刘隶放 – 无限接近“不间断”可用性和无与伦比的可扩展性
08/05/2010

无限接近“不间断“可用性和无与伦比的可扩展性
- DB2 pureScale + POWER7 的高效与节约之道

议程

- DB2 9.8 的竞争特性
- DB2 pureScale 简介和 DB2 pureScale 应用机

议程

- DB2 9.8 的竞争特性
- DB2 pureScale 简介和 DB2 pureScale 应用机

IBM DB2 9.8

追求不同工作负载下高性能优化和最低运营成本的优化

1.高性能,低成本
自动化昂贵的DBA工作，最小化存储需求，且保证高性能

2. 可信赖
经过历史证明的可靠性，可恢复性，可用性，安全性

3. 易用
易于开发,
XML 管理, 以及虚拟化.

TPC

#1 在 TPC-C 性能比较中
比Oracle的性能高出49%然而只使用一半的CPU

#1 在 10TB TPC-H 性能比较中
处于领先地位的时间比所有其他供应商的总和还要长

DOWJONES

"在我们做最后决定之前，我们会比较不同的数据库管理系统，包括Oracle，SQL Server, 以及DB2. 我们最后决定选择DB2基于以下几个原因。一个是可靠性，再一个是性能，或许最重要的是易用性。

—Bashir Khan,
数据管理和商业智能主管

SAP

#1 在 SAP SD 3-tier中
比Oracle的性能高出68%然而只使用了一半的CPU

#1 在 SAP Transaction Banking中

#1 在 SAP BW中

DB2 - 无与伦比的高可用和无限能力扩展

秉承了DB2 for z/OS Coupling Facility 传统血脉

共享磁盘架构的DB2 pureScale 技术

无限能力

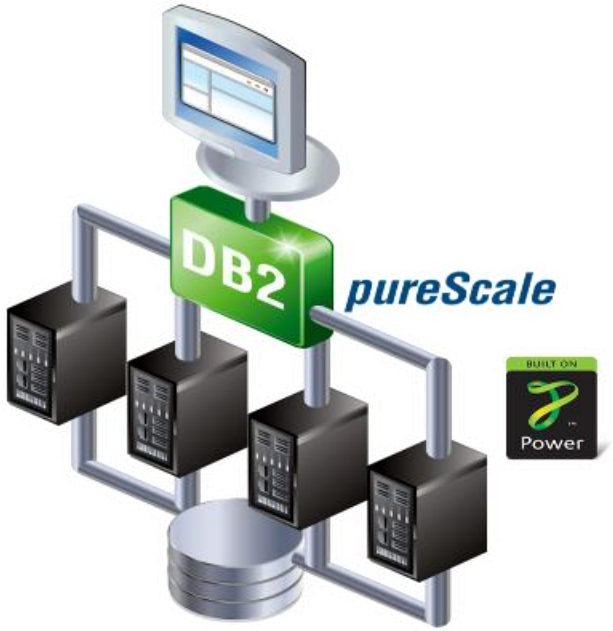
DB2 pureScale 可以为任何事务性工作负荷提供近乎无限的产能。扩展系统只需要连接到新节点并发出两个简单的命令

应用透明性

借助 DB2 pureScale，不需要更改应用代码便可有效扩展多台服务器

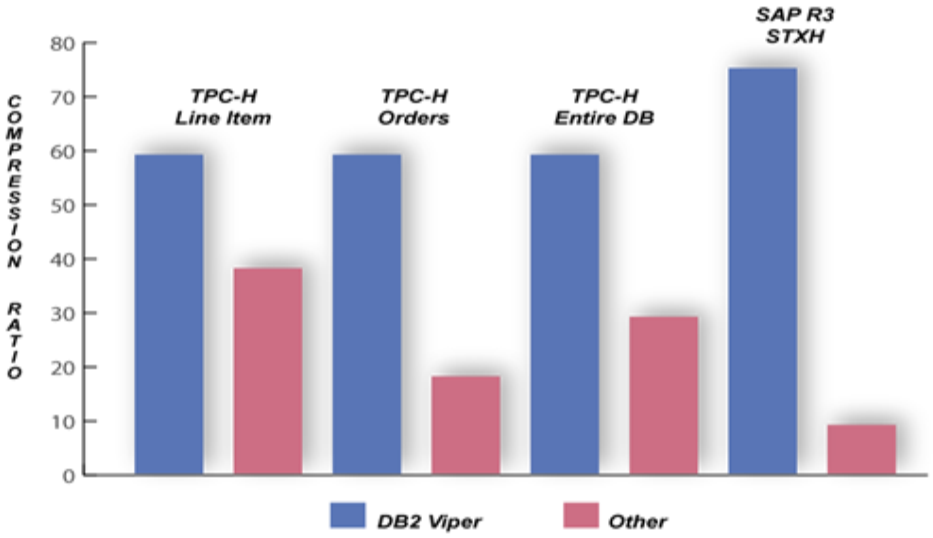
持续可用性

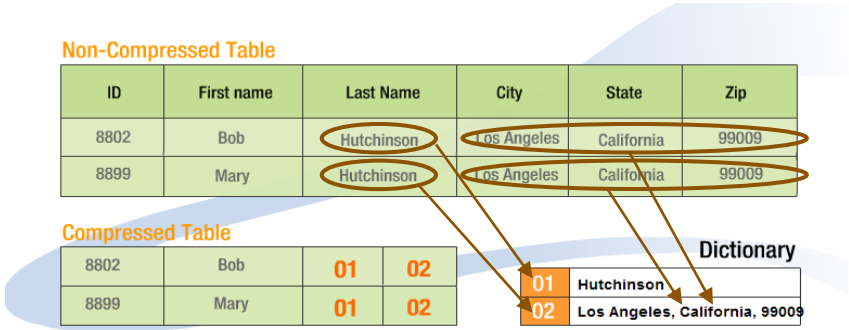
DB2 pureScale 是专为需要持续可用性进行系统设计。系统将瞬间从故障中恢复，同时仍然能保证事务处理不中断



DB2 - 使用压缩获得更低的存储成本

DB2 提供了强大的数据压缩能力，使用字典表实现的数据行压缩方法，大大减少存储空间和成本，并提高了I/O效率





国内测试结果:

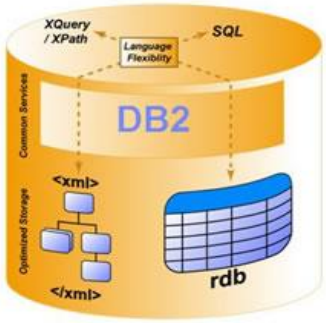
- **某软件公司套装软件: 数据存储减少60%以上, 同时交易性能提升10%**
- **某电信公司: 数据存储减少50%**
- **某银行: 数据存储减少50%**

相比竞争对手节省30%以上存储空间

DB2 - 业界领先的 pureXML 存储

DB2业界领先的XML数据处理能力

- **易于开发与集成**
  无需复杂的关系模式
  无需抽取时解析
- **高效的存储**
  在1TB的XML Benchmark测试中，
  只需要440GB的裸设备空间
- **卓越的性能**
  在1TB的XML Benchmark测试中，
  每秒可处理6,763条XML事务。



基于XML的商业智能

- 使XML数据的分析更加快速
- 易于在数据仓库中应用XML数据
- XML可以存在于数据分区、表分区、数据库视图和物化查询表中
- 改进的XML数据索引和压缩支持

SKYTIDE

由于DB2具有处理pureXML的能力，我们客户的性能得到了5到10倍的提高。”

—Keith Feingold, CEO, Skytide

DB2 - 海纳百川的兼容性

- 客户的应用可轻松方便的从 Oracle 数据库迁移到 DB2
- 可充分利用现有有人员技能，而不需要重新培训
- 迁移到 DB2 的应用可以完全在本地高效快速执行
- 客户不再受 Oracle RAC 限制

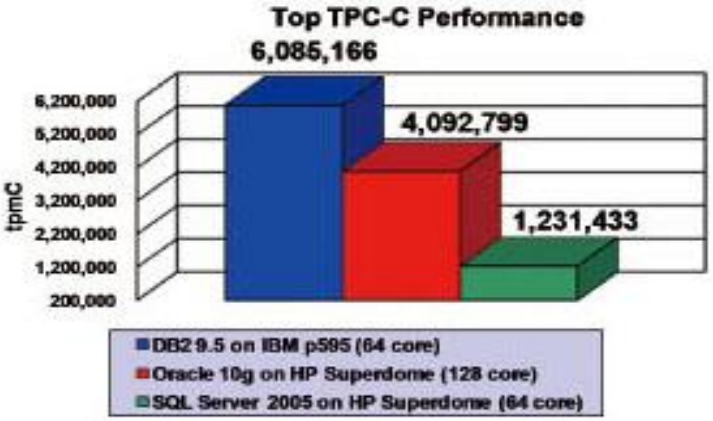
在一些客户和ISV测试中， DB2对Oracle数据库的兼容性超过95%，可轻松支持基于Oracle开发的应用，应用从Oracle迁移到DB2只需1周时间！！

