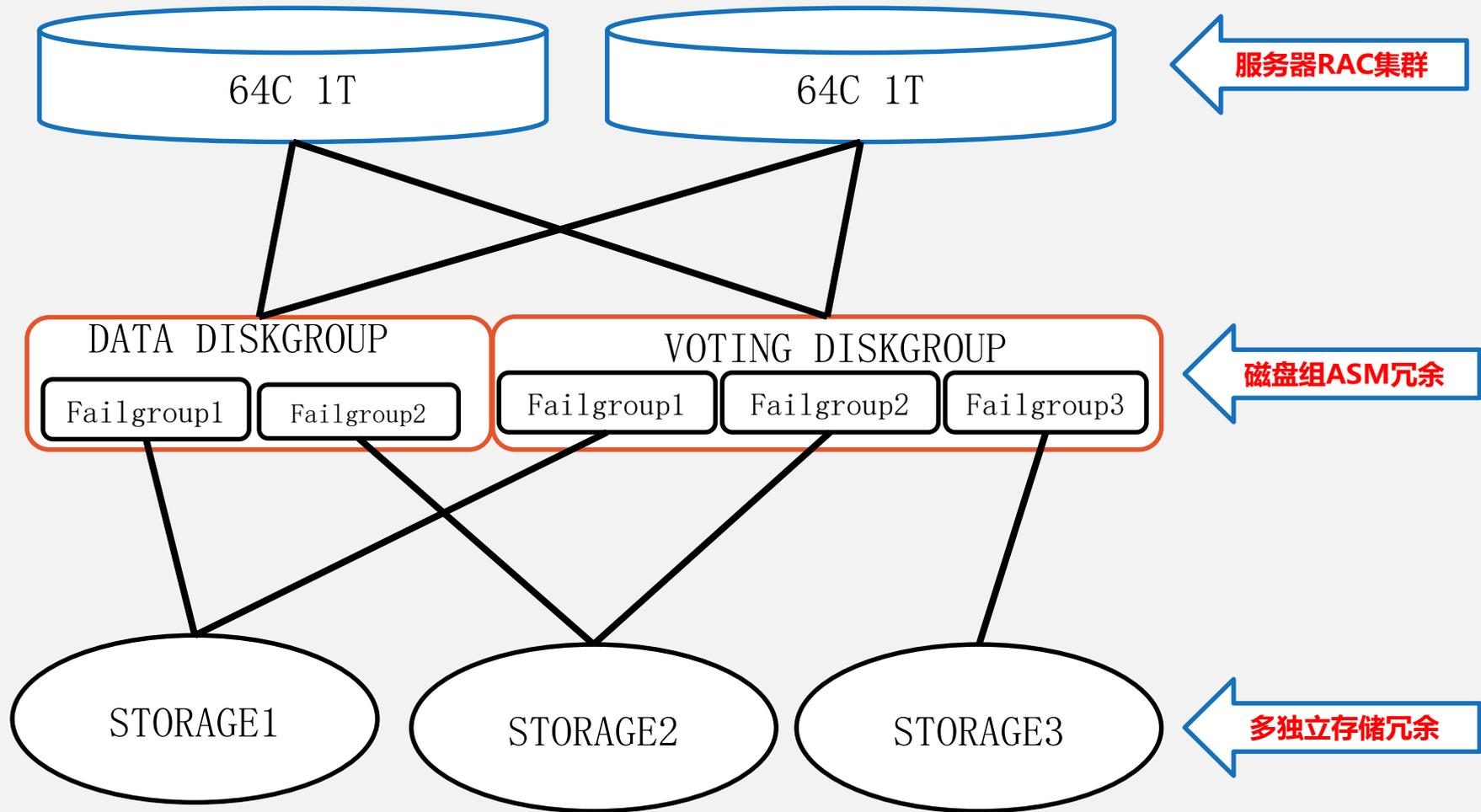
	Snap Id	Snap Time	Sessions	Cursors/Session
Begin Snap:	35976	02-9月 -14 07:00:22	4384	5.8
End Snap:	35991	02-9月 -14 12:00:18	4392	5.8
Elapsed:		299.94 (mins)		
DB Time:		3,425.69 (mins)		

