

# 降低成本、提升速度 开放式超高性能数据库存储平台实践

YUNHE ENMO (BEIJING) TECHNOLOGY CO.,LTD

云和恩墨 成就所托

Kamus@Enmotech

# About Me

新浪微博：@小事儿爹

- **Technical Director @ Enmo Tech**
- **ACOUG Co-founder, President**



**ORACLE**  
ACE Director

**ACOUG**  
All China Oracle User Group

MEMBER

- <http://www.enmotech.com>
- <http://www.acoug.org>
- <http://www.dbform.com>



云和恩墨致力于以技术服务客户，以技术为用户创造价值，在技术分享和传播领域不断推动行业技术进步，迄今已经编译著作出版了12本技术书籍；



# ACOUG

All China Oracle User Group  
中国 Oracle 用户组

云和恩墨一贯支持和创立了ACOUG（中国Oracle用户组），已经成功组织了数十次大型技术活动，影响和帮助了上万人次的技术分享。



李真旭

**2015年4月16日 16:20-17:10**

**专场4：数据库迁移与升级**

《奇思妙想 - Oracle数据库跨平台迁移升级最佳实践》



盖国强

**2015年4月17日 8:50-9:40**

**主会场2 《风云再起—后IOE时代的Oracle架构变迁与创新》**



张乐奕

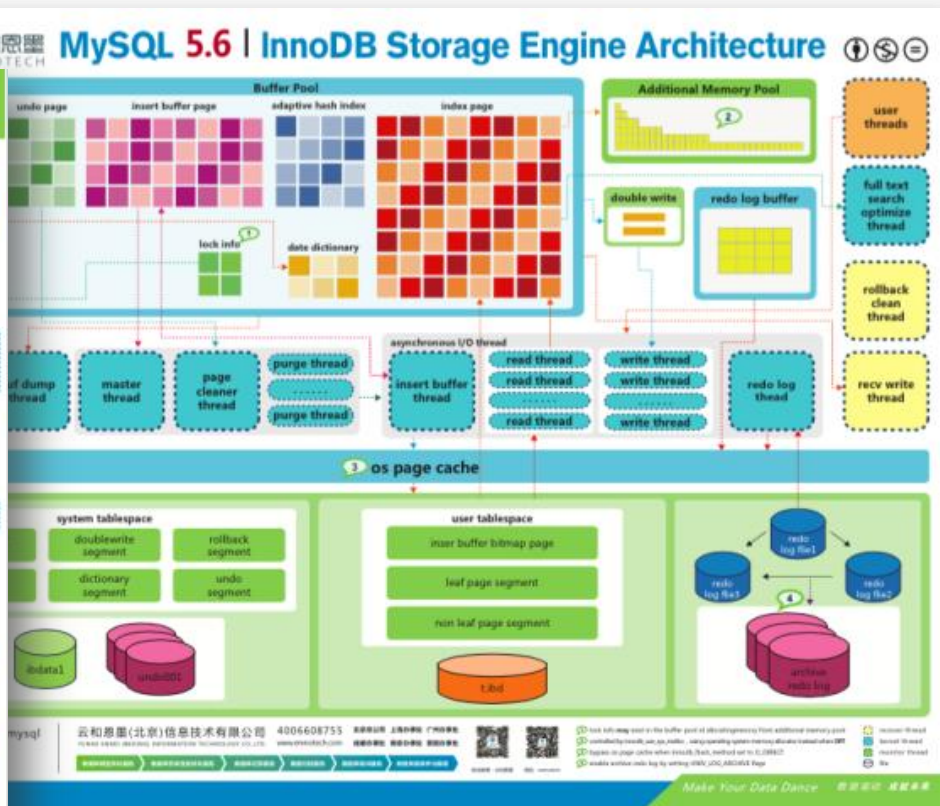
**2015年4月18日 15:10-15:50**

**专场18：存储与文件系统**

《降低成本、提升速度-开放式超高性能数据库存储平台实践》



## 2015 DTCC首发 限量精美技术海报免费领取！



在ACOUUG主办的年度大型Oracle技术分享活动-Oracle 技术嘉年华中，  
云和恩墨总会贡献最多的演讲嘉宾。

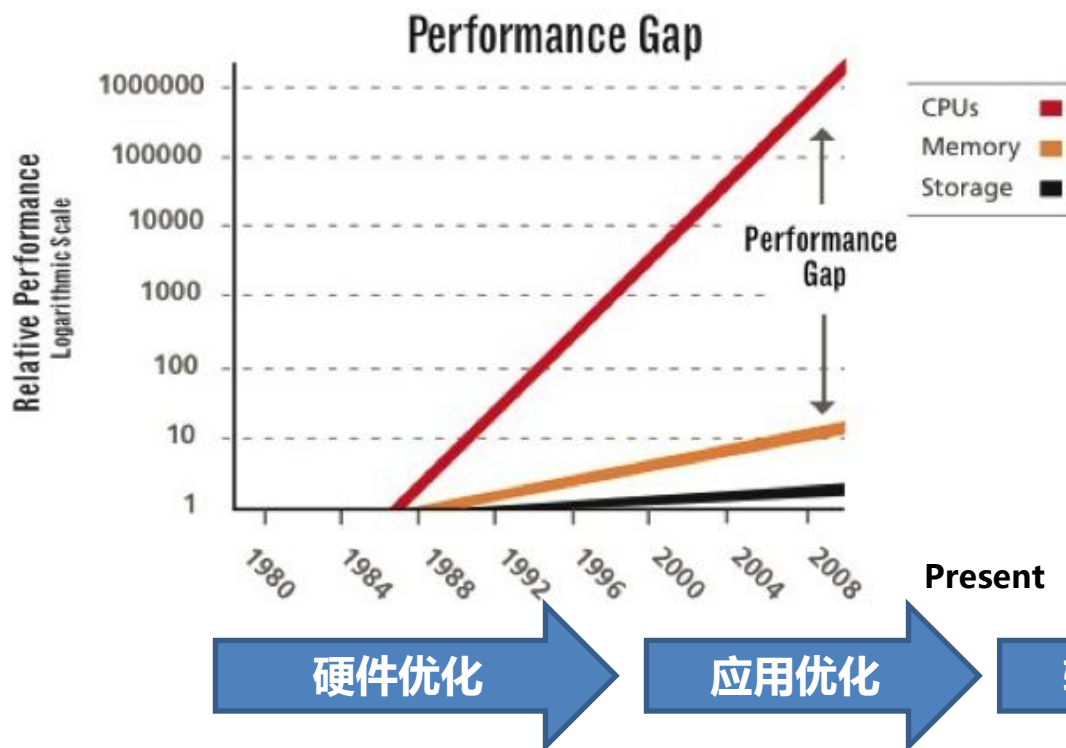


## 去IOE了？Oracle还有前途吗？

- Oracle Database标准版是企业版价格的1/3
- Oracle Database标准版功能远超MySQL
- 任何起因是钱的问题的问题都不是问题
  - Oracle不笨，Larry不因循守旧
  - 2015财年Q3，Oracle的云业务增长30%，达3.72亿美金
  - Oracle每年研发投入高达50亿美金
  - Oracle从5年前开始计划并投入研发In-Memory Option