**** 2010 3季度 DB2 支持 Sybase 兼容性 ****

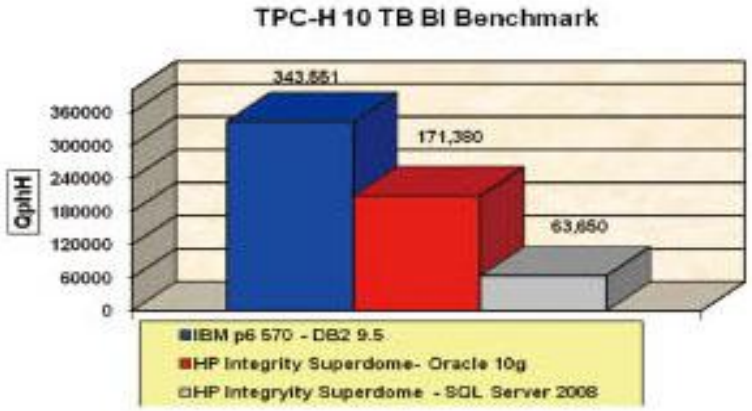
DB2 - 举世无双的卓越性能

DB2 on IBM Power 是无可争议的性能领袖

- 只需更少的处理器就能够完成相同工作
- 更低的软硬件成本（需要获取授权的处理器更少）和维护成本



- 比Oracle快50%
- 比SQL Server快5倍
- 为业务系统降低服务器成本
- 通过solidDB可获得的急速运算能力

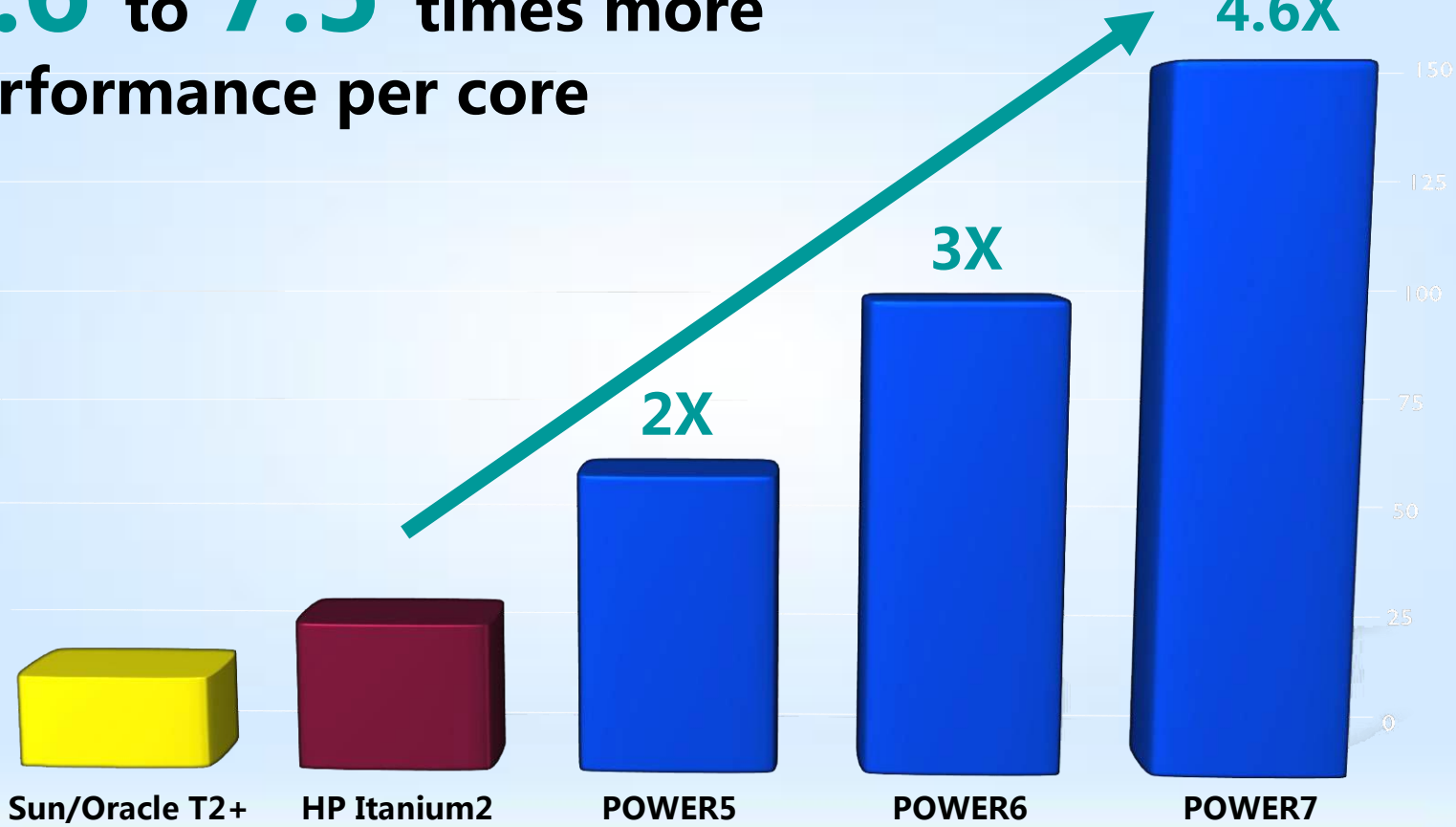


- 比Oracle快65%
- 比SQL Server快5倍
- 为BI系统降低服务器成本
- 为InfoSphere Warehouse提供了强大动力

更低的服务器成本 ⟶ 更低的软件许可费用 ⟶ 更低的维护成本

DB2 on IBM Power 单核 TPC-C 性能远远优于于其它任何平台

4.6 to 7.5 times more performance per core



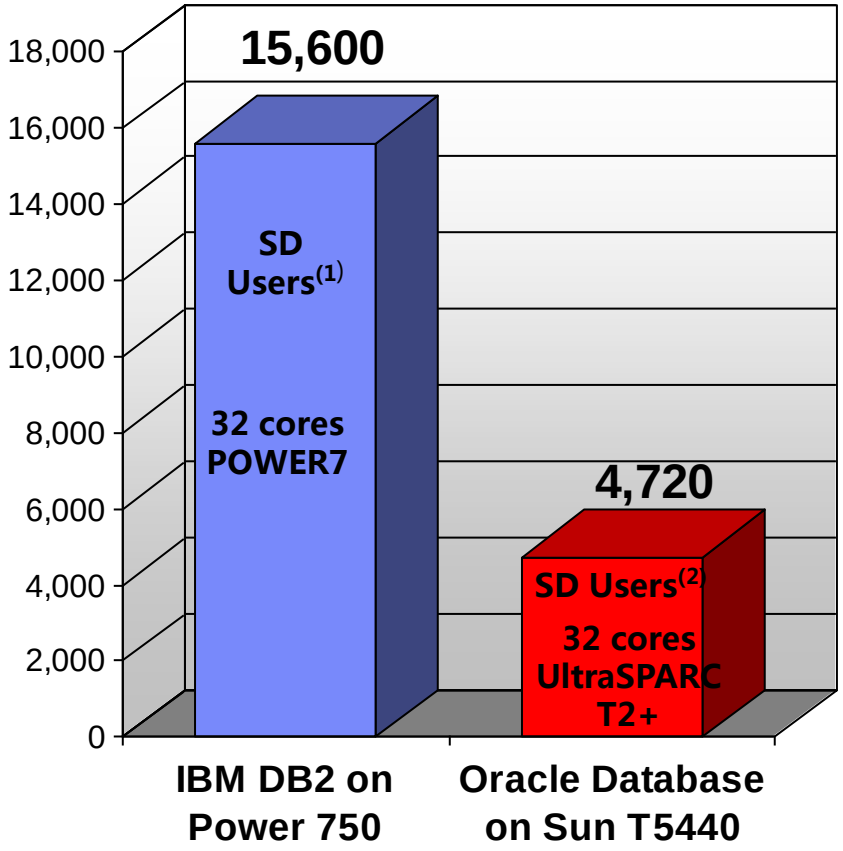
Best results listed for IBM POWER, HP, and Sun/Oracle systems over 1M tpmC
Source: http://www.tpc.org as of 4/1/08. See Power 780 benchmark details for specific results.

单核处理能力 DB2 on IBM Power 可支持 3.3 倍以上的用户请求

SAP and DB2 on Power 750

3.3x

More SD users on IBM DB2 and Power 750 than Oracle Database on Sun T5440



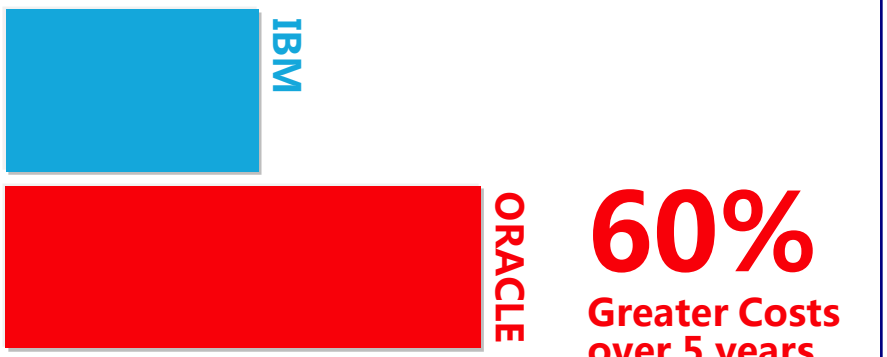
(1) IBM SAP 2-Tier SD result of 15,600 SD (Sales & Distribution) users (Average dialog response time: 0.98 second), running DB2 9.7 on AIX 6.1 and SAP enhancement package 4 for SAP ERP 6.0 on the IBM Power System 750 with 4 POWER7 3.55 GHz processor chips (32 cores, 128 threads) and 256 GB main memory, certification Number: 2010004. For more details, see http://www.sap.com/benchmark.

(2) Sun Microsystems SAP 2-Tier SD result of 4,720 SD (Sales & Distribution) users (Average dialog response time: 0.97 second), running Oracle 10g on Solaris 10 and SAP enhancement package 4 for SAP ERP 6.0 (Unicode) on the SPARC Enterprise T5440 with 4 UltraSPARC T2 Plus 1.6 GHz processor chips (32 cores, 256 threads) and 256 GB main memory, certification Number: 2009026. For more details, see http://www.sap.com/benchmark.

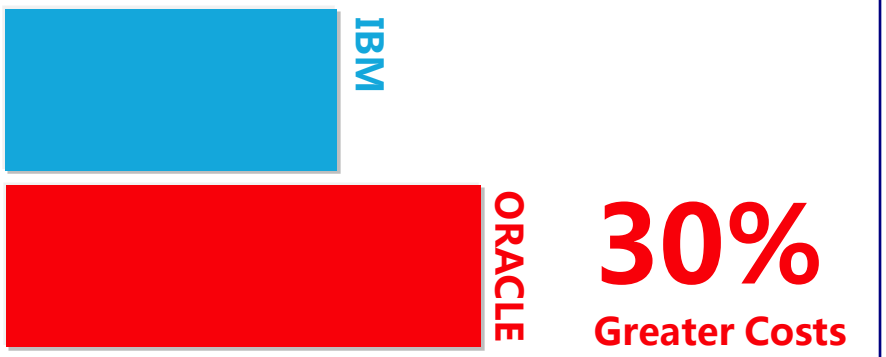
Results as of 4/02/2010

DB2 on IBM Power 优化交易处理帮助客户获取最大的回报

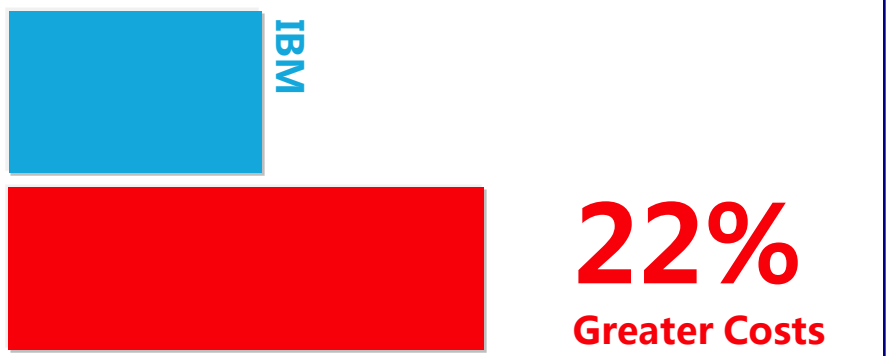
5-year costs for WebSphere Application Server and Oracle



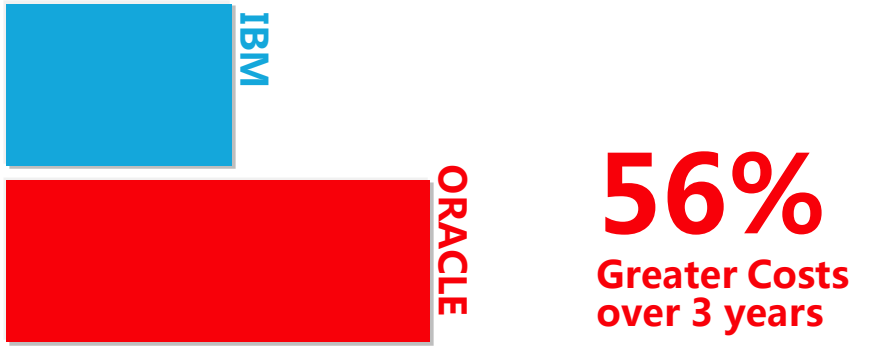
Storage costs for IBM vs. Oracle[2]