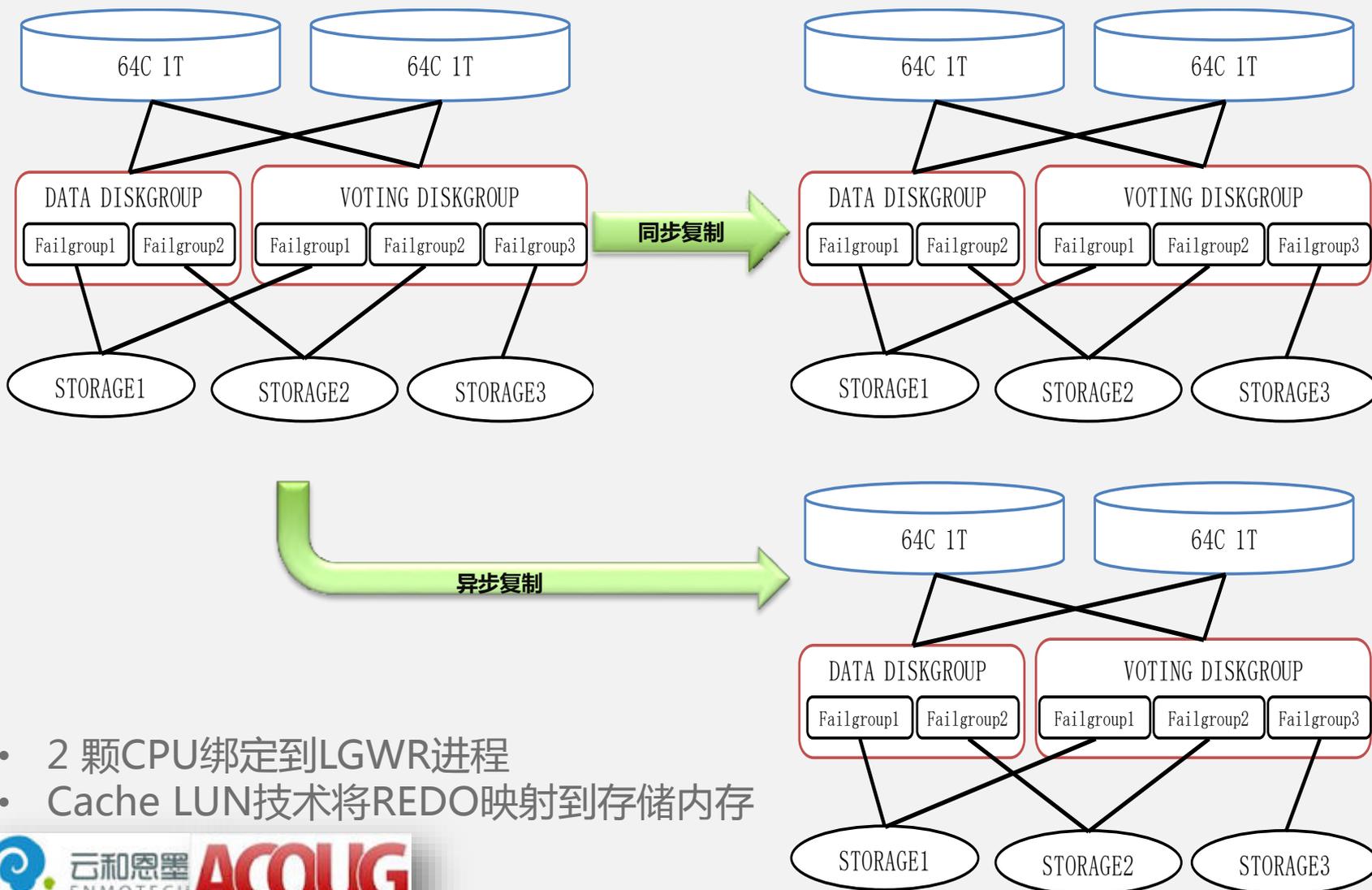
← 高并发

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
log file sync	24,298,813	96,728	4	47.06	Commit
DB CPU		79,401		38.63	
gc current block 2-way	11,621,914	12,780	1	6.22	Cluster
db file sequential read	907,707	6,682	7	3.25	User I/O
gc cr block 2-way	2,278,022	2,076	1	1.01	Cluster

← 高等待

# 高可用设计：主机存储与容灾架构





- 2 颗CPU绑定到LGWR进程
- Cache LUN技术将REDO映射到存储内存

# 分析诊断：数据库的瓶颈来自于何处

## 高并发的事务处理

- Redo Size : 2, 006, 801.5 Bytes/s
- Transactions: 1, 349.4 Trans/s

	Per Second	Per Transaction
DB Time(s):	11.4	0.0
DB CPU(s):	4.4	0.0
Redo size:	2,006,801.5	1,487.2
Logical reads:	90,649.3	67.2
Block changes:	10,321.3	7.7
Physical reads:	91.6	0.1
Physical writes:	431.6	0.3
User calls:	11,527.6	8.5
Parses:	28.2	0.0
Hard parses:	0.4	0.0
W/A MB processed:	0.2	0.0
Logons:	1.5	0.0
Executes:	6,011.8	4.5
Rollbacks:	0.2	0.0
Transactions:	1,349.4	