  - Research发布Openstack收入分析预测，2014年8.83亿美金，2018年33亿美金
  - Hadoop发行商Cloudera 2014年营收1亿美金，Hortonworks 2014年营收8700万美金
- 关系型数据库到底为什么存在？ACID
- 任何领域前20%的技术人员都能生存得不错。

# 什么是数据库系统的优化

- 物尽其用
- 平衡





# IT架构趋势 - 传统SAN面临的挑战



- 两层的计算-存储架构：计算与存储完全独立分层，通过FC或iSCSI互联
- 中心化的存储系统成为I/O存取的瓶颈，扩展成本高昂
- 复杂的系统带来部署及操作、运维和管理的复杂性

传统的SAN架构: 缺乏对性能及大数据处理的灵活性  
昂贵的部署及运维成本

## • 开放架构下网络与存储的巨大进步

存储设备	SAS磁盘	SSD	PCIe Flash
IOPS	100~150	50,000+	200,000+
Throughput	150MB/s	500MB/s	4.0GB/s
Latency	10ms	100us	30us



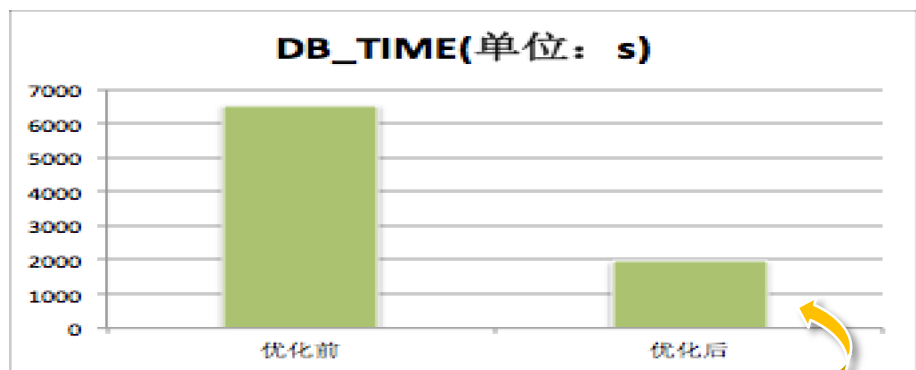
互联设备	Ethernet	SAN	Infiniband
Bandwidth	1~56Gb/s	2~16Gb/s	40~56Gb/s
Latency	10us	2us	200ns

- 
- The diagram consists of three overlapping octagonal shapes. The top-left octagon is teal and contains a white icon of a computer monitor and base, with the text '架构规划' (Architecture Planning) below it. The bottom-right octagon is green and contains a white icon of a storage device (like a SSD or HDD) with the text '存储优化' (Storage Optimization) above it. The bottom-left octagon is yellow and contains the text '生产系统' (Production System) in black. The octagons overlap such that the yellow one is at the bottom, and the teal and green ones are positioned above and to the right of it, respectively.