Price/Performance for WebSphere Application Server vs. competitive App Server[3]



3-year costs for DB2 and Oracle Database[4]



1 Source: Oracle technology global price list. Based on comparison of US Prices of single processor core, equivalent of 100 PVU's.
2 Source: IBM surveys of IBM clients using IBM DB2 and compression
3 Source: IBM testing: IBM WebSphere Application Server 7 - 1 JVM , AIX TL4, 64 bit, 16 threads vs. Competitive Application Server - 1 JVM, Windows 64 bit, 16 threads
4 Source: ITG whitepaper: VALUE PROPOSITION FOR IBM DB2 9.7 - Cost Savings Potential Compared to Oracle Database 11g

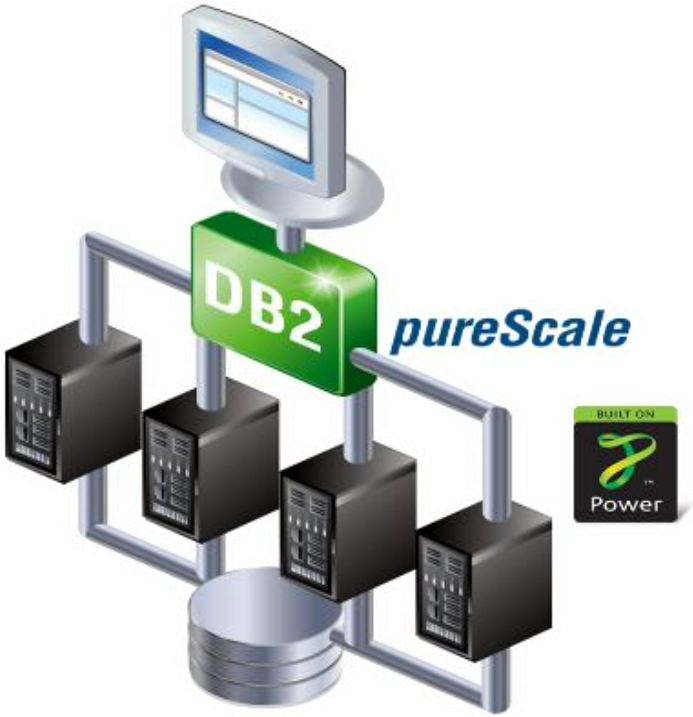
议程

- DB2 9.8 的竞争特性
- **DB2 pureScale 简介和 DB2 pureScale 应用机**

DB2 pureScale 从何而来？

DB2 pureScale 的竞争特色

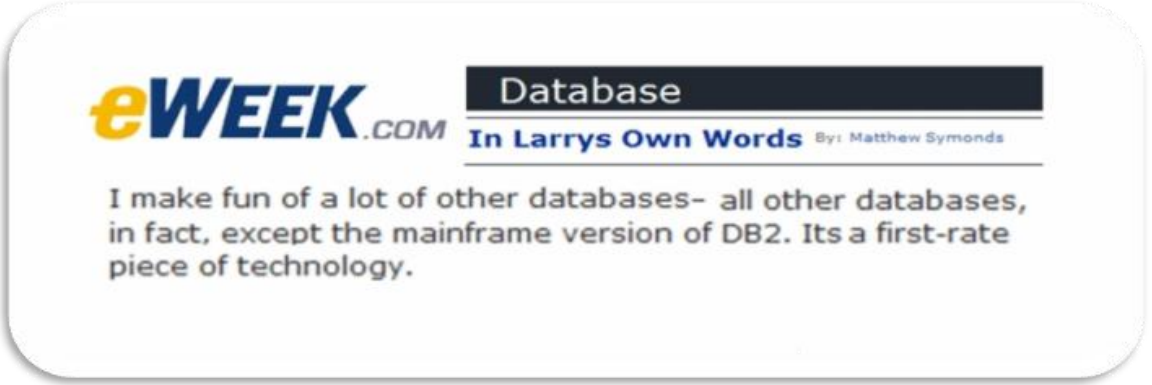
- **无限产能**
  - 仅购买所需要的设备，随需提高产能
- **应用透明性**
  - 避免应用变更带来的风险和成本
- **持续可用性**
  - 交付不中断的数据访问，确保性能一致



借鉴自无可争议的黄金标准……System z

DB2 for z/OS 数据共享是”黄金标准“

- **每个人都认可 DB2 for z/OS 是可伸缩性和高可用性的“黄金标准“**
- **甚至 Oracle 也同意：**

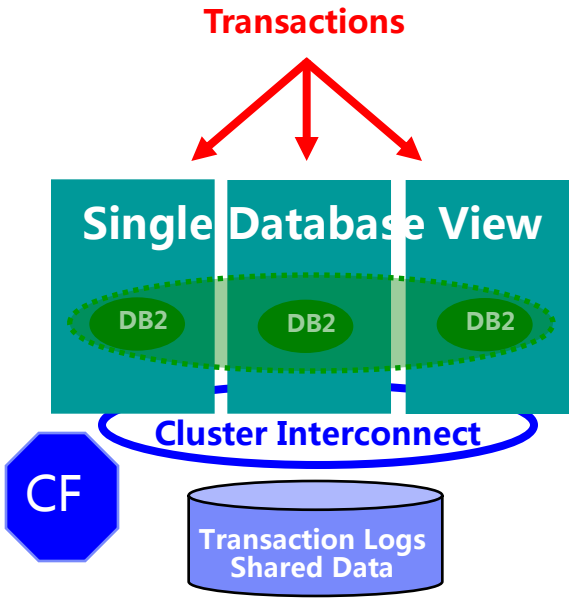


- **为什么？**
  - Coupling Facility！！
    集中锁定、集中缓冲池交付了优异的可伸缩性和优异可用性
  - z/OS 上的整个环境都可用使用 Coupling Facility
    CICS、MQ、IMS、Workload Management 等

DB2 pureScale 的目标

- **24*7的可用性**
  - 无论是针对计划内还是非计划内事件
- **简单扩展**
  - 不需要显著的程序修改
  - 不需要复杂的管理工作
- **快速响应工作负载变化**
  - 在机器和资源增加或减少的情况下，根据动态工作负载进行调整

低管理成本

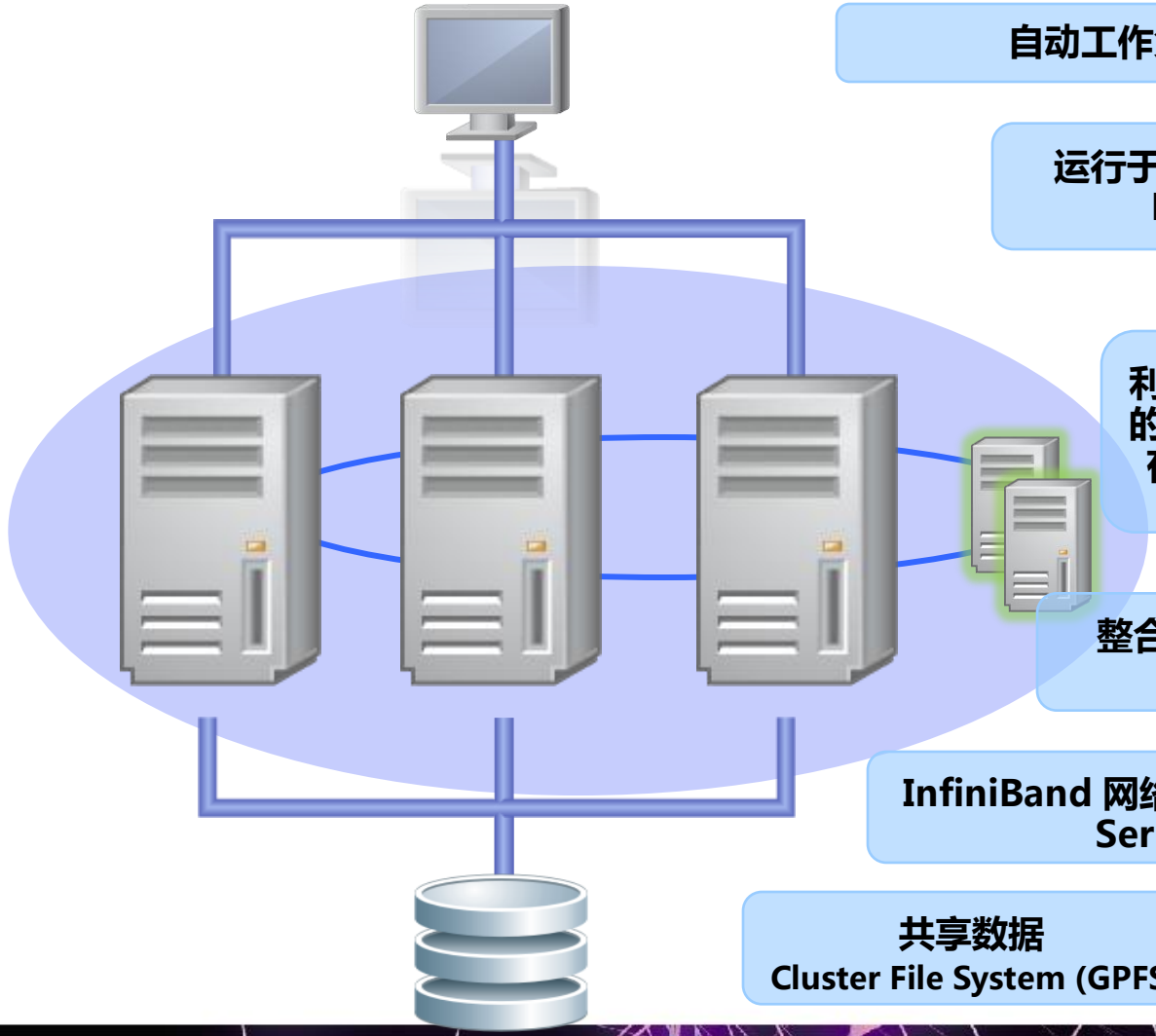


在分布式平台最接近 z/OS “黄金标准”的解决方案

基于 Z Sysplex 模型，使用 COTS 组件

和竞争对手的区别在于超强的高可用性和可扩展性

DB2 pureScale 的架构



自动工作负载均衡

运行于 Power 服务器上的 DB2 成员集群

利用 z/OS 提供的全局锁定和内存管理器技术

Primary PowerHA pureScale

CF

DB2 Cluster Services

CF

DB2 Cluster Services

Secondary PowerHA pureScale

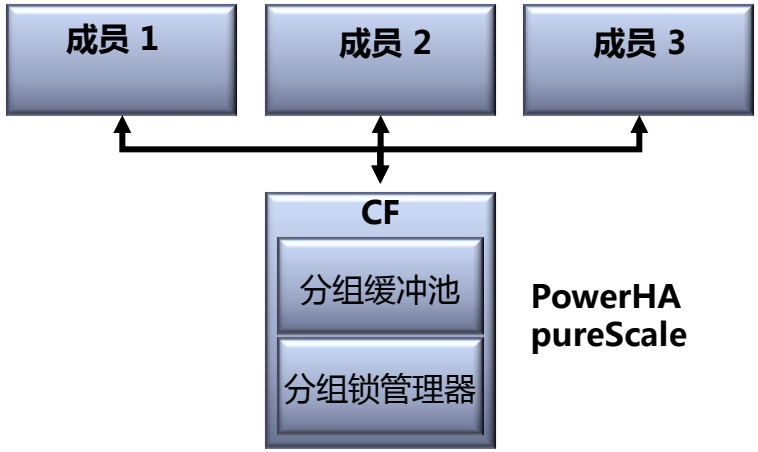
整合的 Tivoli System Automation

InfiniBand 网络和 DB2 Cluster Services

共享数据
Cluster File System (GPFS)

