IBM DB2 数据库介绍

尤祖喜
Information Management Partner Ecosystem
youzuxi@cn.ibm.com

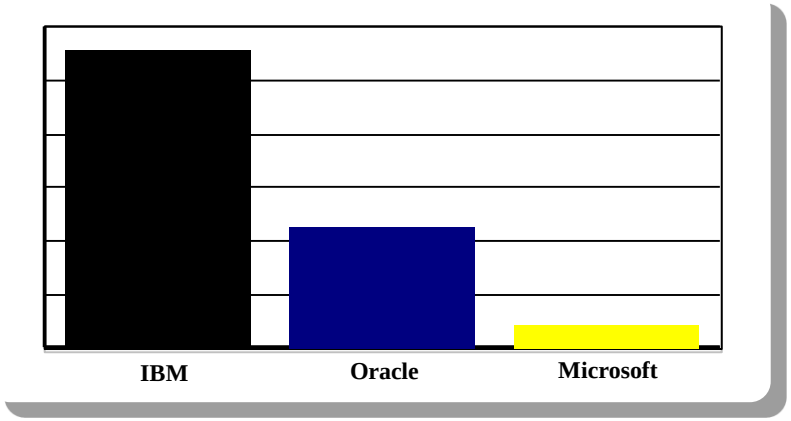
议程

- DB2架构及技术特点
- DB2集群技术
  - DB2 DPF集群
  - DB2 pureScale集群
- DB2列式存储及内存计算
  - DB2 BLU
- DB2客户案例

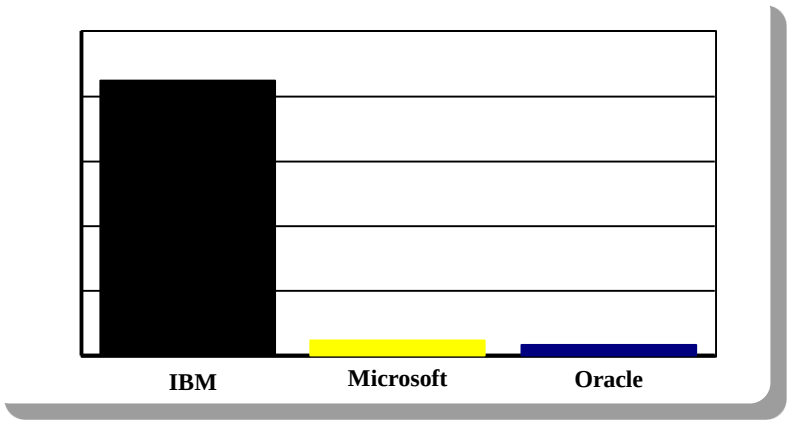
IBM在数据管理领域的实力——技术与标准的领导者

- **关系型数据库发明者 —— Dr. E.F.Codd**
- **SQL语言创造者 —— Dr. Camberlin**

被接受的SQL标准建议



技术专利



IBM 拥有7倍于行业对手的专利技术，拥有2倍于行业对手的标准！

IBM在数据管理30年核心技术的强大基础研发力量。

IBM在数据库领域的技术实力

- 三十年来从理论研究到应用产品， IBM公司在数据库管理系统的研究和发展中作出了巨大的贡献
- 发明关系数据库
  - 1970年，IBM 研究中心E.F.Codd 博士提出了关系型数据库模型，紧接着IBM研究中心发明了第一个关系型数据库SYSTEM R 和 SQL语言
- 数据库发展——发明SQL语言
  - 80年代开始，基于SQL的关系型数据库逐渐成为数据库管理系统的主流，IBM的关系型数据库DB2主宰了大型机上的数据库应用,IBM DB2 提出了分布式数据库DRDA 架构
- 2000年IBM发布开放平台DB2 UDB V7 -- 电子商务数据库
- 2001年收购Informix，与Informix进行技术互补
- 2003年IBM发布DB2 UDB V8.1
- 2006年IBM发布DB2 for LUW V9.1 – 全球第一个混合数据模型数据库
- 2009年IBM发布DB2 for LUW V9.7 – 海纳百川的兼容性
- 2009年IBM发布DB2 for LUW V9.8 – 提供高可用无限扩展能力的PureScale
- 2012年IBM发布DB2 for LUW V10.1 – 大幅提升数据仓库性能
- 2013年IBM发布DB2 for LUW V10.5 – 提供了极高性能的列存储引擎 BLU

DB2 UDB —— 真正的“通用”数据库

代码特色：

- > 90% 共同代码
- < 10% 独特代码（专注于：性能，系统管理等）

zSeries
iSeries

Enterprise
- Linux 32-bit (Intel, AMD)
- Linux 64-bit (Intel, AMD)
- Linux pSeries
- Windows
- Windows 64-bit
- AIX
- HP-UX
- Solaris

Workgroup,
Workgroup
Unlimited
- Linux
- Windows
- AIX
- Solaris
- HP-UX

Express
- Linux
- Windows

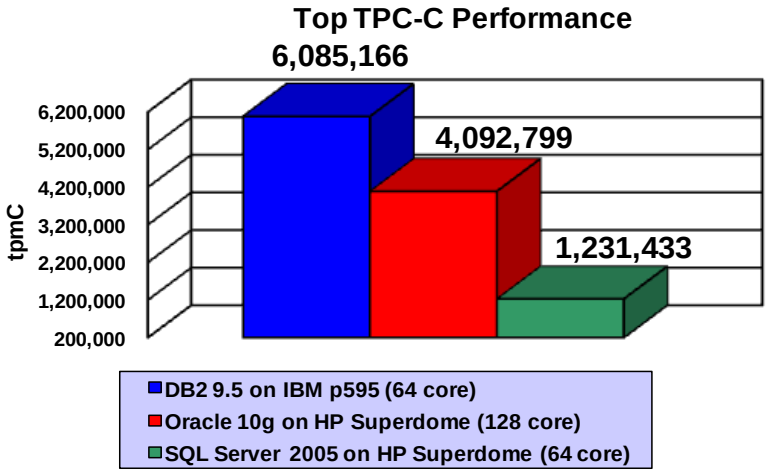
Everyplace
- Linux
- Windows
- PalmOS
- EPOC-32
- Neutrino

卓越性能

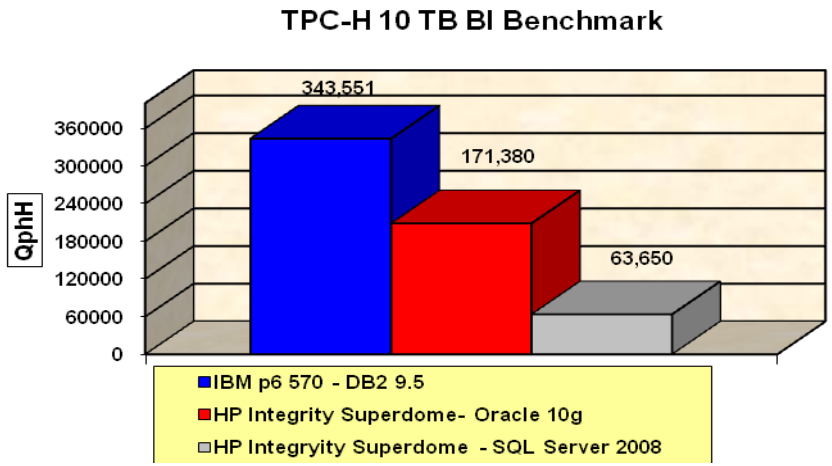
ZURICH

“在与竞争厂商比较测试中，IBM DB2始终显示出更佳的性能。DB2的质量令人惊叹。”

—Benjamin Simmen, 苏黎世金融服务集团



- 比Oracle 快50%
- 比SQL Server 快5倍
- 为联机交易降低服务器成本



- 比Oracle快65%
- 比SQL Server快5倍
- 为数据仓库提供了强大动力

更低的服务器成本 → 更低的软件许可费用 → 更低的维护成本

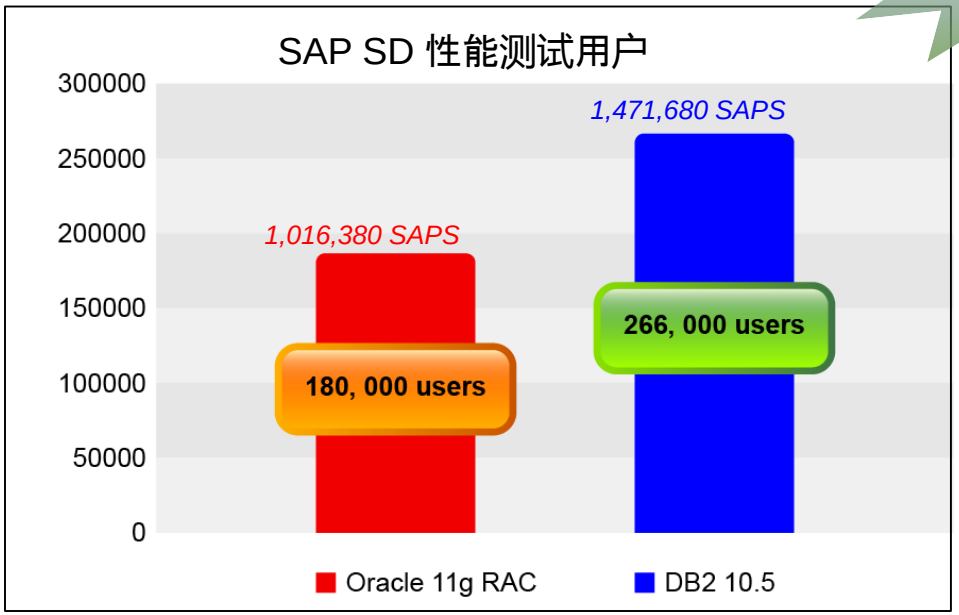
IBM 创造了新的世界纪录(DB2 10.5)

3层架构SAP SD 性能评测, 26.6万 SAP 用户！

64核 IBM Power® 780 AIX® 7.1 及 DB2® 10.5

是Oracle最佳成绩的1.47倍！

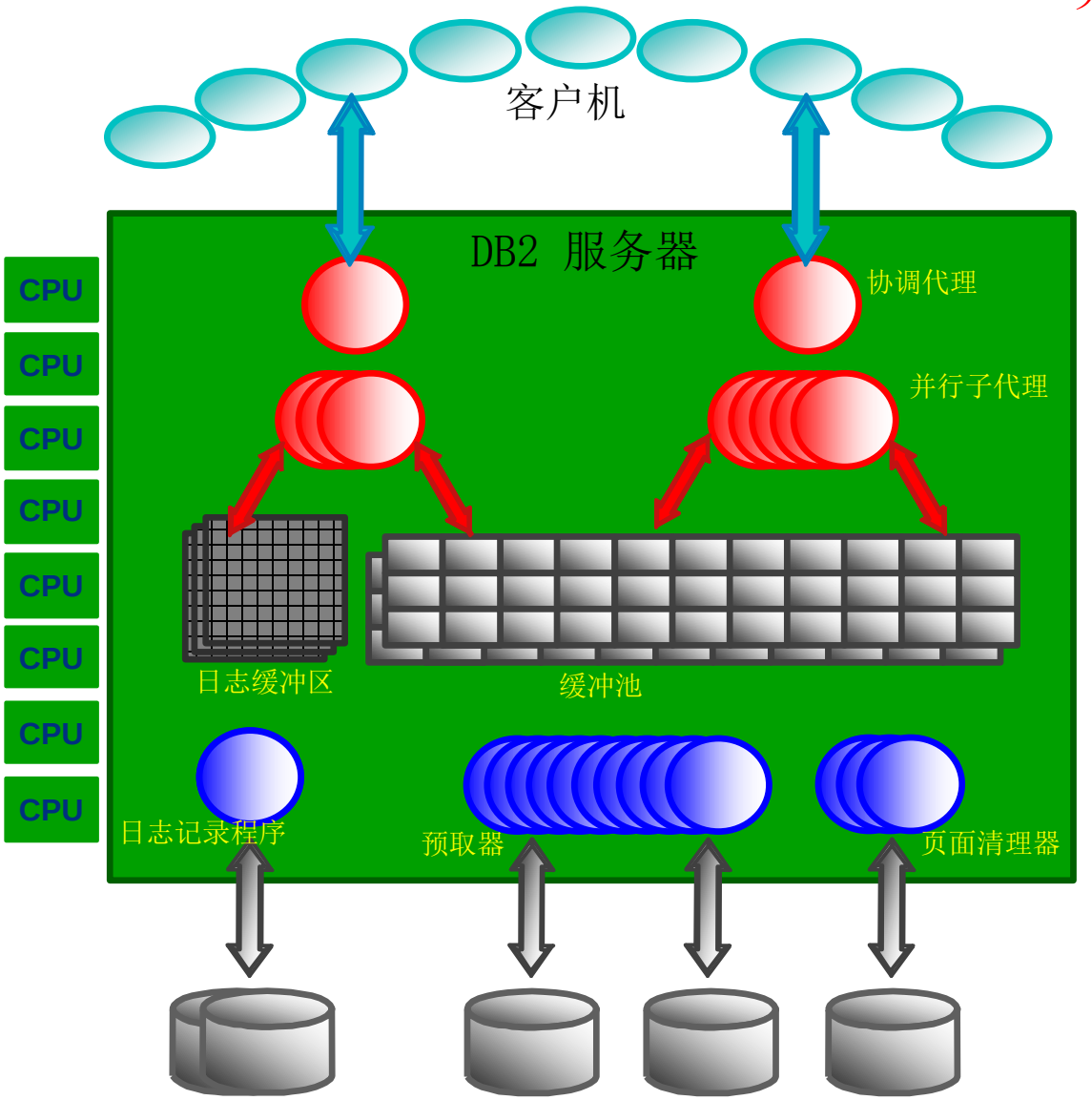
Power平台的DB2在三层架构的SAP SD标准应用评测中取得最高SAP SD用户数成绩已经超过7年！



1) Results of DB2® 10.5 on IBM Power 780 on the three-tier SAP SD standard application benchmark on SAP enhancement package 5 for SAP ERP 6.0, achieved 266,000 SAP SD benchmark users, certification # 2013010. Configuration: 8 processors / 64 cores / 256 threads, POWER7+ 3.72 GHz, 512 GB memory, running AIX 7.1
2) Results of DB2® UDB 8.2.2 on IBM eServer p5 Model 595 on the three-tier SAP SD standard application benchmark running SAP R/3 ® Enterprise 4.70 (ERP) software, achieved 168,300 SAP SD benchmark users, certification # 2005021. Configuration:32-core SMP, POWER5, 1.9 GHz, 256 GB memory, running AIX 5.3
3) Results of Oracle 11g Real Application Clusters (RAC) on SAP sales and distribution-parallel standard application benchmark running the SAP enhancement package 4 for SAP ERP 6.0, achieved 180,000 SAP SD benchmark users, certification # 2011037. Configuration: 8 x Sun Fire X4800 M2 each with 8 processors / 80 cores / 160 threads, Intel Xeon Processor E7-8870, 2.40 GHz, 8 x 512 GB memory, running Solaris 10

Source: http://www.sap.com/benchmark

DB2 架构概览



并行

- SQL 和实用程序
- 分区内与分区间并行
- 具有查询重写功能的基于成本的优化器
- 基于负载的动态调节

对 SMP 的利用

- 通过操作系统线程和进程利用所有 CPU

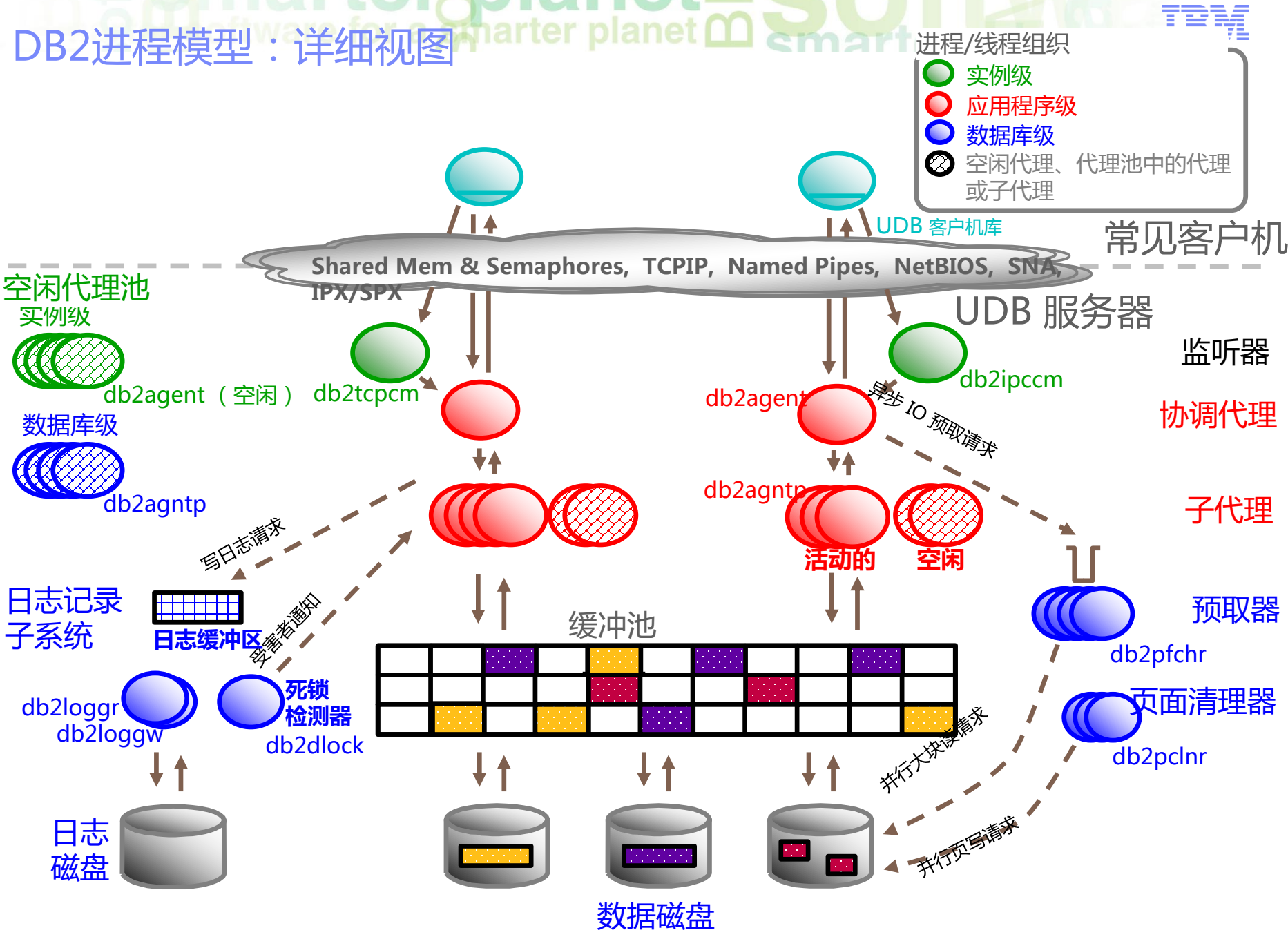
对内存的大量使用

- 64 位支持
- I/O 缓冲
- 多个缓冲池

I/O 子系统

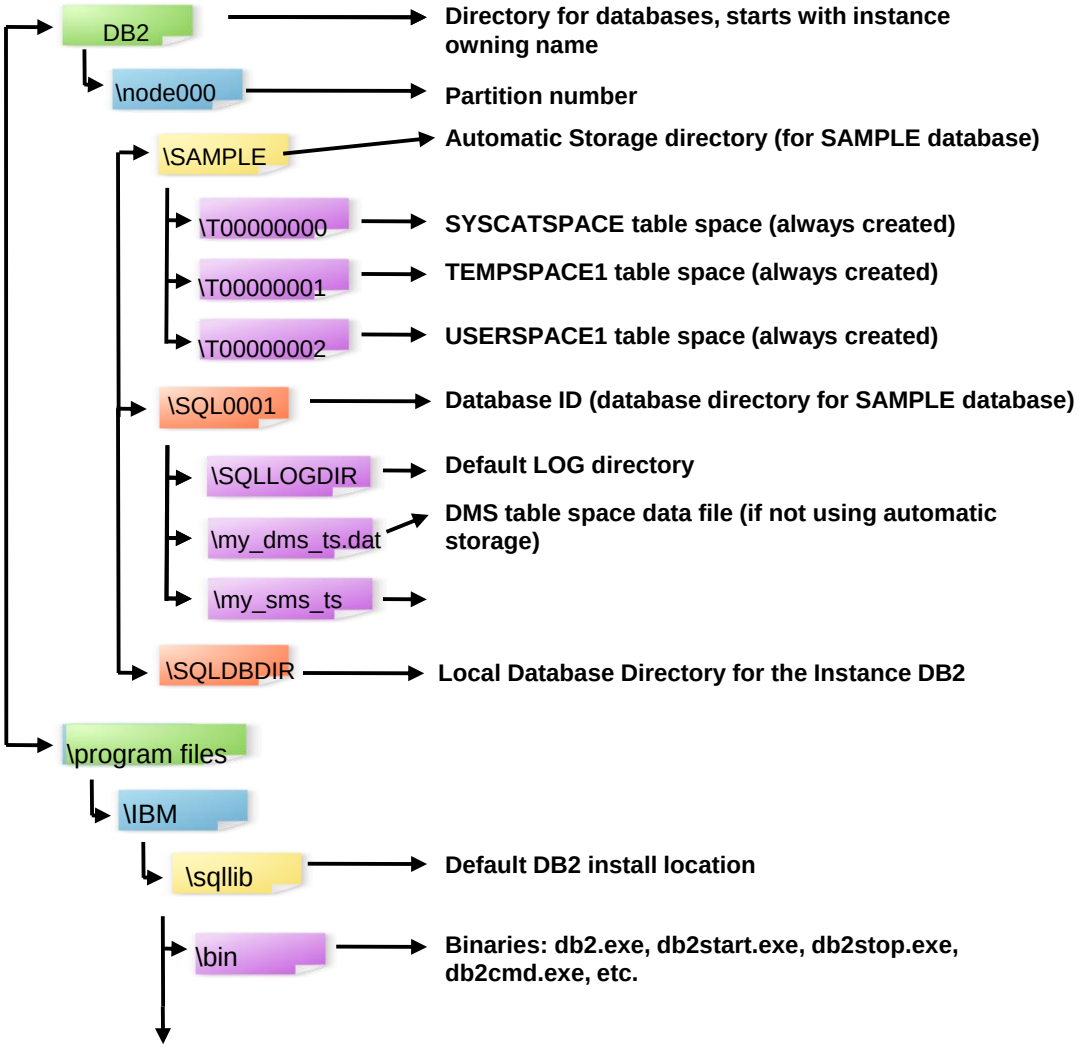
- 异步并行 I/O
- 使用并行 I/O 的自动化智能数据分割
- 大块 I/O
- 分散/收集 I/O

DB2进程模型：详细视图



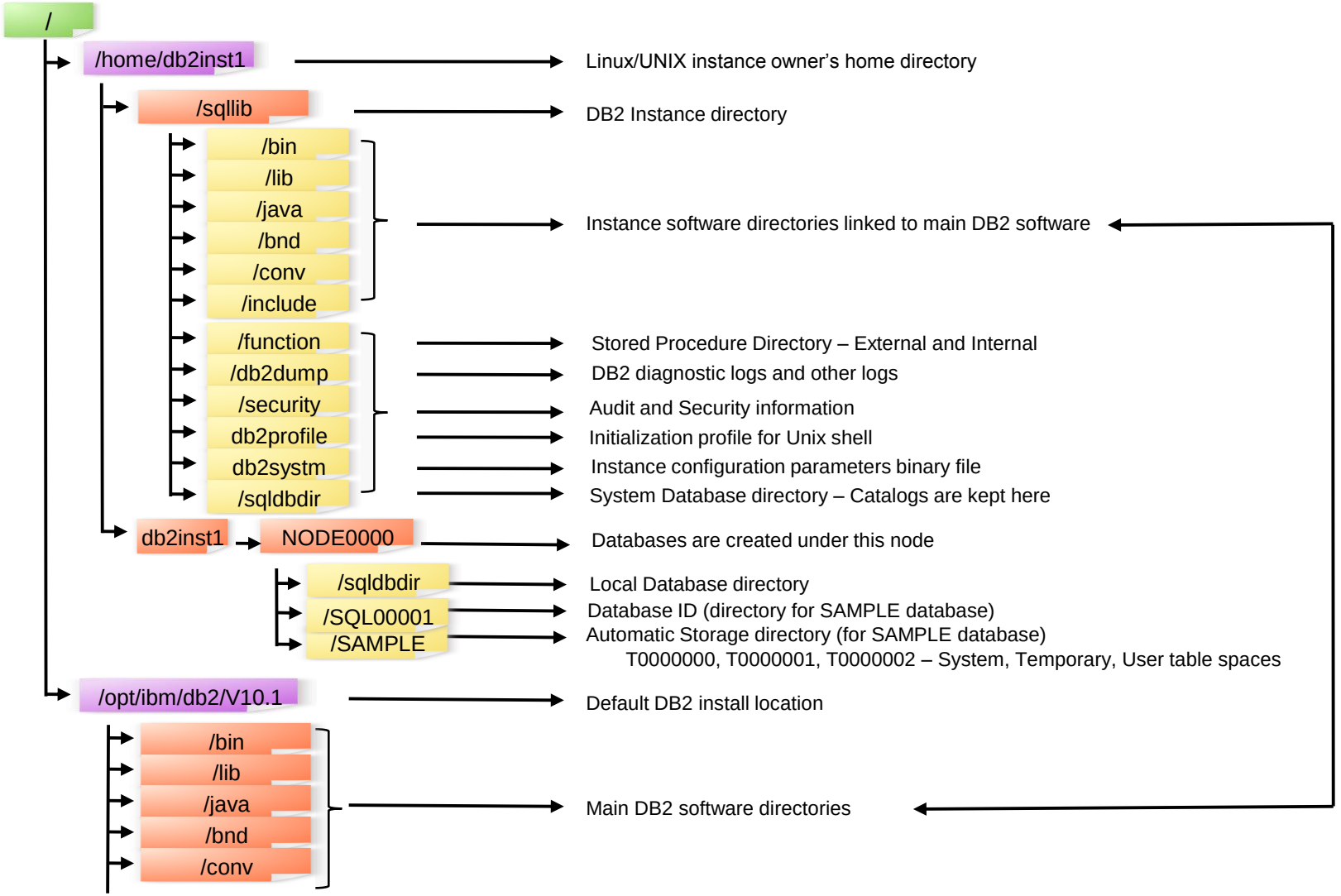
DB2 Installation – Directory Structure

- Windows



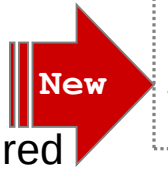
DB2 Installation – Directory Structure

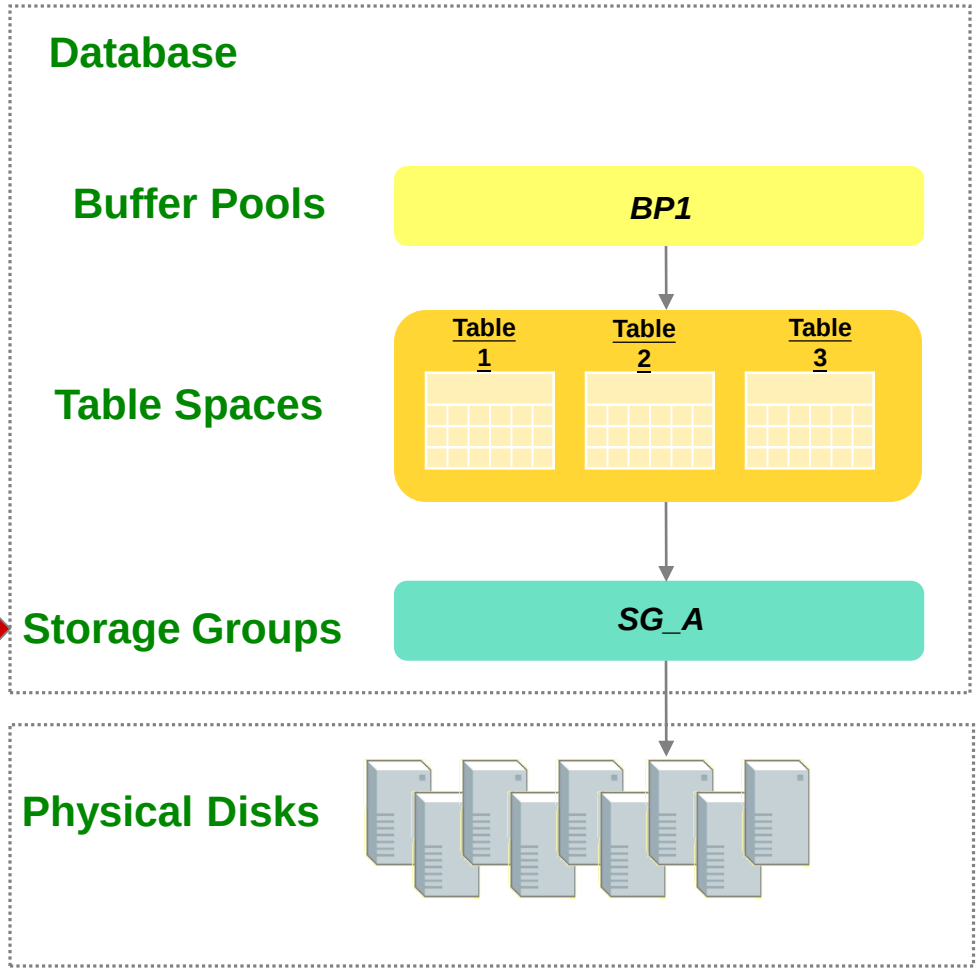
- Linux / UNIX (Automatic Storage)



DB2 Storage Model

- **Database**
  - Contains a set of objects used to store, manage, and access data
- **Buffer Pool**
  - Area of main memory for the purpose of caching data as it is read from disk
- **Table Space**
  - Logical space used to store data objects such as tables and indexes
- **Storage Group**
  - Set of storage paths configured to represent different classes of storage in the database system, where table spaces are stored
- **Physical Disk**
  - Physical location used to store data





Storage and I/O Architecture

Support for BOTH RAW devices (DMS) and File Systems (SMS)

```
create tablespace MYTABLESPACE
    managed by database
    using (device '/dev/rhd7' 100000,
           device '/dev/rhd8' 100000,
           device '/dev/rhd9' 100000)
create table T1 (C1 INT,...) in MYTABLESPACE
insert into T1 values (..., ..., ...,  ..... )
alter tablespace MYTABLESPACE
    add stripe set (device 'dev/rhda' 100000,
                    device 'dev/rhdb' 100000)
```

-OR-

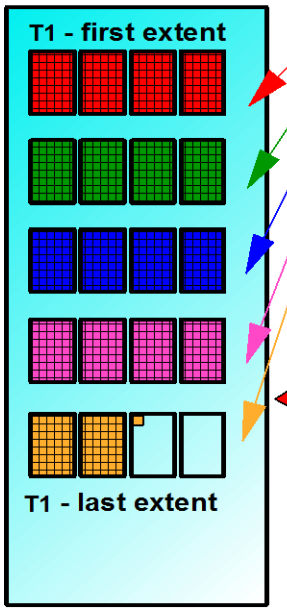
```
create tablespace MYTABLESPACE
    managed by system
    using ('/dir1', '/dir2', '/dir3' )
```

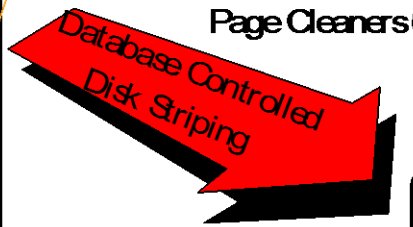
Database Managed Space (DMS)

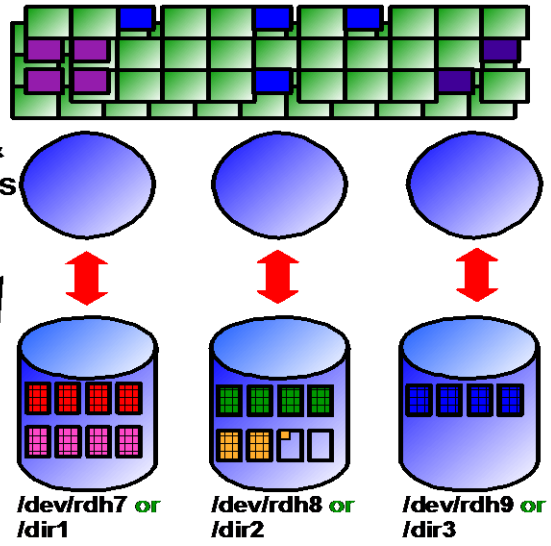
- DB2 takes control of entire device
- Optimum Performance

System Managed Space (SMS)

- DB2 works through the file system
- Excellent Performance
- Ultra-simple Administration