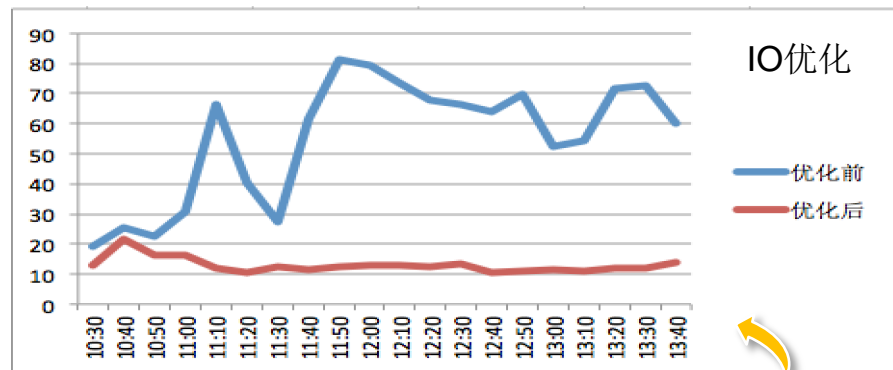
# PCIe Flash : 去大型存储简化IT架构

- 整合后的主要指标性能提升比率 > 100%



相同时间段  
1个小时内的AWR数据

指标	优化前	优化后	优化提升率
日志量	7,627,295	17,957,028	235.40%
逻辑读	71,058	236,944	333.50%
修改块	30,242	97,636	322.80%
SQL执行	4,234	5,235	123.60%



单块读指标	优化前	优化后	优化提升率
等待次数	5,948,198	15,212,847	255.76%
等待时间	340,638	31,918	1000.00%
平均等待	57	2	2850.00%
响应时间	86.87	27.29	318.50%