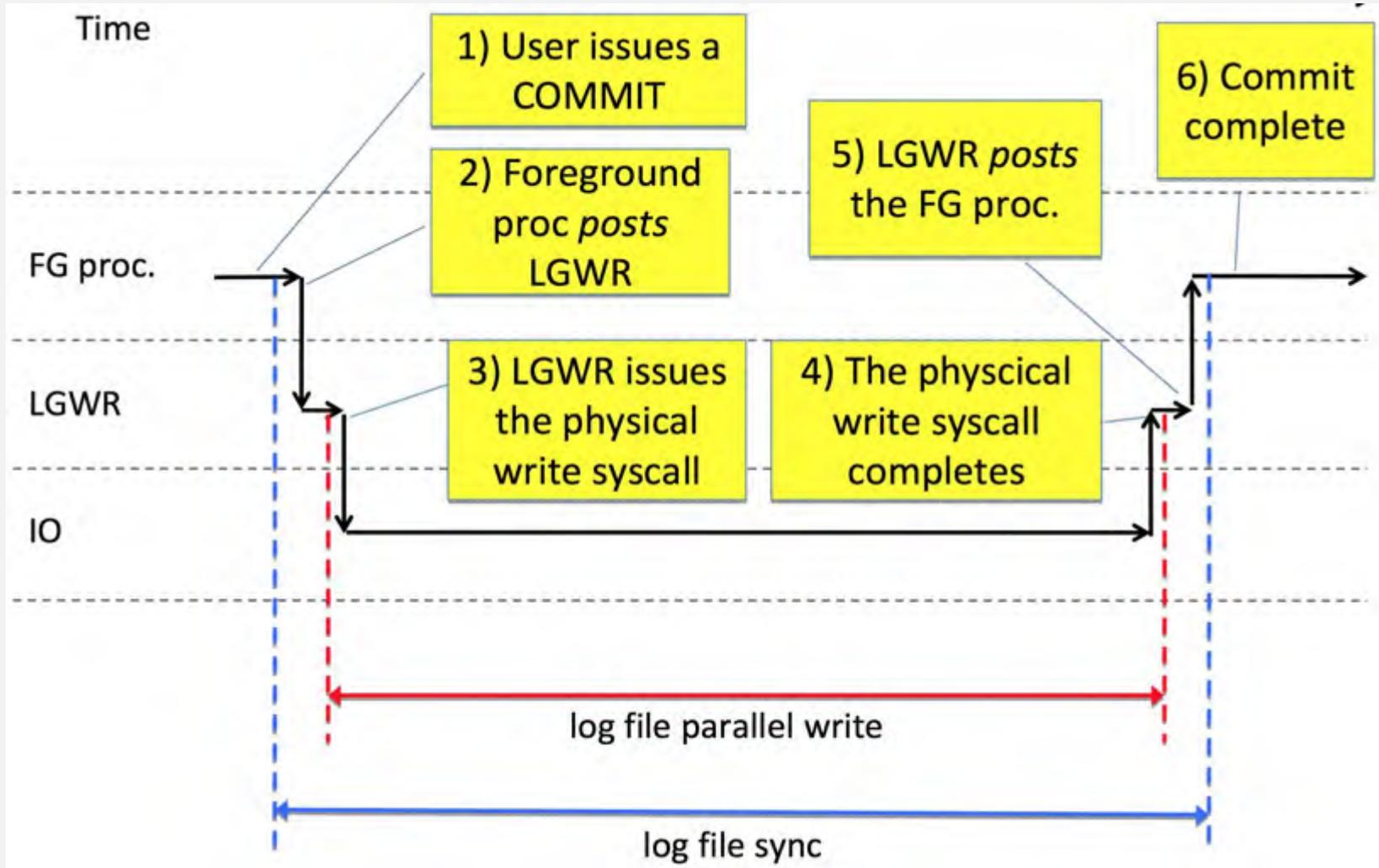
Snap Time	Load	%busy	%user	%sys	%idle	%lowait
02-9月 07:00:22	0.29					
02-9月 07:20:28	0.32	26.43	21.11	5.32	73.57	0.14
02-9月 07:40:38	0.36	29.16	23.37	5.79	70.84	0.16
02-9月 08:00:48	0.42	32.70	26.45	6.25	67.30	0.32
02-9月 08:20:53	0.50	38.35	31.38	6.97	61.65	0.42
02-9月 08:40:03	0.55	41.67	34.30	7.36	58.33	0.34
02-9月 09:00:10	0.60	45.27	37.61	7.68	54.73	0.26
02-9月 09:20:16	0.67	48.62	40.61	8.01	51.38	0.26
02-9月 09:40:24	0.70	52.86	44.30	8.58	47.14	0.26
02-9月 10:00:32	0.72	51.63	43.29	8.34	46.37	0.28
02-9月 10:20:41	0.69	52.05	43.51	8.55	47.95	0.27
02-9月 10:40:10	0.67	50.83	42.21	8.62	49.17	0.28
02-9月 11:00:21	0.68	51.93	43.19	8.73	48.07	0.27
02-9月 11:20:05	0.67	49.45	40.81	8.63	50.55	0.30
02-9月 11:40:10	0.59	46.92	38.84	8.09	53.08	0.30
02-9月 12:00:18	0.59	46.06	38.13	7.93	53.94	0.41

## 数据库的容量规划

- 支付类业务最高每小时约3000万事务（全球）
- 金融业务最高每小时约100万事务（全国）
- 证券类业务最高每小时约50万事务（单省）

支付宝2012.11.11								
指标	全天总额	全天总额数字	每秒指标	每数据库量	事务数	每小时事务数	SQL执行	CPU
事务数	41亿	4100000000.00	47453.7037	2372.685185	1,914.59	14,092,524.00	18,783.36	64c
执行SQL	285亿	28500000000.00	329861.1111	16493.05556	59.6	214,560.00	2,837.60	48c200g
日志量	15T	15000000000000.00	173611111.1	8680555.556	40.69	146,484.00	124.43	80
内存块访问	1931亿	193100000000.00	2234953.704	111747.6852	430.81	1,550,916.00	11,722.71	
物理读	13亿	1300000000.00	15046.2963	752.3148148	754.3	2,715,480.00	6214.26	22c94g
	总笔数	无线支付			294.65	1,060,740.00	36,435.99	126c
业务交易笔数	1亿580万笔	900万笔			78.22	281,592.00	734.16	96c
	总交易额	天猫交易额	淘宝		6580.48	23,689,728.00	62,484.20	
	支付宝交易额	191亿	132亿	59亿	658	2,368,800.00	8,630.83	128c512g
	总独立用户访问	第一分钟			29.19	105,084.00	699.78	24c128c
用户访问	2.13亿独立用户	1000万独立用户			381.47	1,373,292.00	11,221.06	128c512g
	预计总量	11日申通完成量	11日中通完成		25	90,000.00	415.20	8c32g
快递	8000多万件	600万件	330万件		17.43	62,748.00	221.95	16c32g
		12日申通预计			46.84	168,624.00	1,038.44	10c28g
		720万件			1121.52	4,037,472.00	3,487.65	48c92g
					176.01	633,636.00	1,413.06	64c100g
					72.8	262,080.00	7,632.40	60c96g
					4070.8	14,654,880.00	57,065.20	256c512g