可伸缩性和高可用性的关键

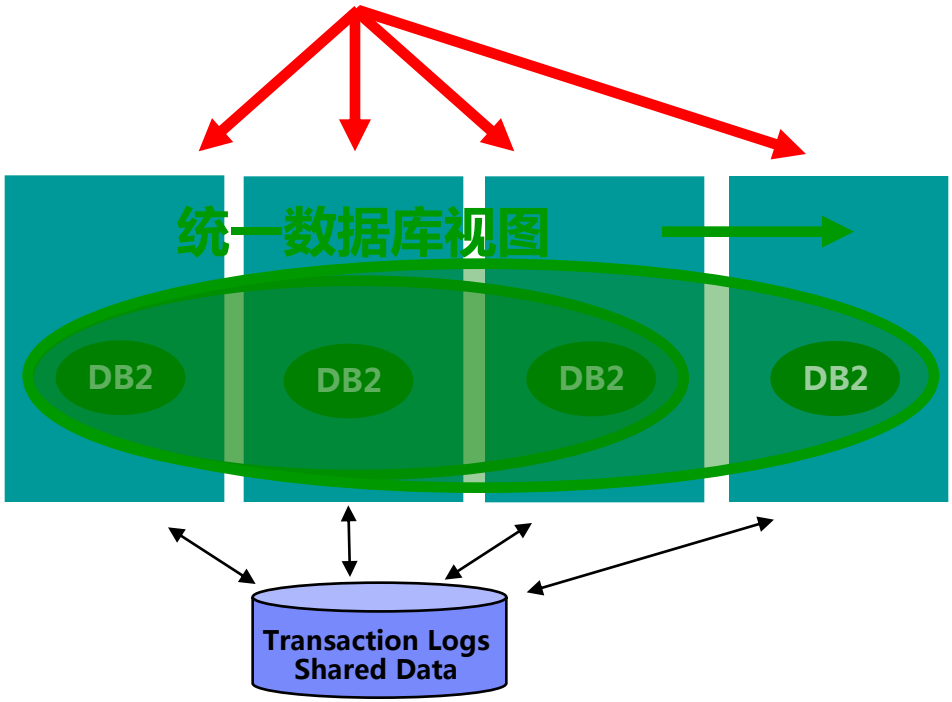
- **有效的集中锁定和缓存**
  - 随着集群的不断增长，DB2 会始终在 CF 维护锁定信息和共享页面
  - 针对超高速访问而优化
    - DB2 pureScale 使用 Remote Direct Memory Access (RDMA) 与 PowerHA pureScale 服务器通信
    - 没有 IP 套接字调用、没有中断、没有上下文切换
- **结果**
  - 为大量服务器提供接近线性的可伸缩性
  - 持续感知各成员当时的工作状态
    - 如果其中一个成员出现故障，不会造成
      - 其他成员 I/O 阻塞
      - 以内存速度恢复运行



易扩展

扩展

- ✓ 不需应用程序显著修改的完美扩展
- ✓ 对于数据所属节点没有限制
- ✓ 灵活适应工作负载路由



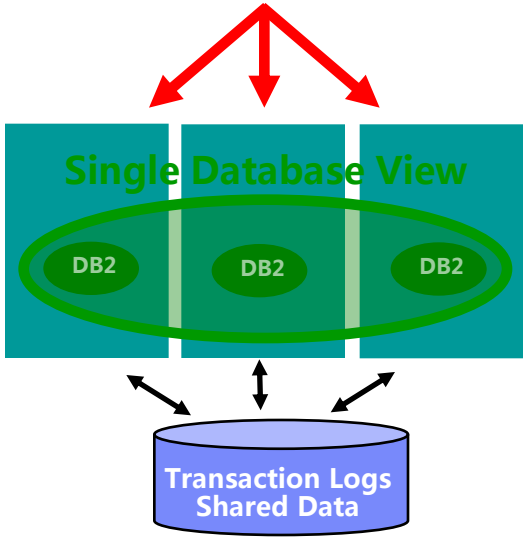
快速部署新成员

- ✓ 不需要数据重新分布

易维护和升级

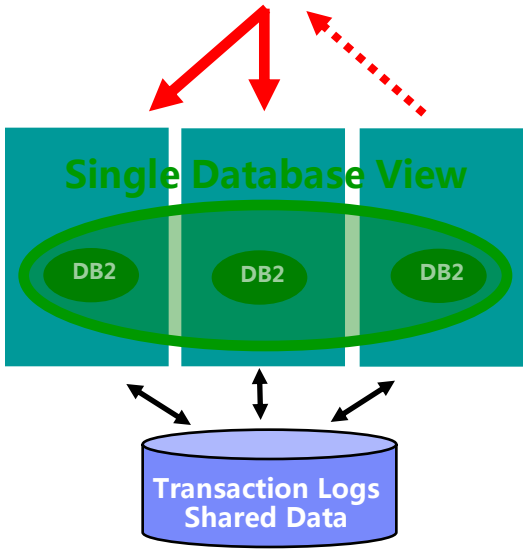
1)运行系统

- 设定目标节点
- （可选）增加一个新的节点以保证整个系统的整体能力



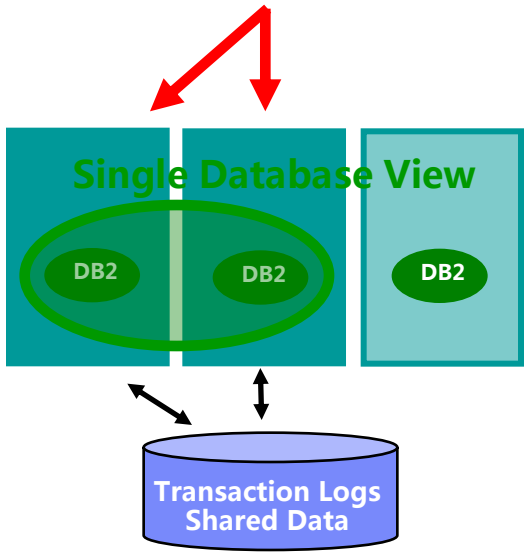
2）排干（Drain）目标节点

- 停止新的路由
- 允许已有交易完成



3) 执行维护工作

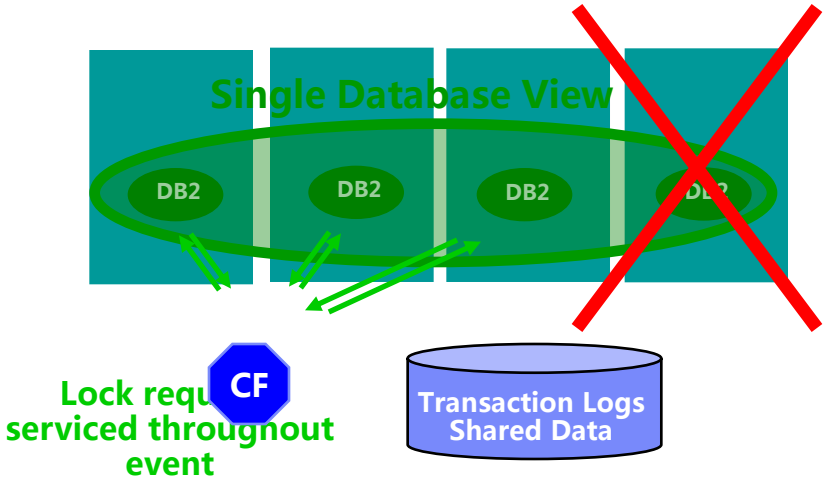
- 排干完成后

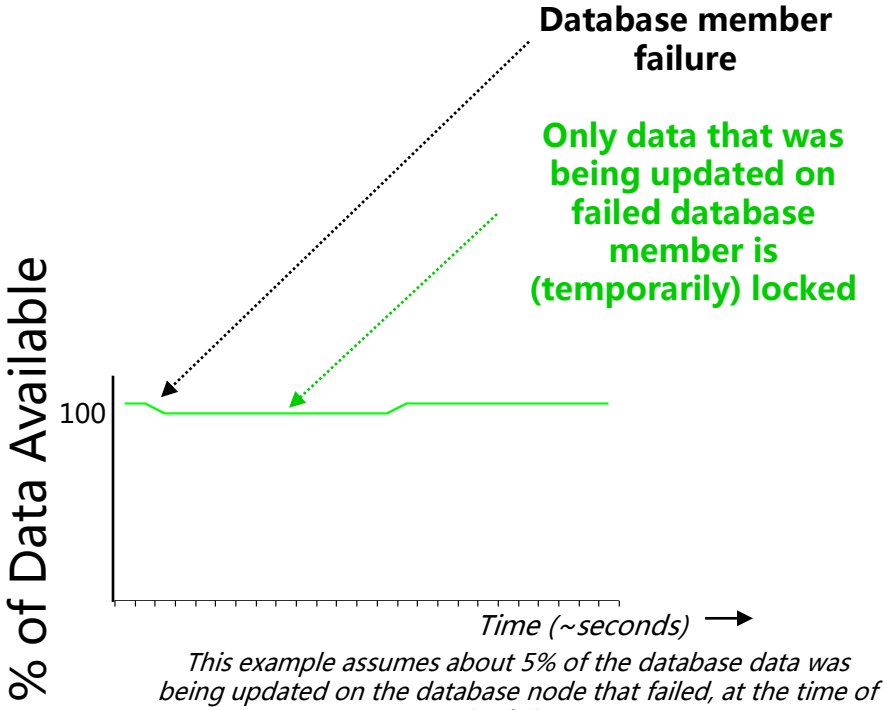


在系统可用性上无断点；无Quiesce时间; 不需要对已有工作强制回滚

最小化非计划宕机时间

- **DB2 pureScale 的设计重点就是最大化成员在非正常宕机的情况下的可用性**
  - 当数据库成员失败的情况下，只有”in-flight“的数据在成员恢复完成前被锁定
    In-flight = 在成员失败时在该成员上参与交易的修改的数据
  - 目标成员恢复时间： 10-15 秒
    失败成员上的只读数据在这段时间不被锁定