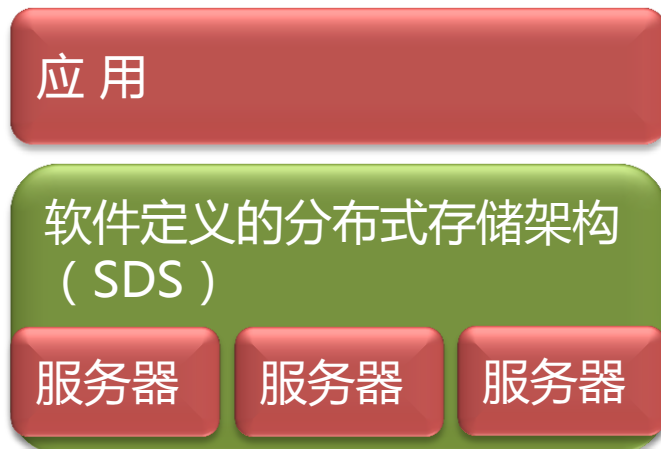
# IT架构趋势 – 软件定义存储

## 传统SAN架构



- 两层的计算-存储架构
- 中心化的存储系统
- IO存取边界，扩展成本高昂
- 部署与运维复杂
- IO系统孤立，无法了解应用的变化

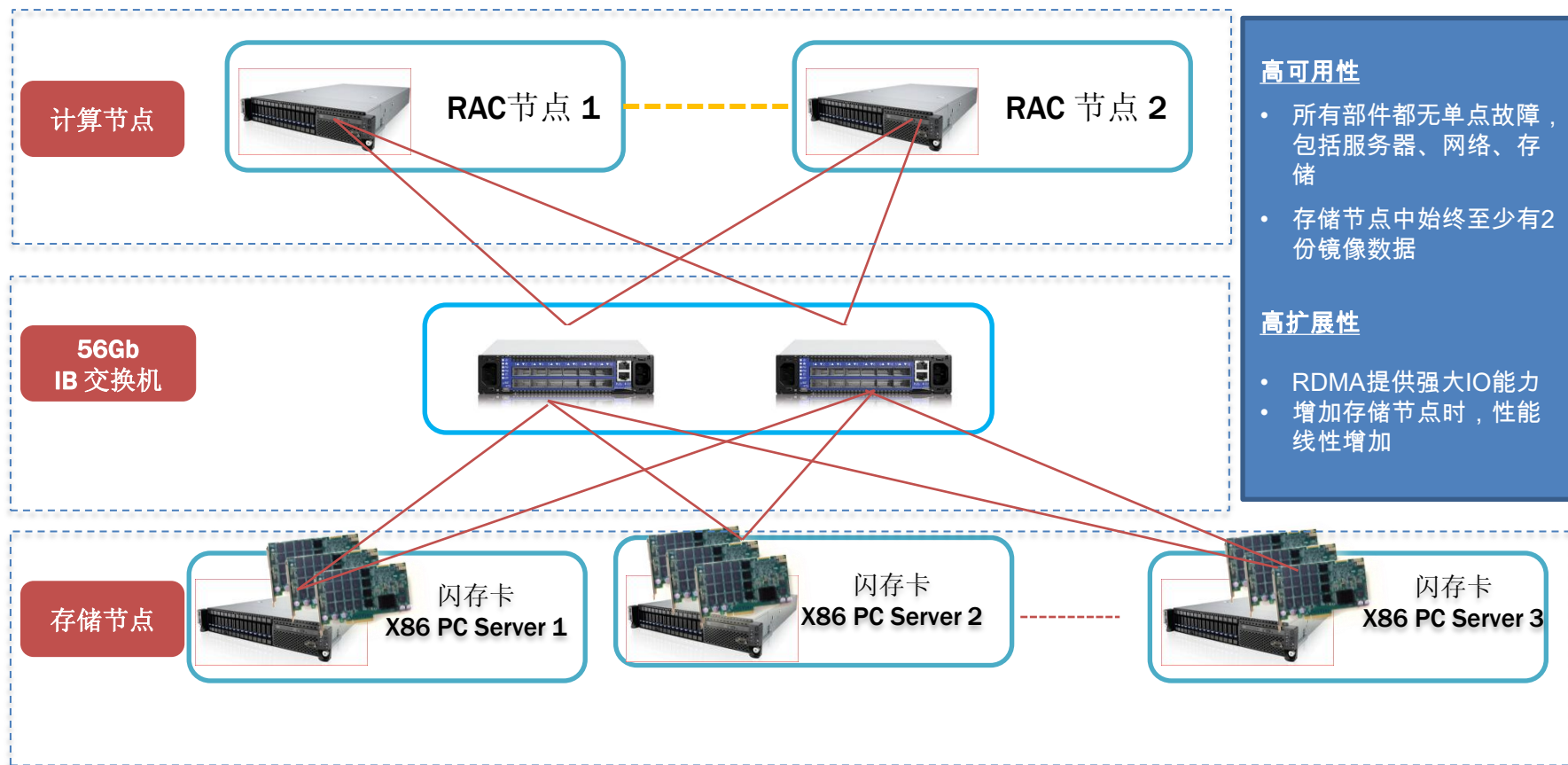
## 超级数据中心/云计算



- 软件定义的分布式基础架构: 弹性, 可灵活扩展
- 部署及维护: 开放的x86服务器



# zData Light 架构



— 56Gb Infiniband

# zData Light架构 - 开放性

开放

Database