# 由表及里 - 透过表征看本质



# 由前至后 - 数据库的瓶颈来自于何处

## 深入分析 - 从前台到后台

### Background Wait Events

Event	Waits	%Time-outs	Total Wait Time (s)	Avg wait (ms)	Waits /txn	% bg time
log file parallel write	7,922,482	0	15,731	2	0.33	40.56
db file parallel write	3,456,187	0	8,566	2	0.14	22.09
gcs log flush sync	1,187,491	1	1,368	1	0.05	3.53
reliable message	3,403	8	401	118	0.00	1.03
Log archive I/O	40,593	0	160	4	0.00	0.41

	Snap Id	Snap Time	Sessions	Cursors/Session
Begin Snap:	35976	02-9月 -14 07:00:22	4384	5.8
End Snap:	35991	02-9月 -14 12:00:18	4392	5.8
Elapsed:		299.94 (mins)		
DB Time:		3,425.69 (mins)		



- 提高存储响应降低Log File Parallel Write等待时间;
- 业务端合并短小事务，降低每秒处理的事务数量;
- 初始化参数COMMIT\_WAIT=nowait;
- 增加实例，分担单个实例每秒处理事务数量;
- 数据库级切分，增加新的数据库来分担每秒处理事务数量。
- 数据库版本升级到12c

## 同时段的对比报告

	Snap Id	Snap Time	Sessions	Cursors/Session
Begin Snap:	36048	03-9月 -14 07:00:30	4393	5.8
End Snap:	36063	03-9月 -14 12:00:48	4402	5.8
Elapsed:		300.30 (mins)		
DB Time:		2,642.16 (mins)		

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
DB CPU		76,590		48.31	
log file sync	23,949,904	46,365	2	29.25	Commit
gc current block 2-way	11,665,593	15,464	1	9.75	Cluster
db file sequential read	1,611,454	8,906	6	5.62	User I/O
gc cr block 2-way	2,328,981	2,447	1	1.54	Cluster

## 从后台看前台

### Background Wait Events

Event	Waits	%Time -outs	Total Wait Time (s)	Avg wait (ms)	Waits /txn	% bg time
log file parallel write	7,922,482	0	15,731	2	0.33	40.56
db file parallel write	3,456,187	0	8,566	2	0.14	22.09
gcs log flush sync	1,187,491	1	1,368	1	0.05	3.53
reliable message	3,403	8	401	118	0.00	1.03
Log archive I/O	40,593	0	160	4	0.00	0.41

Event	Waits	%Time -outs	Total Wait Time (s)	Avg wait (ms)	Waits /txn	% bg time
log file parallel write	18,294,289	0	6,842	0	0.76	26.77
db file parallel write	3,500,665	0	2,799	1	0.15	10.95
gcs log flush sync	580,316	2	603	1	0.02	2.36
reliable message	3,622	12	536	148	0.00	2.10
Log archive I/O	45,838	0	181	4	0.00	0.71

## 从串行到并行

	Snap Id	Snap Time	Sessions	Cursors/Session	Pluggable Databases Open
Begin Snap:	358	23-Jan-14 16:07:37	735	12.0	64
End Snap:	359	23-Jan-14 16:14:28	738	11.9	64
Elapsed:		6.84 (mins)			
DB Time:		3,977.44 (mins)			

	Per Second	Per Transaction	Per Exec	Per Call
DB Time(s):	581.6	0.0	0.00	0.01
DB CPU(s):	222.7	0.0	0.00	0.00
Redo size (bytes):	349,857,547.5	5,673.2		
Logical read (blocks):	6,247,707.1	101.3		
Block changes:	1,802,463.4	29.2		
Physical read (blocks):	31,355.8	0.5		

## Background Wait Events

- ordered by wait time desc, waits desc (idle events last)
- Only events with Total Wait Time (s) >= .001 are shown
- %Timeouts: value of 0 indicates value was < .5%. Value of null is truly 0

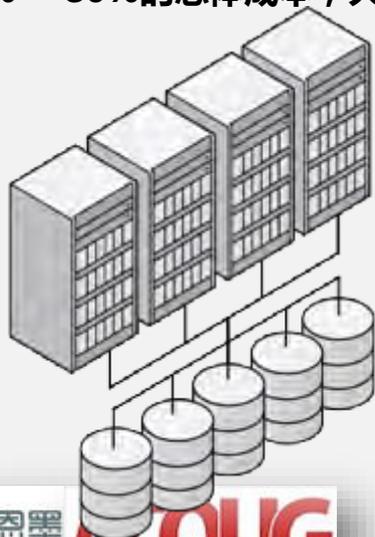
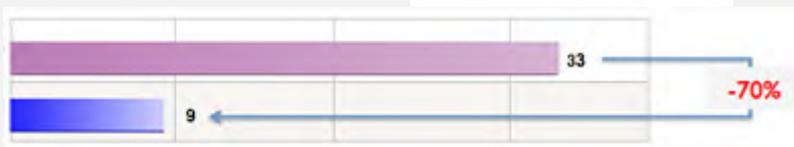
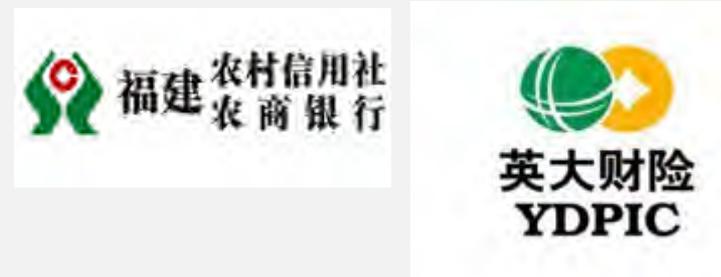
Event	Waits	%Time-outs	Total Wait Time (s)	Avg wait (ms)	Waits /txn	% bg time
db file parallel write	1,514,307	0	1,906	1.26	0.06	14.96
log file parallel write	910,059	0	1,701	1.87	0.04	13.36
LGWR intra group sync	788,563	0	842	1.07	0.03	6.61
target log write size	113,581	1	352	3.10	0.00	2.77