Multiple Buffers & Page Sizes within Single Database

h

Intelligent Striping enables Full Parallel I/O

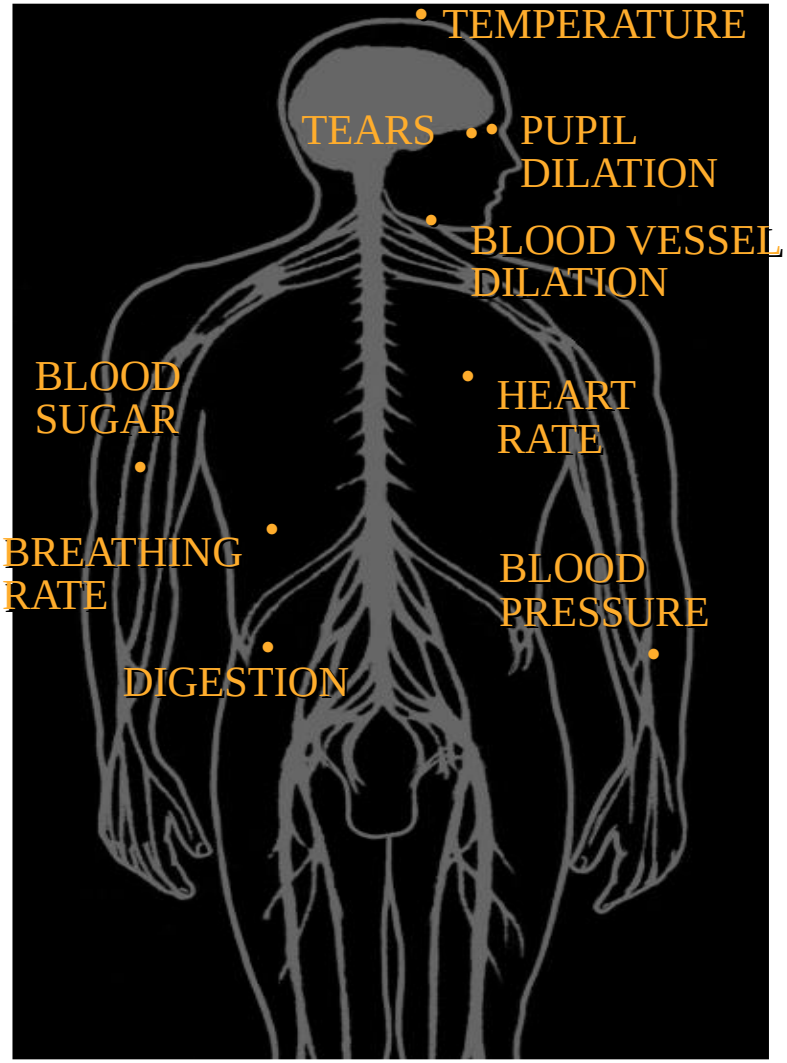
DB2的自主管理 —— 数据库管理员技能不再是问题

自主管理和资源调整(SMART)数据库

- 减少数据库日常运行过程中需要的人为干预，包括高质量的自动化和提供专家建议
- IBM“电子蜥蜴”计划的一部分

SMART DB计划的主要内容

- 安装、配置和运行
- 专家设计
- 自我维护
- 自我治愈
- 自我恢复

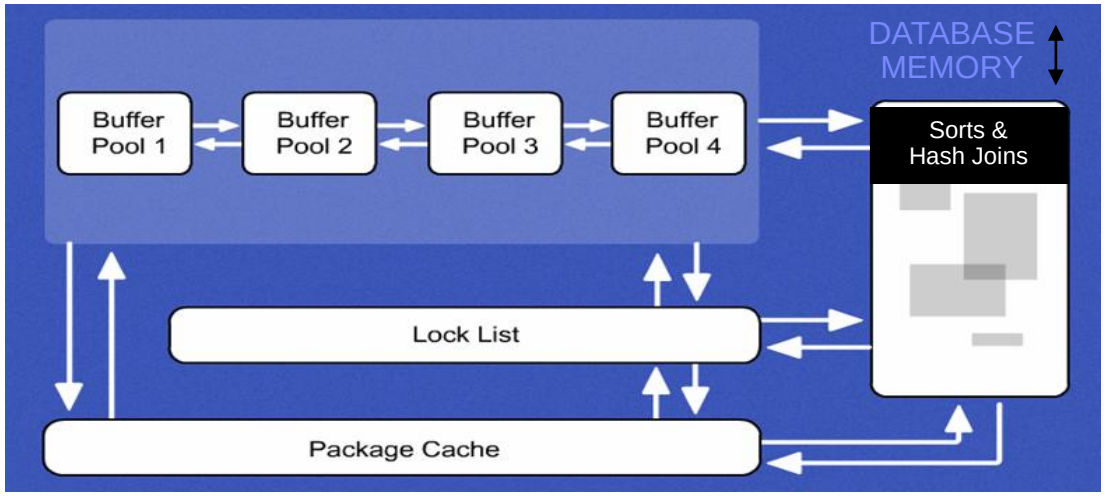


- 2008 Solitaire Interglobal Study
  - 250家客户实际情况调查，对比DB2和其他数据库产品的人工需求：
    - DB2 平均总人员数：15.6个全职员工
    - 其他厂商平均总人员数：22.1个全职员工 (多42%)

”DB2涉及的管理工作比任何其他数据库要少得多，英国商业银行运行着50到100 个SAP系统，仅由12个人来维护，简直太不可思议了！”
- Sheila Moran at Bank Of Commerce in UK

自动优化

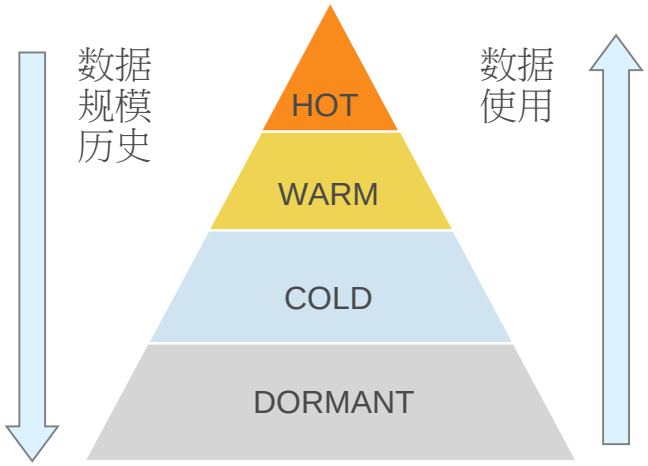
- DB2 自动调优——the Self Tuning Memory Manager (STMM)
  - 自动控制DB2主要的内存对象：
    - Sort, locklist, package cache, buffer pools, and total database memory
  - 无需人工干预的内存自我在线调优
  - 自我感知工作负载、按需调整内存大小
  - 能够迅速适应工作负载的突然变化，自动重新划分内存区域



(SELF_TUNING_MEM) = ON
(DATABASE_MEMORY) = AUTOMATIC
(DB_MEM_THRESH) = 10
(LOCKLIST) = AUTOMATIC
(MAXLOCKS) = AUTOMATIC
(PCKCACHESZ) = AUTOMATIC
(SHEAPTHRES_SHR) = AUTOMATIC
(SORTHEAP) = AUTOMATIC

“多温度”存储

- 同一个表的数据按照使用“热度”分布在相应的存储上
- 充分发挥SSD、SAS、SATA的不同特点
- 存储成本与查询性能的智慧平衡
- 新增Storage Group对象，对存储分类



自动时间版本管理

- 支持“系统时间旅行”和“业务时间旅行”
- 自动维护数据的历史版本
- 数据库支持基于“时点”快照的查询
- 简化代码开发，优化查询性能
- 支持 “业务时间切片”管理

d_employee

EmpID	Dept	System_start	System_end
12345	M15	05/31/2000	12/31/9999

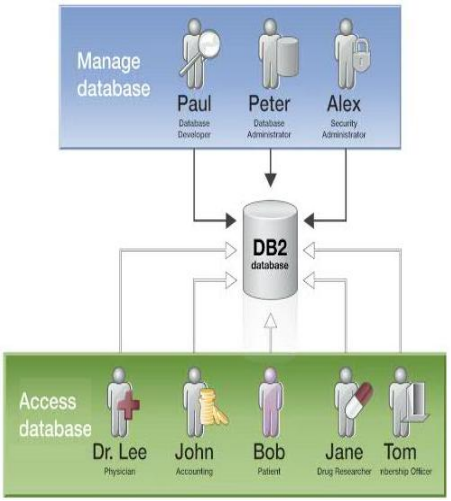
d_employee_history

EmpID	Dept	System_start	System_end
12345	J13	11/15/1995	01/31/1998
12345	M24	01/31/1998	05/31/2000
67890	K25	11/15/1995	03/31/2000

```
--Step 1
CREATE TABLE d_employee_history like travel in hist_space;
--Step 2
ALTER TABLE d_employee
ADD VERSIONING USE HISTORY TABLE d_employee_history;
```

行列级权限控制

- 行列级权限控制
- 简化架构设计
- 通过授权即可简单实现



同一张表

Account	Name	Income	Branch
1111-2222-3333-4444	Ana	22,000	A
2222-3333-4444-5555	Bob	71,000	B
3333-4444-5555-6666	Celia	123,000	B
4444-5555-6666-7777	Dinesh	172,000	C

张三
看到的

Account	Name	Income	Branch
2222-3333-4444-5555	Bob	71,000	B
3333-4444-5555-6666	Celia	123,000	B

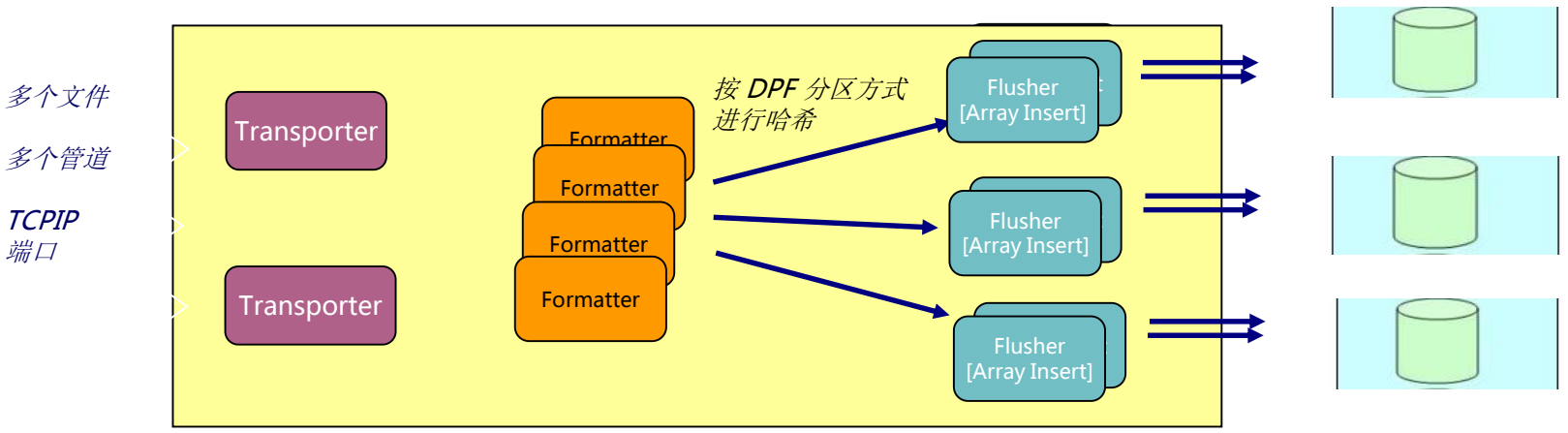
李四
看到的

Account	Name	Income	Branch
xxxx-xxxx-xxxx-4444	Ana	22,000	A
xxxx-xxxx-xxxx-5555	Bob	71,000	B
xxxx-xxxx-xxxx-6666	Celia	123,000	B
xxxx-xxxx-xxxx-7777	Dinesh	172,000	C

Continue Data Ingest – 新一代数据导入工具

一种新的DB2 客户端工具，将包括文件、管道和TCP/IP客户端等数据高速、持续的导入到目标表

支持“实时”数据仓库，提高数据可用性，支持并行数据读写，是DB2 Load实用程序的一种备选方案.



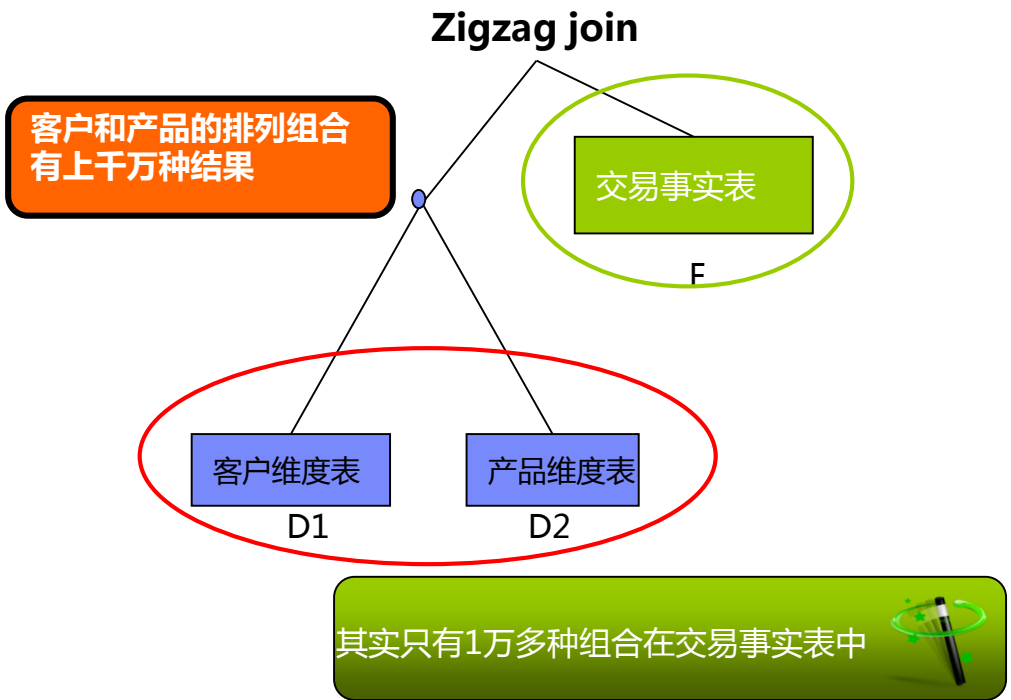
客户端并行运行的线程

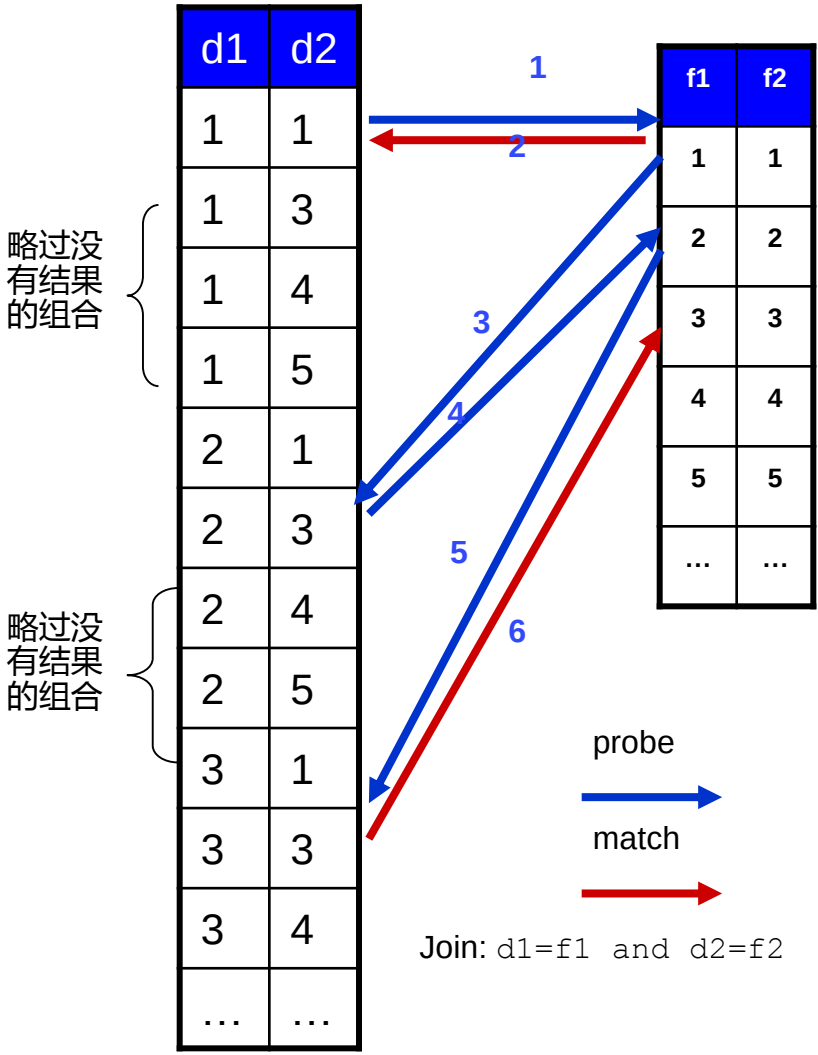
DB2 DPF

“星型”模型优化

- "事实-维度" 是典型的星型模型
- ZZJOIN优化查询效率
- 优化器自动识别“星型”查询

那些产品卖给了40-55年龄段的客户？

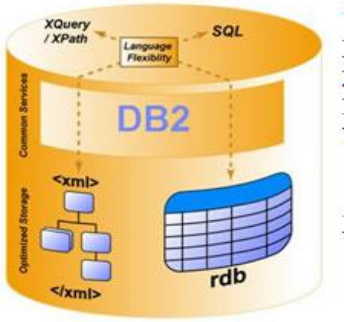




原生XML支持 — pureXML

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

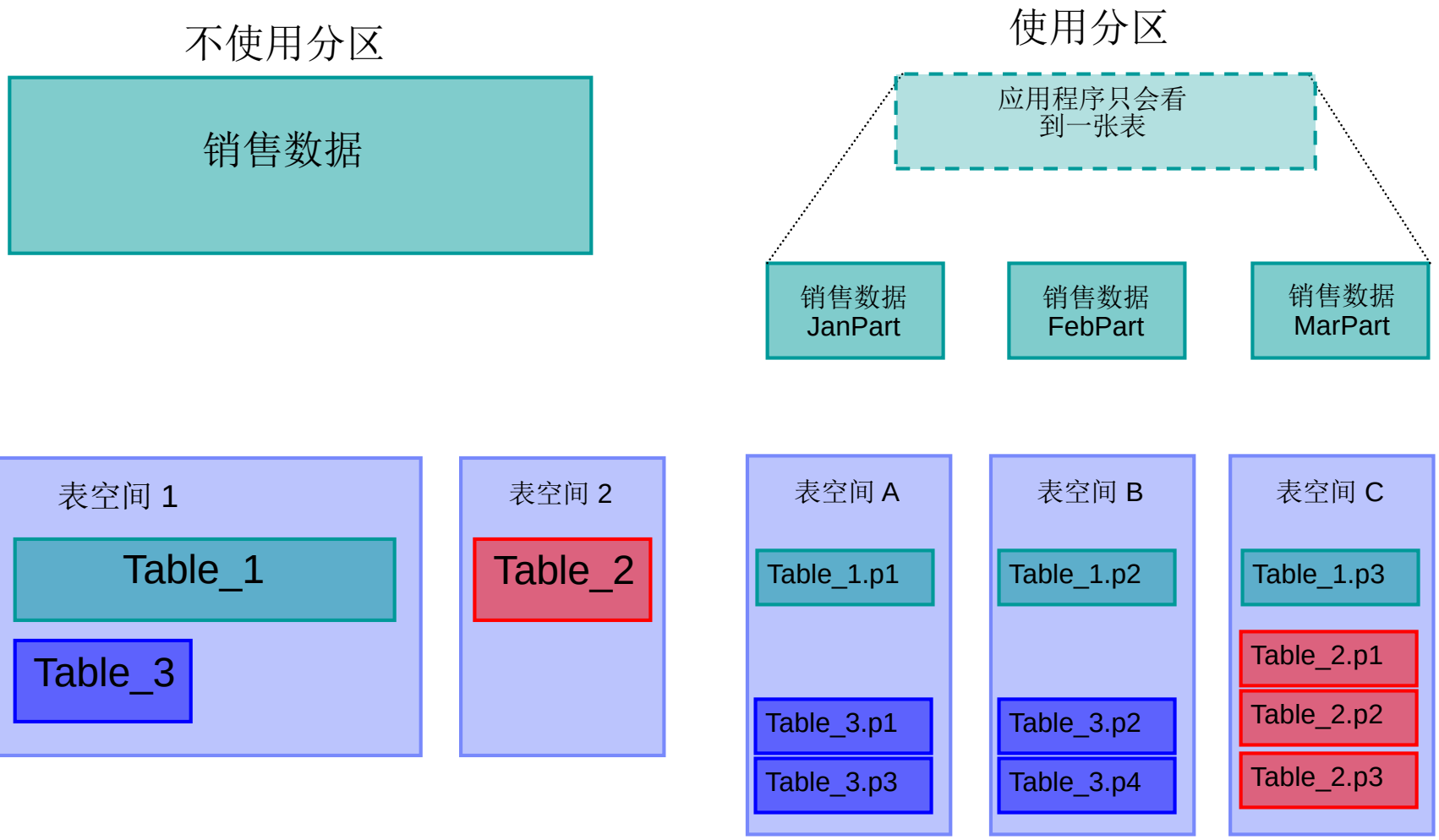
- 使XML数据的分析更加快速
- 易于在数据仓库中应用XML数据
- XML可以存在于数据分区、表分区、数据库视图和物化查询表中
- 改进的XML数据索引和压缩支持

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。”

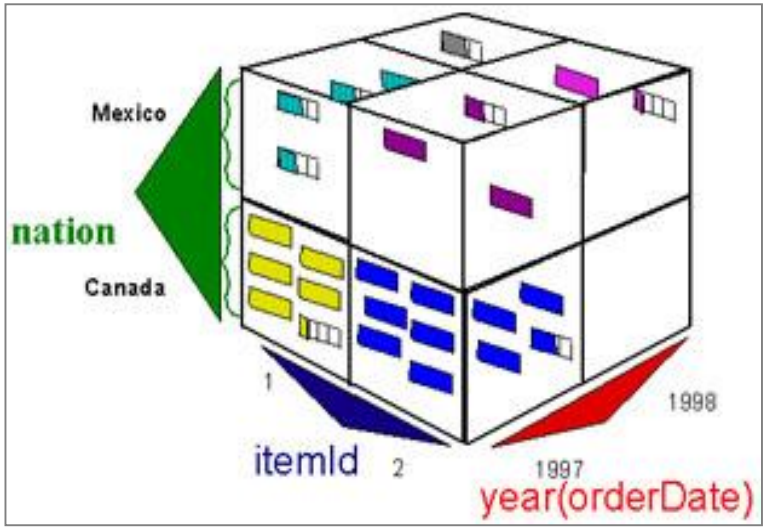
—Keith Feingold, CEO, Skytide

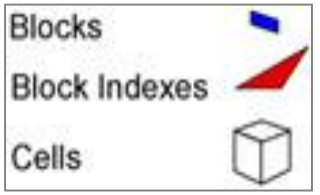
表分区


不使用分区
销售数据
使用分区
应用程序只会看
到一张表
销售数据
JanPart
销售数据
FebPart
销售数据
MarPart
表空间 1
Table_1
Table_3
表空间 2
Table_2
表空间 A
Table_1.p1
Table_3.p1
Table_3.p3
表空间 B
Table_1.p2
Table_3.p2
Table_3.p4
表空间 C
Table_1.p3
Table_2.p1
Table_2.p2
Table_2.p3

多维聚簇索引MDC

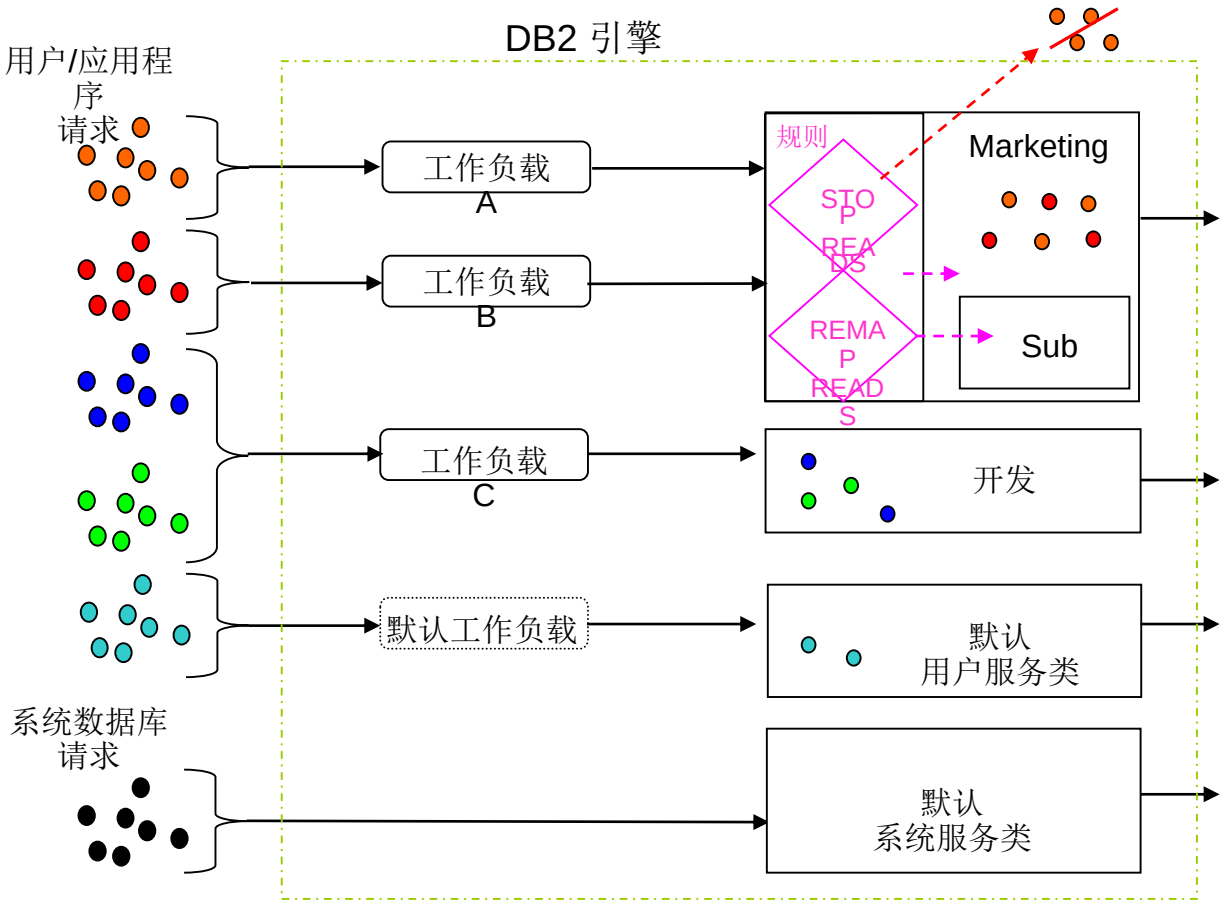
- 块索引
  - –指向一整块页面的索引
- 块
  - –DB2将拥有相同列值的记录放置在相邻的物理地址
- 单元
  - –有同样维值的块组合在一起
- 即使经常执行INSERT操作的话，也可以保证集群长期的有效性





工作负载管理能力

- 防止系统资源过度消耗
  - 并发作业数等控制
  - 对作业运行的总时间进行限制
  - 阻止”垃圾“SQL、”流氓“查询
- 实现不同服务级别
  - 对作业进行优先级管理
- 资源公平使用
  - 分组资源总量控制
  - …



Oracle兼容性

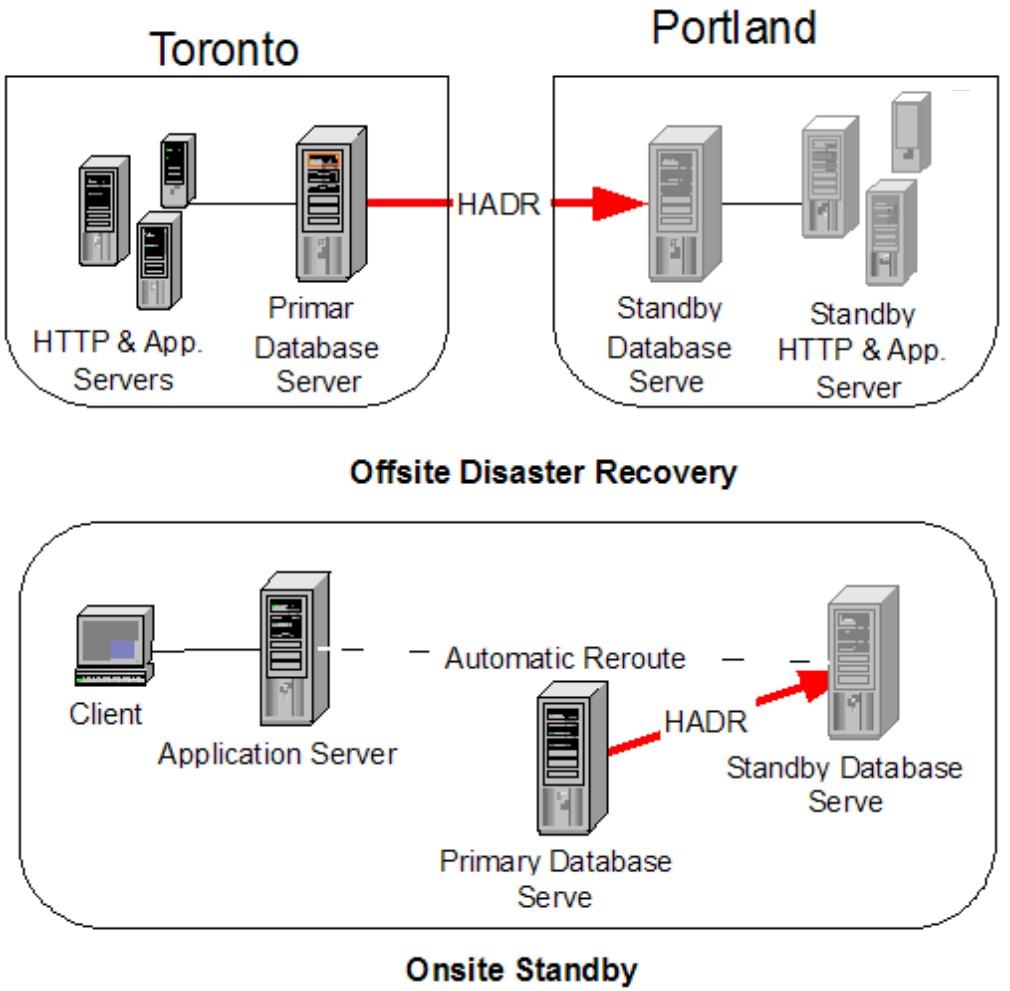
Oracle	→	DB2
Concurrency Control	→	No change
Oracle SQL	→	No change
PL/SQL	→	No Change
Packages	→	No Change
Built-in packages	→	No Change
JDBC	→	No Change
SQL*Plus Scripts	→	No Change

- DB2 9.7对Oracle的并发性控制、SQL 专用语言、PL/SQL、Package、带有扩展的 JDBC 客户端和SQL*Plus 脚本都做到了直接支持
- 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

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！！

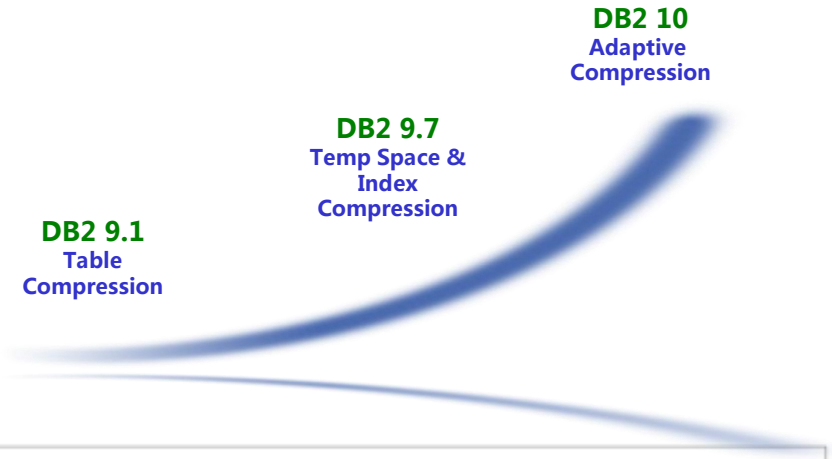
高可用灾难恢复技术 (HADR)

- 目标定位
  - –针对在线交易
- 需求
  - –24 x 7 可用
- 解决方案
  - –离线灾备:
  - –在线热备:
- 价值
  - –业务不被中断
  - –容易使用
  - –客户端自动切换