ORACLE  
DATABASE



Light Storage

开放



Any Server  
with SSD  
or PCIe Flash



inspur 浪潮

IBM

lenovo

LSI 

开放



# zData Light架构 – 极致性能

存储节点支持基于Infiniband的RDMA协议

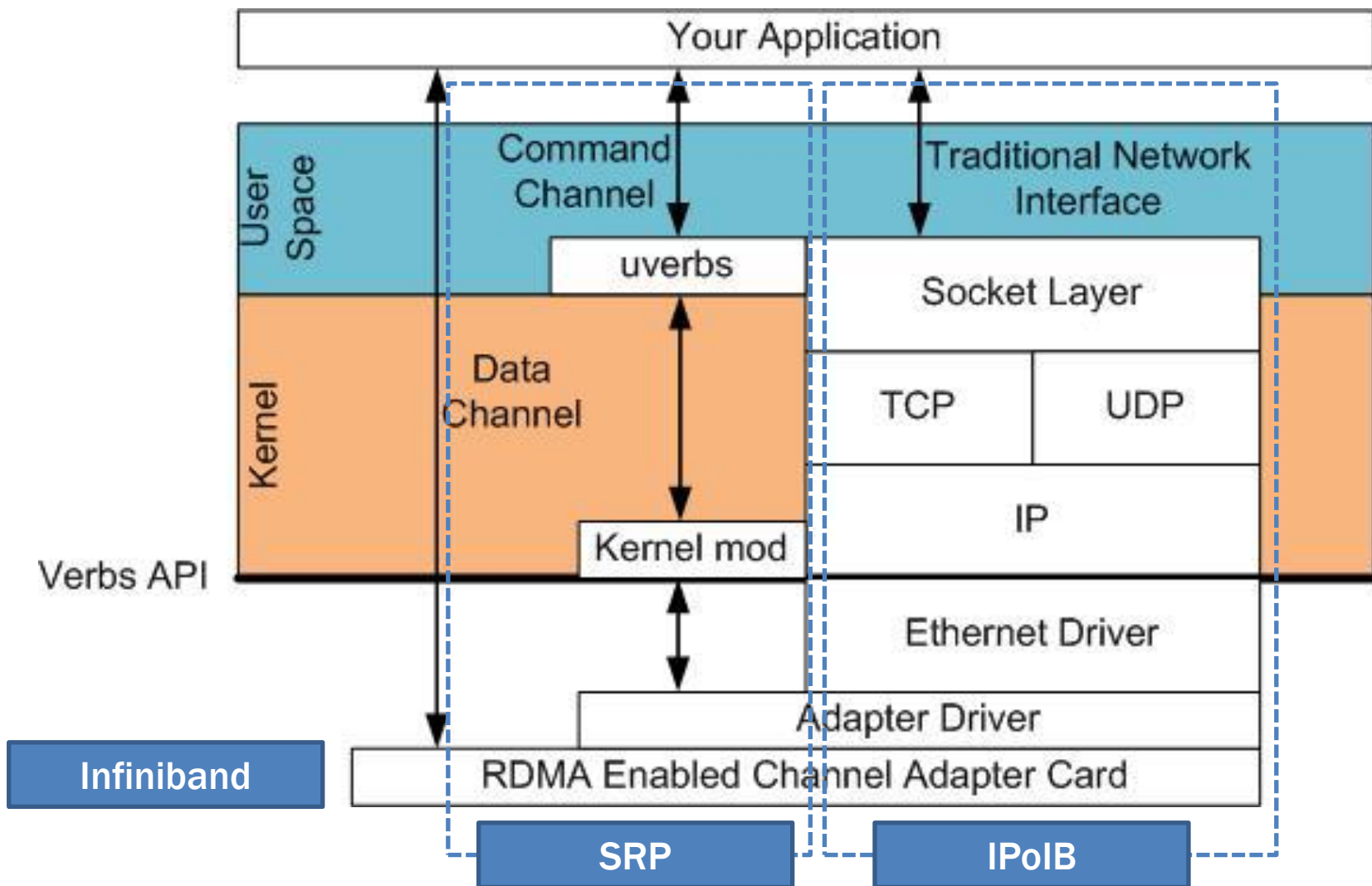
数据库节点的互联网络支持基于Infiniband的RDS协议

**IOPS**  
**>50万**

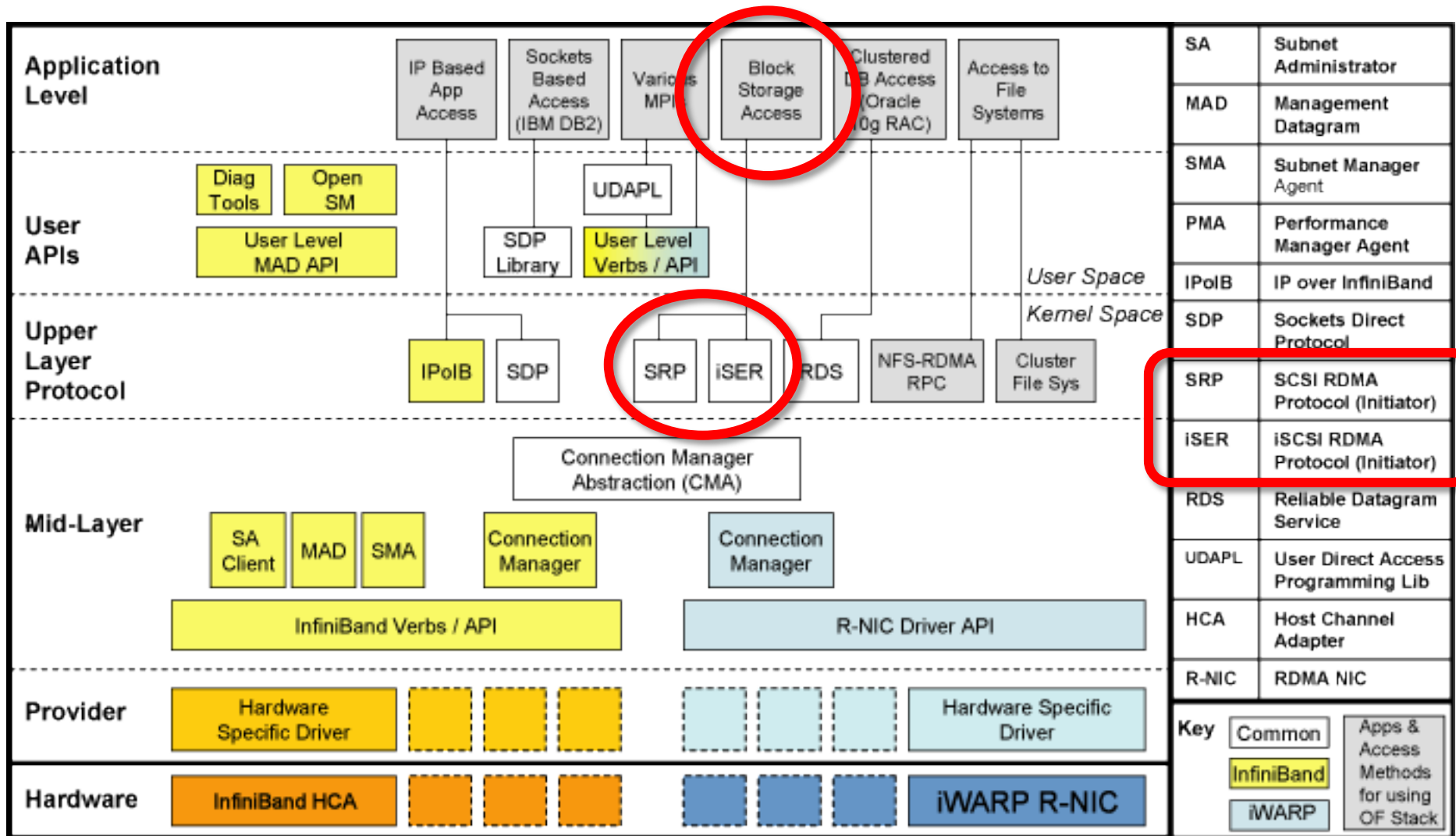