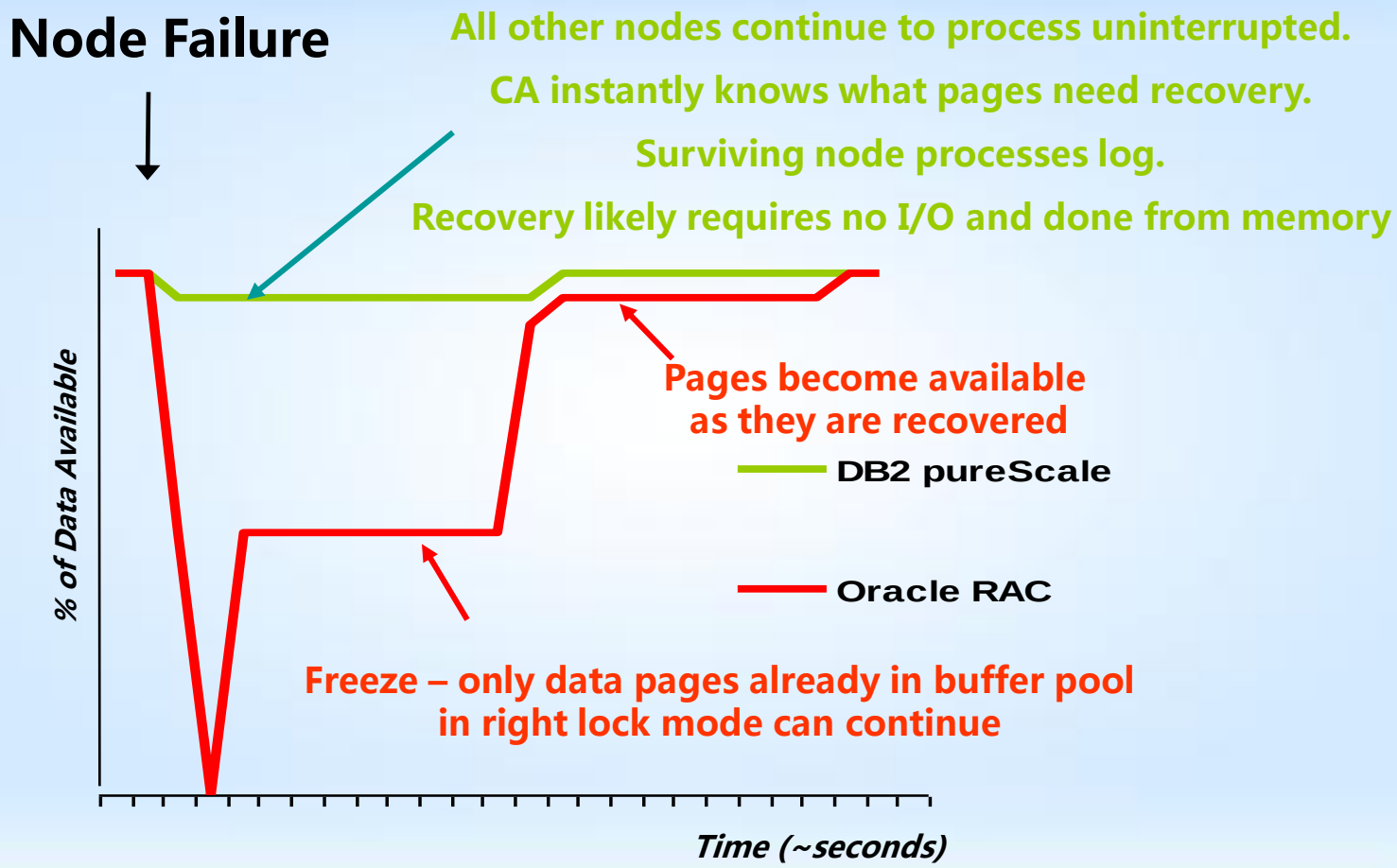




This example assumes about 5% of the database data was being updated on the database node that failed, at the time of the failure.

IBM智慧系统全球行2010

DB2 pureScale 和其它集群技术的崩溃恢复比较



什么是 PowerHA pureScale?

- **PowerHA pureScale 是 DB2 pureScale Feature 的一个集成组件**
  - **秉承 z/OS CF 技术，有 AIX 实验室开发**
- **协调多个成员对共享数据的访问**
  - 为所有成员提供锁定和数据缓存一致性服务
  - DB2 使用它来保证数据在所有的节点上都是一致的
- **RDMA capable fabric**
  - **直接修改内存不消耗 CPU 资源**
  - 为 zSeries Sysplex 发明
- **包括3个主要部件**
  - Group Buffer Pool (GBP)
    确保所有成员都能读到最新提交的数据页
  - Global Lock Manager (GLM)
    提供给成员以能够序列访问对象
  - Shared Communications Area (SCA)
    提供 DB2 控制数据的一致性机制
- **PowerHA pureScale 应该配置一对以避免单点故障**

由 IBM AIX 实验室参与开发了 PowerHA pureScale 技术

Workload-Optimizing Systems

+

AIX - the future of UNIX

Total integration with i

Scalable Linux ready for x86 consolidation

PowerVM™

Virtualization without Limits

- ✓ **Drive over 90% utilization**
- ✓ **Dynamically scale per demand**

POWER7™ BUILT ON Power™

Dynamic Energy Optimization

- ✓ **70-90% energy cost reduction**
- ✓ **EnergyScale™ technologies**

PowerHA™

Resiliency without Downtime

- ✓ **Roadmap to continuous availability**
- ✓ **High availability systems & scaling**

IBM® Systems Director

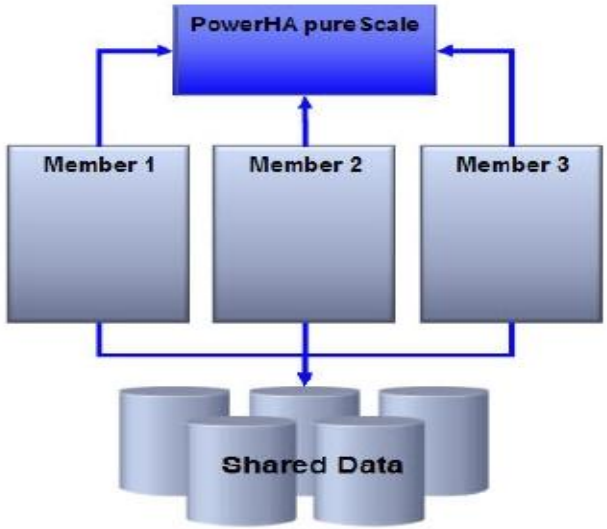
Management with Automation

- ✓ **VMControl to manage virtualization**
- ✓ **Automation to reduce task time**

Smarter Systems for a Smarter Planet.

DB2 pureScale 使用了 PowerHA pureScale 技术

- **实现最有效率和连续的操作**
- **减少节点间通信以减少系统额外开销**
  - 集中的数据库锁定和缓冲减小节点间通信成本，最大化计算机生产能力
- **通过直接内存访问减少系统通讯成本**
  - RDMA（Remote Direct Memory Accesses）事实上消除了处理器在系统内 IP 网络通讯的 Context Switching
- **减小节点失败影响维持业务连续性**
  - 所有节点可以马上访问到数据和锁的状态以保证应用程序性能



PS：我们的 InfiniBand 实现有何特殊之处

- **DB2 利用 Remote Direct Memory Access (RDMA) 直接对远程服务器的内存执行写操作**

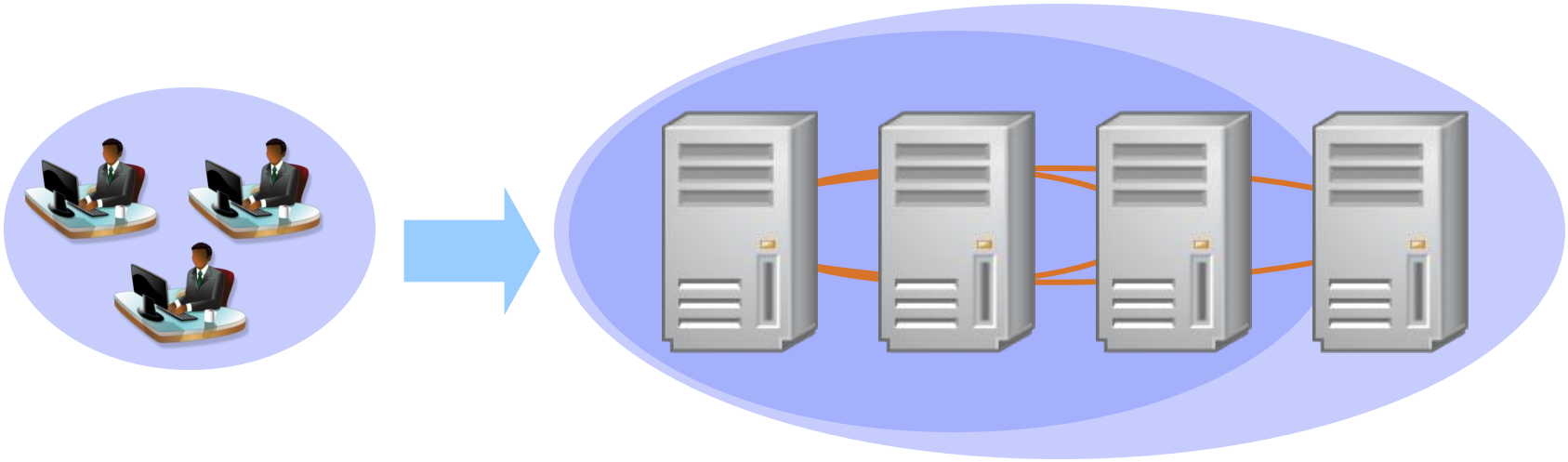
- **可以将 InfiniBand 应用于 Oracle RAC 集群**
  - Oracle RAC 可以使用 Reliable Datagram Sockets”（RDS) over InfiniBand