## 解决方案

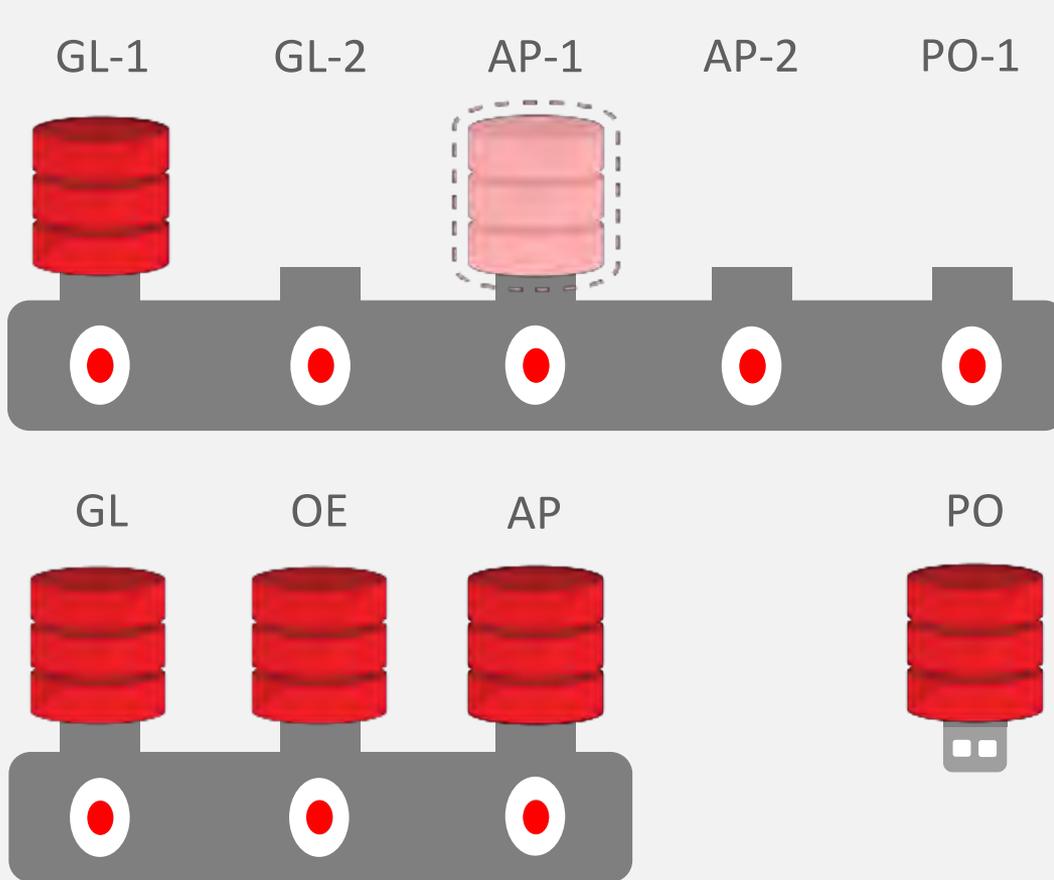
- 通过全面的数据采样与分析，进行数据整合规划，容量评估，安全和高可用评估；
- 结合业务现状、数据库现状、软硬件资源现状，为用户制定数据整合方案；
- 指导用户实施数据整合和集中；

## 成果

- 通过整合模型，进行科学的容量规划和性能推演，制定了科学的整合方案；
- 将客户现有的数十套数据库，整合为1、2、3套数据库环境；
- 降低了70% ~ 80%的总体成本，大大改善了用户的运维管理；