**Latency**  
**<0.6ms**

**MBPS**  
**>10GB/s**

# Infiniband/RDMA/SRP/IPoIB



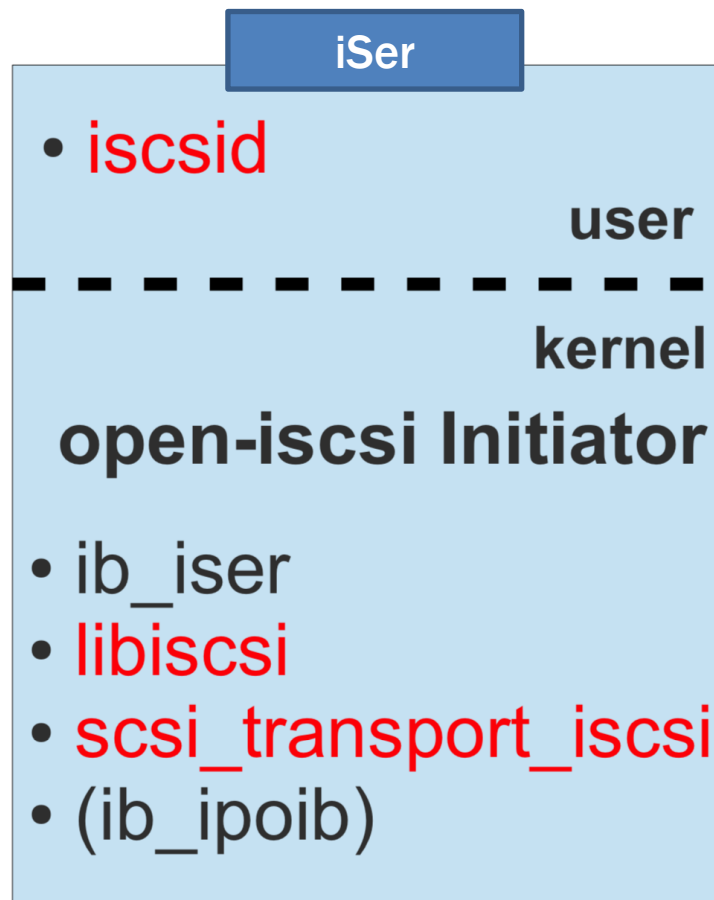
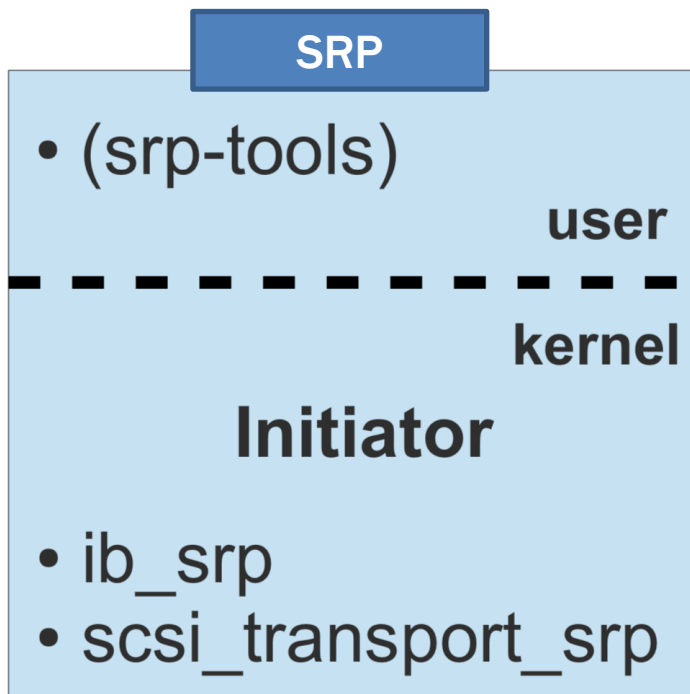
# 为什么选择SRP – 支持Infiniband有哪些协议



<https://fs.hlr.de/projects/craydoc/docs/books/S-2393-31/html-S-2393-31/chapter-qxr64y45-n14202-openfabricsinterconnectdriversforsystems.html>