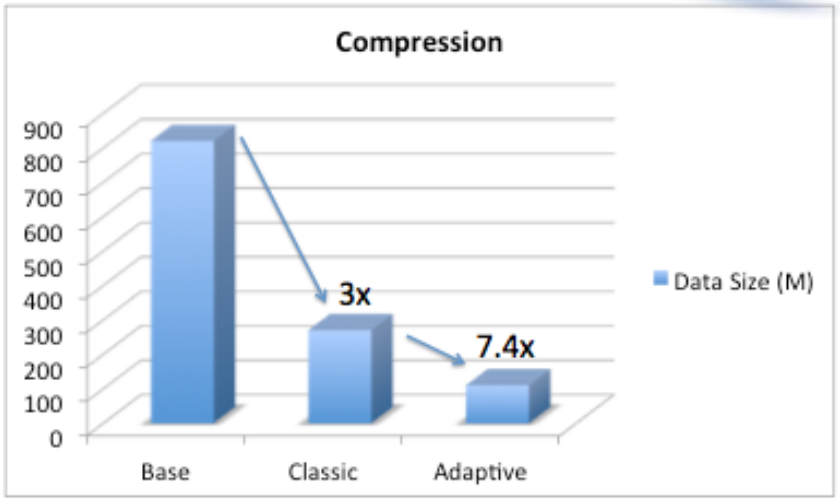


适应性数据压缩
-Adaptive Compression

结合使用页面级压缩技术，压缩效果更好



- 平均7倍压缩比
- 节省大量存储空间
- 大幅降低IO瓶颈
- 提高整体性能
- 透明压缩，实时压缩，无需额外动作

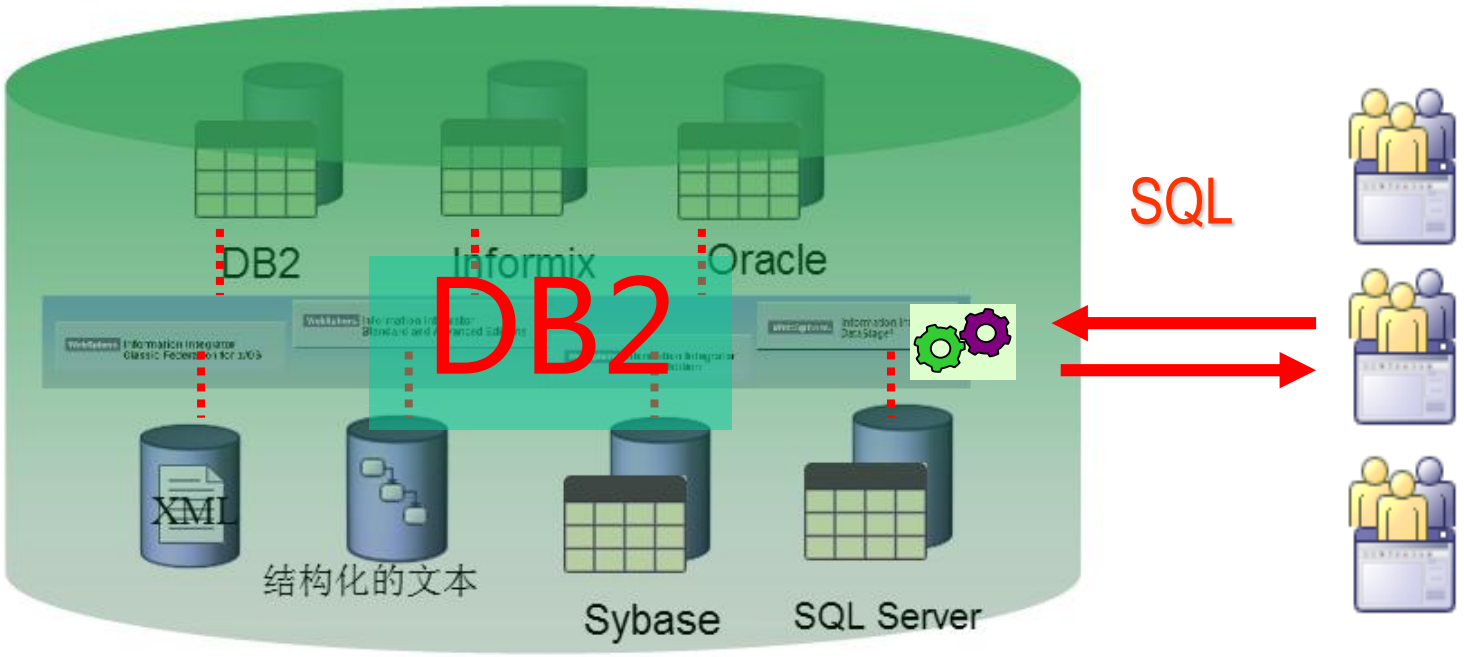


降低存储成本并提供更好的性能

联邦访问 —异构数据虚拟整合

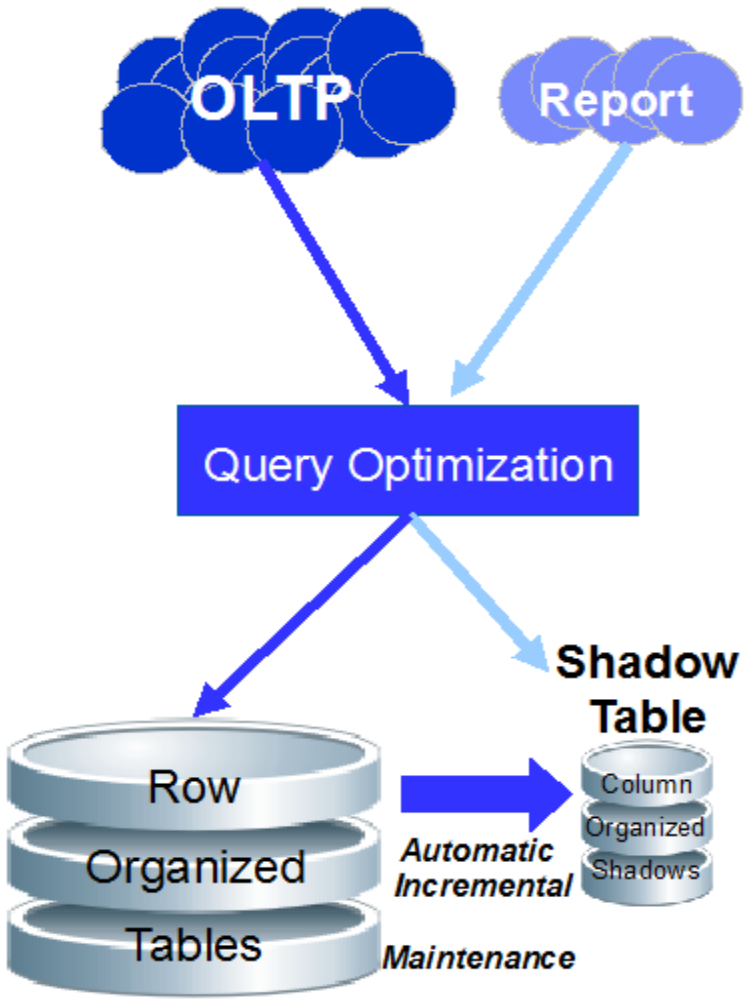
联邦技术可以组合多种不同的数据源，实现虚拟整合，对于最终用户、数据管理者、开发人员来看，所有数据如同在同一个DB2数据库中，可以方便的进行查询和处理。

- 独立性，几乎对原有系统应用没有影响，无需在原有系统上安装程序、软件, 无需特有的接口
- 高性能，优化分布式访问，SQL转换下推
- 安全性，与被访问数据源的用户建立映射，双层用户认证/授权机制

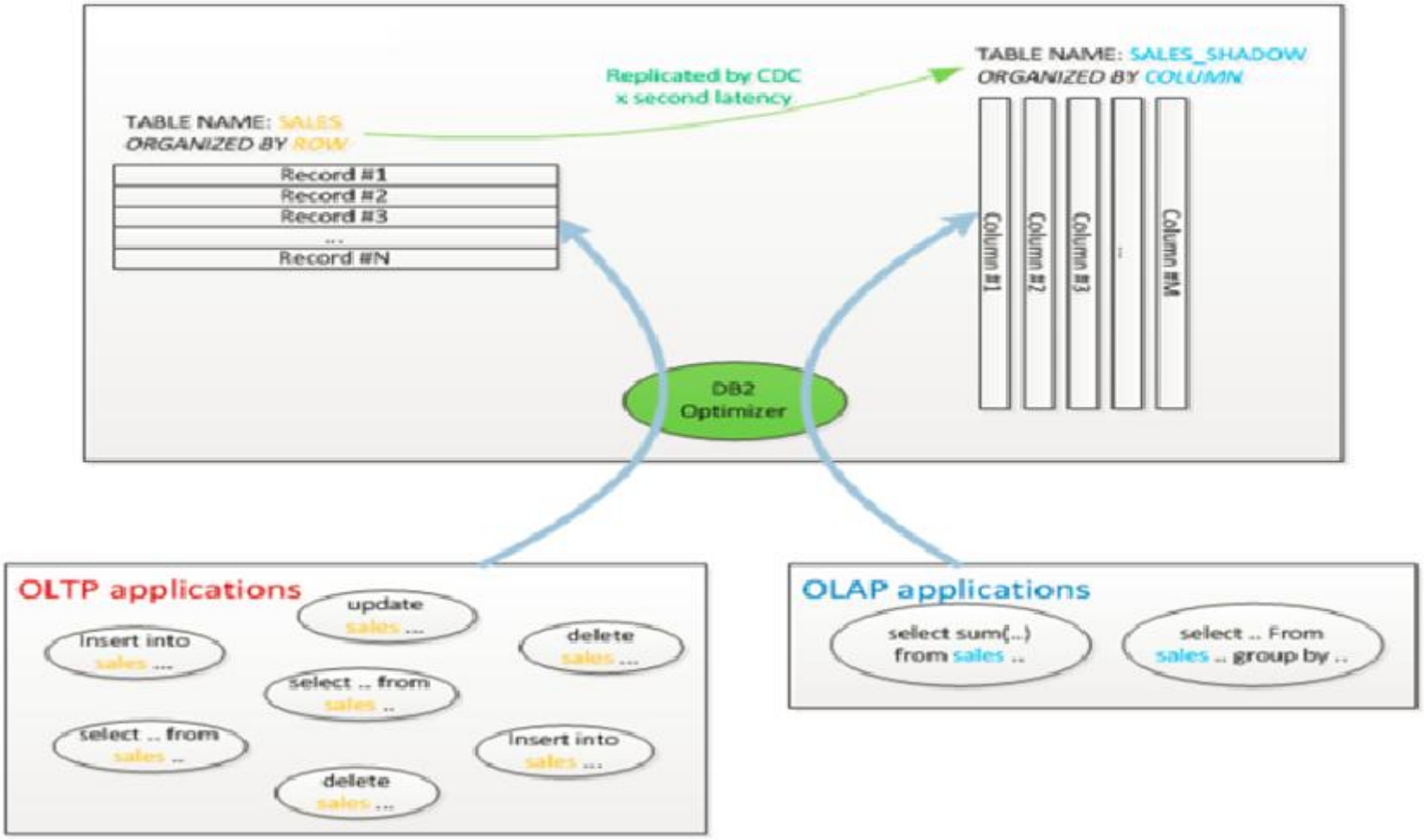


What Is a Column Organized Shadow Table

- **Transparent BLU "Shadow Table"**
  - A new way to implement MQT as columnar data store
- **Powered by DB2 BLU Acceleration**
  - Queries only perform I/O on the columns and values that match query. Work performed directly on columns
- **Smart: Analytical queries issued against the normal row based table automatically diverted to shadow tables to take advantage of BLU Acceleration**
- **Improved performance vs. traditional MQT**



High Level Architecture: Shadow Tables – DB2 View



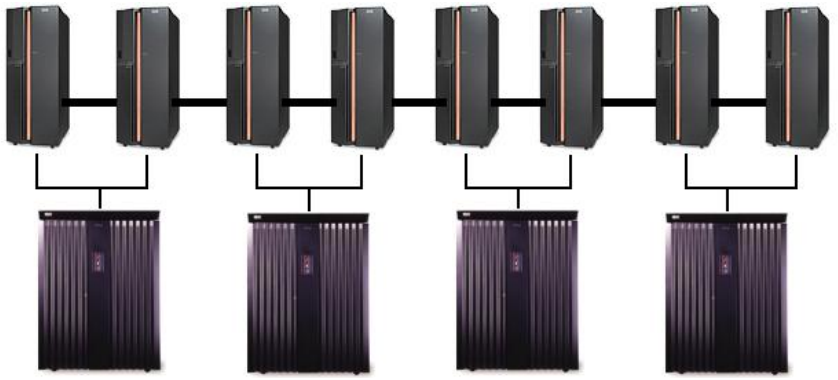
议程

- DB2架构及技术特点
- DB2集群技术
  - DB2 DPF集群
  - DB2 pureScale集群
- DB2列式存储及内存计算
  - DB2 BLU
- DB2客户案例

DB2 数据库分区特性(DPF)

- 分区数据库分布在多个服务器上
- 为什么需要分区?
  - **规模超大, 性能需要, …**
- 优势
  - 对用户和应用程序透明
  - 并行性
    - 工作负载分散在所有结点
  - 通过增加更多服务器来增加伸缩性
- 适用于大型数据库:
  - 数据仓库
  - 数据挖掘
  - 在线分析处理
- DPF 属于IBM InfoSphere数据仓库的一部分

DB2 with DPF

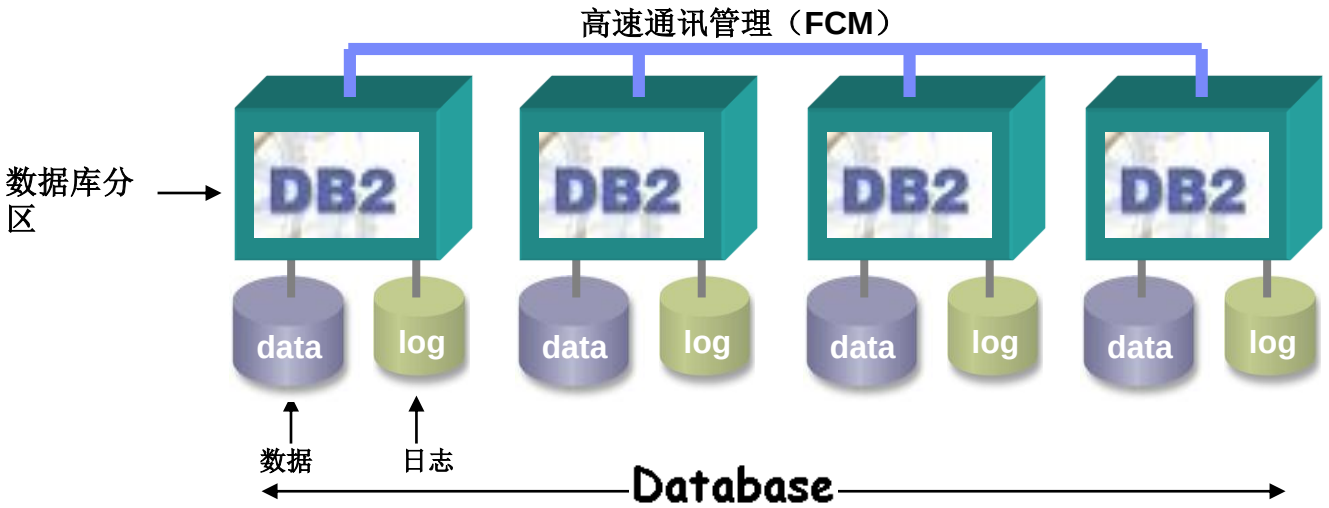


- DB2 核心超尺寸构架基于并行性，又称为 Shared Nothing
  - Inter and intraNode/partition Parallelism… Inter Query Parallelism.. Intra Query…
  - 性能, divide and rule, 没有限制的规模
  - 基于开销的优化器和查询重写器
  - SQL和实用工具完全平行运行
  - 根据负载动态分流
  - 异步I/O 平行I/O

DB2 DPF – 非共享体系架构,无限扩展性

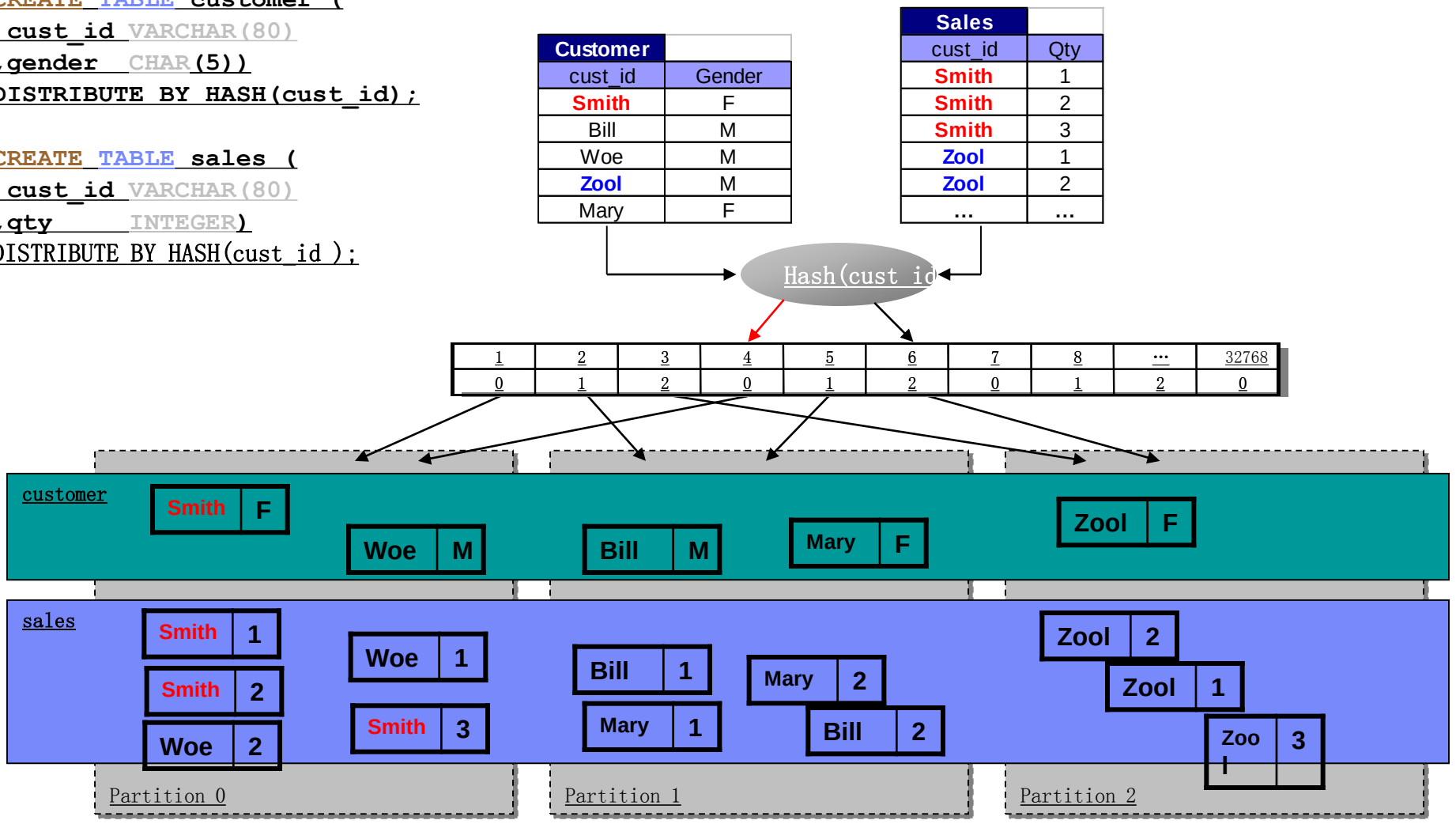
分区数据库模型

- ✓数据库被分为多个分区
- ✓数据库分区可运行在不同的节点上
- ✓每个数据库分区有独立的资源（引擎、日志管理、锁管理、缓存管理等）
- ✓所有分区并行处理，由数据库系统进行统一协调和管理
- ✓对用户和应用来讲是*单一系统映象*



使用哈希（Hashing）方式自动分布数据

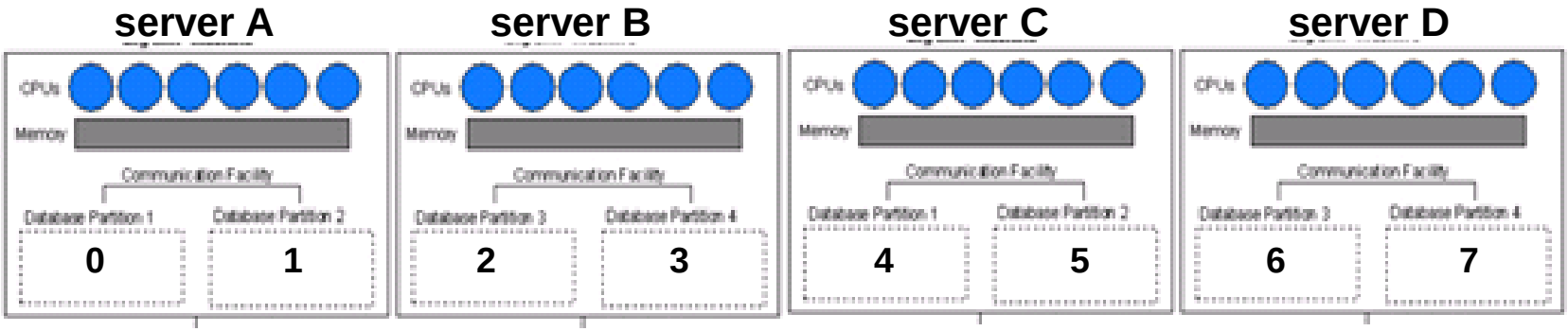
```
-CREATE TABLE customer (
- cust_id VARCHAR(80)
-,gender  CHAR(5))
-DISTRIBUTE BY HASH(cust_id);

-CREATE TABLE sales (
- cust_id VARCHAR(80)
-,qty     INTEGER)
-DISTRIBUTE BY HASH(cust_id );
```



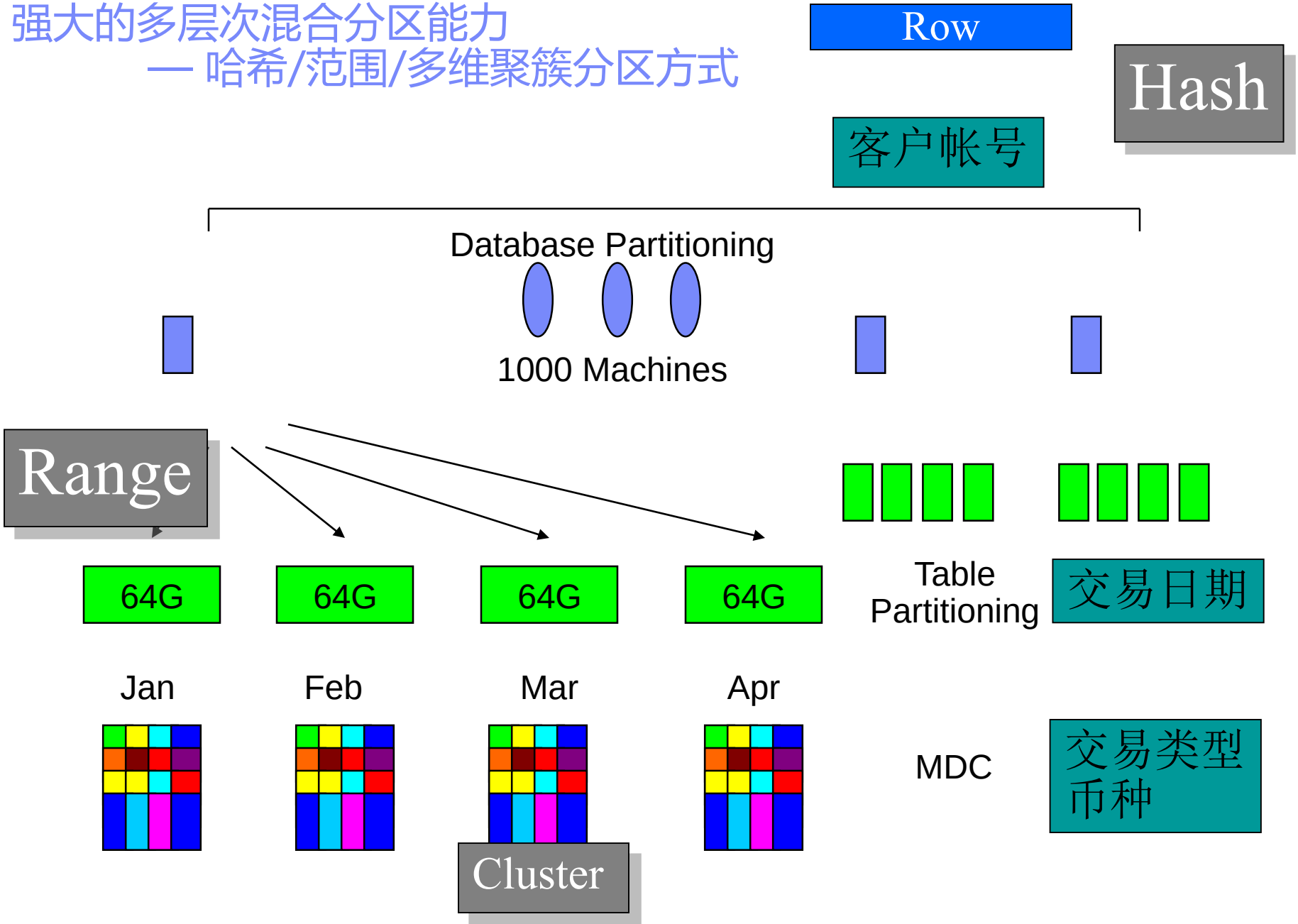
DPF – 节点配置

分区	服务器名	逻辑端口
0	serverA	0
1	serverA	1
2	serverB	0
3	serverB	1
4	serverC	0
5	serverC	1
6	serverD	0
7	ServerD	1

- 在db2nodes.cfg文件中定义
- 必要的参数:
  - **数据库分区代号: 独特的数据库分区ID**
  - **主机名称: 机器名称或者IP地址**
  - **逻辑端口: 机器内逻辑分区ID**



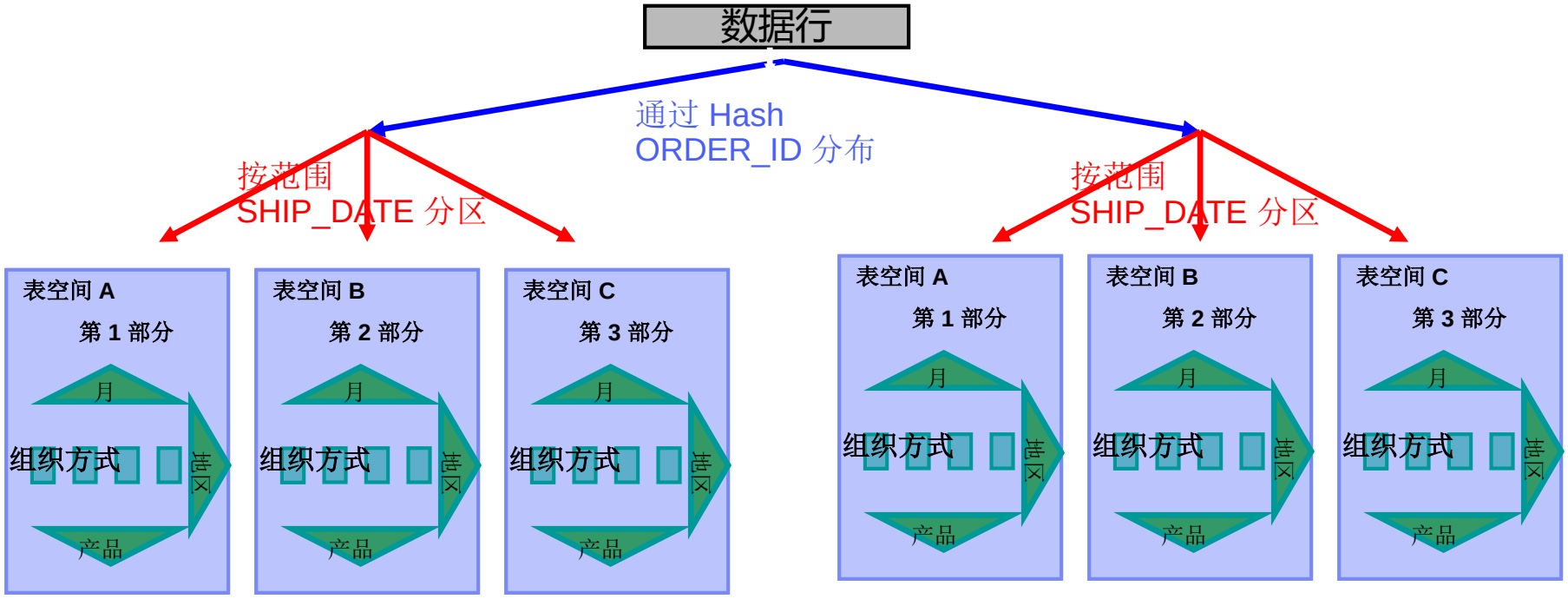
强大的多层次混合分区能力
— 哈希/范围/多维聚簇分区方式



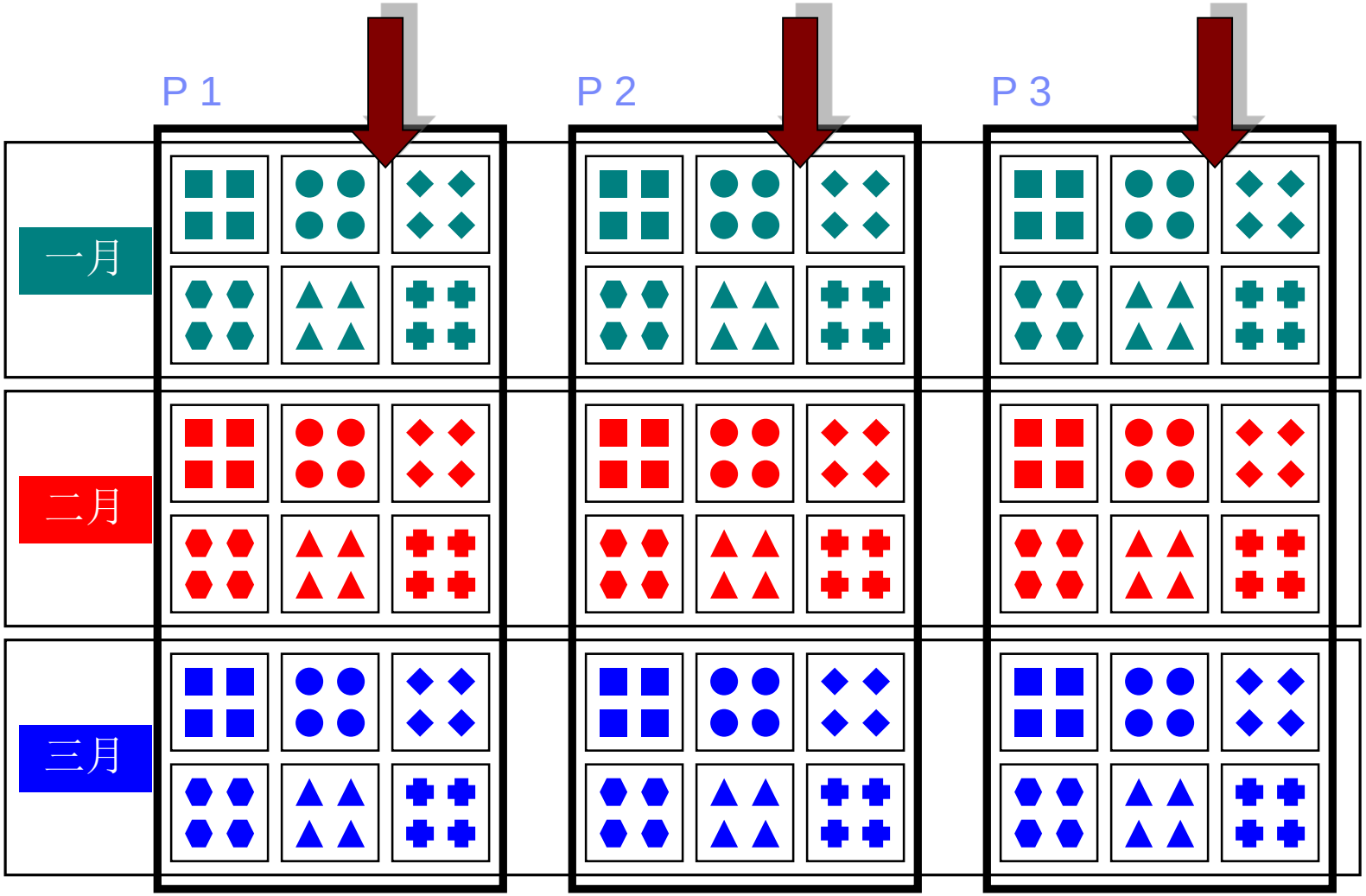
哈希分布，范围分区，多维聚簇

```
CREATE TABLE ORDERS
  (ORDER_ID, SHIP_DATE, REGION, CATEGORY)
   IN TBSP1, TBSP2, TBSP3, TBSP4
   DISTRIBUTE BY (ORDER_ID)
   PARTITION BY (SHIP_DATE)
   (STARTING FROM('01-01-2007')  ENDING ('9-31-2007') EVERY (3 MONTHS))
        MONTH GENERATED ALWAYS AS ((INTEGER(SHIP_DATE)/100)
        ORGANIZE BY (MONTH, REGION, PRODUCT)
```

世界上最丰富的 Slice & Dice 容量



分布式 + 分区 + 按维度组织

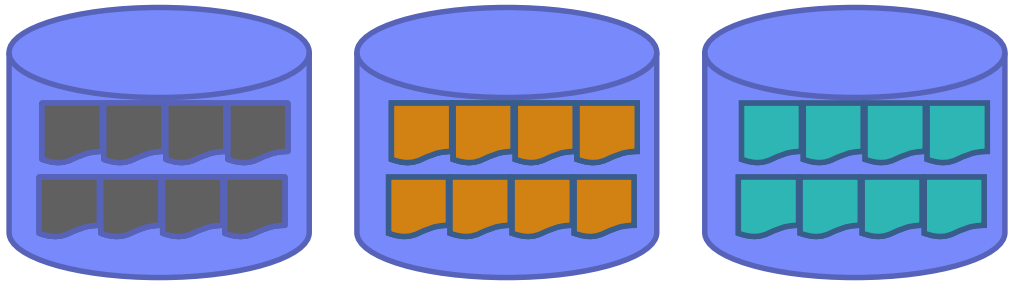


扫描共享

缓冲池

重新读未读取的页面

第2个扫描从第1个扫描当前位置开始扫描

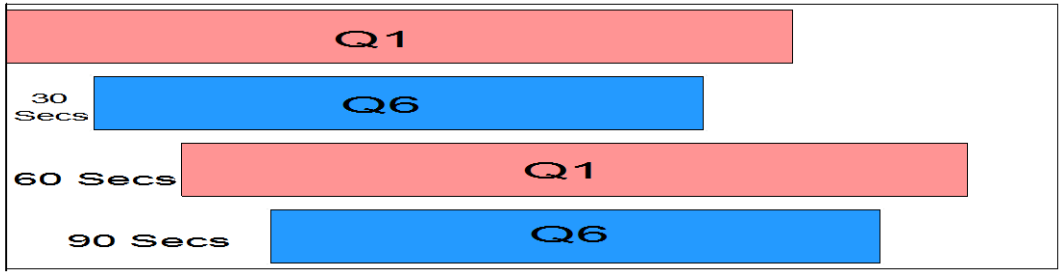


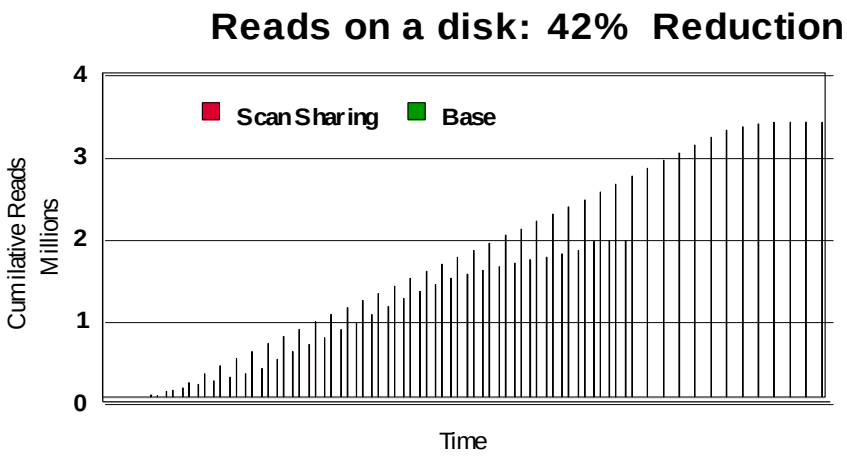
用户1扫描数据

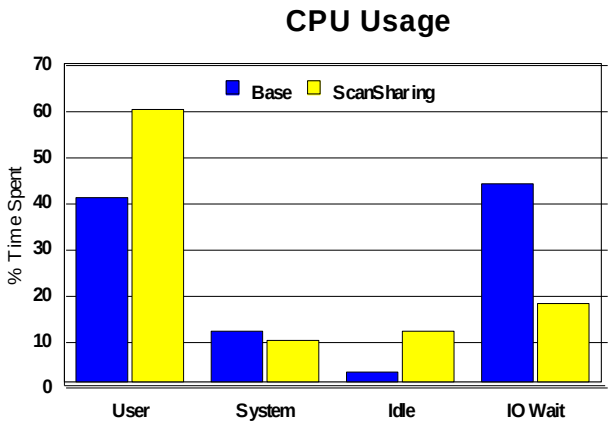
用户2扫描数据

扫描共享性能测试

- **TPCH 查询1 ： CPU 繁忙，Lineitem 表扫描，慢查询**
- **TPCH 查询6 ： IO 繁忙，Lineitem 表扫描，快查询**
  - 测试场景： 查询按如下顺序并行执行







效果 ：端到端性能提升34%!

