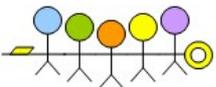


Teradata系统架构及特性

DW项目组 赵世辉



10010010100001010010100101001010



Teradata系列培训

基础培训

1. Teradata软硬件体系架构原理
2. Teradata数据库对象介绍
3. Teradata工具集介绍

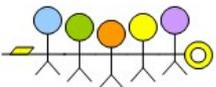
中级培训

1. Teradata数据库设计规范
2. Teradata SQL规范
3. 数据仓库Teradata平台管理规范

高级培训

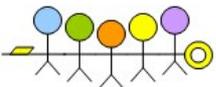
1. Teradata工具使用方法和技巧
2. Teradata程序设计与开发
3. 数据库高级管理
4. 数据库调优
5.



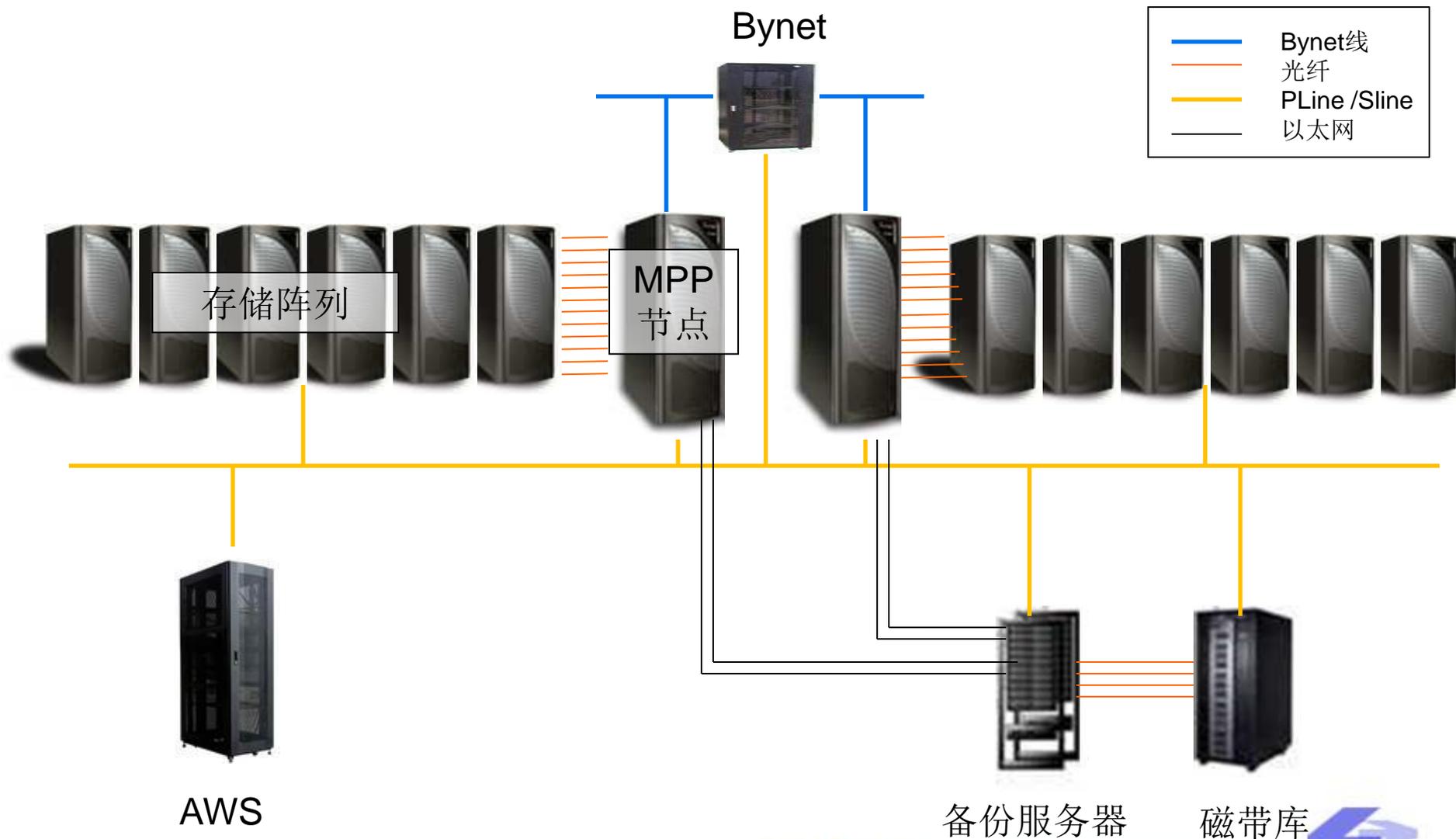


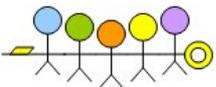
- **Teradata**软硬件体系结构
- **Teradata**数据库原理及特点
- **Teradata**数据保护机制
- **Teradata**系统访问配置及连接方式
- **Teradata**使用中的一些问题及案例分析