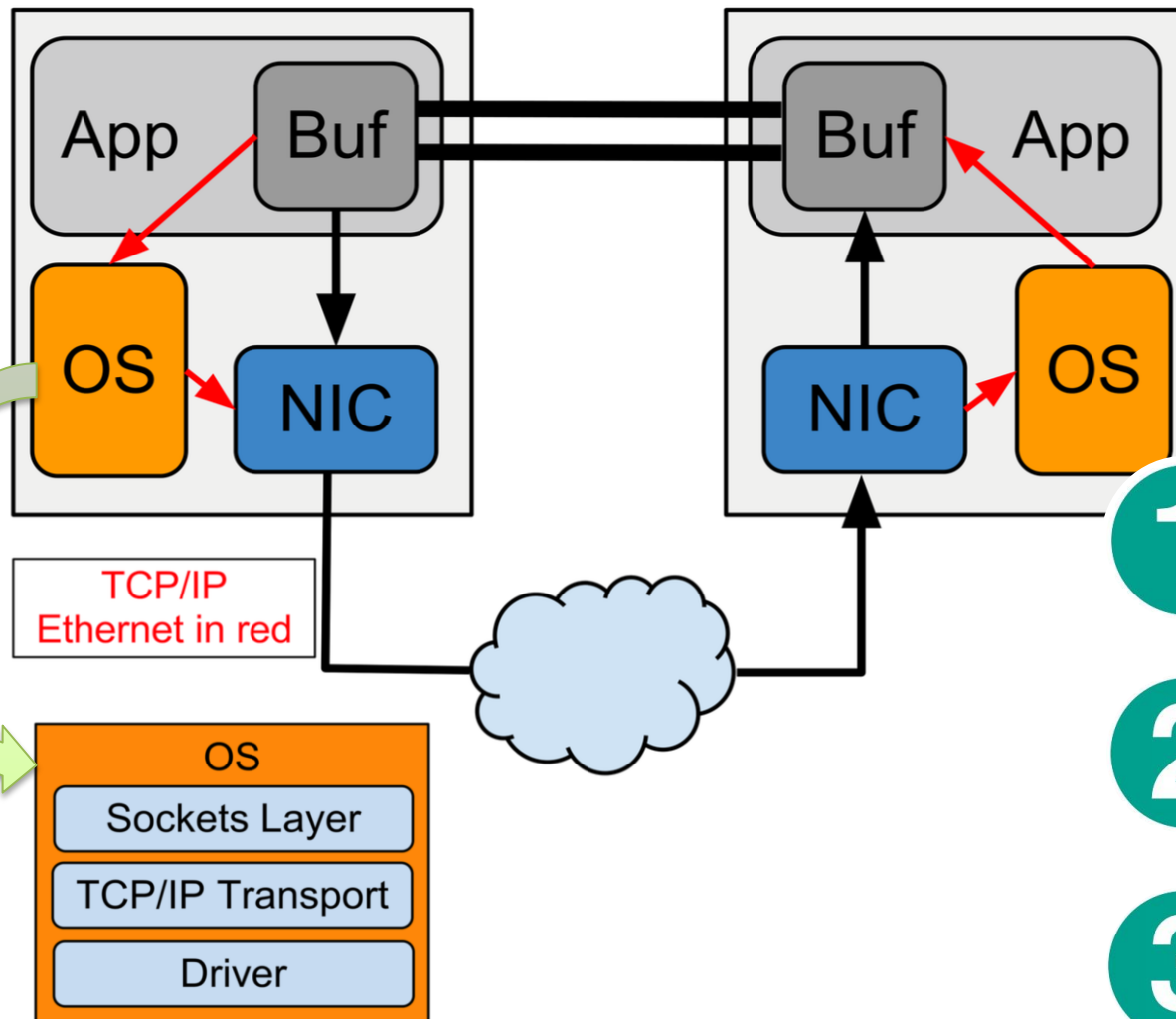


# 为什么选择SRP - SRP vs. iSer



Simple = Stable  
简单 = 稳定

# zData Light架构 – RDMA



1

减少中断

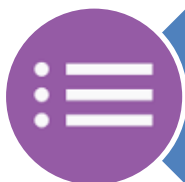
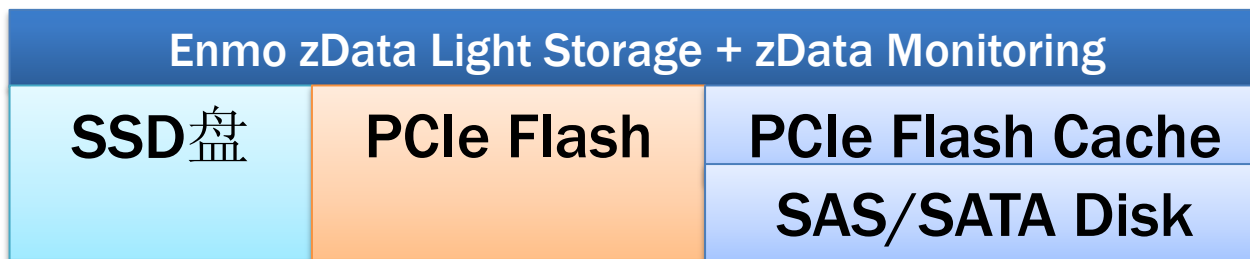
2