DB2 Load实现并行数据装载-批量

协调程序 - Coordinator

- 创建并监控其他代理

预分区代理 - Pre-partitioning agent

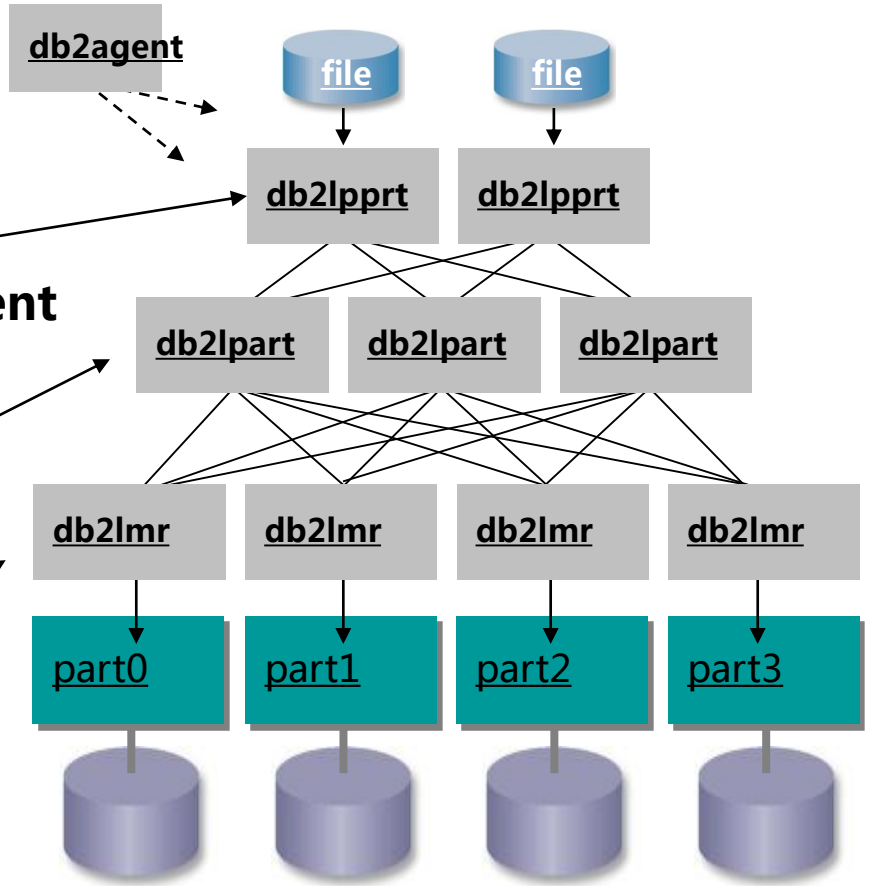
- 每个输入源一个代理
- 运行在协调分区上

分区代理 - Partitioning agent

- 代理数量和运行的分区可配置

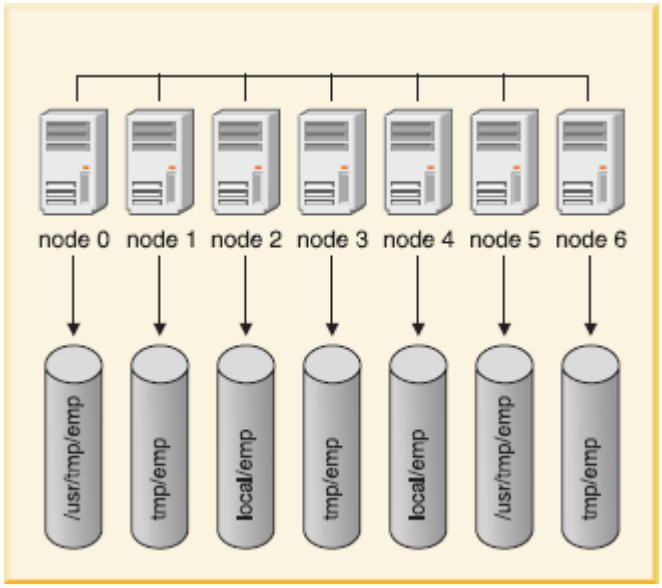
介质读程序 - Media reader

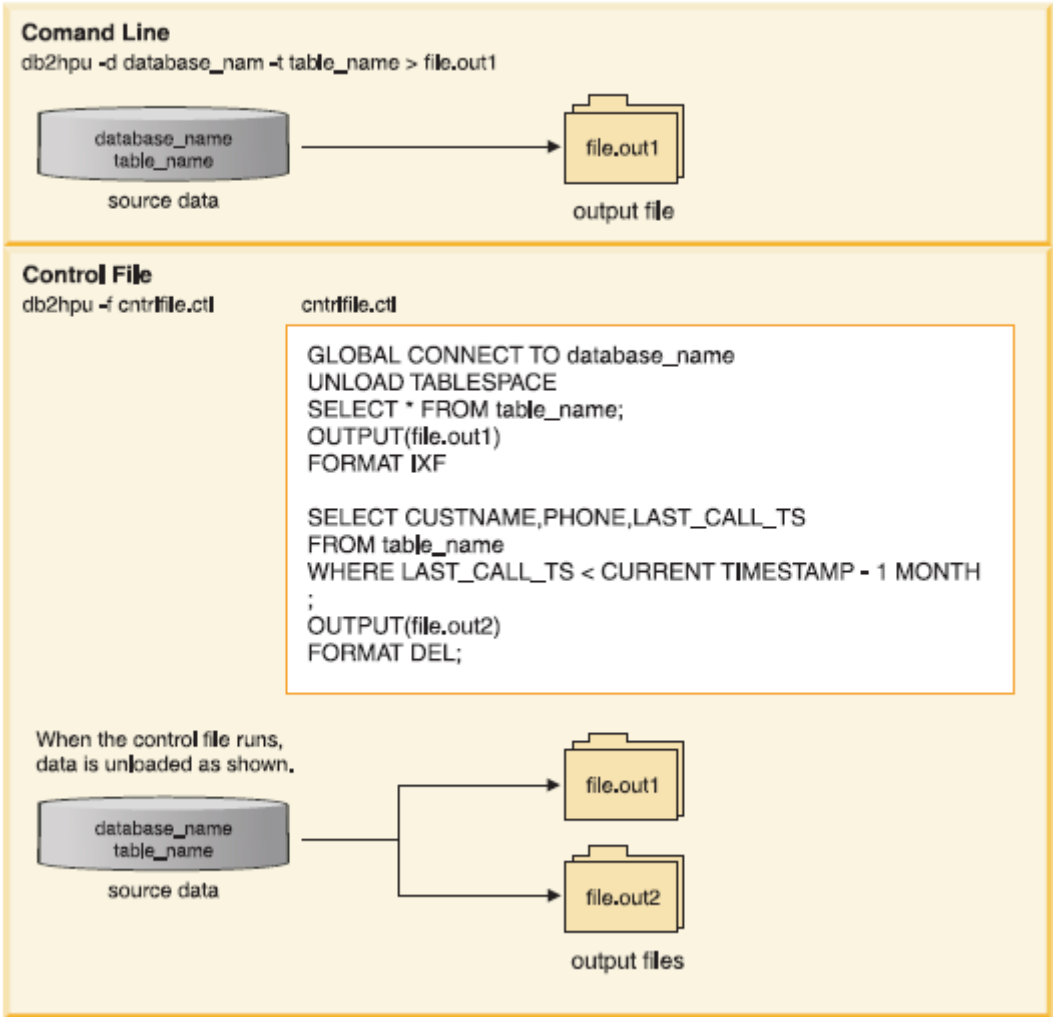
- 每个目标分区一个



DB2-HPU高性能数据卸载

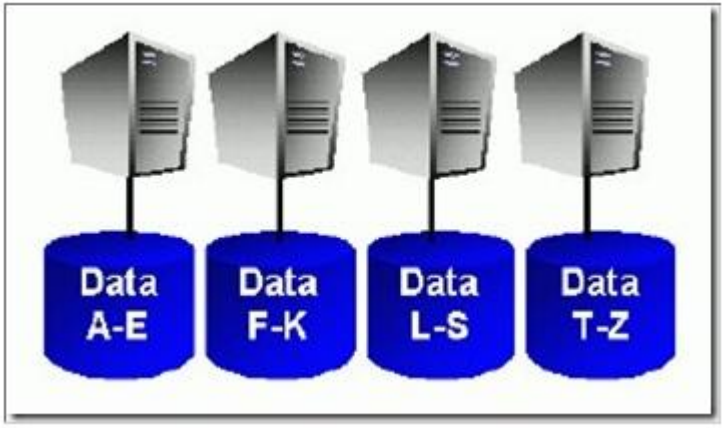
- 所有数据库节点并行导出出数据
- 直接读取数据页面，绕过SQL引擎，效率高
- 可读取数据库备份文件，避免恢复数据库
- 可按照目标数据库分区方式卸载数据，方便数据高速加载到目标库





什么是MPP数据库？

- **数据仓库世界里面MPP（massively parallel processing）定义**
  - 数据分散在多个节点上，每个节点管理各自的数据
  - 将任务并行的分散到多个节点上，每个节点分别处理各自的数据，计算完成后，将各自部分的结果汇总在一起得到最终的结果
  - MPP数据库采用非共享（Share Nothing）架构，而不是共享磁盘（Share Disk）架构
  - MPP数据库不是特指“PC+内置磁盘”，传统的“服务器+外置磁盘”非共享架构数据库也是MPP数据库



非共享架构(Share Nothing, MPP)

共享磁盘架构(Share Disk)

典型的基于SAN存储的DB2 DPF系统

- DB2 DPF 是强大的、具有高扩展性的大型数据仓库和数据集市解决方案
- 典型的硬件配置为Power服务器集群和SAN存储
- SAN 存储可以方便解决服务器失效的问题，通常采用HACMP (仅限于AIX)或者TSA集群软件进行故障监控和失效转移
- DB2 实例用户目录通常为NFS（如ISAS 5600）或GPFS文件系统 （如 ISAS 7700）
- SAN 存储（如 IBM DS8800、V7000等） 特别适合联机交易处理（OLTP ），比较适合数据仓库，因为：

✓ 价格比内置磁盘昂贵，也需要额外的硬件，如HBA卡，SAN交换机等

✓ 安装配置相比内置磁盘复杂的多

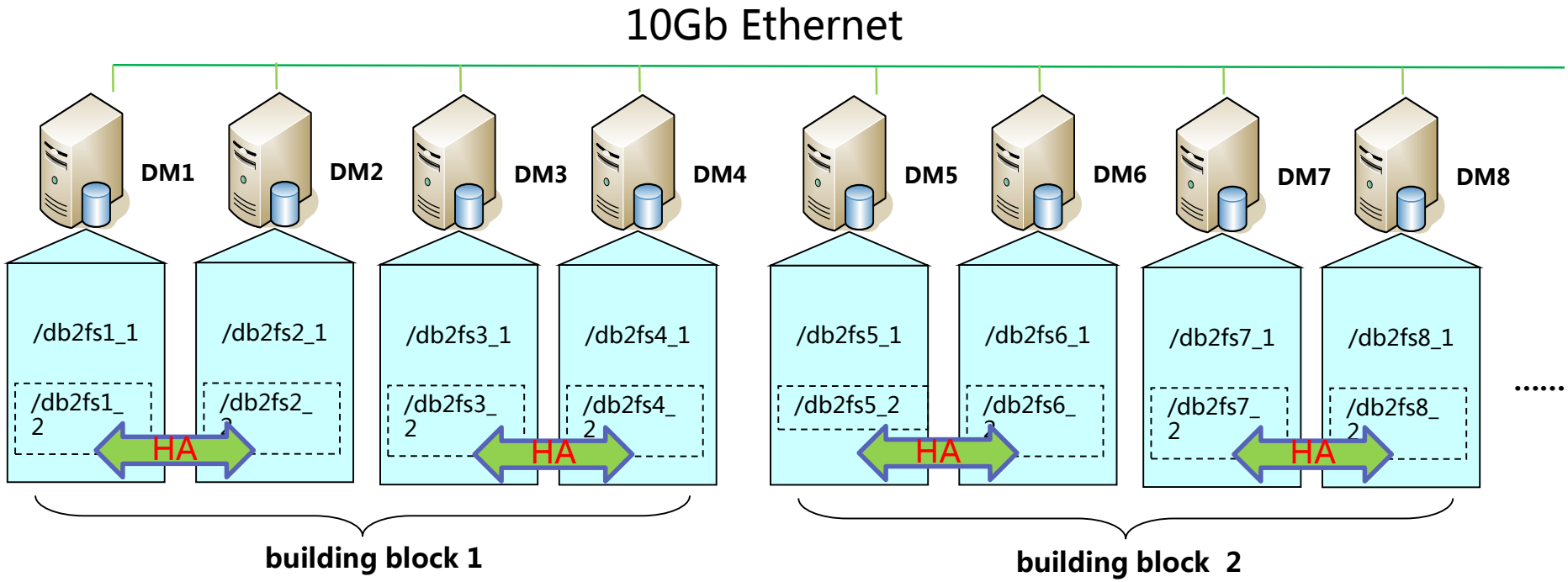
✓ 偏向顺序读的操作（如数据仓库），内置磁盘的性价比更好

✓ SAN存储磁盘由RAID技术保护，但是存储系统整体失效的隐患依然存在

基于内置磁盘的DB2 DPF 系统

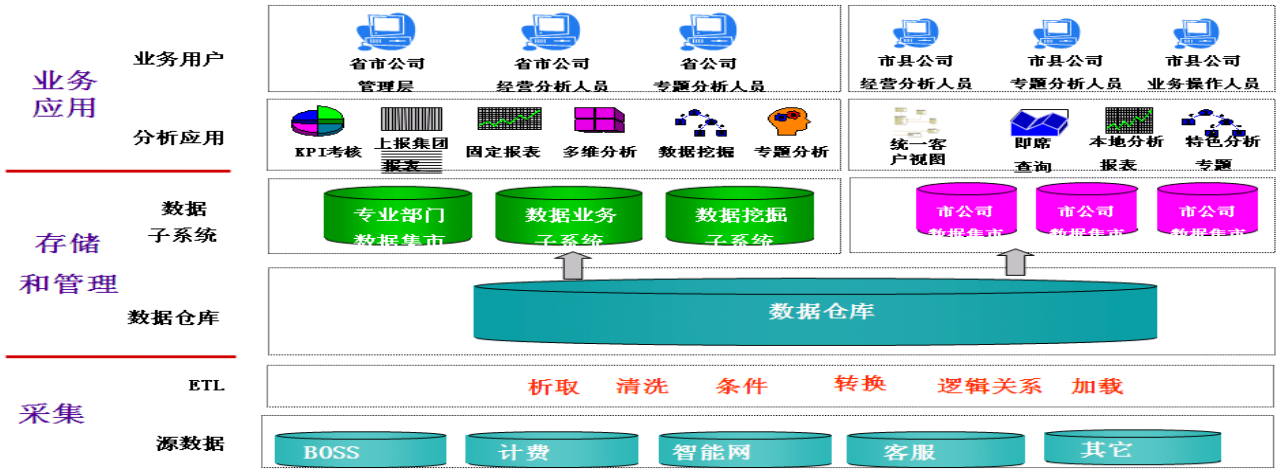
- 越来越多的客户开始考察在PC内置磁盘集群上部署数据仓库或集市的可行性，因为：
  - ✓ 更便宜
  - ✓ 安装配置更简单
  - ✓ 良好的顺序读性能
  - ✓ 云概念
- DB2 DPF是MPP架构数据库，可以很好支持内置磁盘
- DB2 DPF内置磁盘方案需要解决的主要问题是服务器失效，服务器失效会导致其所有内置磁盘不可访问
- GPFS FPO 能够很好解决这个问题
  - ✓ 所有数据库对象（临时表空间除外）复制到另外一台服务器，解决集群单点故障问题
  - ✓ 数据库主目录（如/db2home）单独放在一个GPFS文件系统上，可以与3个数据副本
- 临时表空间考虑
  - ✓ 为获得复杂查询的最佳性能，不建议将临时表空间部署在GPFS FPO上用2个数据副本进行保护
  - ✓ DMS 裸设备性能最优, 基于本地文件系统的DMS或SMS亦可
  - ✓ 基于SSD的临时表空间(不管是 DMS 还是SMS) 是最佳性能选择，但是需要支付额外的大笔费用.

基于PC服务器内置磁盘的DB2 DPF解决方案

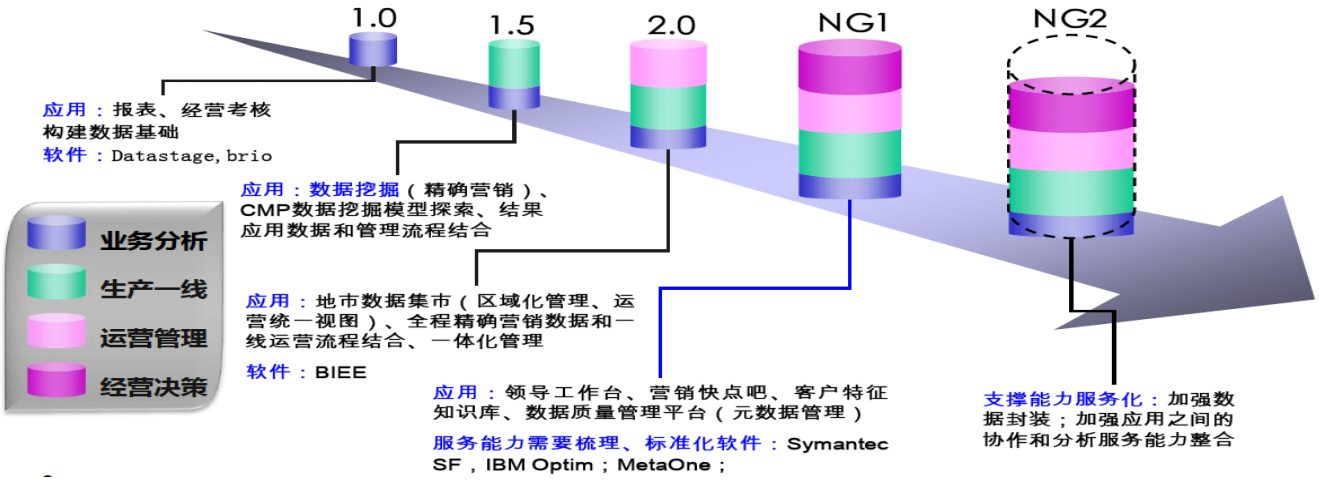


- DB2 MPP数据库够建于X86服务器（或PowerLinux服务器）集群之上
- 采用内置磁盘作为数据库存储，基于GPFS文件系统
- 数据库数据通过GPFS软件进行管理，并通过GPFS实现数据同步镜像，保证数据强一致性
- 集群互联采用万兆以太网络，或Infiniband网络
- **构建单元**（Building Block）是1个独立的GPFS集群，一个大集群采用多个构建单元、或多个小集群的方式进行“分而治之”的管理；每4个数据节点为一个“构建单元”，它是水平扩展的最小单位；每个构建单元也可以是3个数据节点
- **高可用：**简单起见，采用2台服务器互备的高可用模式；3节点构建单元情况下，采用2接1的模式
- **数据模块**（Data Module）：是运行 DB2 DPF的数据库物理服务器，它运行1个或多个DB2数据库分区，数据库分区数量取决于服务器CPU内存及内置磁盘的个数等
- **GPFS 文件系统（/db2fsx_x）**：跨1个构建单元的4个数据模块，2个数据拷贝，1个构建单元中4个数据模块两两之间实现数据镜像
- **DB2HOME：**第一个构建单元4个数据模块构成的GPFS集群上创建数据库实例用户的主目录(/db2home)，其他构建单元采用GPFS 客户端方式远程挂载/db2home

中国最大的数据仓库系统（广东移动DPF） – 800TB！



一线营销、管理创优、决策支持



主要挑战

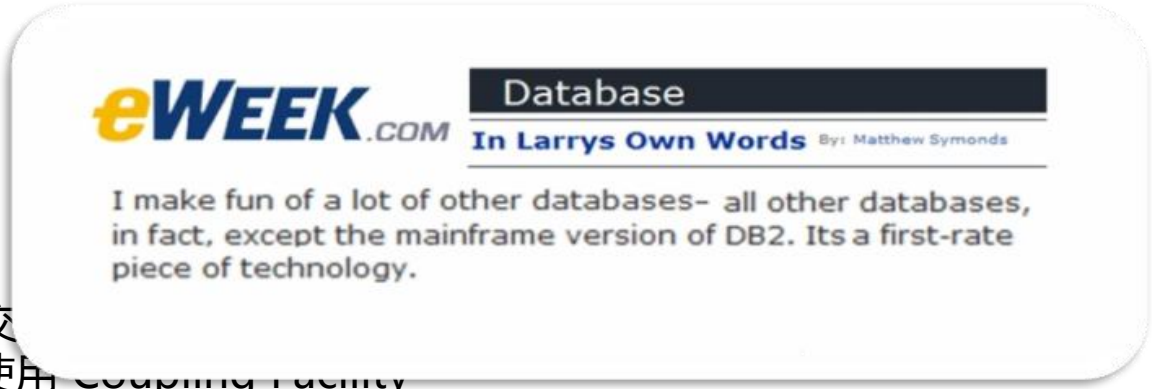
- 海量数据存储
- 时间窗口(对一线的支撑)
- 业务连续性(运维问题、备份等)
- 标准化

议程

- DB2架构及技术特点
- DB2集群技术
  - DB2 DPF集群
  - DB2 pureScale集群
- DB2列式存储及内存计算
  - DB2 BLU
- DB2客户案例

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
  - z/OS 上的整个环境都可用使用
    - CICS、MQ、IMS、Workload Management 等

DB2 PureScale 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

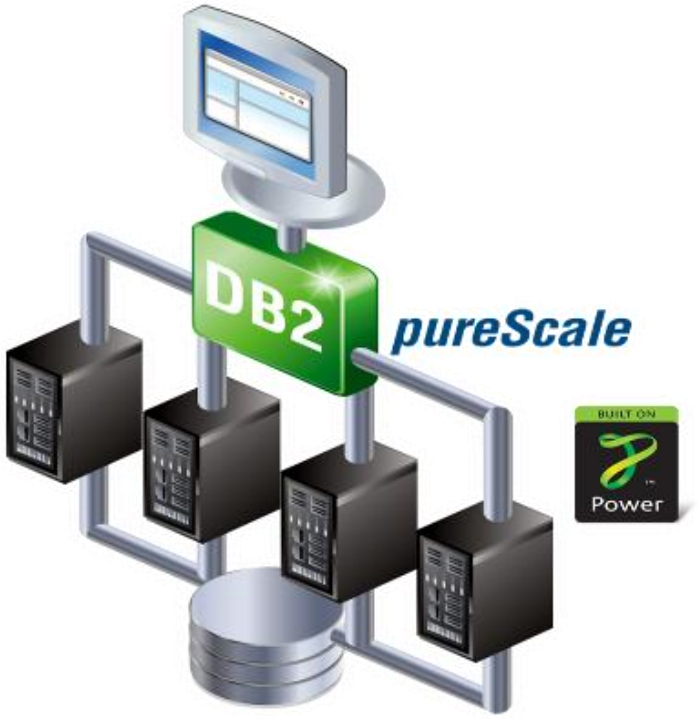
DB2 PureScale 可以为任何事务性工作负荷提供近乎无限的产能。扩展系统只需要连接到新节点并发出两个简单的命令

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

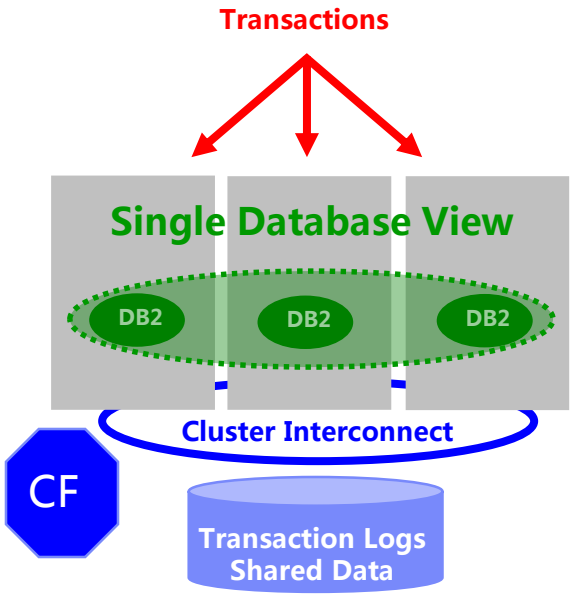
DB2 PureScale 是专为需要持续可用性进行系统设计。系统将瞬间从故障中恢复，同时仍然能保证事务处理不中断