    Oracle RDS 运行在操作系统内核中，需要 CPU 中断来执行 Socket通讯，相对于 DB2 来说，仍有比较高的通讯开销

DB2 pureScale 技术优势

DB2 pureScale 的应用透明性

- **立即利用额外的产能**
  - 不需要修改您的应用代码
  - 不需要调优数据库基础设施



管理员可以增加产能，而不需要重新调优或重新测试

开发人员甚至不需要知道增加了更多节点

透明的应用可伸缩性

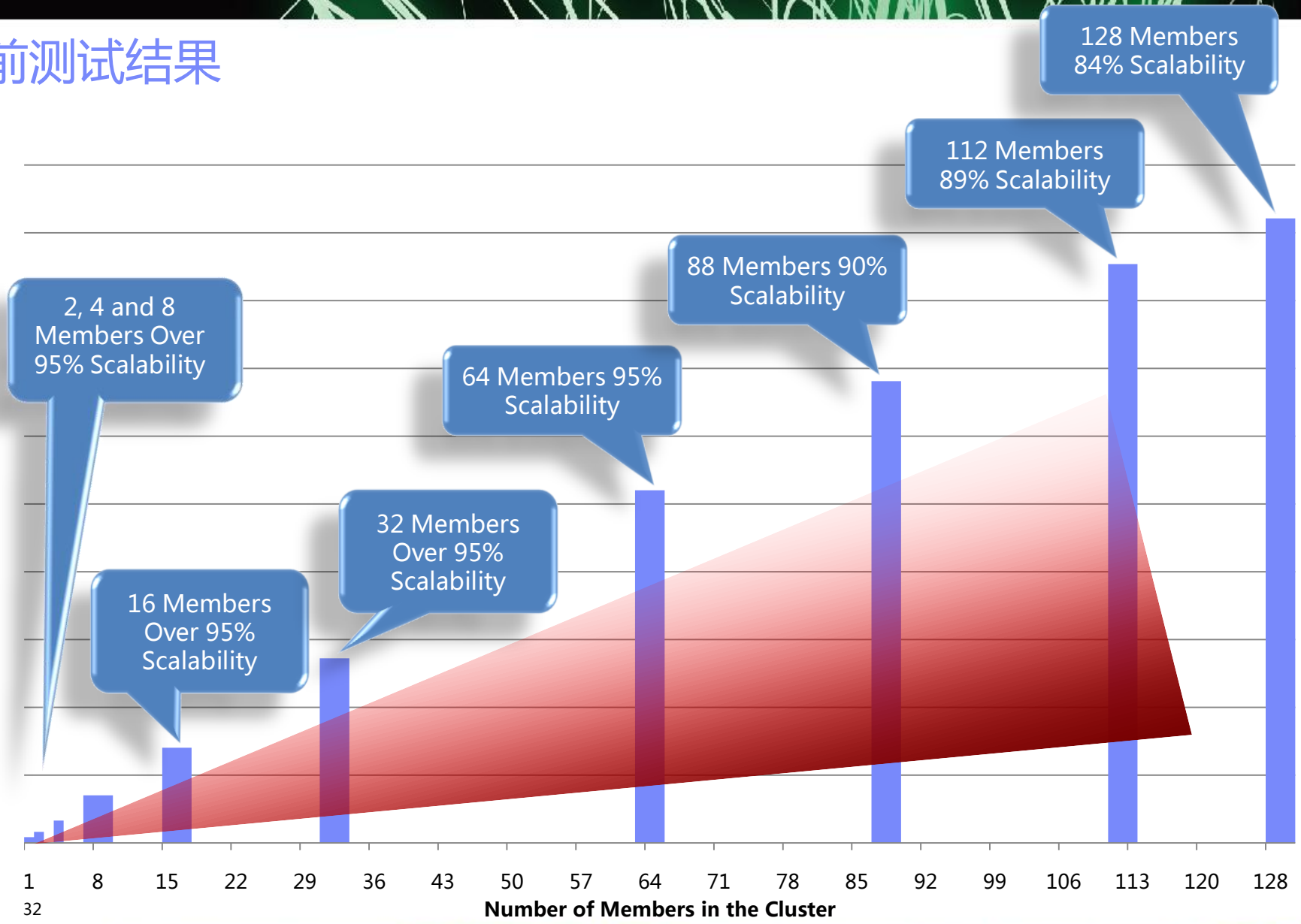
- **无需应用或数据库分区的可伸缩性**
  - 支持 RDMA 访问的集中锁定和全局缓冲池可以带来高可伸缩性，而不会让应用集群感知到
    - 数据页面的共享将在实际共享的缓存中通过 RDMA 来实现
    - 不需要通过应用或数据分区来实现可伸缩性
      - 服务器之间的进程中断造成访问无法同步
      - 降低了管理和应用开发成本

- **其它集群技术中分布式锁定会增加开销并降低可伸缩性**
  - Oracle RAC 最佳实践建议
    - 每个页面使用较少的行（避免热页面）
    - 通过数据库分区来避免热页面
    - 通过应用分区来获取一定水平的可伸缩性
    - 所有这些都会造成管理和开发成本增加

DB2 pureScale 架构可扩展性的佐证

- **可伸缩的程度？**
- **以 Web Commerce 工作负载为例**
  - 大多数可读，但并非只读
- **不会让应用集群感知到**
  - 无需将事务发送给成员
  - 访问数据库中的随机行的事务
  - 实现透明的应用伸缩
- **持续扩展到超过 100 位成员**

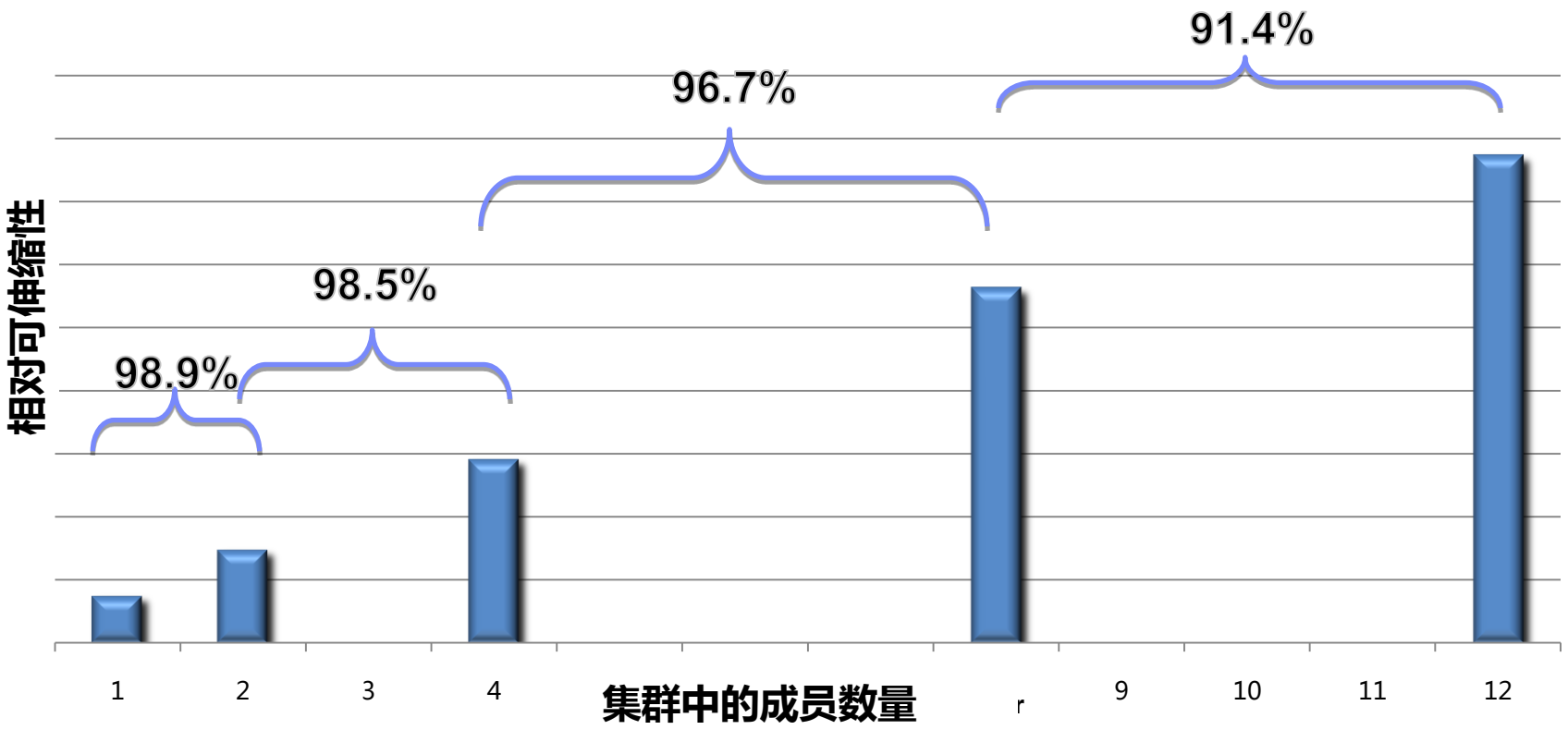
当前测试结果


128 Members
84% Scalability
112 Members
89% Scalability
88 Members 90%
Scalability
64 Members 95%
Scalability
32 Members
Over 95%
Scalability
16 Members
Over 95%
Scalability
2, 4 and 8
Members Over
95% Scalability
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
128
Number of Members in the Cluster

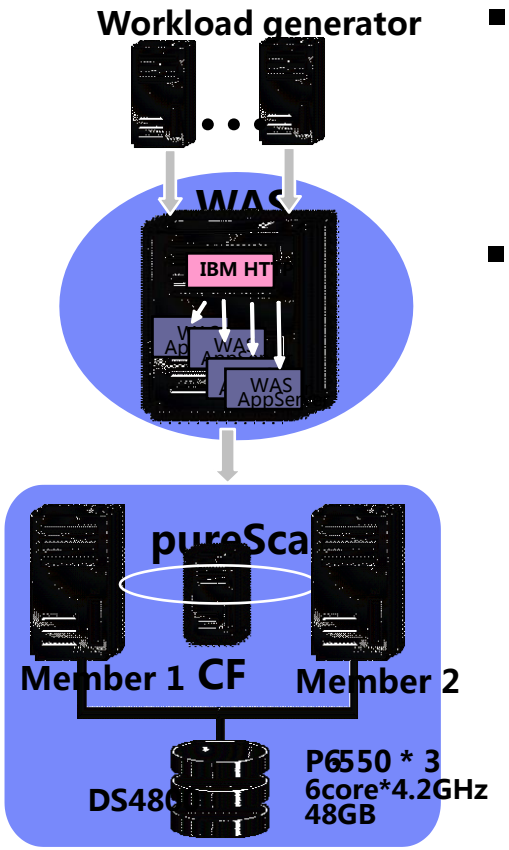
12 成员集群深入分析

- **查看更具挑战、更多更新的工作负载**
  - 每 4 个读事务中就有 1 个更新事务
  - 许多 OLTP 工作负载都具有典型的读/写速率
- **应用中没有集群感知性**
  - 无需将事务发送给成员
  - 实现透明的应用伸缩
- **冗余系统**
  - 14 个 8 核 P550 Express，包括 duplexed PowerHA pureScale™
- **可伸缩性仍然保持在 90% 以上**