# 12c 多租户 – 快速的数据库分分合合

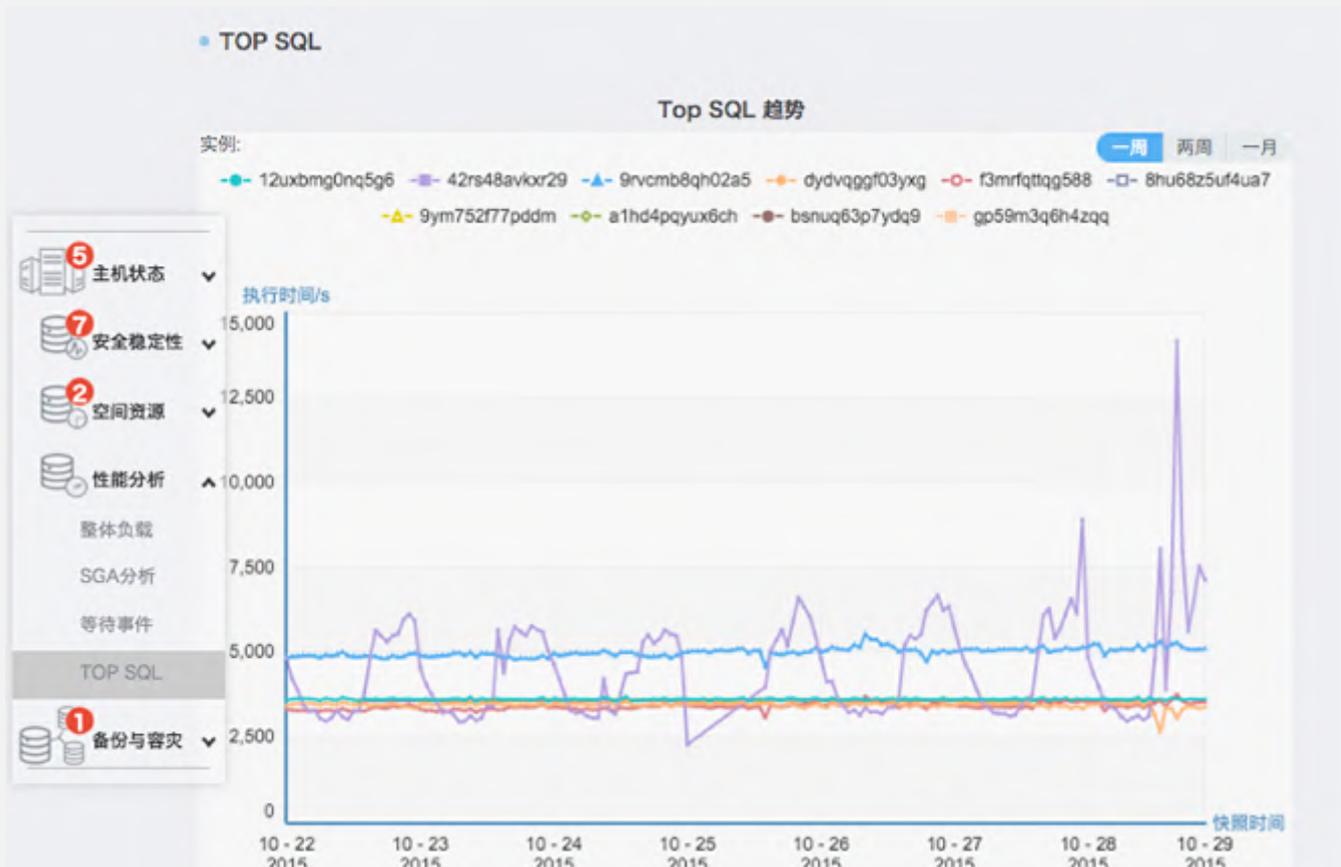


- PDBs 可以在 CDBs 内部进行快速的克隆

```
CREATE PLUGGABLE DATABASE  
newpdb FROM salespdb
```