DB2 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

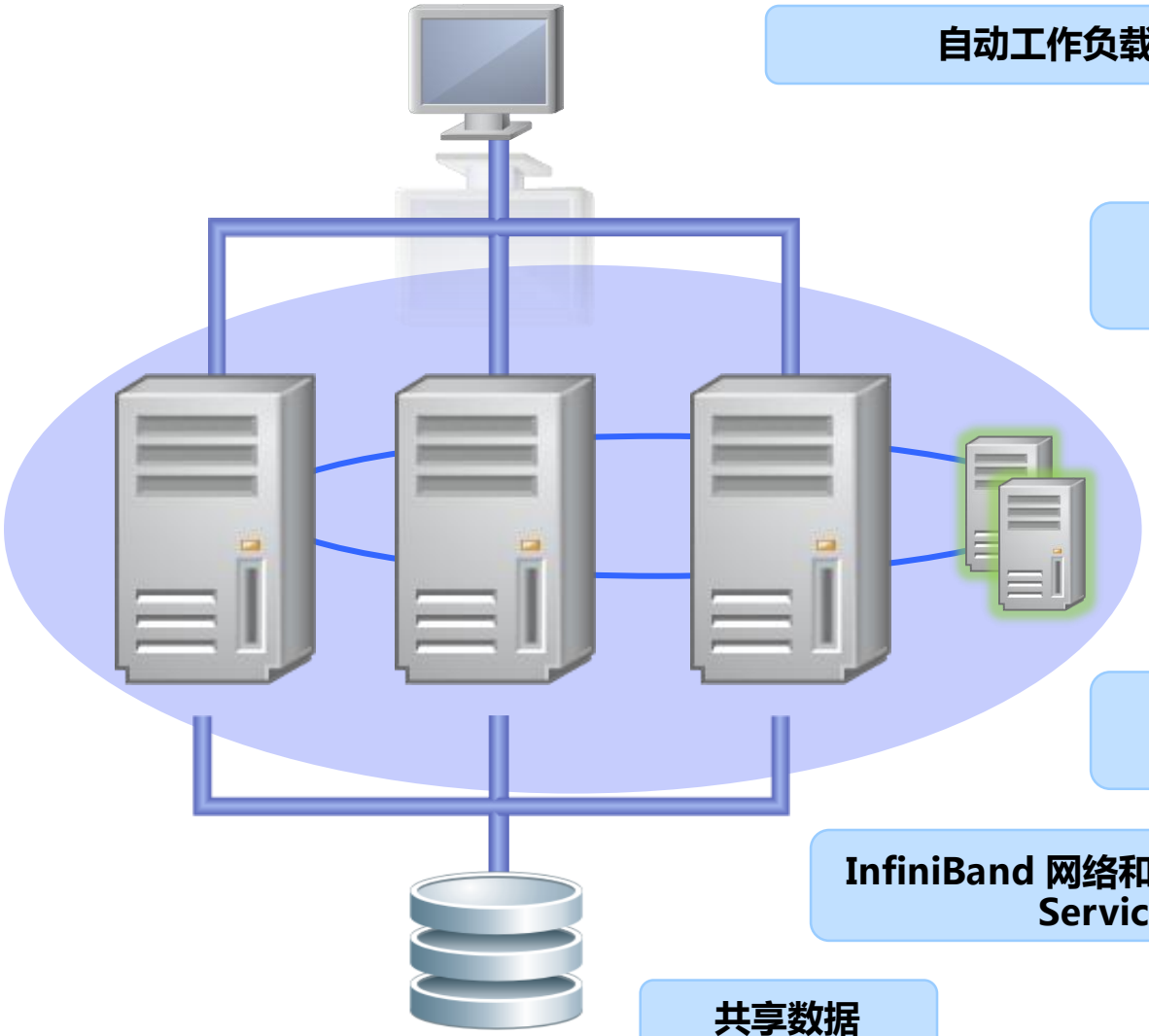
低管理成本



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 的体系架构



自动工作负载均衡

运行于 Power 服务器上的 DB2 成员集群

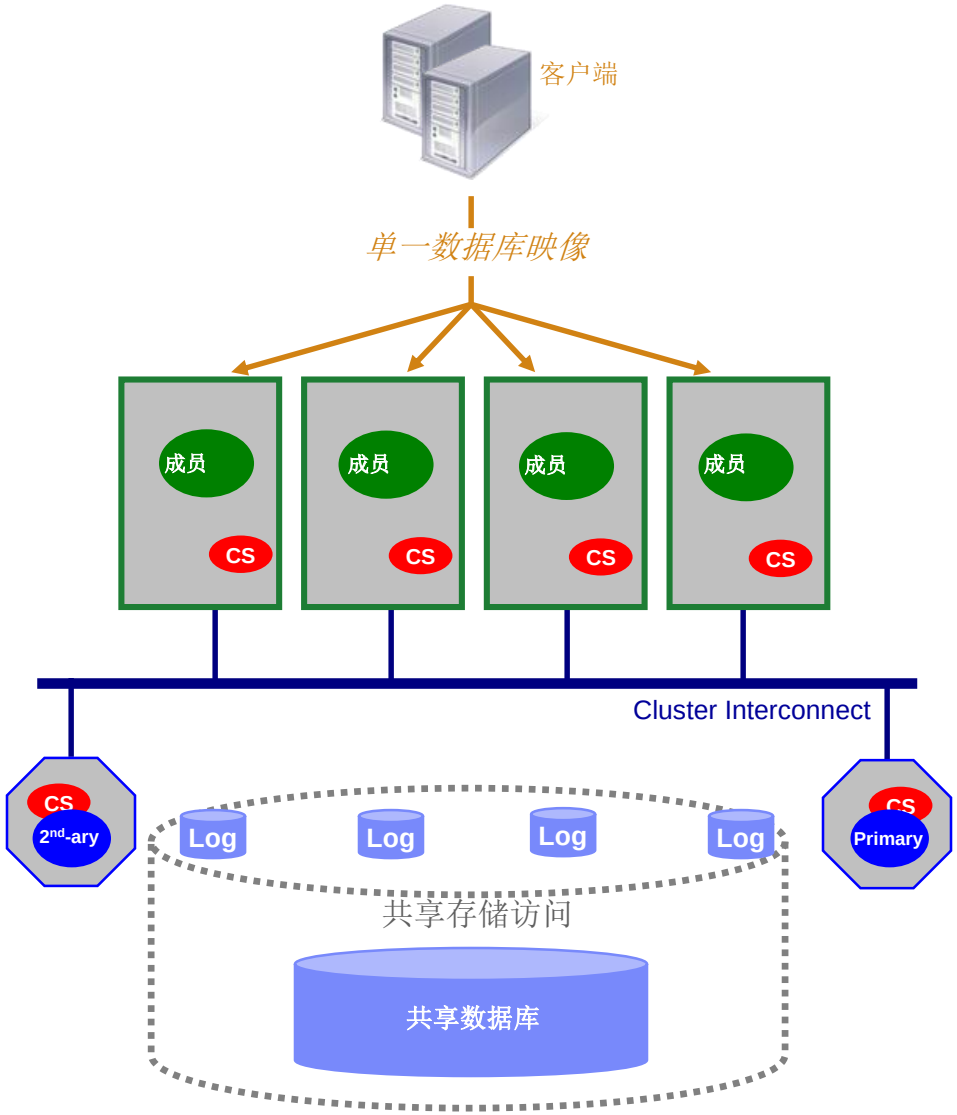
利用 z/OS 提供的全局锁定和内存管理器技术

整合的 Tivoli System Automation

InfiniBand 网络和 DB2 Cluster Services

共享数据

DB2 pureScale系统架构



客户端随处连接,...
... 看到同一个数据库

- 客户端连接到任何一个成员
- 自动负载管理和客户端重新路由功能可以改变客户端所连接的成员服务器

DB2引擎在多台服务器上运行

- 相互协作提供来自任何成员服务器对数据库的一致访问

集成的集群服务

- 错误检测，自动化恢复，集群文件系统
- 使用STG和Tivoli的产品

低延迟、高速互联

- 采用基于InfiniBand交换技术的远程直接内存访问协议(RDMA)并特别优化，提供最佳互联方式

PowerHA pureScale技术

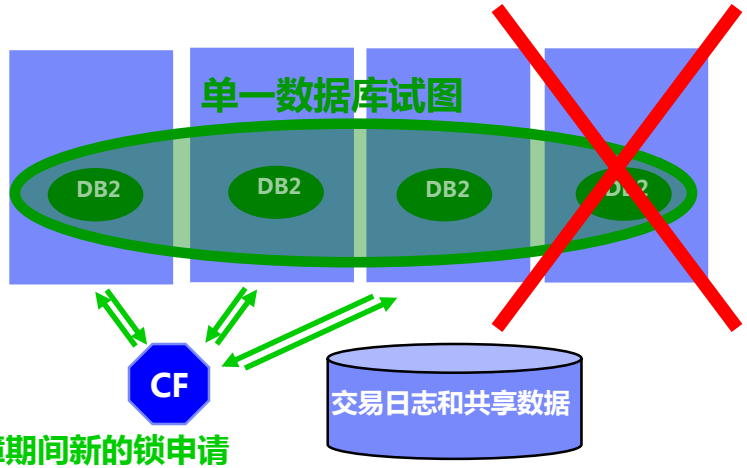
- 高效全局锁和内存管理
- 提供同步双工方式写备份CF提高可用性

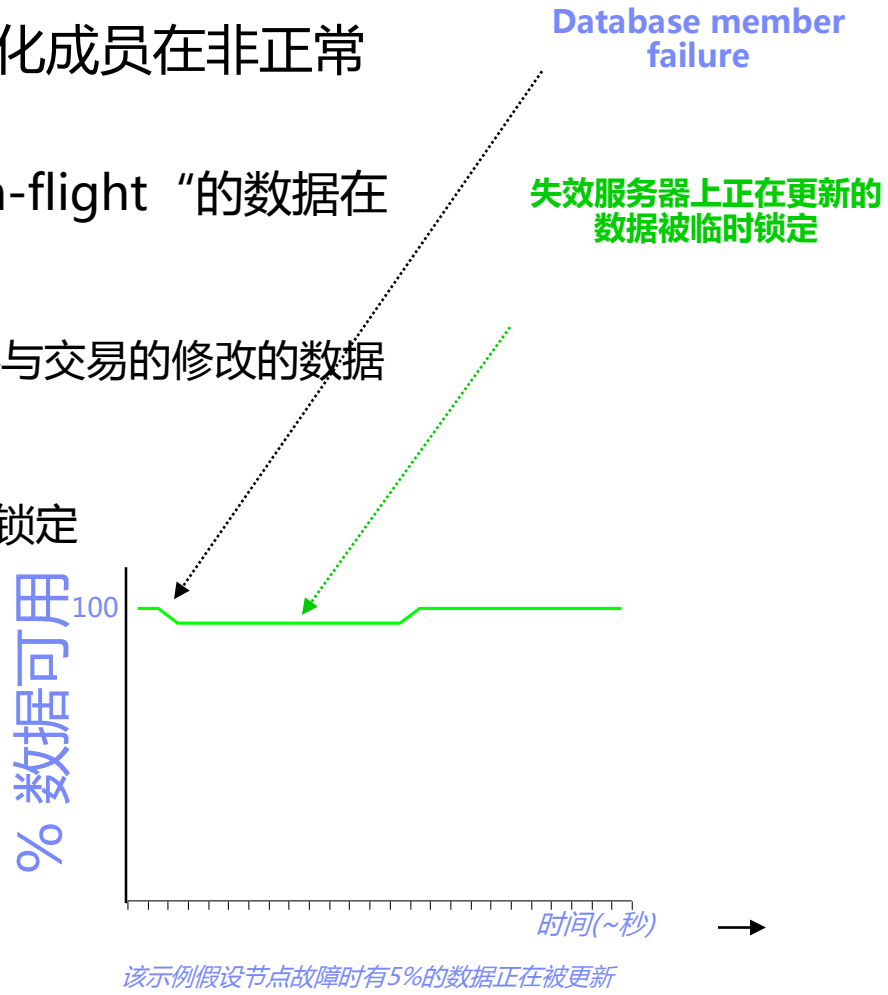
数据共享架构

- 共享数据库访问
- 成员服务器将日志写到共享磁盘
- 日志在故障恢复期间被其他服务器可见

DB2 PureScale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

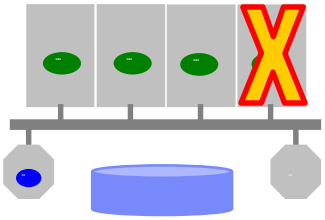
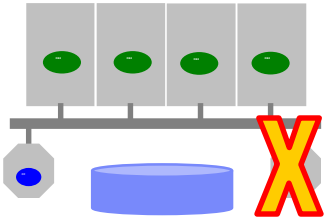
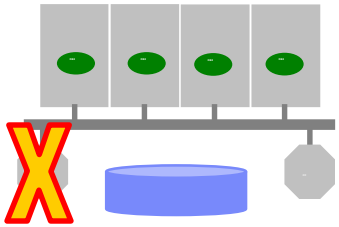




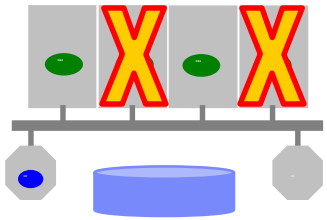
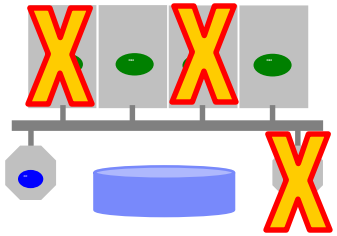
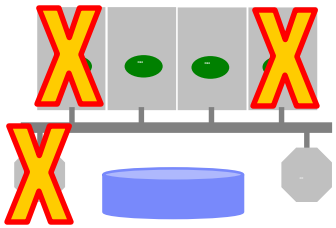
节点故障类型和恢复行动

故障类型	恢复行动	典型预期恢复时间	可用性影响
DB2 成员故障	**Member Restart** **(aka member crash recovery)**	~ 15 秒	故障成员上更新的数据被锁定 (其他数据完全可访问) 连接自动路由到其他成员，未决交易可重试
主 CF 故障	**Notify members to fail-over to secondary CF**	~ 5 秒	应用透明 CF响应时间瞬间受影响
从 CF 故障	**Notify members to stop duplexing**	~ 3 秒	应用透明 CF响应时间瞬间受影响

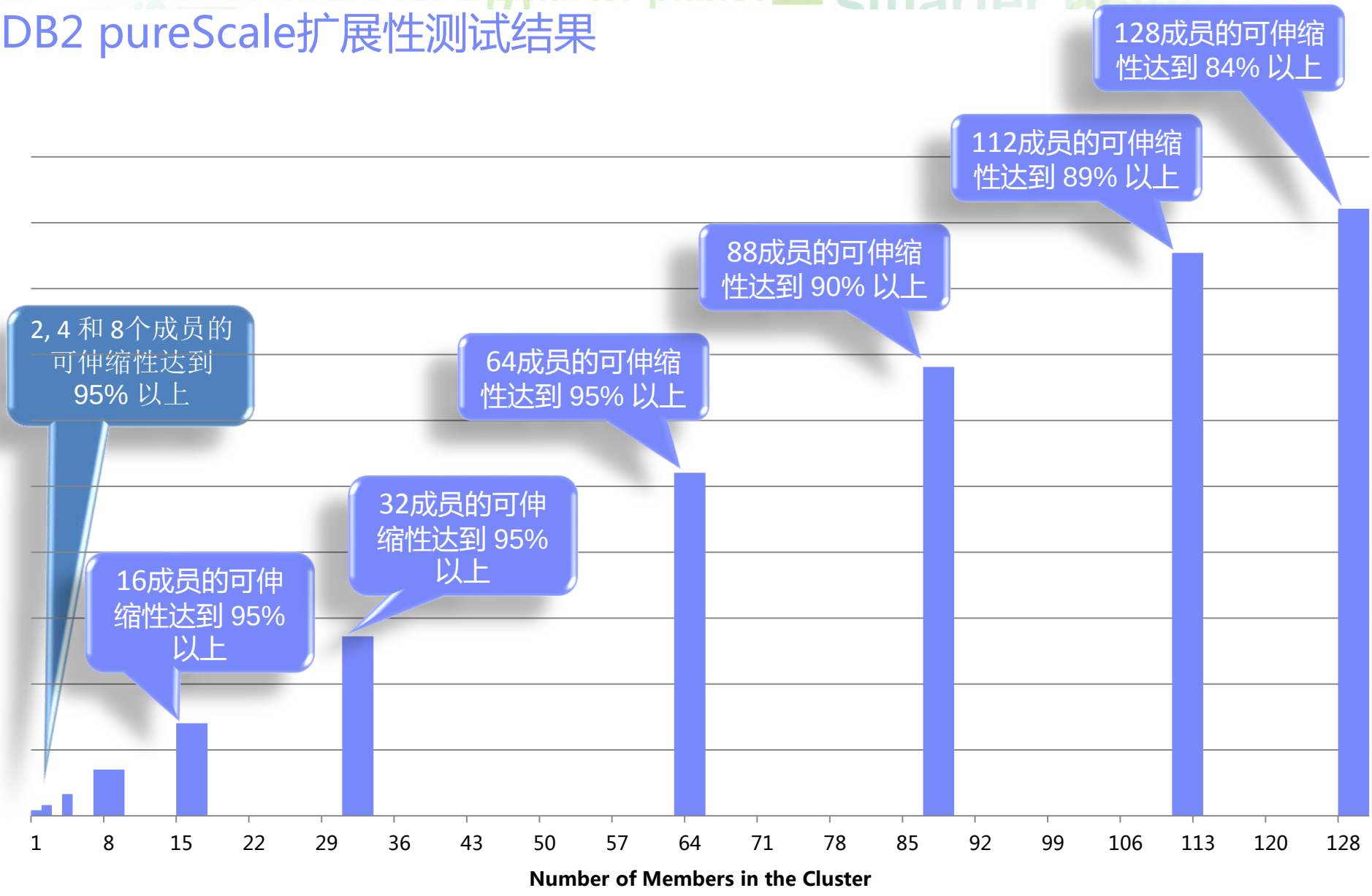
单点故障

故障类型		其他成员状态?	自动透明？	备注
成员			 到故障成员的连接自动透明路由到别的成员	仅仅故障成员上更新的数据临时不可用
主CF				CF服务瞬时不可用 对成员透明(故障时对CF的服务请求能够正常完成，仅需额外多几秒时间)
从CF				CF服务瞬时不可用 对成员透明(故障时对CF的服务请求能够正常完成，仅需额外多几秒时间)

多点故障

故障类型	其他成员状态?	自动透明？	备注
	✓	✓ 到故障成员的连接自动透明路由到别的成员	仅仅故障成员上更新的数据临时不可用 各成员并行恢复
	✓	✓ 到故障成员的连接自动透明路由到别的成员	仅仅故障成员上更新的数据临时不可用 CF服务瞬时不可用
	✓	✓ 到故障成员的连接自动透明路由到别的成员	仅仅故障成员上更新的数据临时不可用 CF服务瞬时不可用

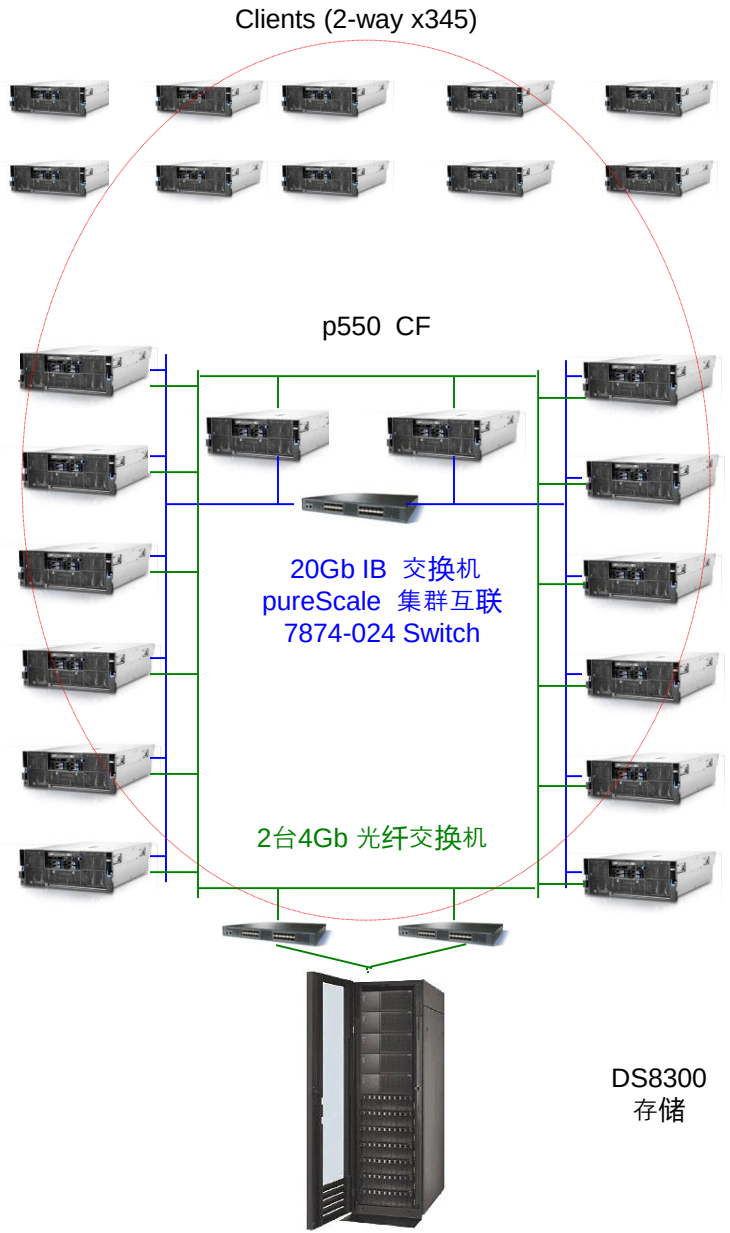
DB2 pureScale扩展性测试结果



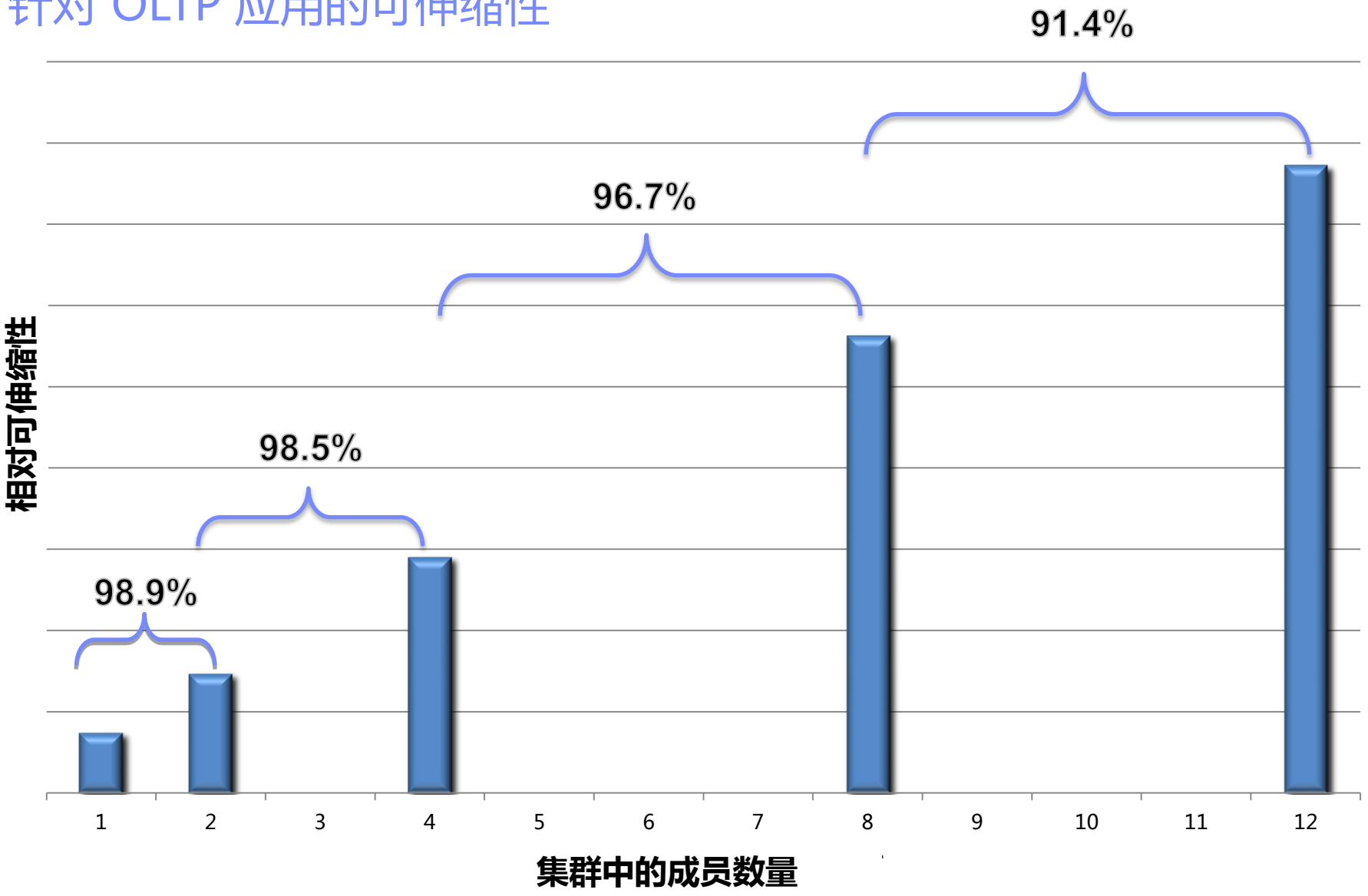
Validation testing includes capabilities to be available in future releases.

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/写比例
- **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 2个CF节点，12个成员节点™
- **可伸缩性仍然保持在 90% 以上**

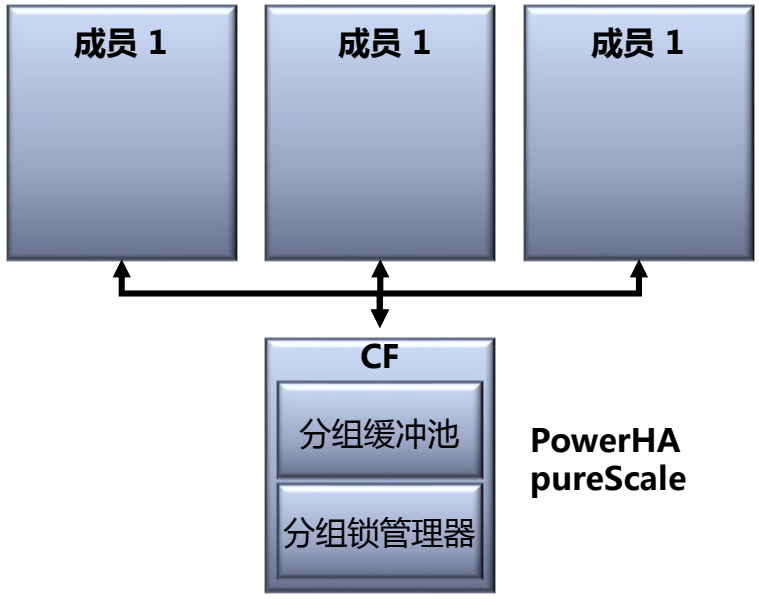


针对 OLTP 应用的可伸缩性



DB2 pureScale高伸缩性和高可用性的关键

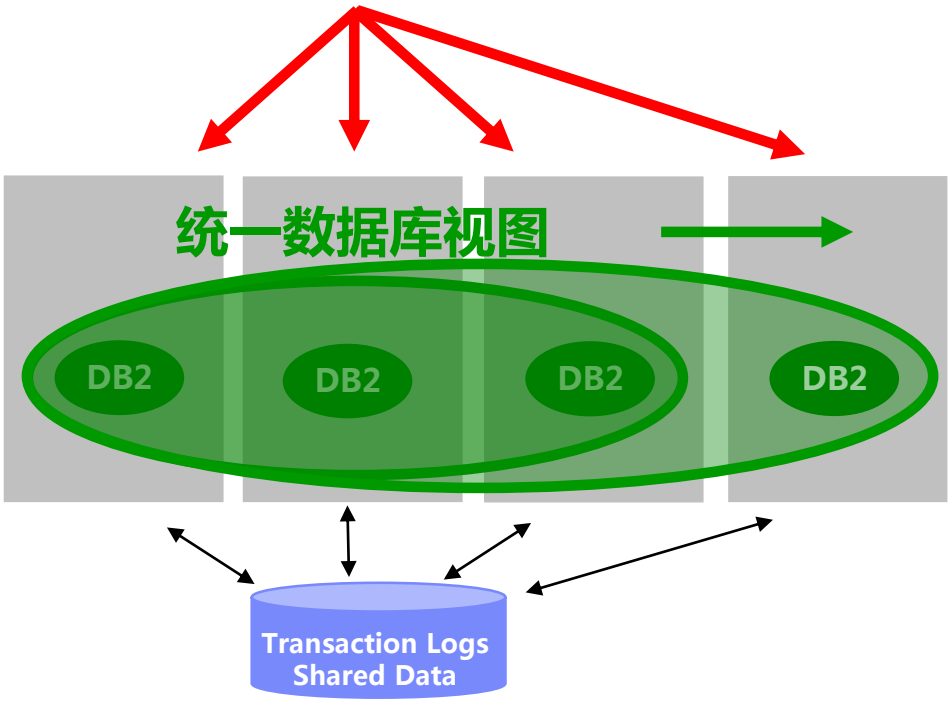
- **集中锁定和缓存**
  - 随着集群的不断增长，DB2 会始终在 CF 维护锁定信息和共享页面
  - 针对超高速访问而优化
    - DB2 pureScale 使用 Remote Direct Memory Access (RDMA) 与 PowerHA pureScale 服务器通信
    - 没有 IP 套接字调用、没有中断、没有上下文切换
- **结果**
  - 为大量服务器提供接近线性的可伸缩性
  - 持续感知各成员当时的工作状态
    - 如果其中一个成员出现故障，不会造成
      - 其他成员 I/O 阻塞
      - 以内存速度恢复受影响页面



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



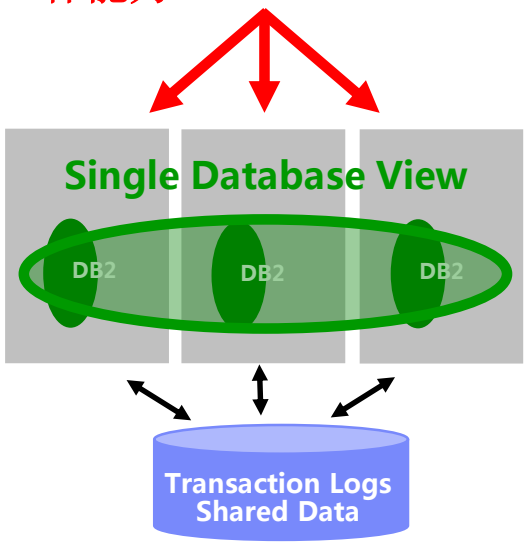
快速部署新成员

- ✓ 不需要数据重新分布

DB2 pureScale易维护和升级

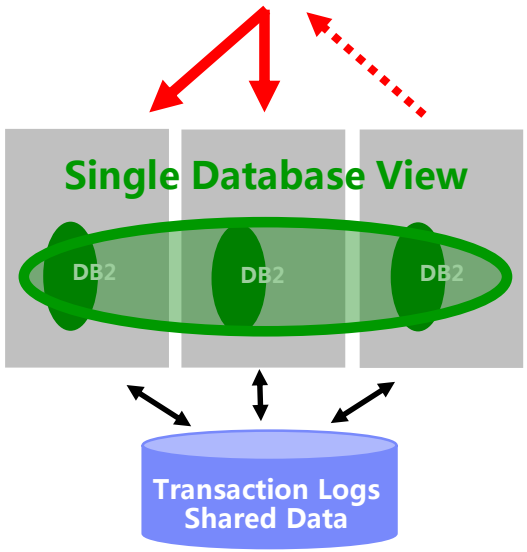
1)运行系统

- 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



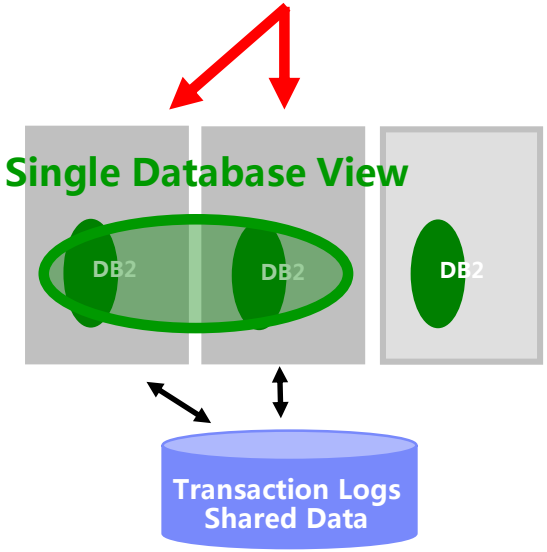
2) 排干（Drain）目标节点

- 停止新的路由
- 允许已有交易完成



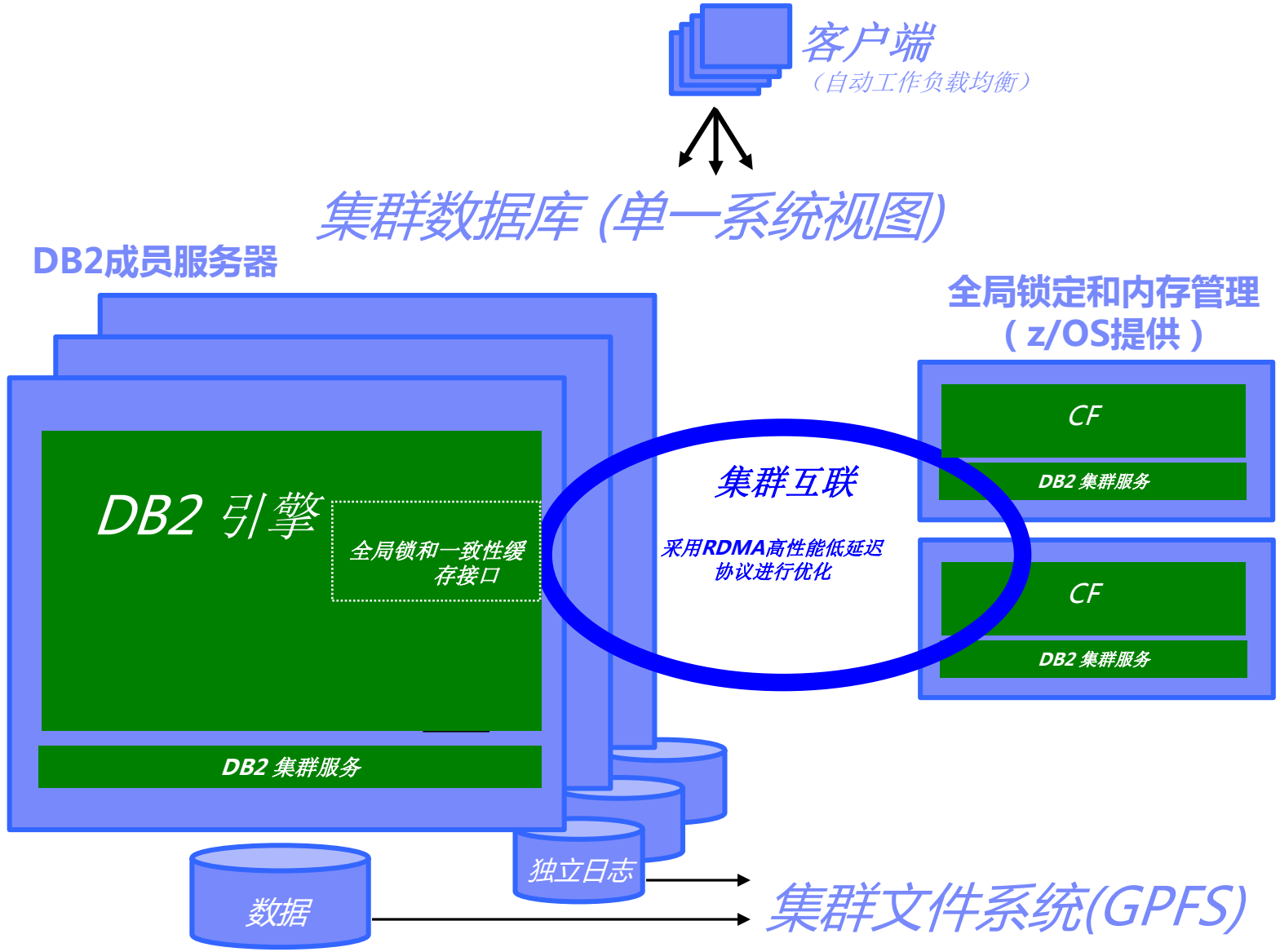
3) 执行维护工作

- 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；
支持数据库大补丁(非小的PatchSet)滚动升级!