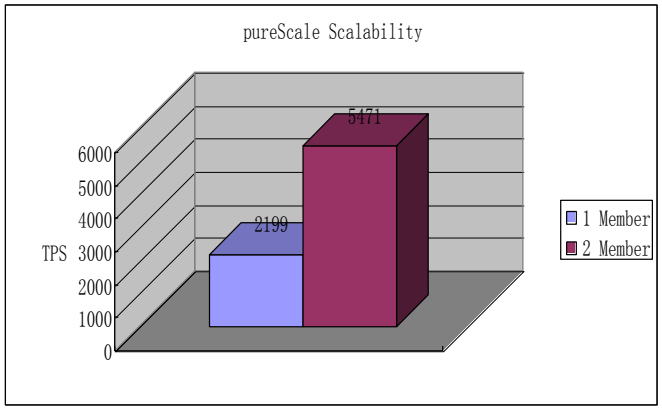
针对 OLTP 应用的可伸缩性

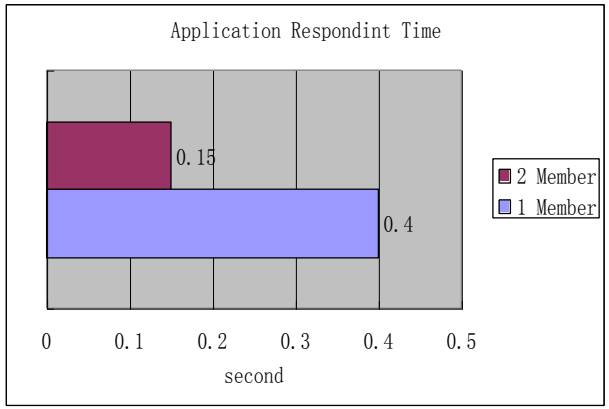


中国实际案例

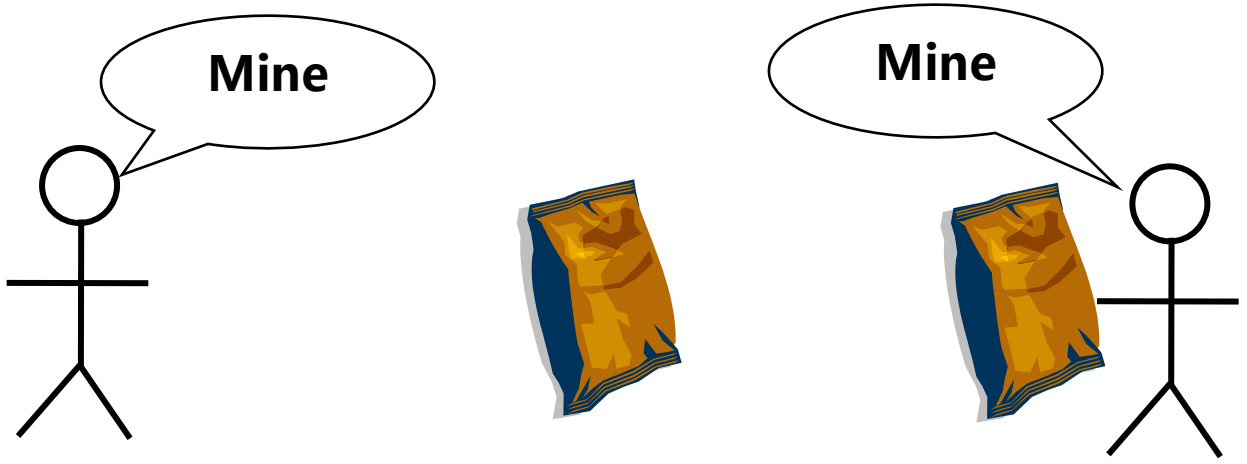


- 测试案例
  - 1000万用户数据
  - 1000 并发用户
  - 在 Oracle 兼容性模式下执行并使用工作负载均衡
- 扩展性测试结果
  - pureScale 在增加成员的情况下扩展能力很好
  - 工作负载管理器可以自动地调整成员间的工作负载

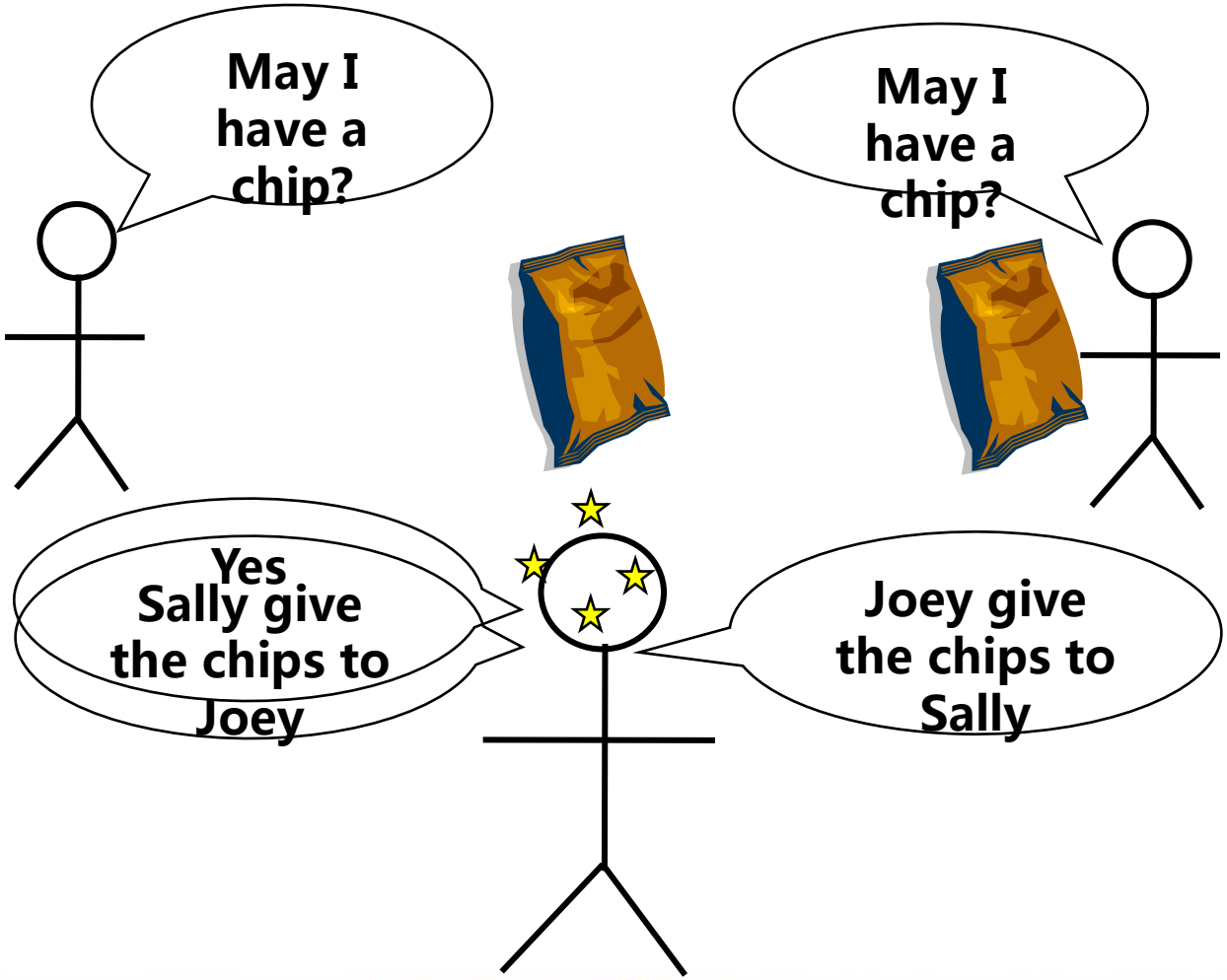




简单概括没有集中管控技术的伸缩性 - 2 个节点



简单概括没有集中管控技术的伸缩性 - 3 个节点



市场信息 vs 真实情况

ORACLE

PRODUCTS AND SERVICES | INDUSTRIES | SUPPORT | PARTNERS | COMMUNITIES | ABOUT

PRODUCTS AND SERVICES
Oracle Database
- Editions
  - Overview
  - Enterprise Edition
  - Enterprise Options
    - Real Application Clusters
    - RAC One Node
    - Real Application Testing
    - Active Data Guard
    - Advanced Compression

Oracle Real Application Clusters

Lower the Cost of Computing
Oracle Real Application Clusters (Oracle RAC), with Oracle Database 11*g* Enterprise Edition, enables a single database to run across a cluster of servers, providing unbeatable fault tolerance, performance, and scalability with no application changes necessary. Analysts are taking note of Oracle RAC's growing importance as large numbers of customers across all industries consolidate their transaction processing and data warehousing applications.

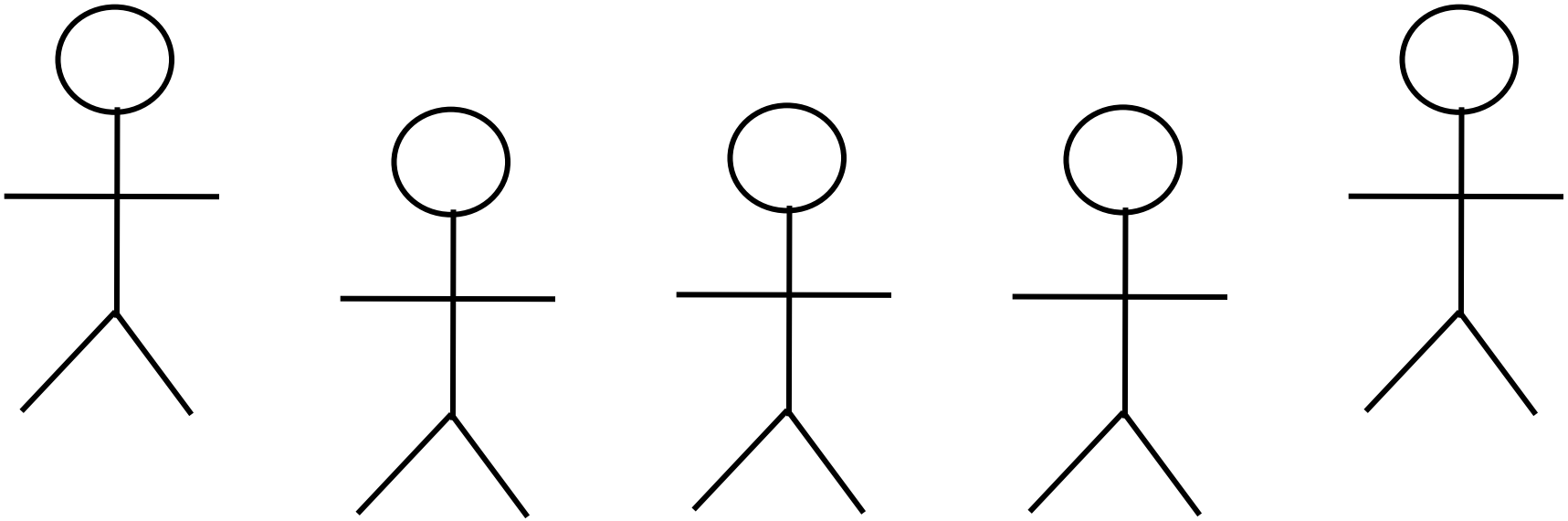
- Webcast: Consolidate on the Grid with Oracle Database 11*g* Release 2
- Gartner: Oracle RAC Moved to Mainstream Use