- PDBs 可以从远程 CDBs 进行克隆
- PDBs 可以从 non-CDBs 进行克隆
- 通过 *snapshot* 进行秒级快速克隆

- 自动化巡检 - 让DBA去完成那20%最有价值的工作



## — 背景

- 一年业务增长超过100倍
- 预计在今后一年内，业务量仍然会有数十倍增长

## 架构

- Exadata 1/8配一体机
- 已采购Exadata X5满配一体机

## — 困惑

- 当前数据库架构所能支撑的业务极限

## CPU资源主要消费

逻辑读

硬解析

闪锁

排序

## 现有环境CPU资源

1/8配Exadata 2节点18CPU36核

满配Exadata 8节点36CPU72核

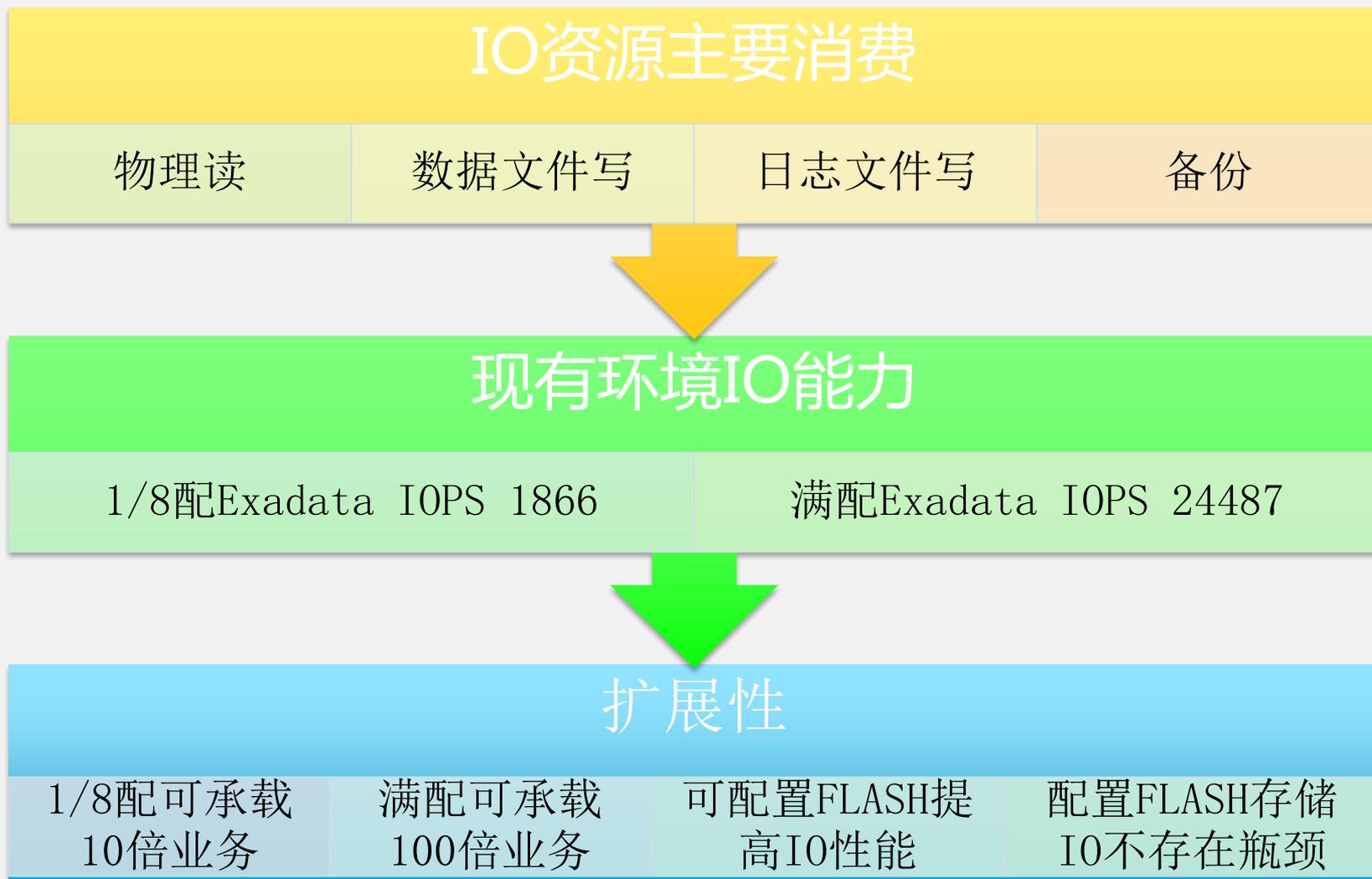
## 扩展性

1/8配单节点可  
承载45倍业务

满配单节点可  
承载90倍业务

可通过多节点  
扩展CPU资源

SQL优化前提下  
CPU不存在瓶颈



## 心跳网络资源主要消费

节点间心跳

GC数据传输

数据库后台进程通信

## 现有环境网络带宽

Exadata Infiniband 40G网络带宽

## 扩展性

可承载7000倍业务

多节点承载业务心跳  
压力会成倍增长

心跳网络不存在瓶颈

## LGWR时间消耗

log file parallel write

CPU调度

远程节点通信

## 当前配置下资源消耗

LGWR进程目前使用率6%

## 扩展性

单节点可承载15  
倍业务

升级到满配不带  
来扩展性提升

可通过部署FLASH  
提升性能

可通过多节点业  
务进行扩展

# 互联网金融爆发性业务增长案例

1/8配

- IO为主要瓶颈
- 可承载业务压力10倍
- 通过配置FLASH可提高IO处理能力

满配

- LGWR单进程为主要瓶颈
- 可承载业务压力15倍
- 通过配置FLASH可提高LGWR处理效率

多节点

- 单个数据库的锁、热点等内部机制为主要瓶颈
- 可承载业务50到100倍
- 需要对现有业务模块改造，避免多个实例间的争用和冲突

Sharding

- 可承载业务压力超过100倍，具备良好的扩展性
- 需要对现有业务架构进行彻底重构

## Automating sharding for custom-designed OLTP applications

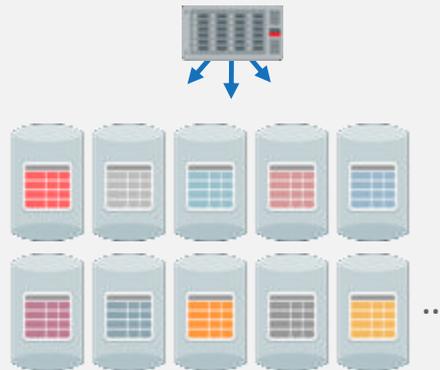
Geographic Distribution



水平扩展最高达到1,000个数据分片，每个分片包含一个数据子集；

用户可以基于性能、灾备等原因自定义数据的摆放；

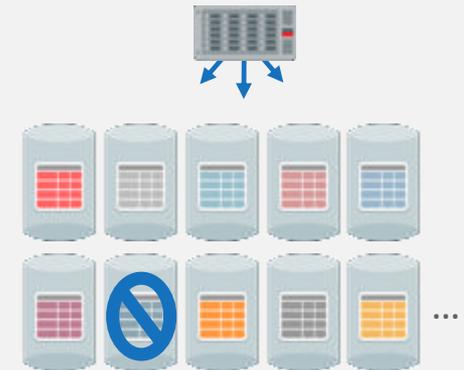
Linear Scalability



可以通过在线增加Shards扩展数据库的吞吐能力；

支持数据分片的在线 Split 和 rebalance；

Fault Tolerant



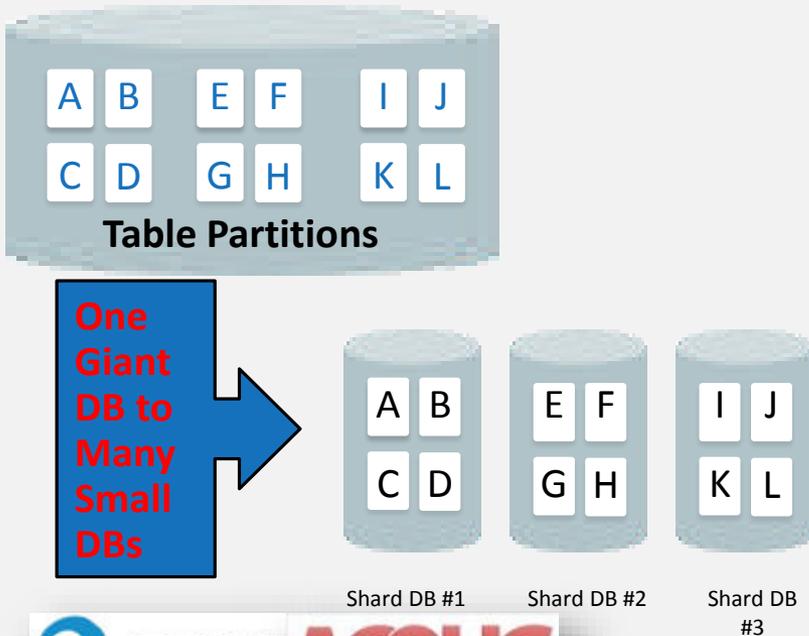
分布式软硬件隔离故障的影响度；

不同的Shards可以运行不同的Oracle数据库版本；

# 合久必分 - 由分区而Sharding

```
CREATE SHARDED TABLE Customers
( CustId    VARCHAR2(60) NOT NULL,
  FirstName VARCHAR2(60),
  LastName  VARCHAR2(60),
  ...
  PRIMARY KEY(CustId),
)
PARTITION BY CONSISTENT HASH (CustId)
```