DB2 PureScale 架构概览



什么是“成员”（member）？

- **一个通过集群文件系统访问共享数据的数据库引擎**
  - 一个 db2sysc 进程的实例
- **拥有自己的内存，缓冲池，交易日志和锁机制，且自行编译和执行**
- **可以是物理机器或逻辑节点**
  - 可以是:
    - 一个成员每台主机 – 生产推荐
    - 多台逻辑成员运行在一台主机上 – 开发和 QA 推荐
  - 和 DPF 中的“node”和数据库分区类似

什么是 CF?

- **CF 是 DB2 pureScale Feature 的一个集成组件**

- **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
- **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，包括 control blocks, log sequence numbers (LSN) 等
- **CF 应该配置一对以避免单点故障**
  - 只配置一个的做法是支持但不推荐的，只适用于测试环境

注: GBP 和 GLM 并不能替代成员在本地维持本地缓冲池和锁管理的需求

双 CF 保证连续可用性

- **如果主 CF 失败，辅 CF 可以接管避免造成整个系统宕机**
  - 无单点故障
- **辅 CF 由 DB2 保持可用状态**
  - 更新会同时发给主辅 CF，信息请求只送给主CF
  - 辅 CF 也复制 GBP, GLM, SCA，但不是全部信息都复制
- **当启动实例时**
  - 指定的 CF 会被启动为主 CF
    - 辅 CF 保持 peer 状态
    - 且由复制保持 peer 状态

DB2 pureScale集群互联

- **需求**
  - RDMA capable fabric
    - 直接修改内存不消耗 CPU 资源
    - 为 zSeries Sysplex 发明
  - 高速，大容量成员间交换，且使用主辅 CF
- **解决方案**
  - 使用 InfiniBand (IB) 和 uDAPL (User Direct Access Programming Library) 解决性能问题
    - InfiniBand 支持 RDMA 且支持高速，大容量网络交换
    - uDAPL 降低 AIX 的 CPU Kernel 时间

DB2 pureScale 集群文件系统

- **需求**
  - 共享磁盘和共享文件系统
  - 失败成员上文件系统的 fencing

- **解决方案**
  - General Parallel File System – GPFS
    - 由 DB2 提供许可证、安装和配置
    - 同时，客户自己预先配置的 GPFS 文件系统也可以接受

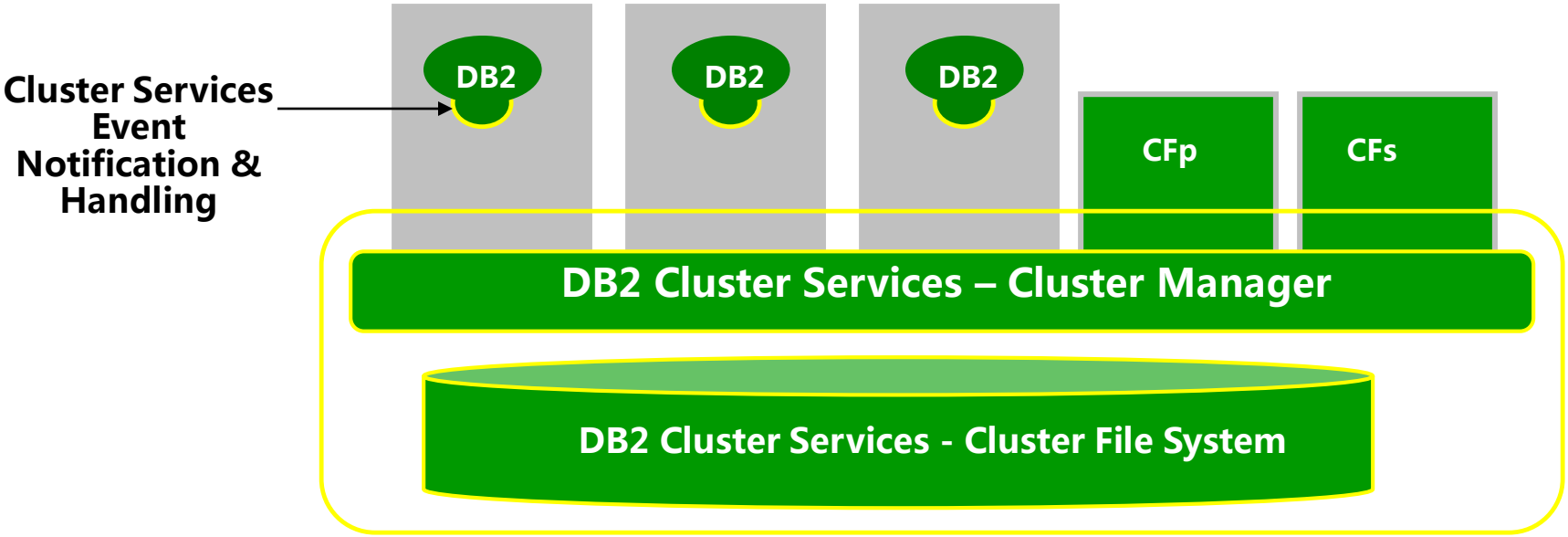
  - “SCSI-3 永久保留”(Persistent Reservations)实现快速 I/O Fencing

DB2 pureScale 集群服务

- **协奏 (Orchestrate)**
  - 非计划事件提示保证无缝恢复和系统可用性
    - DB2 成员检测到失败将自动地触发在本机或其它成员上的成员恢复
    - 主 PowerHA pureScale 检测到失败将会把辅 PowerHA pureScale 初始化为主 PowerHA pureScale
    - 主机检测失败将初始化 I/O Fencing 并重启主机
  - 计划事件
    - ”秘密“ （Stealth）维护
    - 增减成员和 PowerHA pureScale
- **整合**
  - TSA (Tivoli System Automation) 提供集群管理
    - 监控、失败检测，重启和恢复成员和 PowerHA pureScale
  - GPFS (General Parallel File System) 提供集群文件系统
  - 作为 Feature，TSA 和 GPFS 由 DB2 pureScale 提供介质，安装和配置

DB2 pureScale 集群服务

- 集成的 DB2 组件
- 在 DB2 安装过程中安装
- 由 DB2 补丁升级维护
- 为数据库管理员而不是系统管理员设计 (SQL, administrative views, db2pd etc.)



DB2 pureScale 集群服务

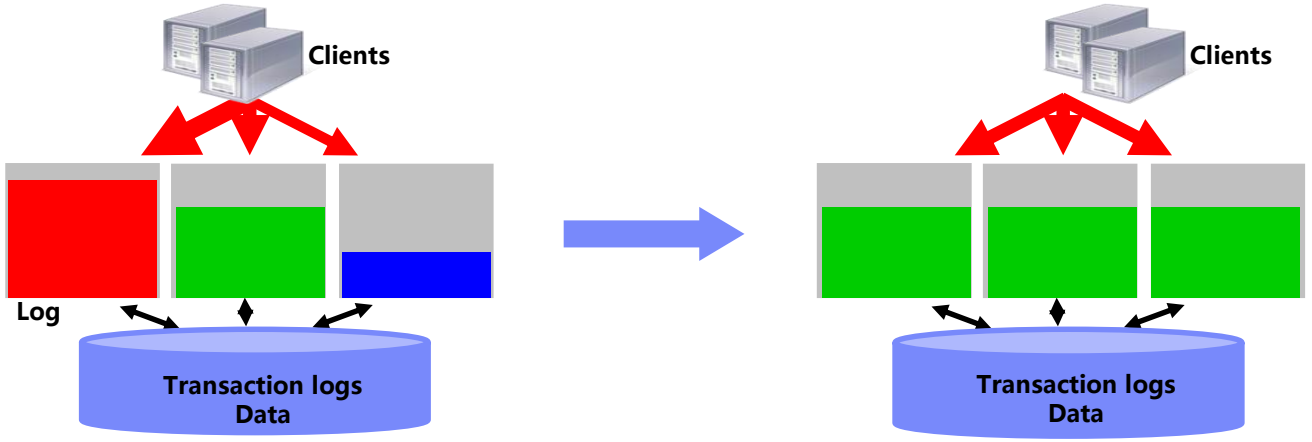
- 强壮的心跳和失败检查算法
  - 防止因网络阻塞和高 CPU 使用率造成的错误判断
- SCSI-3 Persistent Reserve I/O Fencing 防止脑裂（split-brain）
  - 当1个或多个错误的主机从网络分裂情况下对数据提供保护
  - 比其它业界技术更健壮 (Self Initiated Reboot based algorithms、STONITH - shoot the other node in the head).
  - 允许分裂的主机在网络恢复正常的情况下，无需重启重新加入集群
- 在3秒中，主机完成失败检查和 I/O Fencing

DB2 pureScale 集群服务和可用性

- **自动软硬件失败恢复，无需配置，命令和内部命令**
- **检查所有组件的健康并作出反应 (DB2, PowerHA pureScale (CFs), 集群文件系统, 网卡等)**
  - 例如， IB HCA 在 DB2 成员上失败将导致 DB2 成员的失败，并在其它主机上进行恢复 (Restart Light)
- **在 DB2 启动时检查资源模型（resource model）和依赖，避免 DB2 在资源模型不一致的情况下启动**
  - 例如，主机被停止做维护，网卡被移走或替代

工作负载平衡和自动路由

- **运行时负载信息用于负载平衡 (和 Z Sysplex 一样)**
  - 每个成员记载自己的工作负载
  - 回复给访问的客户端
  - 对下一个连接或者可选下一个交易进行路由
  - 路由对应用程序是透明的
- **Failover： 失败成员上的工作负载平均地分布到其它存活的成员上**
  - 一旦失败的成员恢复，恢复的成员重新承担负载



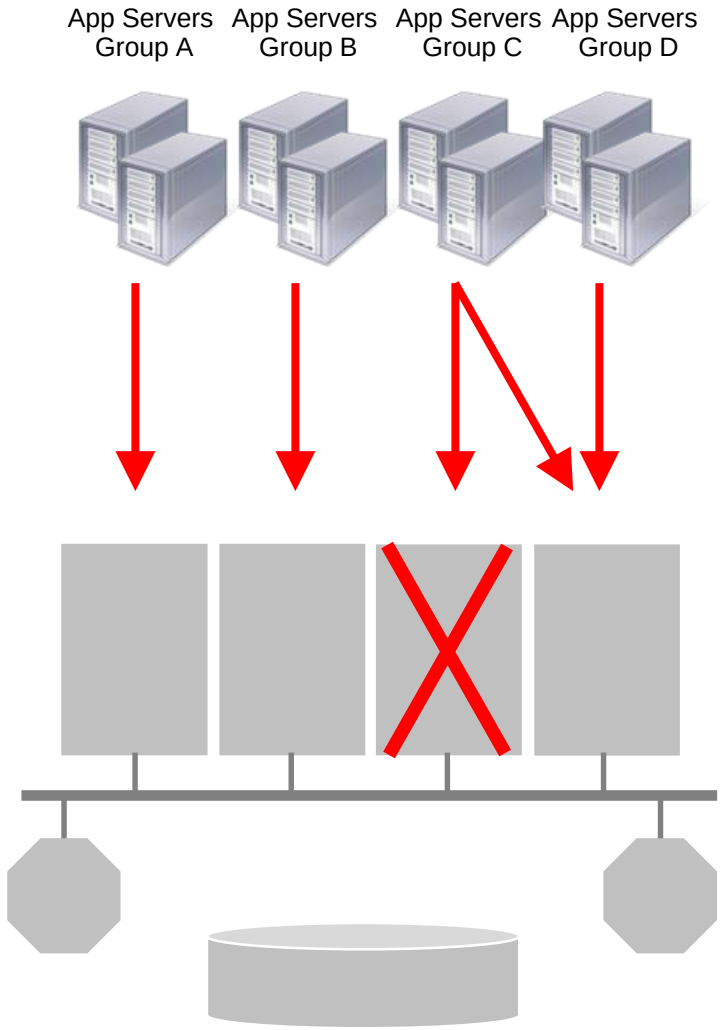
客户端路由

- **工作负载平衡**
  - 服务器驱动
  - 基于成员的工作负载
  - 路由发生在：
    - Connection Level 或 Transaction Boundary
  - 平衡机制
    - 连接路由到工作负载最低的成员
- **也可以采用 Affinity 方式驱动**
  - 客户端驱动
  - 不考虑工作负载
  - 2种方式
    - 指定的
    - Round robin

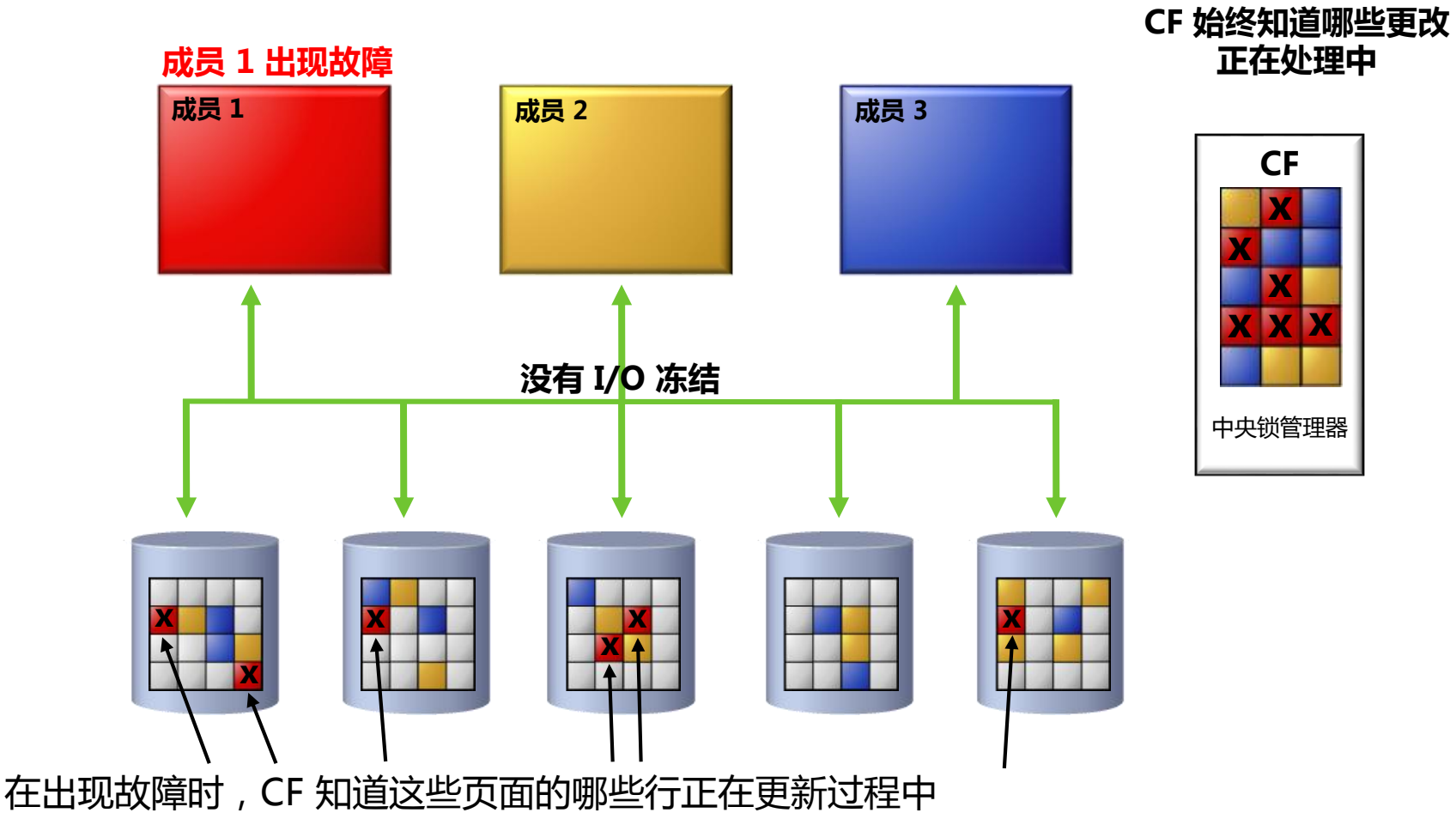
可选的Affinity负载路由方式

db2dsdriver.cfg 文件:

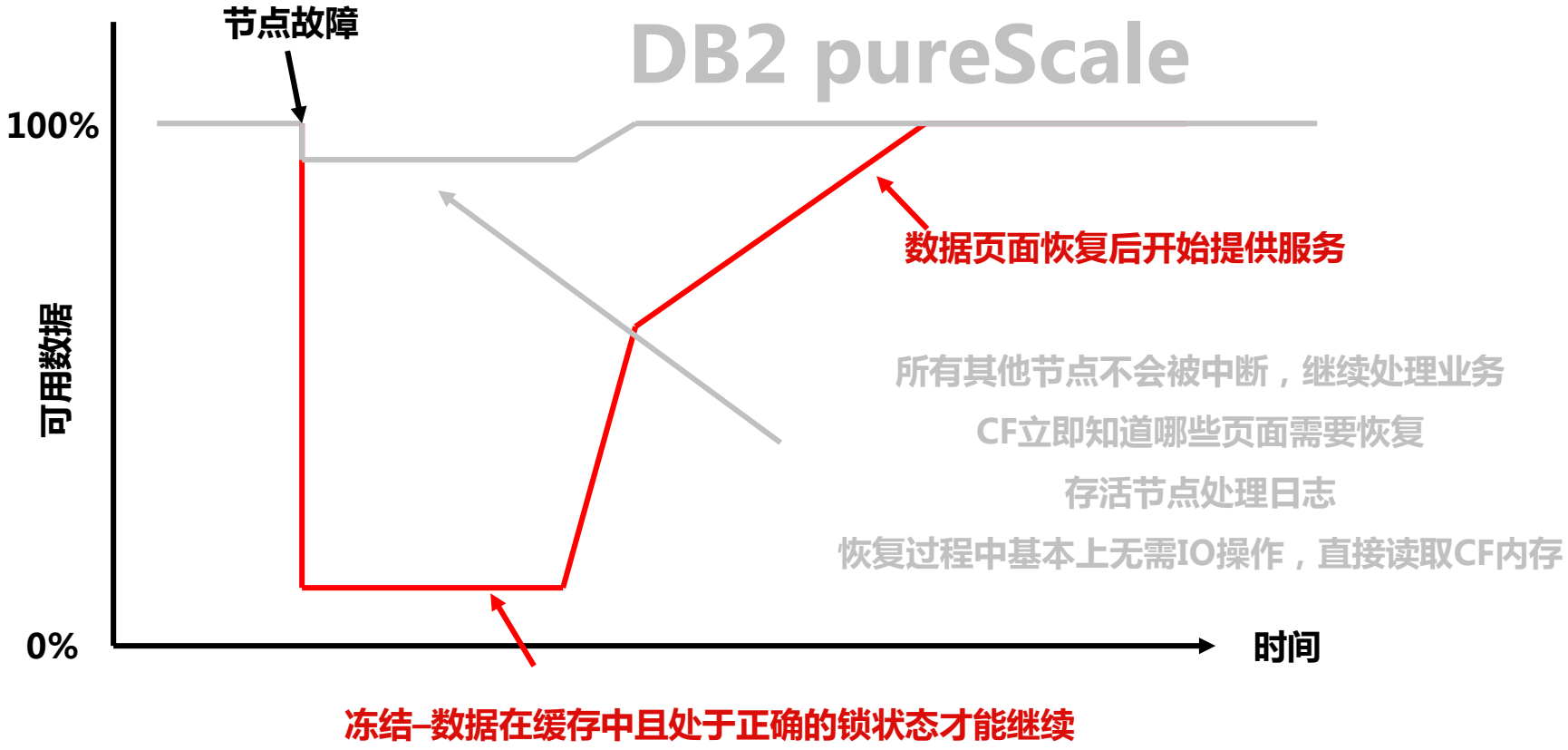
```
<affinity_list>
 <list name="list1"
                serverorder="member1,member2,member3,member4"
></list>
  <list name="list2"
                serverorder="member2,member3,member4,member1"
></list>
  <list name="list3"
                serverorder="member3,member4,member1,member2"
></list>
  <list name="list4"
                serverorder="member4,member1,member2,member3"
></list>

</affinity_list>



<client_affinity_defined>

  <client name="groupA" hostname="appsrv1.ibm.com" listname="list1" >
  <client name="groupB" hostname="appsrv2.ibm.com" listname="list2" >
  <client name="groupC" hostname="appsrv3.ibm.com" listname="list3" >
  <client name="groupD" hostname="appsrv4.ibm.com" listname="list4" >

</client_affinity_defined>
```



DB2 pureScale – 完全没有冻结

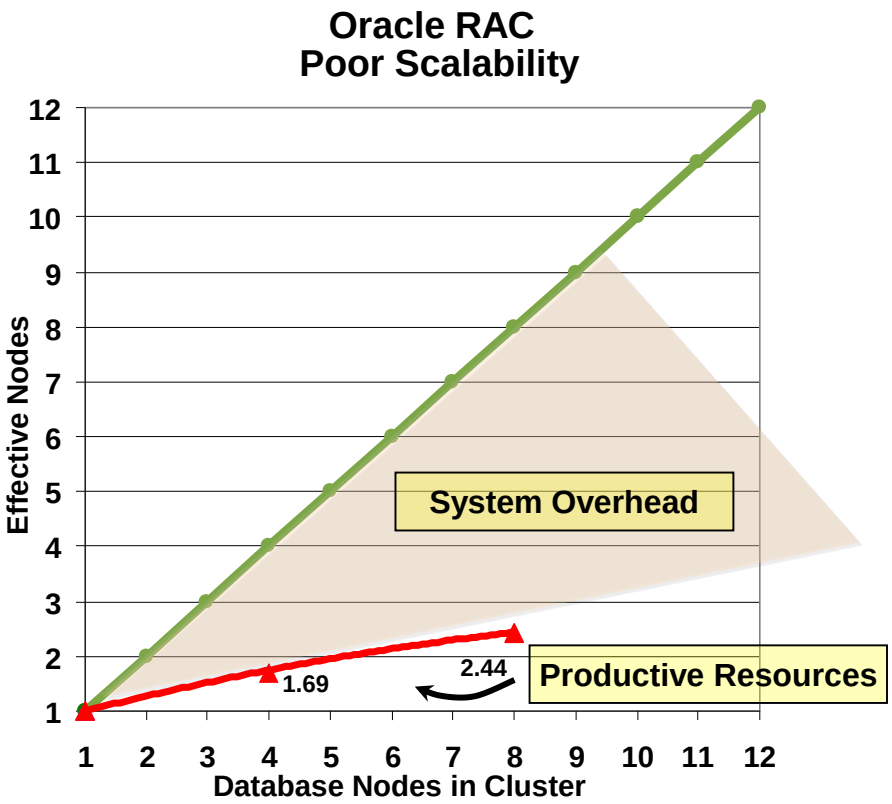


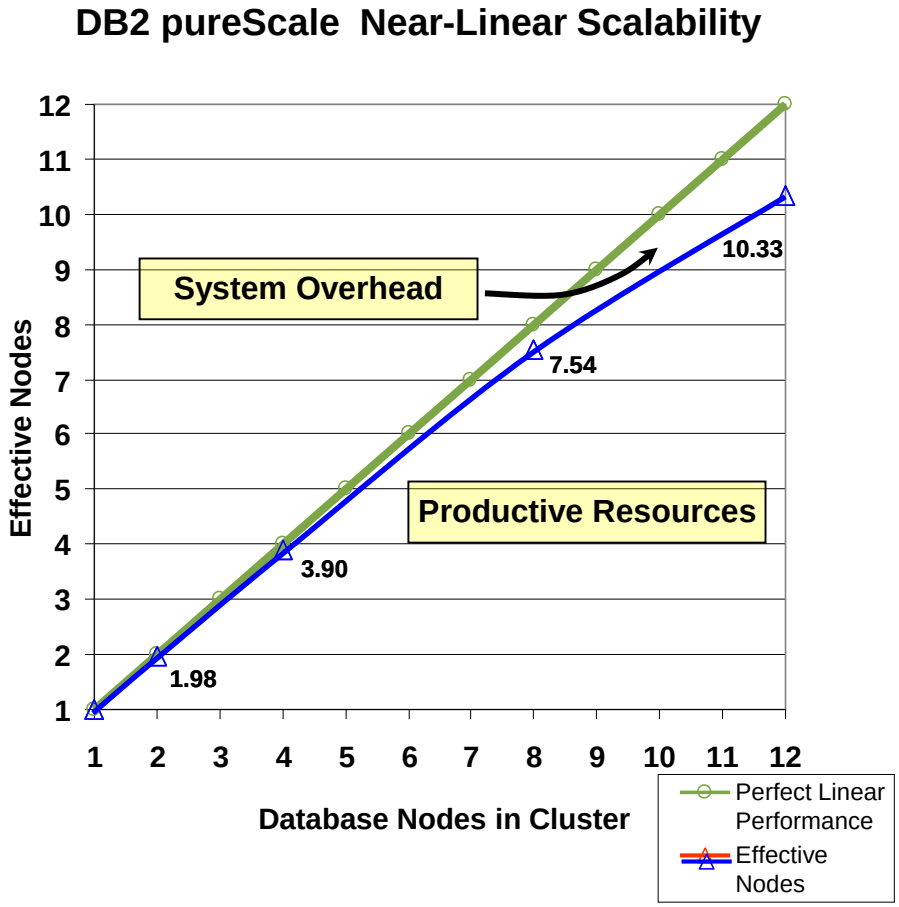
DB2 pureScale与某集群数据库崩溃恢复比较



某集群数据

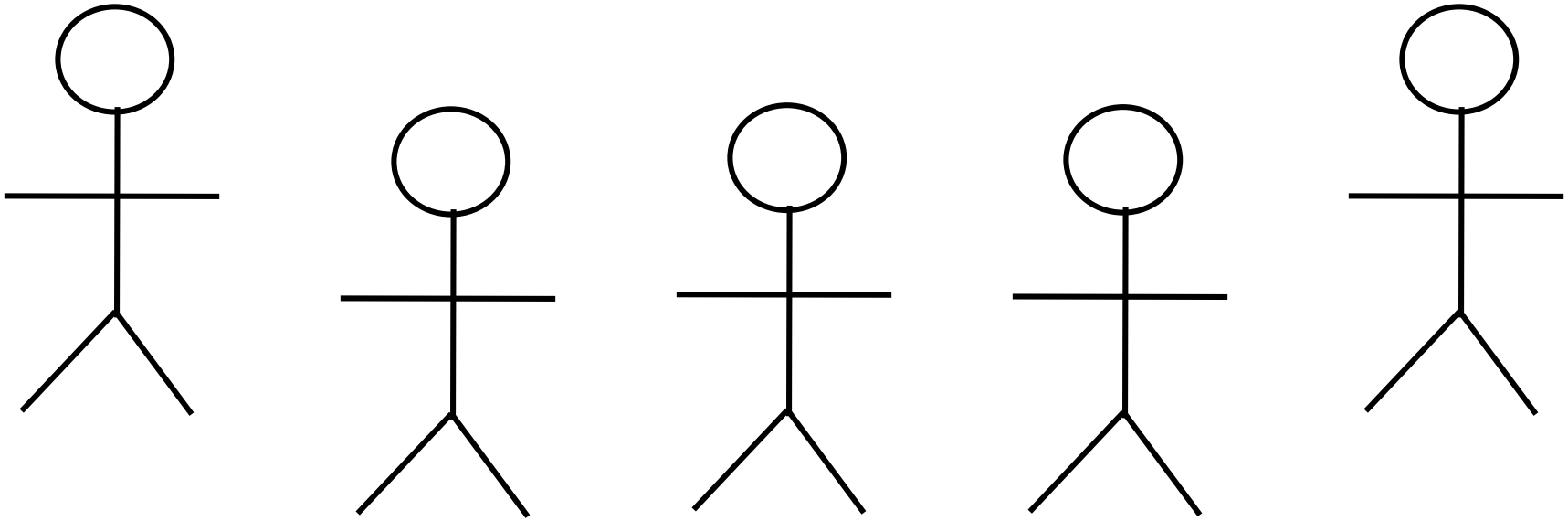
DB2 pureScale 与 某集群数据库的扩展性对比
某集群数据库随节点的增加, 系统吞吐量的效率降低
DB2 pureScale –近线性的扩展效率





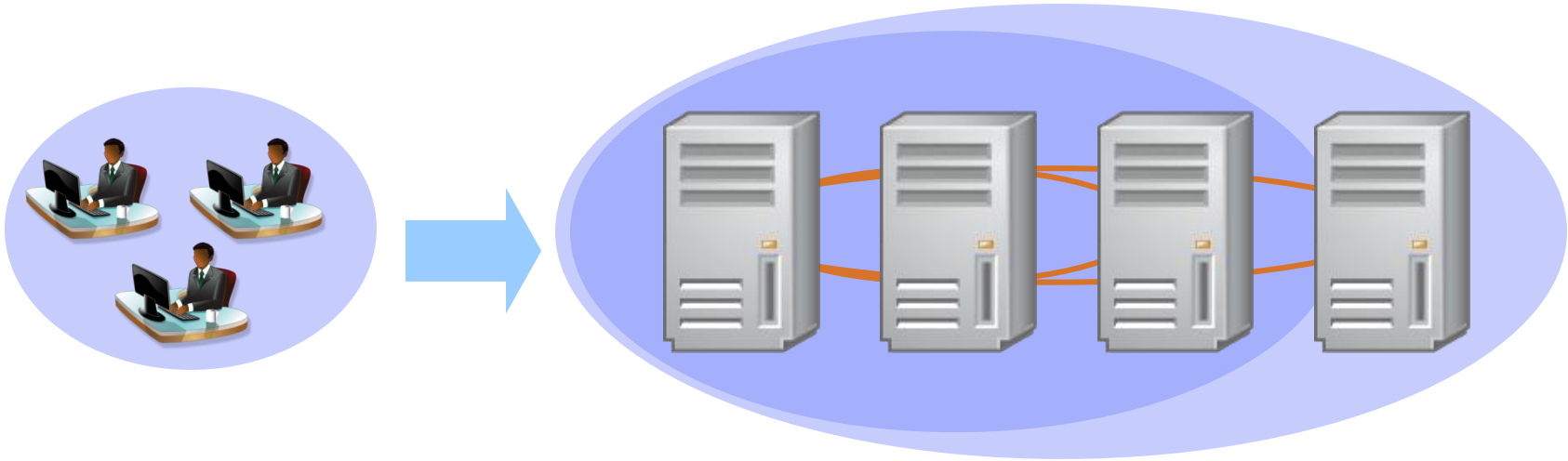
Source: Characteristics as published in Dell test results - http://www.dell.com/downloads/global/power/ps2q07-20070279-Mahmood.pdf
Source: DB2 pureScale characteristics as shown in IBM published results from internal tests on Oct 2009

简单概括 DB2 pureScale 的扩展性
DB2 Data Sharing is like an adult dinner party



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

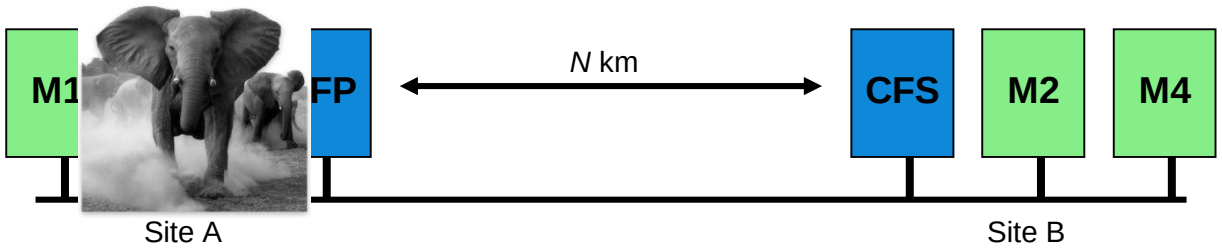
开发人员甚至不需要知道增加了更多节点

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，而不需要应用感知集群架构
    - 数据页面的共享将在实际共享的缓存中通过 RDMA 来实现
      - 不会出现服务器之间的进程中断调用造成的访问无法同步问题
    - 不需要通过应用或数据分区来实现可伸缩性
      - 降低了管理和应用开发成本
  - RAC 中的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

通过GDPC实现同城“双活”灾备

- **扩展的或地理位置分离的pureScale 集群跨2个站点，间距可达数十公里**
  - 目标：提供双活解决方案，访问一个或多个公用的集群数据库
  - 对各种灾难事故提供灾难恢复保护