减少CPU占用

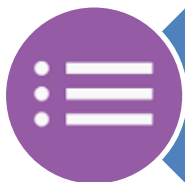
3

减少延迟

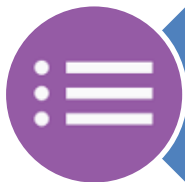
# zData Light架构 – 存储节点



SSD盘用于主存提供性能和成本的平衡

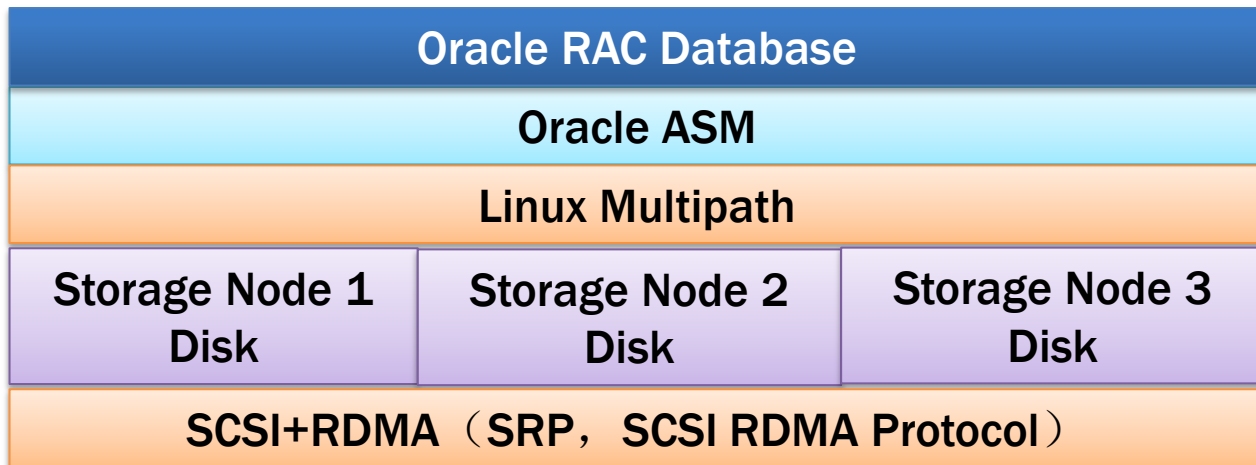


PCIe Flash用于主存提供极致性能



zData Flash Cache将闪存用于磁盘随机访问缓存

# zData Light架构 – 计算节点



- ☰ Oracle ASM提供数据冗余和分布式
- ☰ Linux Multipath提供链路冗余的高可用性
- ☰ 多个存储节点的盘全部映射到计算节点上
- ☰ 标准SCSI协议，使用RDMA传输，保持兼容性

## 典型案例—某省电信数据库仓库

- 原环境

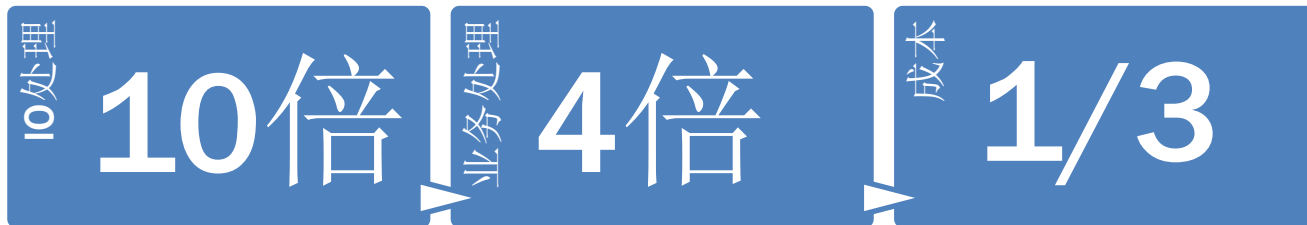
项目	描述
主机	Power 780, 24 CPU, 320G内存
存储	DS8300系列
数据库	单实例 Oracle 11.2.0.3
最大IO吞吐量	1GB/s左右



## 典型案例—某省电信数据库仓库

- zData

项目	描述
数据库服务器	2台IBM x3850，每台4路60核CPU，512G内存
存储节点	12存储节点，DELL R720，每节点N*7200转3TB SATA硬盘+1.2TB闪存卡。
存储容量	硬盘裸容量：242TB；闪存裸容量：14TB
数据库	Oracle 12.1.0.2 RAC
最大IO吞吐量	单节点10GB/s，两节点同时15GB/s



## 我在怎样的团队工作？

- 团队中的每个人都有我力所不及的长处
- 每天我总能从其他人那里学到新的东西
- 总有激烈的碰撞
- 但总在最终能够达成共识
- 最牛逼的团队欢迎你 [hr@enmotech.com](mailto:hr@enmotech.com)





云和恩墨  
ENMOTECH

数据驱动 成就未来  
*Make Your Data Dance*