- Sharding 技术基于分区技术演进而来;
- Sharded Table 关键字指定分片数据表;
- 建表语句创建执行后会自动将数据按照分片方式分布;
- 支持多重数据分片方式:
  - System managed - consistent hash
  - Composite - range-hash, list-hash
  - User defined - range, list (12.2.0.2)
- 连接池通过 Shard Key 来分发查询请求, 支持跨Shard的查询;
- 常规参考表自动在Shard之间复制同步;



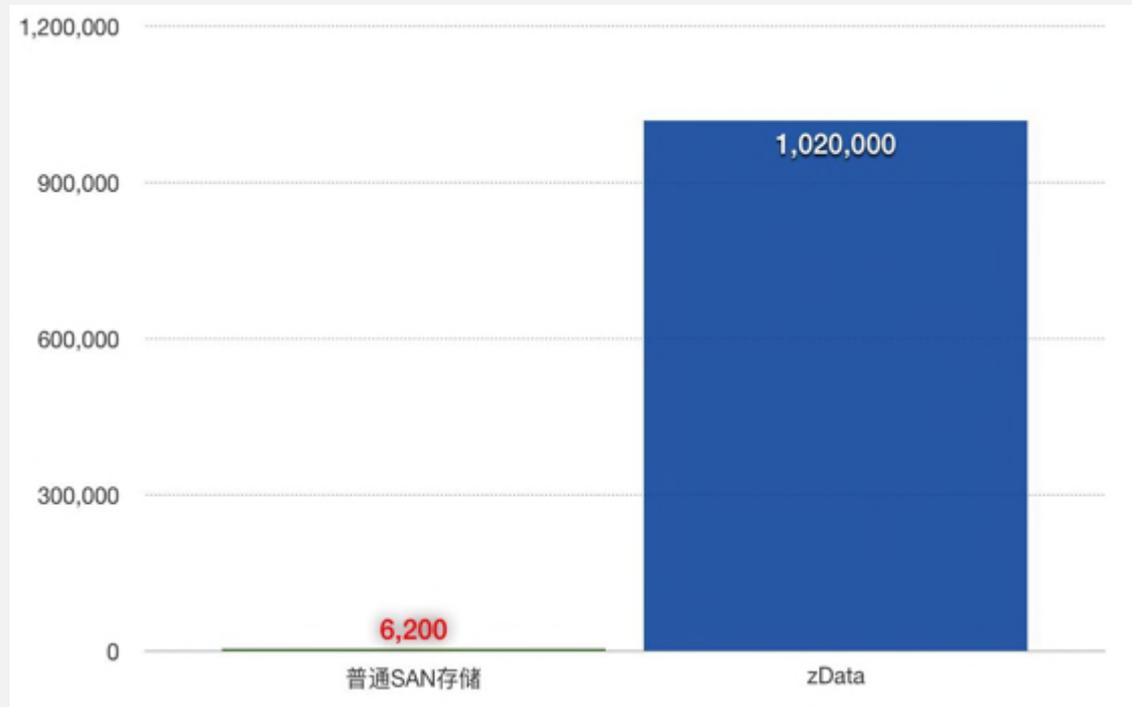
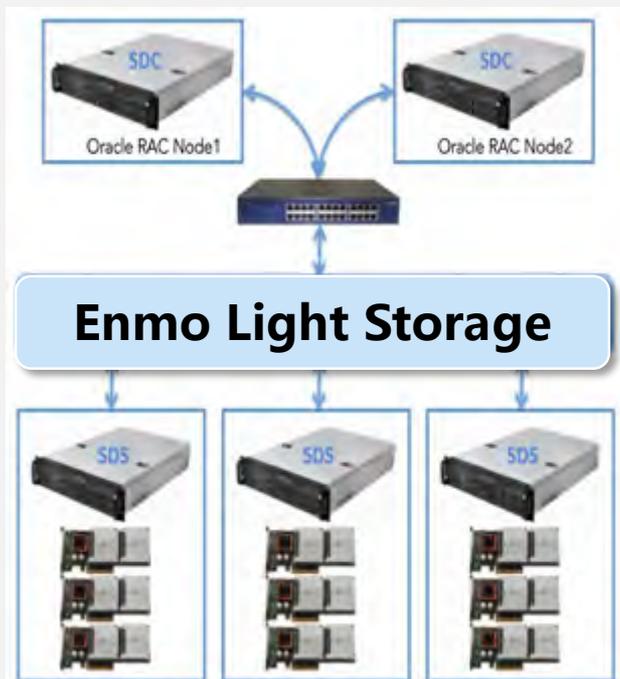
# 分布式 - 弹性分布式存储解决方案

大数据整合与集中面临的平台压力

- 去“IE”架构通过Virtual SAN替代FC SAN
- 同样的成本获得 20倍+ 的IO性能
- 动态扩展、高性能的存储解决方案



**Solutions and Services**  
**Integrated to Work Together**



# 数据为王 - 数据库性能数据决断未来

明确业务增长预期并转换为数据库指标

深入分析发现当前系统瓶颈

综合考虑可用性、性能和扩展性

根据业务预期和瓶颈进行容量预测



关注『Oracle』公众号



关注『Eygle』个人信息

**THANK YOU**