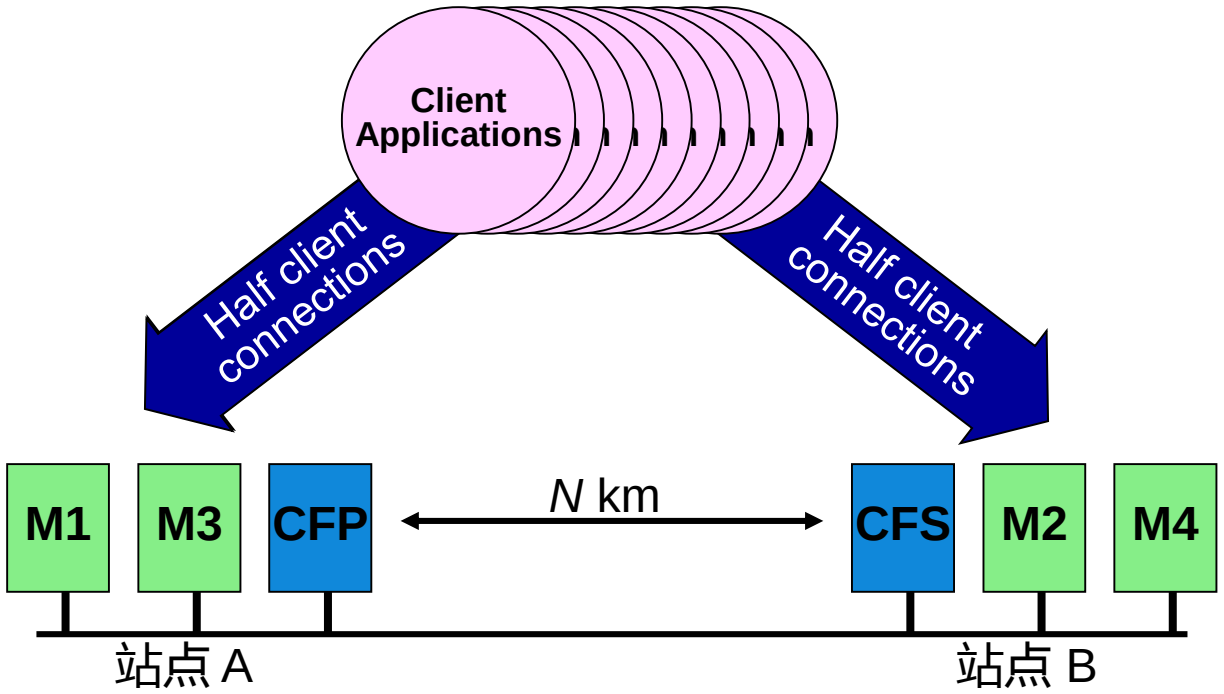
–

- 如- **DB2/z Geographically Dispersed Parallel Sysplex (GDPS)**

http://www-03.ibm.com/systems/z/advantages/gdps/index.html

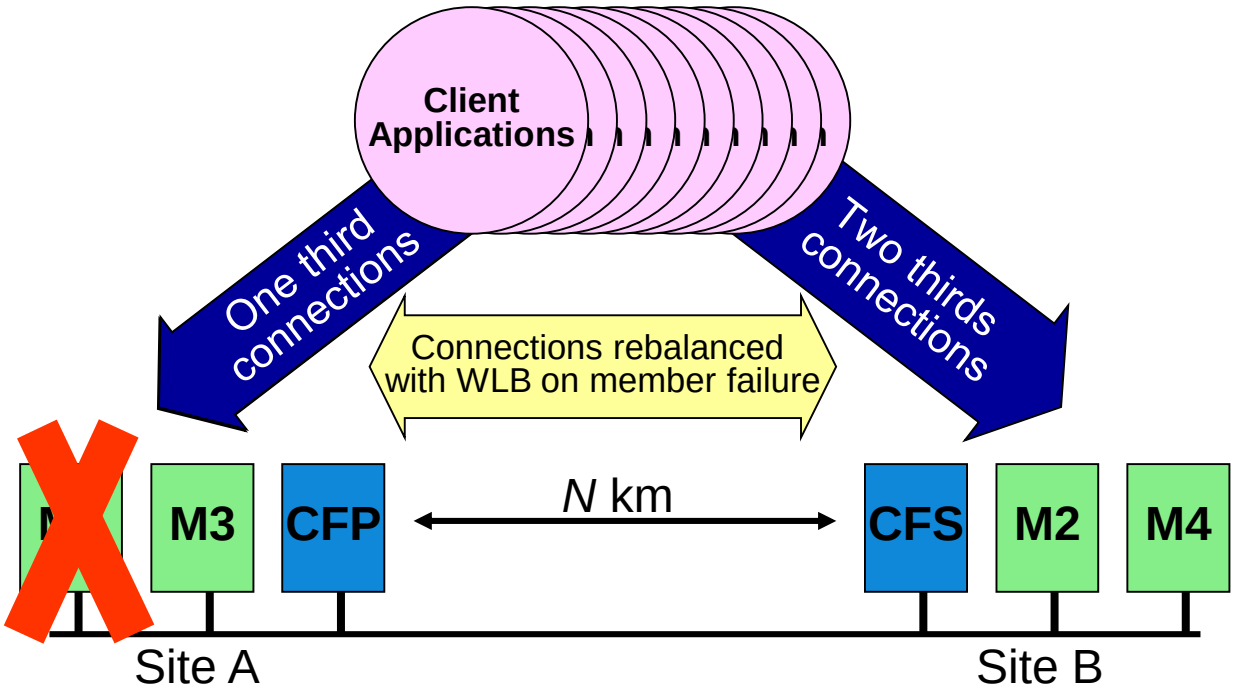
GDPC 目标场景

- 正常时，两个站点均活动并提供交易处理
- 失效时，客户端连接自动重定向到存活成员
  - 无论是站点内成员失效，还是整个站点失效



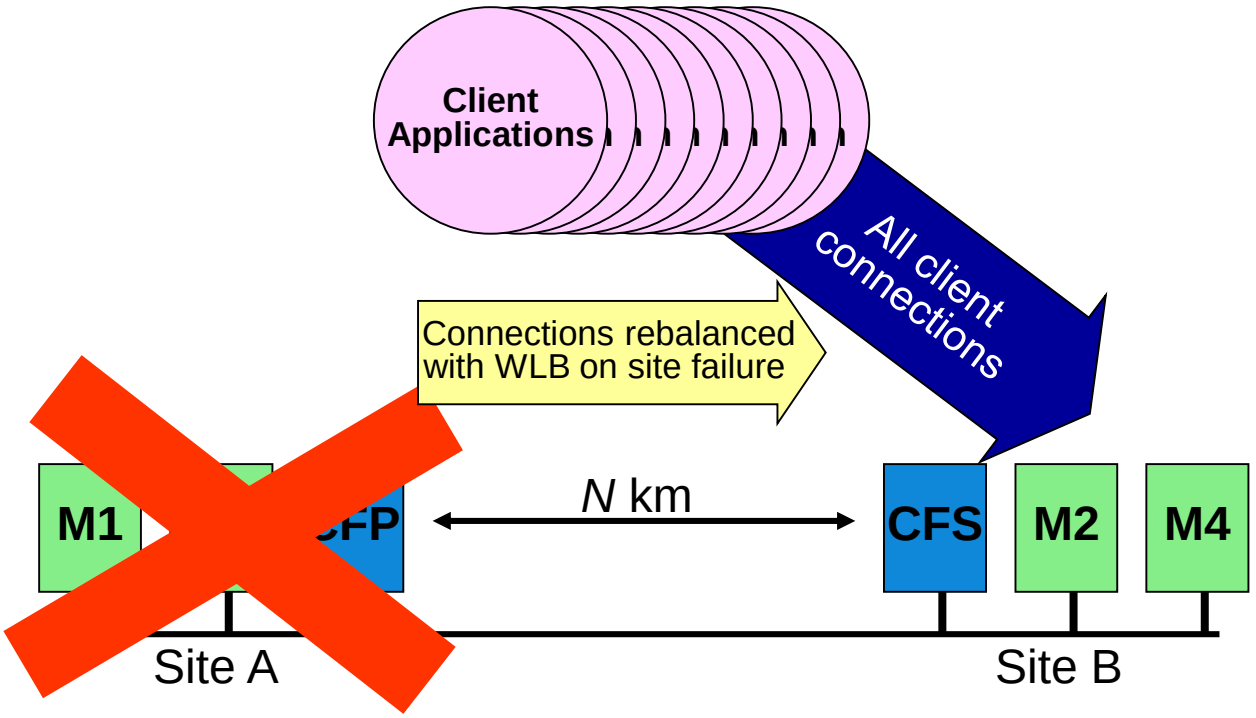
GDPC 目标场景

- 正常时，两个站点均活动并提供交易处理
- 失效时，客户端连接自动重定向到存活成员
  - 无论是站点内成员失效，还是整个站点失效



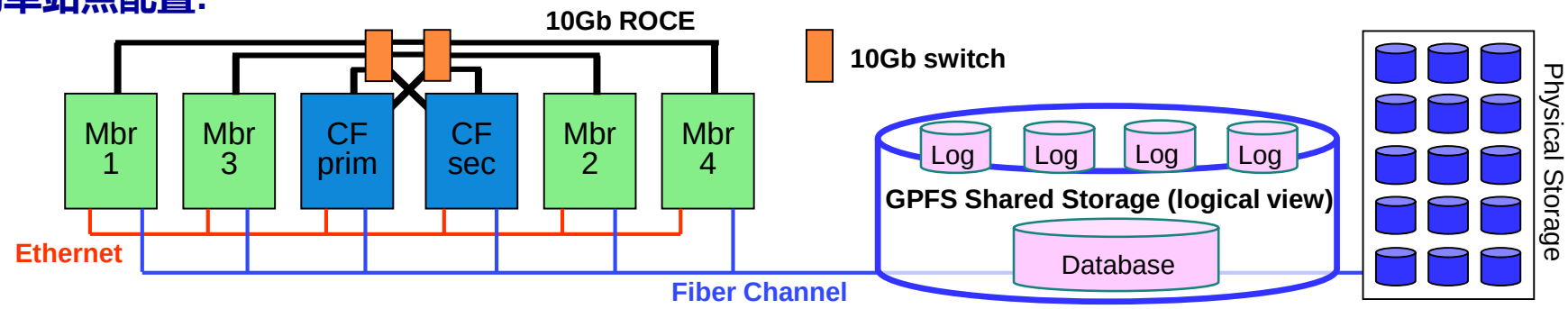
GDPC 目标场景

- 正常时，两个站点均活动并提供交易处理
- 失效时，客户端连接自动重定向到存活成员
  - 无论是站点内成员失效，还是整个站点失效

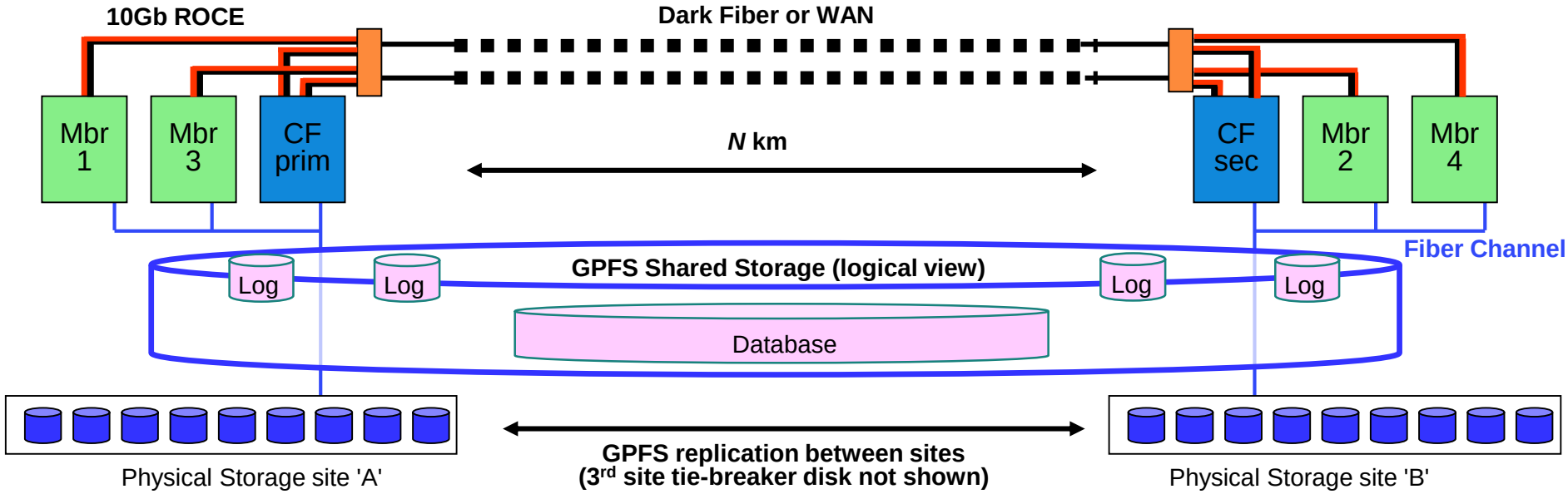


推荐的pureScale GDPC配置

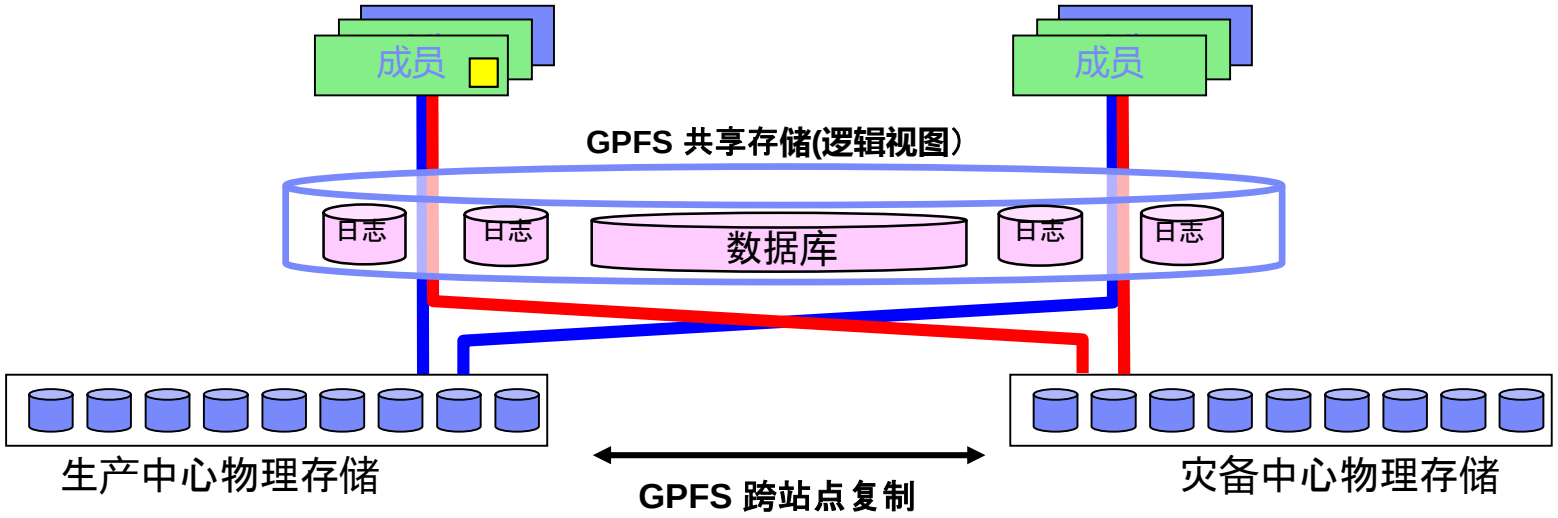
典型的单站点配置:



推荐的 pureScale GDPC 配置:



GDPC使用的磁盘存储



- GPFS 复制协调跨站点的同步写操作
  - 生产中心上任何写操作都复制灾备中心，反之亦然
- 生产和灾备中心的所有主机都能访问所有存储
  - 每台服务器上的GPFS 进程都直接写两个站点的存储，无需通过其他GPFS进程来写变更的数据页面
- 数据复制及站点间距均对增加了数据库交易日志和表空间容器的写操作时间
- 读操作使用本地的数据拷贝以提升性能

分布式pureScale 集群典型特征

- **信号在光纤中传输每公里增加延时5微秒**
  - 3公里距离CF通讯交换需要30微秒, 10公里需要100微秒
  - 实际的延迟取决于网络实现方式以及信号中继器的数量
  - 消息延迟越大对性能的负面影响越大
- **利用GPFS 同步复制功能保证分布式集群文件系统数据的一致性**
  - 广泛认可的方式 (参阅GPFS 管理指南SC23-5182-03)
  - 存储配置为2个失效组，依靠第3个站点进行仲裁
    - 不需要共享磁盘，但需要TCP/IP及本地磁盘
  - 所有更新操作在两个站点的存储中均执行
- **对于读比例较高的工作负载（如80%或更多的读操作）更合适**
  - 读写比对性能的影响随着2个站点间距离增加而增大
- **利用工作负载均衡(WLB) 和客户端自动路由 (ACR) 机制实现故障发生时客户端重定向到健康成员服务器**
- **单个成员的失效恢复时间在60到80秒之间**
- **站点失效恢复时间约120秒**

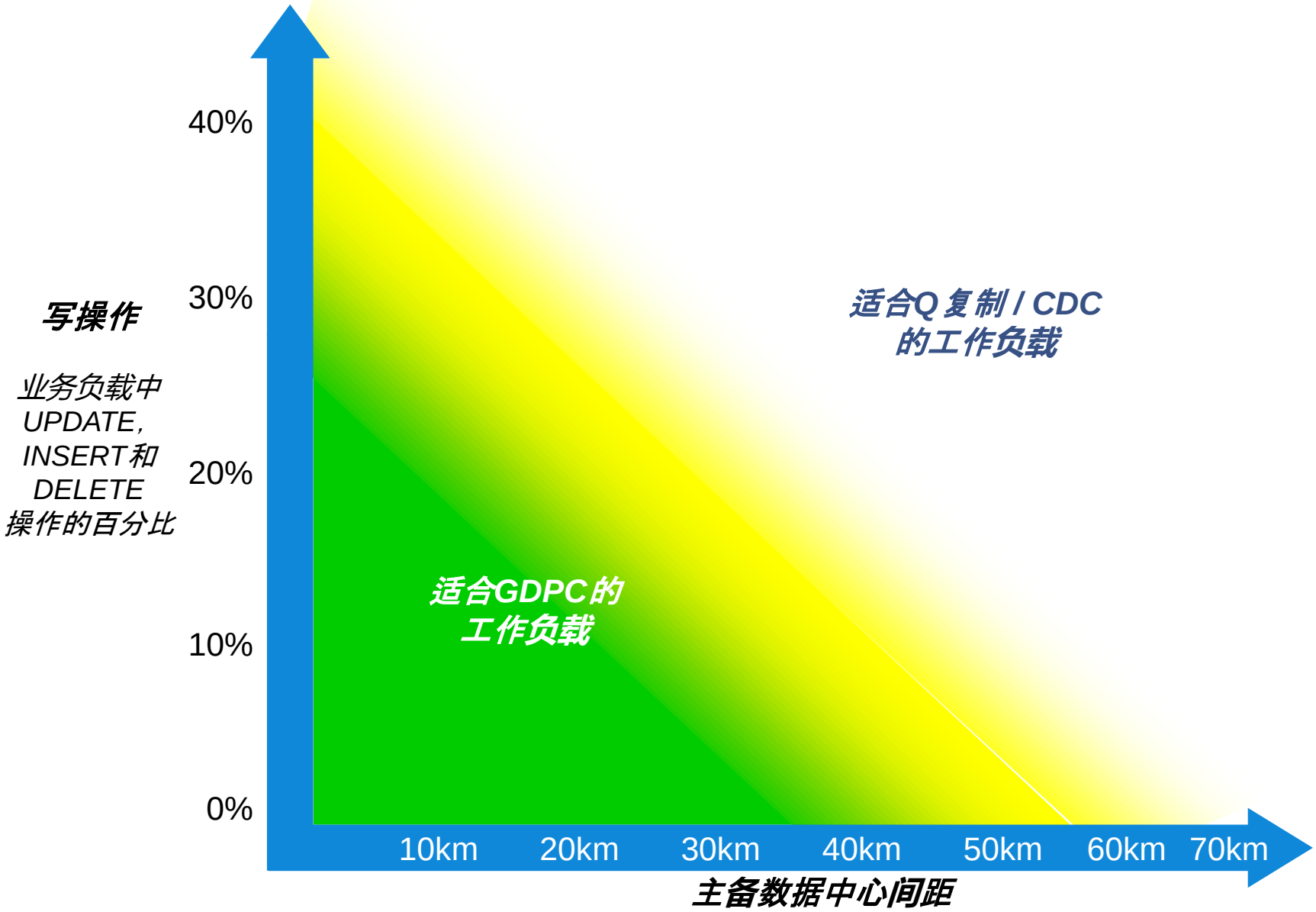
其他GDPC 要求

1. **两个站点之间需要有裸光纤(密集波分复用DWDM) 连接或WAN连接**
   - 建议10 Gb/s 可用带宽
   - 为避免单点故障，建议使用2条链路

2. **远程SAN 结构以支持跨2个站点的GPFS 复制**
   - 长距离需要SAN Router，走IP网络，运行FCIP协议
   - 所有存储必须同时对2个站点的服务器“可见”，保证站点失效后数据仍然可以访问
   - 第3个站点的仲裁盘用于避免在2个站点间通讯故障造成的“脑裂”情况的出现
   - 细节请参阅GPFS 红皮书获取GPFS复制相关信息

3. **客户端站点需要能够同时访问到主站点和灾备站点**

GDPC 适应性



GDPC性能一瞥-- 实验室测试

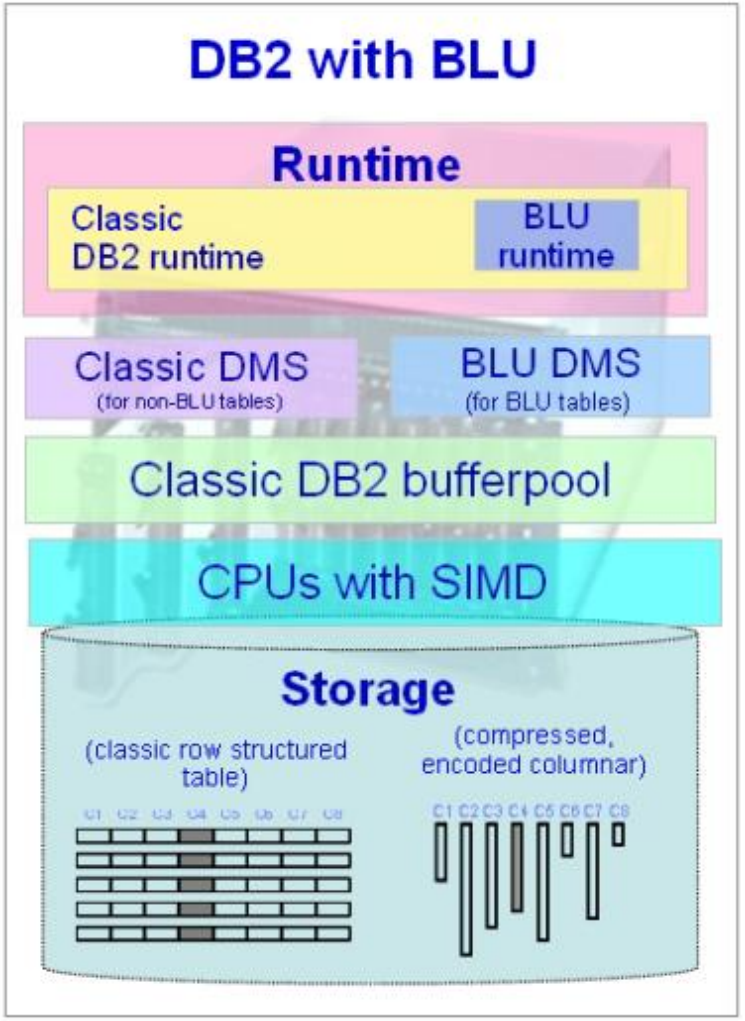
- **IB 带宽**
  - AIX Galaxy 2 DDR
  - SDR 扩展器@ 10 Gb/s (Obsidian E-100)
- **SAN 带宽**
  - 8 Gb/s 光纤通道直连
  - 基于实际生产部署中使用光纤交换机，且两个站点间使用相同的广域网 / 裸光纤
- **站点间距**
  - 模拟距离为0到80公里以上
  - 距离和交易读写比高度相关
  - 通常建议的距离为40-50公里
- **交易读写比**
  - 70/30
  - 80/20
  - 90/10
- **吞吐量指标**
  - 2000-7000 笔交易/秒
  - 依赖于集群中节点数，交易中读写比等（和压力负载密切相关）

议程

- DB2架构及技术特点
- DB2集群技术
  - DB2 DPF集群
  - DB2 pureScale集群
- DB2列式存储及内存计算
  - DB2 BLU
- DB2客户案例

DB2 下一代分析技术 BLU

- DB2 BLU是什么？
  - DB2 BLU 是面向分析的内存计算数据库
  - 直接内置在 DB2 内核中的列计算引擎
  - 按列方式存储和管理表
- DB2 BLU的价值
  - 极高性能，BLU针对CPU、内存和IO进行特别优化，如矢量计算引擎、自适应CPU并行、高级页面缓存、按列选取数据、无效数据过滤等
  - 深度智能压缩，大规模节约存储空间，大幅减少运算量
  - 简单易用，无需索引、分区及物化视图，更低的运营成本
  - 和DB2无缝集成，一致的SQL语句、开发接口和管理方式
- 突破的技术
  - 结合并扩展最好的技术
  - 已申请和待批的专利超过25 项
  - 利用跨全球7 个国家的10 个实验室的多年IBM R&D 成果
- 典型的体验
  - 易于实施和使用
  - 10 倍至25 倍的性能提升
  - 和未压缩的数据和索引相比，实现10 倍至20 倍的存储节



DB2 BLU用例1：企业数据仓库负载剥离/数据集市加速

EDW 应用程序

OLAP 应用程序

Cognos BI
，*带有 BLU Acceleration*

Oracle 或
Teradata 仓库的性
能较差

轻松创建和加载
BLU Acceleration **内存中集市**

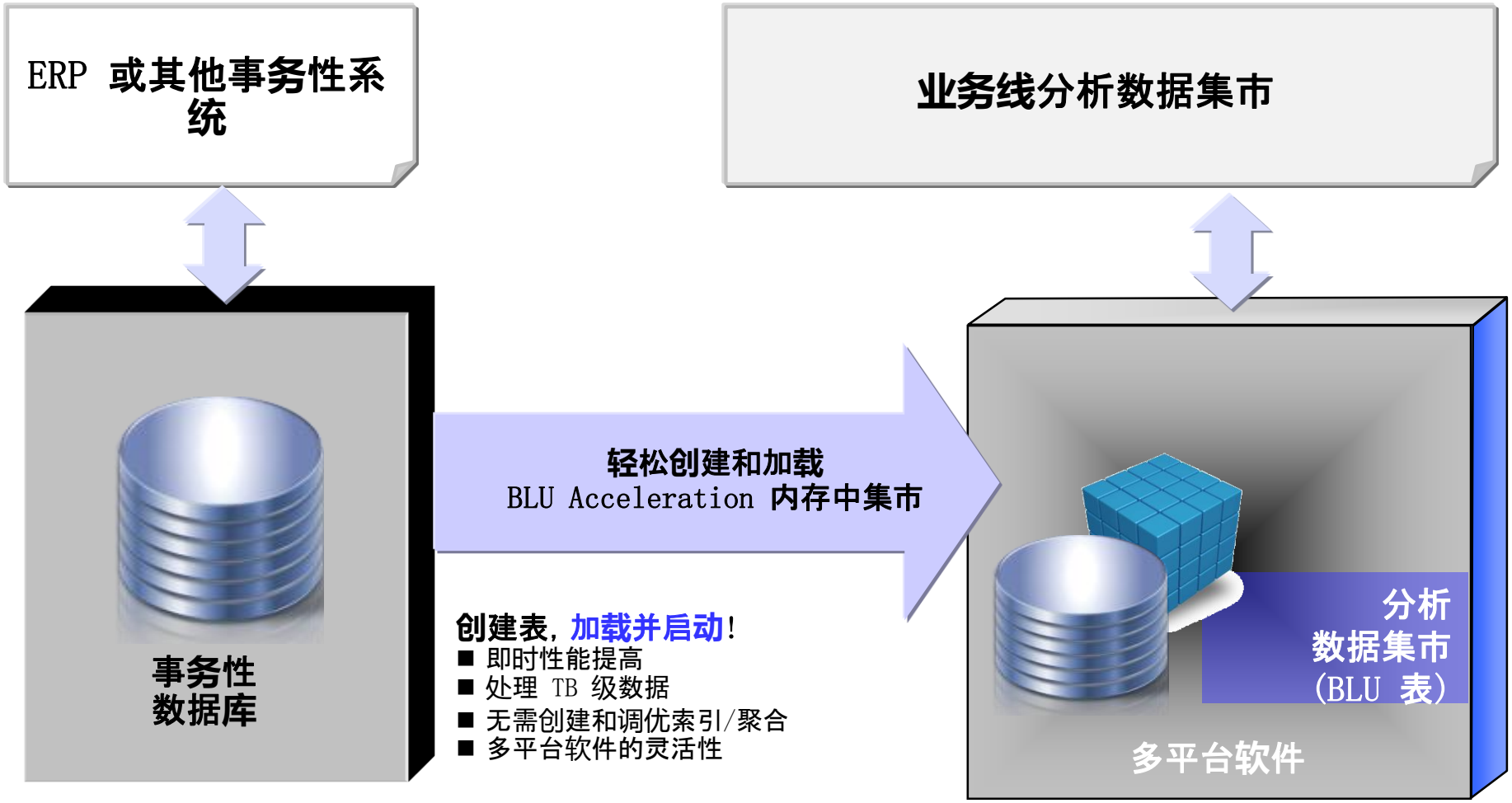
Cognos BI
，*带有 BLU Acceleration*

分析
数据集市
(BLU 表)

多平台软件

创建表，加载并启动！
- 即时性能提高
- 处理 TB 级数据
- 无需创建和调优索引/聚合
- 多平台软件的灵活性

DB2 BLU用例2：分析数据集市 - 数据源自事务性数据库



DB2 BLU的价值

DB2 BLU

面向分析的下一代数据库

- 开箱即用的极高性能
- 大规模存储节省
  - 无需索引
- 更低的运营分析成本

无缝集成

- 无缝内置到 DB2 中
- 一致的 SQL、接口、管理
- 大幅简化
  - 设计更少
  - 调优更少
  - 只需加载就可以启动

硬件优化

- 内存中优化
  - 在内存中压缩
- 现代 CPU 利用
- I/O 优化
  - 只读取感兴趣的列

DB2 10.5 Beta同DB2 10.1查询速度测试比较结果

客户	查询速度提高
某大型金融服务公司	46.8倍
某第三方软件供应商	37.4倍
某分析软件业务公司	13.0倍
某全球零售公司	6.1倍
某大型欧洲银行	5.6倍

分析查询速度
平均提高
10-25 倍

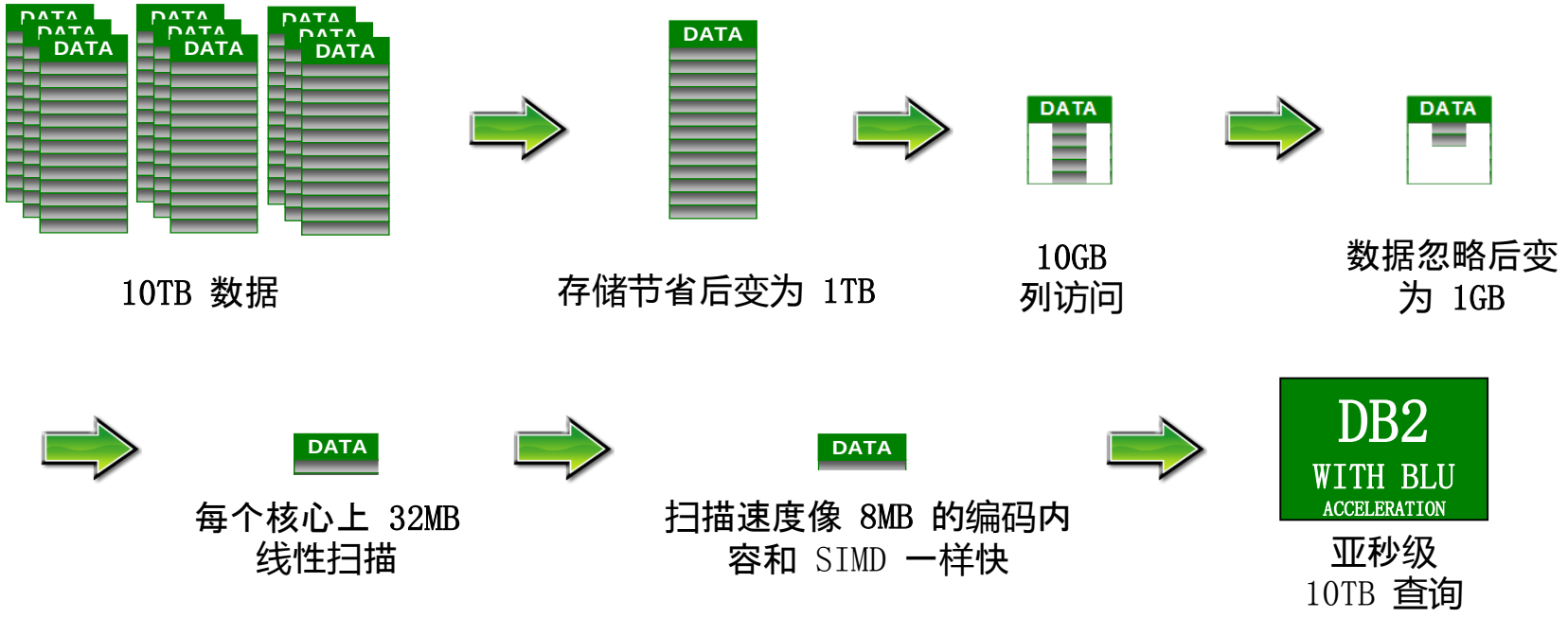
BNSF RAILWAY

"It was amazing to see the faster query times compared to the performance results with our row-organized tables. ***The performance of four of our queries improved by over 100-fold! The best outcome was a query that finished 137x faster by using BLU Acceleration."***