CIO Update

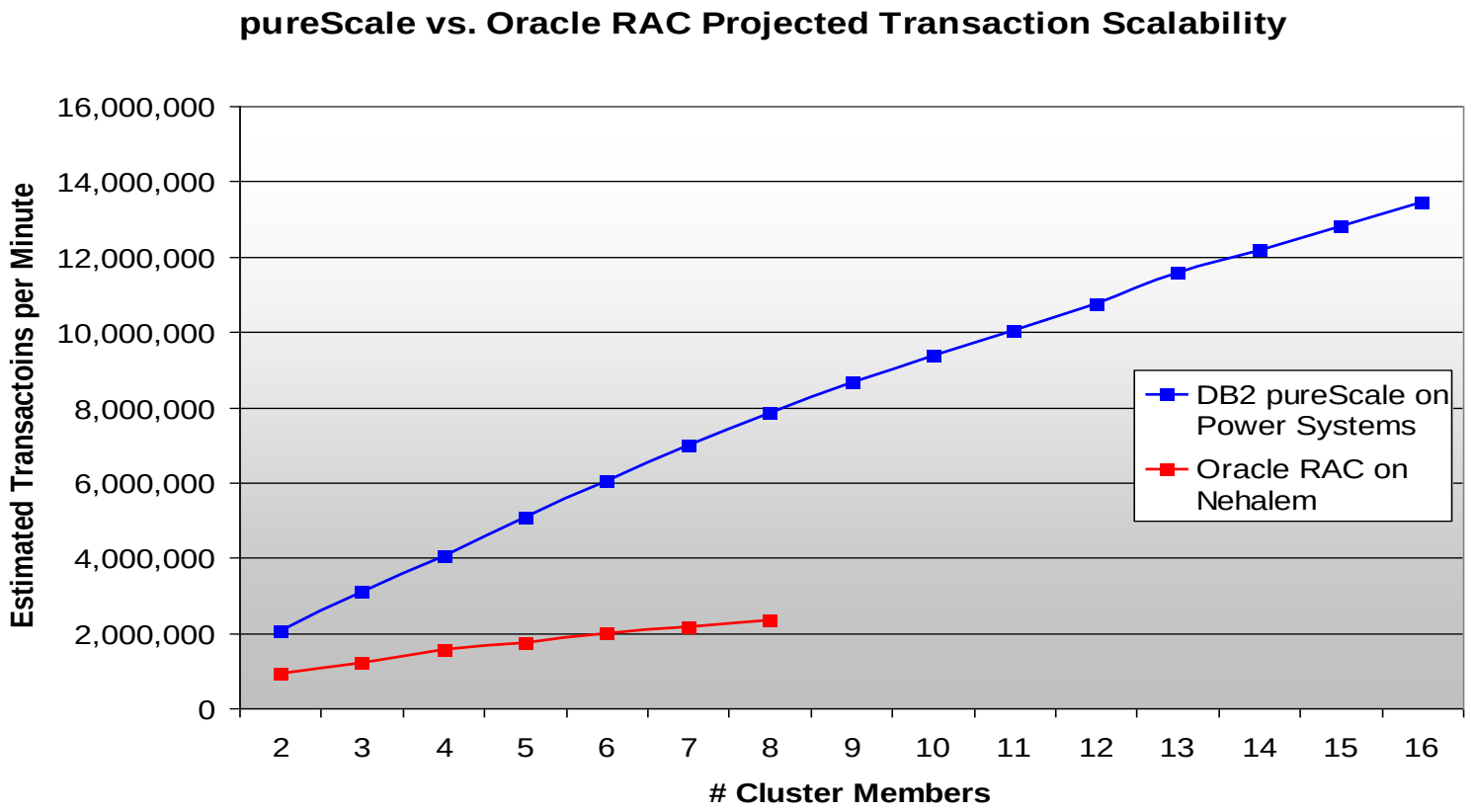
Larry Ellison: Oracle Betting On Linux

"You can afford to use a bunch of these low cost machines," he said, "and it doesn't matter if they fail periodically, because our software conceals the fact that you're having hardware component failures. If one or two or three fail, the rest keep running, and your users don't even notice it."

简单概括 DB2 pureScale 的伸缩性



集群数据库伸缩性比较


pureScale vs. Oracle RAC Projected Transaction Scalability
Estimated Transactoins per Minute
16,000,000
14,000,000
12,000,000
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Cluster Members
DB2 pureScale on Power Systems
Oracle RAC on Nehalem

IBM pureScale Application System 6100

IBM pureScale 应用机

IBM pureScale 应用机

The always-available, scalable transaction processing system

- IBM Power 770, DB2 pureScale, 和 WebSphere Application Server
- 预先安装，预先优化的高效低成本应用机
- 软件充分利用硬件的强大并行处理和浮点运算能力
- 一站性采购软硬件

IBM pureScale 应用机

可以从 8 核无缝高效地扩展到 8192 核

IBM Power 770

IBM Power 770

Scale Up
Expand each member up to 64 cores

Scale Out
Add additional members, up to 128 total [2]

IBM Power 770

Scale Within
Add additional cores as needed with Capacity OnDemand

1 – Assumes 64 cores x 128 members (see note 2)

2 – Architecture proven in lab testing to 128 nodes using DB2 release code

IBM pureScale 应用机的可扩展性

WebSphere Application Server 7

DB2 pureScale

Infiniband

WebSphere Application Server 7

DB2 pureScale

Power 770

Power 770

System and Servers arrive application ready

Scale Out

Add more servers in minutes with DB2 pureScale

WebSphere Application Server 7

DB2 pureScale

Power 770

Arrives fully configured

Scale Up

Add more cores as your needs grow

Scale Within

Add cores and dynamically add//change LPARs to adapt in seconds to needs of each tier

Delivery via Rochester
Integrated support for all components (initially only US & Canada)

Leverages existing standard STG & SWG procurement and compensation systems

Q3: more countries and support for channel partners
Q4: Enhancements and more integrated solution

IBM pureScale 应用机 IpAS 6100 系统组件列表

IpAS 6100 一瞥

- DB2 pureScale
  - Offers superior performance and scaleout efficiency.
- WebSphere Application Server 7
  - On POWER7 provides 73% better performance than a competitive application server on Nehalem.
- Infiniband
  - High-speed interconnect ensures system can scale to most demanding needs.
- PowerVM
  - Dynamically adjusts server to meet hanging workloads demands.
- AIX 6
  - Provide the highest level of performance and reliability of any UNIX operating system
- Power 770 servers
  - Drives up to 90% server utilization with industry-leading virtualization and provides resiliency without downtime.

IpAS 6100 DB2 和 WAS 集群的测试结果

测试执行

DayTrader Benchmark with RPT Tool used
Each DB2 Member Shares 3 Cores with one WAS Member
Scalability Test from 1 WAS-DB2 Member pair to 2 WAS-DB2 Member Pairs
CPU Saturation of 95% per WAS-DB2 Member pair

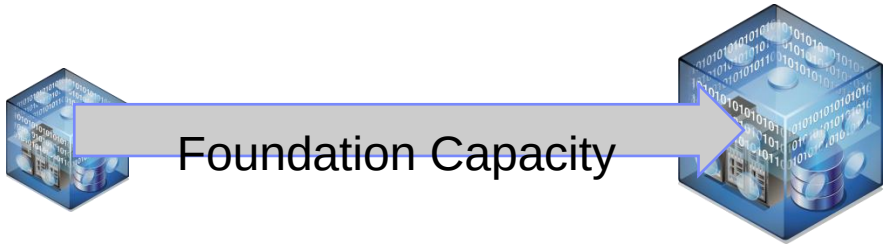
测试结果

Scalability factor is 1.85 times from 1 WAS-DB2 Member Pair to 2 WAS-DB2 Member Pairs

CPU Saturation of 99 % for 1410 business transaction requests per second per WAS-DB2

CONFIG	Transaction per second	CPU Utilization	Cores usage	Comments
First Shared Pool Test: WAS member 1 and DB2 member 1 within 1st Shared Pool	1410 req/s	DB2 Member 1 – 94%, pc=1.75 WAS Member 1 – 95%, pc=1.22	~ 2.81	1 DB2 Member and 1 WAS Member in a Single Shared Pool of 3-Cores
Second Shared Pool Test: WAS member 2 and DB2 member 2 within the 2nd Shared Pool	1490 req/s	DB2 Member 2 – 90%, pc=1.74 WAS Member 2 – 96%, pc=1.23	~ 2.75	1 DB2 Member and 1 WAS Member in a Single Shared Pool of 3-Cores
Two WAS members and two DB2 members in both Shared Pools of 1 and 2	2640 req/s	WAS Member 1– 94%, pc=1.72 DB2 Member 1 – 85%, pc=1.23 WAS Member 2 – 94%, pc=1.26 DB2 Member 2 – 85%, pc=1.63	~ 5.23	DB2 Member 1 and WAS Member 1 in a First Shared Pool of 3-Cores DB2 Member 1 and WAS Member 2 in a Second Shared Pool of 3-Cores

IpAS 6100 硬件配置

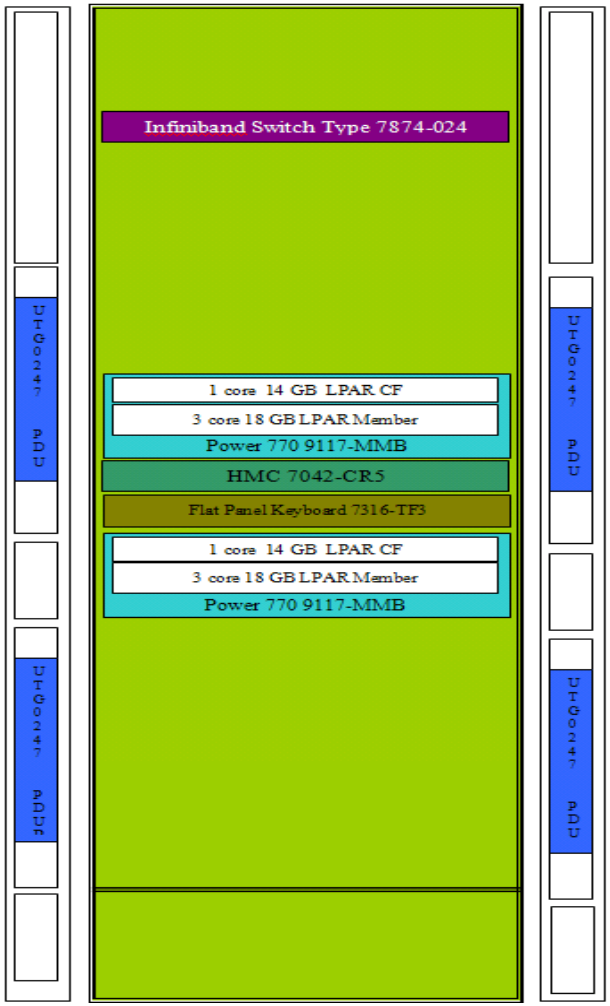


- **两台 IBM Power 770 服务器**
- **三种可选配置**
  - 4, 8, and 16 cores, 3.1GHz proc card, activated (16 cores populated)
- **32/48/64 GB RAM**
- **Infiniband Switch**
- **HMC console**
- **T42 rack**

Size	Small 4 cores	Medium 8 cores	Large 16 cores
# of P770 servers	2	2	2
Memory per server (GB)	32	48	64

IpAS Rack 内部规划

IBM pureScale Application System 7014-T42 Rack



小结 - DB2 pureScale 可以为您带来哪些收益？

- 交付出色的可伸缩性和可用性
- 执行常规操作时的并发性更佳
- 出现成员故障时的并发性更佳
- 降低应用设计的复杂性和返工率，提供良好的可伸缩性
- 改善 SLA 一致性
- 降低对事务性能和可用性要求极高的应用的总体成本

小结 - 为什么 DB2 pureScale 优于 Oracle RAC

- **高可用性**
  - DB2 pureScale 消除了节点在宕机时的冻结时间
- **性能更好**
  - 锁定，缓冲池集中管控
  - DB2 pureScale 使用了 RDMA 的技术
- **易于管理**
  - DB2 pureScale 提供了真正意义上的应用透明性
- **低成本**
  - 相同性能下，DB2 pureScale 配置更低
  - DB2 pureScale 可以更有效地扩展


Thank
You