- Kent Collins, Database Solutions Architect, BNSF Railway

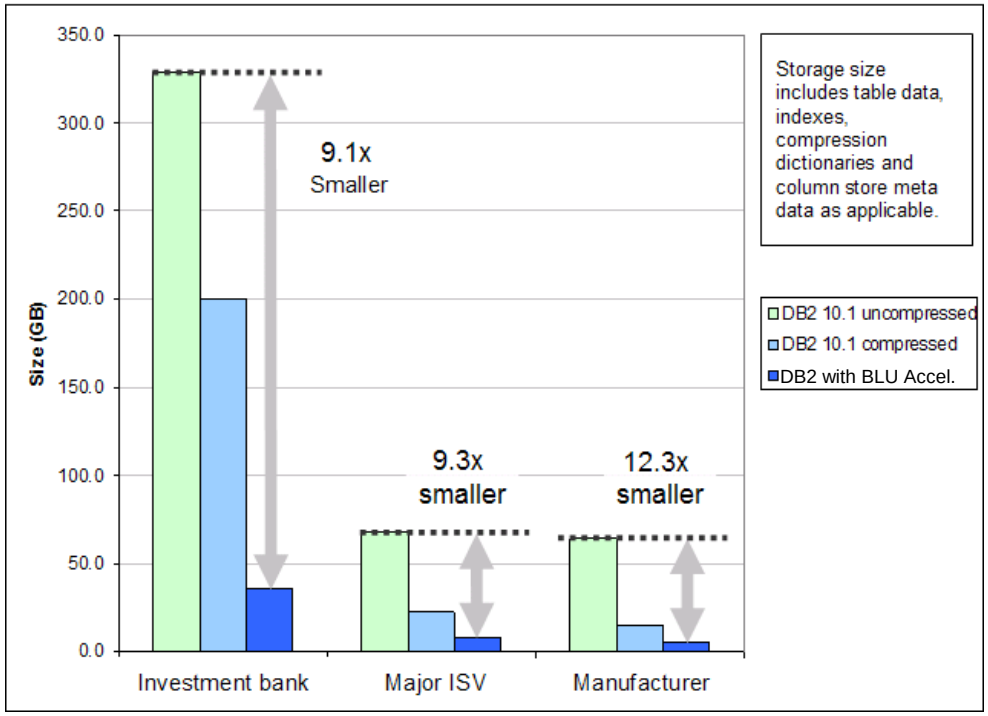
DB2 BLU查询示意图
亚秒级 10TB 查询

- 系统 – 32 核，10TB 的表，含 100 个列，10 年的数据
- 查询：2010 年有多少事务
  - SELECT COUNT(*) from MYTABLE where YEAR = '2010'
- 乐观的结果：亚秒级 10TB 查询！每个 CPU 核心只检查相当于 8MB 的数据



节省存储空间

- **更少的数据存储空间**
  - 同未压缩数据比较，平均可以节省95%的数据存储空间
  - 只需要存储表数据，不需要额外空间来存储索引等其它类型数据
- **应用了多种压缩技术**
  - 所有操作都在压缩数据上进行
- **针对不同数据类型应用最优的压缩算法**



简单易用 —— 建表，装数，查询！

数据库设计和调优

重复

1. 决定分区策略
2. 选择压缩策略
3. 创建表
4. 加载数据
5. 创建辅助性能结构
   - 具体化的视图
   - 创建索引
     - B+ 索引
     - 位图索引
6. 调优内存
7. 调优 I/O
8. 添加优化程序提示
9. 统计信息收集

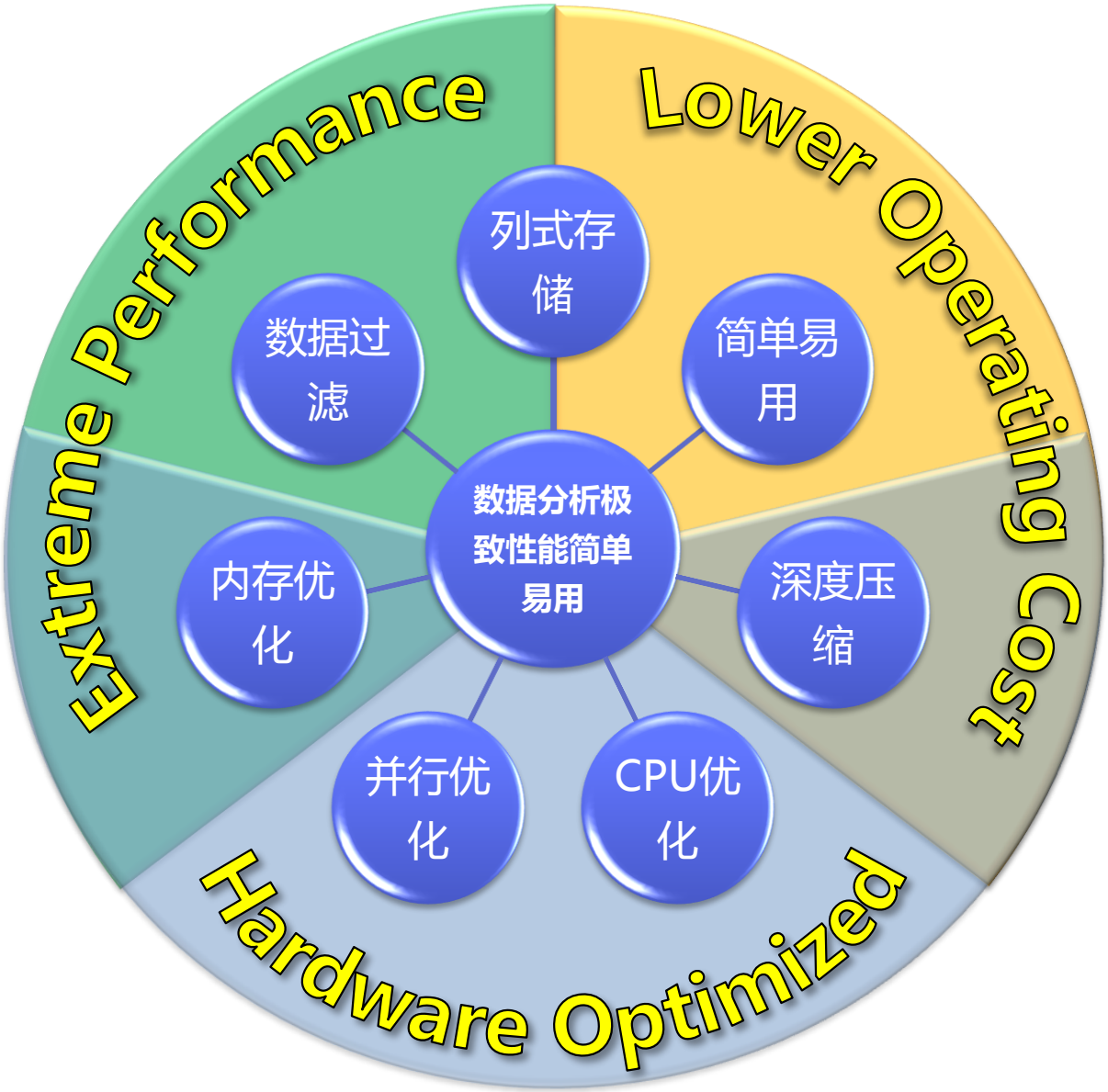
DB2 BLU

1. 创建表
2. 加载数据

DB2 BLU易于部署和操作

- 运营
  - 只需加载数据就可以启动
  - 像所宣称的那样易于评估和执行
- BI 开发人员和 DBA – 更快地交付成果
  - 无需配置或物理建模
  - 无需索引或调优 – 开箱即用的性能
  - 数据架构师/DBA 可专注于业务价值，而不是物理设计
- ETL 开发人员
  - 无需聚合各个表 – 更简单的 ETL 逻辑
  - 更快的加载和转换速度
- 业务分析师
  - 真正的即席查询 – 无调优，无索引
  - 针对大型数据集提出复杂的查询

DB2 BLU 7大技术亮点展示 – 不只是列存储！



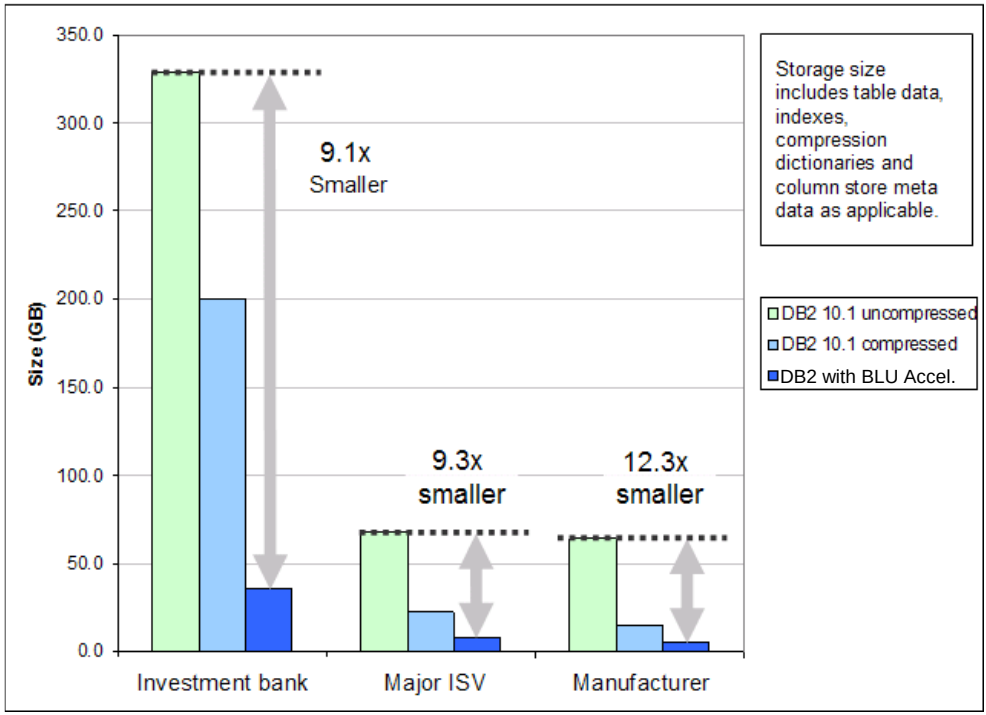
7 大亮点: 1 简单易用

- 数据即装即用
  - 无需索引
  - 无需整理数据
  - 无需更新统计信息
  - 无需分区
  - 无需物化视图
  - 无需提供”SQL优化提示(hint)”

- 与传统DB2无缝集成
  - 相同的：SQL语法, 编程接口（JDBC/ODBC）, 管理命令
  - 相同的：DB2的处理模型、存储管理、各种工具

7 大亮点: 2 智能压缩技术

- 更少的数据存储空间
  - 同未压缩数据比较，平均可以节省95%的数据存储空间
  - 只需要存储表数据，不需要额外空间来存储索引等其它类型数据
- 应用了多种压缩技术
  - 延迟解压，所有操作都在压缩数据上进行
- 针对不同数据类型应用最优的压缩算法
  - 字符型数据采用哈夫曼编码压缩

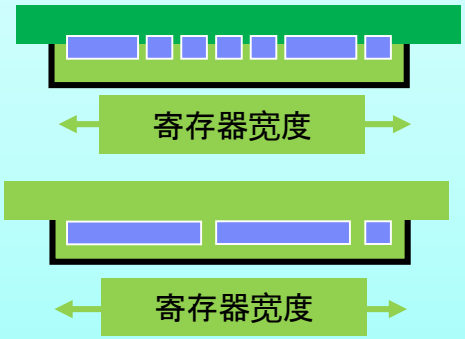


7 大亮点: 2 智能压缩技术

- 哈夫曼编码（Huffman）
  - 基于概率编码，出现频率最高的信息用最精简的编码
- 面向寄存器优化
  - 编码组合按照CPU的寄存器宽度进行组合
  - 更少的IO访问，更好的内存使用率，更少的CPU周期处理

Last Name	编码
Johnson	
Smith	
Smith	
Smith	
Smith	
Johnson	
Smith	
Gilligan	
Sampson	
Smith	

按照CPU的寄存器宽度打包



7 大亮点: 2 智能压缩技术

- 延迟解压，无需解压缩即可计算
  - 条件选择和关联计算直接在压缩编码上进行

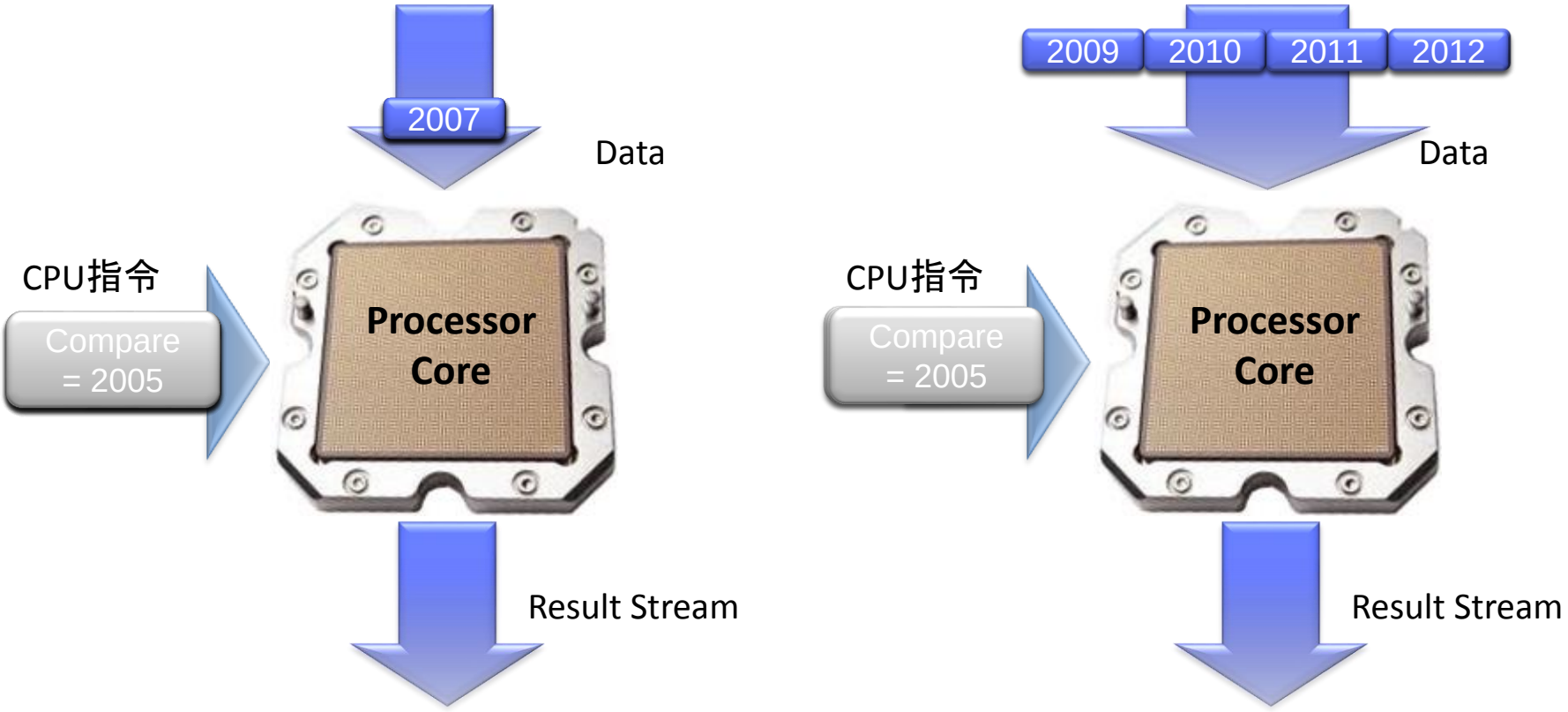
```
SELECT COUNT(*) FROM T1 WHERE LAST_NAME = 'SMITH'
```

LAST_NAME	Encoding
Johnson	
Smith	
Smith	
Smith	
Smith	
Johnson	
Smith	
Gilligan	
Sampson	
Smith	

Encode

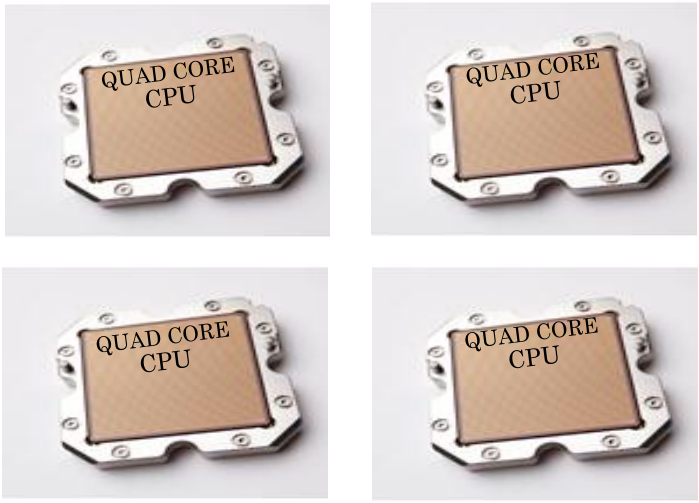
7 大亮点: 3 CPU优化（SIMD)

- 没有SIMD优化的情况下，CPU每个指令只能处理一个数值
- 进行SIMD优化后，性能大幅度提升
- CPU每个指令可以处理多个数值



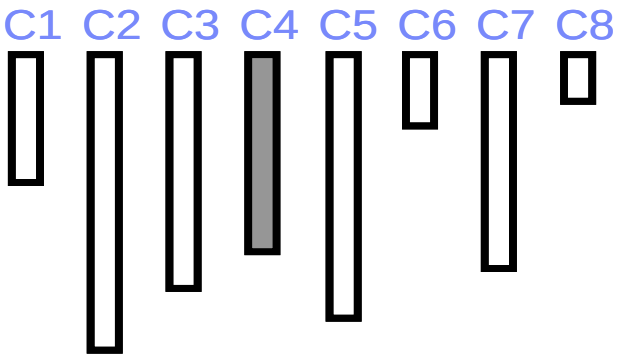
7 大亮点: 4 智能并行

- 自动感知服务器的CPU数量
  - 基于BLU的查询会自动并行执行
- 最大程度利用CPU的缓存



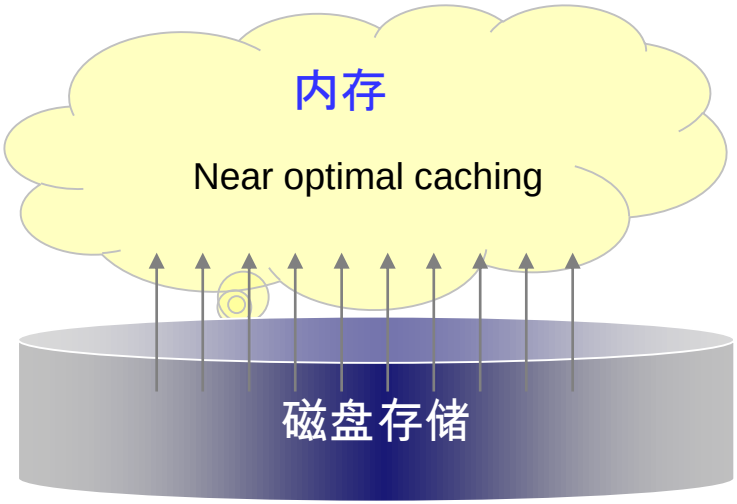
7 大亮点: 5 列式存储

- 最小化 I/O
  - 仅在查询需要用到的列上进行IO操作
  - 大大减少数据页的访问
- 基于列的运算
  - 条件选择、关联、扫描等，所有的计算以列的形式进行
  - 只有在返回结果集的时候才重组为行
- 提高内存数据密度
  - 数据在内存中也保持列压缩的方式
- 深度压缩
  - 在存储和内存中都以压缩形式存放更多数据
- 提高缓存效率
  - 数据按照CPU寄存器的宽度进行打包



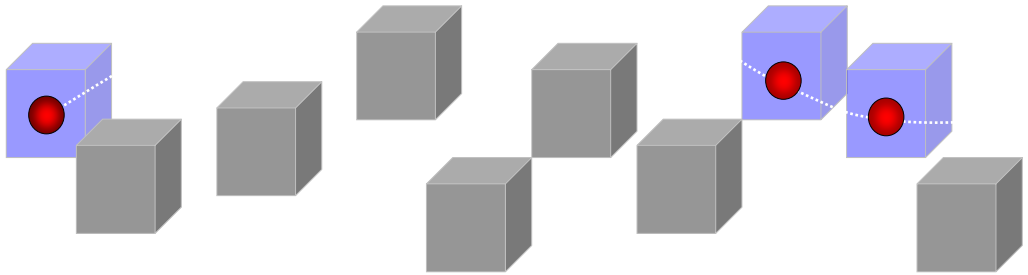
7 大亮点: 6 高效内存缓存机制

- 新的内存缓存算法
- 使用频率最高的数据优先缓存在内存中
  - 基于访问概率的调度算法代替基于时效性的调度算法
- 数据可以比内存大
  - 无需将所有的数据都缓存到内存中
  - 面向内存和IO效率优化

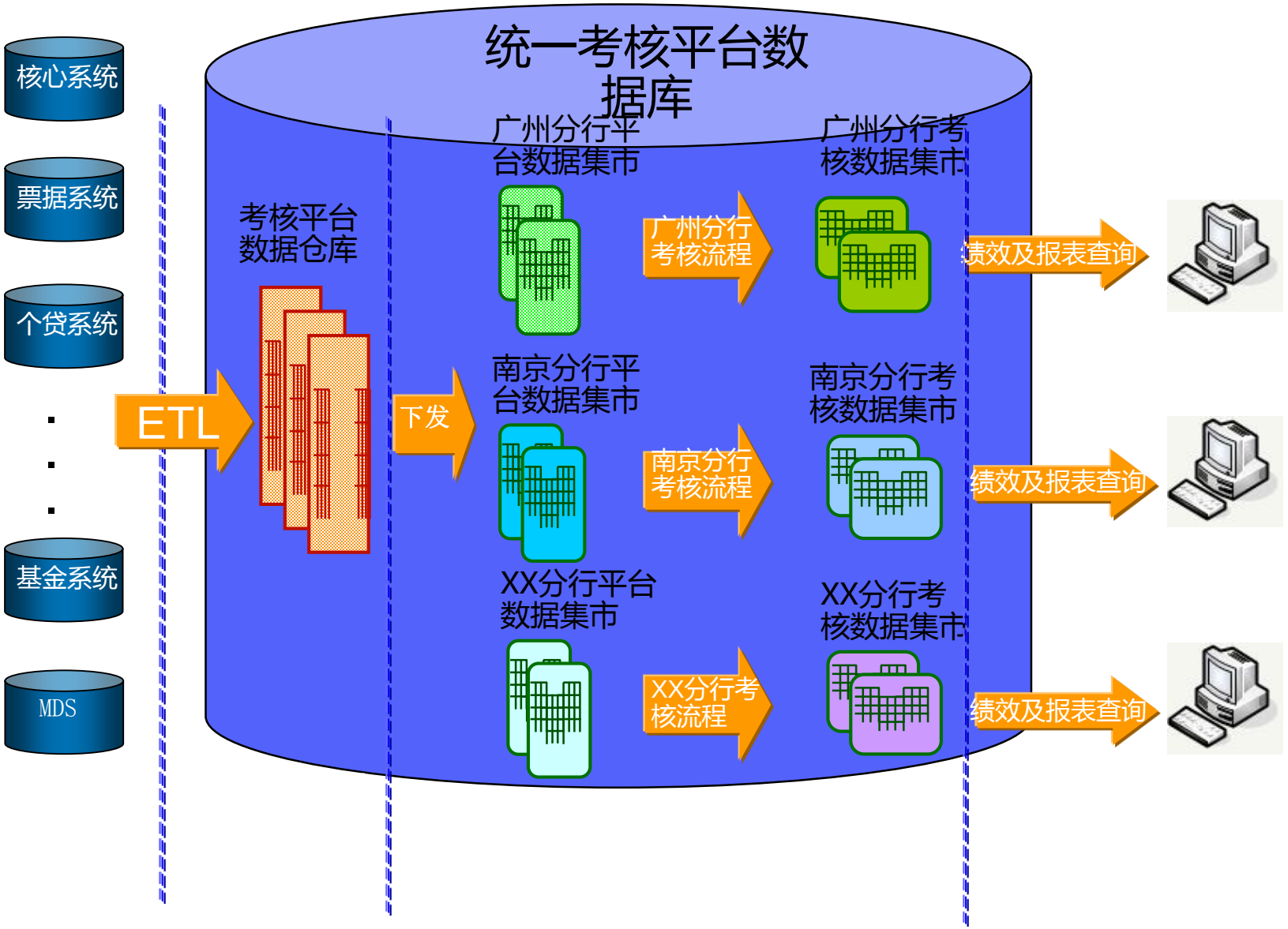


7 大亮点: 7 数据过滤

- 数据扫描时自动过滤不符合条件的数据块
- 自动过滤节省大量IO、内存和CPU开销
- 无需任何DBA的管理工作即可使用 – 完全透明
  - 自动记录存储数据块的最大最小值



某银行DB2 BLU体验 – 考核系统

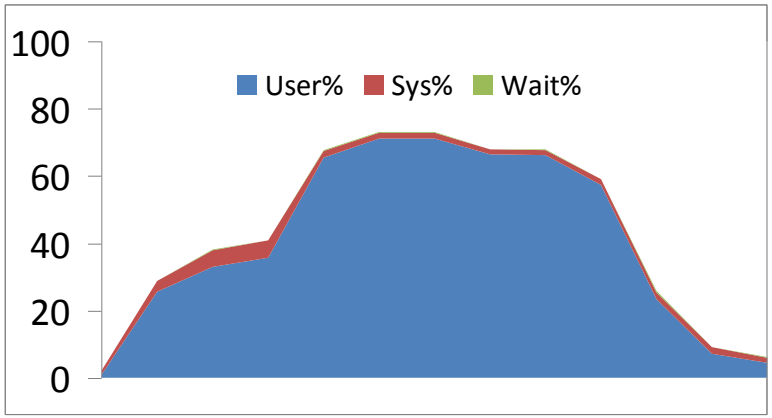


优化方案结果验证

- 选择在生产系统中耗时较长的SQL, 对比优化后的响应时间
- 所选择SQL的特性：
  - 多(13个)表/视图的Join
  - 3张最大的表记录数分别是（8.8亿， 2800万， 670万）
  - 选择的列数目约占到总列数的10%（总列数为260列）

DB2 BLU 优化方案硬件资源利用情况分析

CPU



- 充分利用CPU资源 (~70%)
- 任务可以根据CPU核数充分并行化
- 并行化独立程度很高，wait值很小

IO

- 只访问使用到的列数据，减少IO
- 采用列压缩方式，节省存储空间（平均节省70%以上，最高节省91%），进一步减少IO。
- 智能化的Data Skipping方法跳过不需要访问的数据(不同表跳过20% - 90%以上不需要访问数据)，将IO降到最小。

Memory

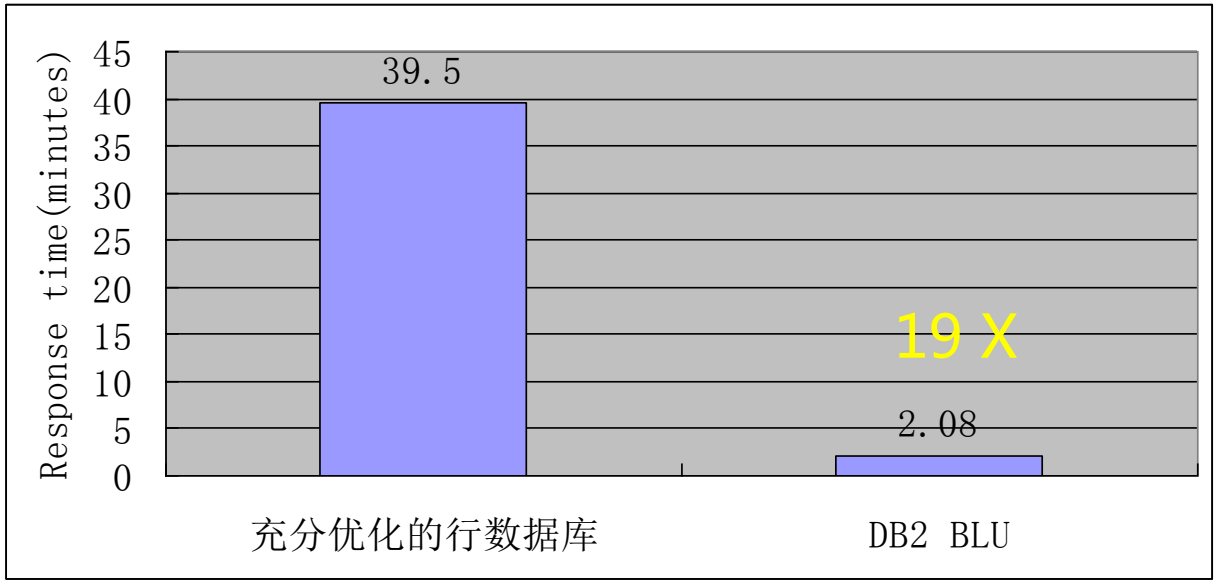
- 数据在内存中保持压缩状态，只在返回最终结果时才进行解压
- 同样大小的Buffer Pool可以缓存更多的原始数据

优化方案结果对比

建立额外的索引
使用range partition分区表
采用页压缩技术
收集统计信息
调优内存配置

VS

DB2_WORKLOAD=ANALYTICS
建库，加载数据, 运行



DB2 BLU无需繁杂的调优过程取得了19倍的性能提升!

IBM DB2 高级企业版 – 集群数据库的全能王

IBM DB2高级企业版的优势

高性能：无论是**OLTP**还是**OLAP**，都能提供无与伦比的优异性能，支持行列混合存储，保持着单机和集群环境环境下单核的各种TPC测试最高性能记录，以及SAP标准测试中保持性能最佳记录

灵活性：支持DPF和PureScale两种集群部署方式，能够根据实际的负载需求灵活选择部署模式，实现负载优化

易管理：提供最大程度的自我管理能力，在性能自动优化、存储自动管理等方面领先竞争对手，减轻DBA工作负担。PureScale集群环境下支持“滚动升级”，无需停机

高可用：PureScale集群部署，最接近IBM大型主机级别的高可靠性。实现集群单点故障对整体完全透明，支持业务7X24小时不间断

扩展性：集群环境中物理节点增加，能实现处理能力获得“线性” 提升

兼容性：支持最广泛的SQL标准，兼容Oracle数据库95%以上

深压缩：提供10倍以上压缩比，大幅度节省存储资源

安全性：通过数据库领域最高级安全评测EAL 4+

议程

- DB2架构及技术特点
- DB2集群技术
  - DB2 DPF集群
  - DB2 pureScale集群
- DB2列式存储及内存计算
  - DB2 BLU
- DB2客户案例

IBM DB2 在国内的部分成功案例

银行业:

- 交通银行
- 中国银联
- 人民银行
- 招商银行
- 民生银行
- 中信银行
- 浙江商行
- 中信实业银行
- 山东农信
- 海南农信
- 广东农信
- 广东中行
- 大连建行
- 山东建行
- 大连农行
- 上海农行
- 江苏农行
- 德州商行
- ……

市场领先！

电信业:

中国移动

- 中国移动前10家运营商，共17家
  如广东移动、北京移动、浙江移动，上海移动等

中国网通

- 网通集团IC卡分析
  智能网分析
- 河南网通，河北网通

中国电信

- 福建电信，重庆电信，陕西电信，安徽电信

中国联通

- 上海联通，山东联通，
- ……

政府:

- 外汇管理局
- 发改委——金宏工程
- 公安部金盾工程数据交换平台
- 国家税务总局金税工程数据交换平台
- 海关总署金关工程数据交换平台
- 国家审计署金审工程
- 中国人民银行支付系统
- 中国人民银行财税库行横向联网系统
- 海南卫生厅
- 中航信
- 国家质检总局金质工程
- 国家烟草专卖局[illegible]号工程
- 云南烟草
- 湖北烟草
- ……

其他:

- 新华社全球采编
- 苏宁物流

DB2是银行、电信对稳定性要求高的客户的首选！