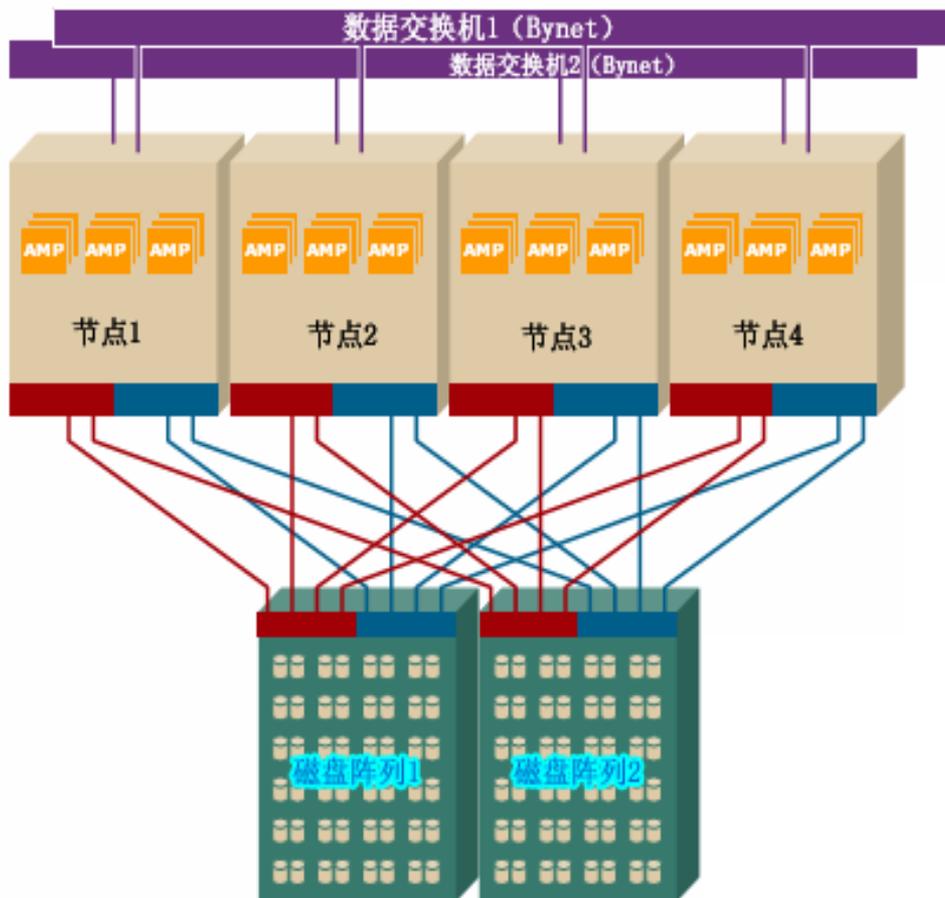


Teradata系统的硬件构成





Teradata主机结构



- MPP系统
 - 工作站集群模式
 - 批量处理优化
 - 底层并行
 - 线性扩展
 - 均衡负载
- 高可用性
 - 热备组件
 - RAID技术
 - Clique技术





AWS

- 收集显示主机、存储、Bynet所有模块运行信息
- 设备管理的统一界面
- 通过TVI进行远程维护和故障通知



备份服务器

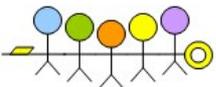
- 基于LAN-BASE备份技术
- 由备份服务器处理备份任务，减轻数据库压力
- 使用Netvault工具，可在AWS上的客户端操作备份恢复



磁带库

- 由机械手+磁带驱动器+磁带槽位+磁带组成
- 根据磁带的条码自动实现磁带的拆卸和装填
- 可远程控制，可多驱动器并行工作和交叉工作



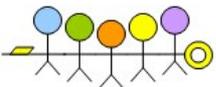


扩展知识：Teradata 主机产品线



	550	1550	2550	5xxx
用途	数据集市或开发测试机	在极端大量数据环境中的分析	企业入门级数据仓库或部门级的数据集市	企业级的数据仓库系统，应用于战略性和操作性的企业智能化的EDW/ADW
扩展性 (支持数据量)	单节点 6 TB	1024节点 50 PB	46节点 140 TB	1024节点 10 PB



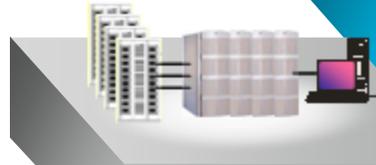
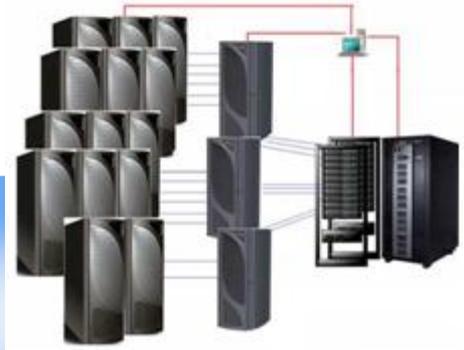


建行数据仓库生产设备的演变

- 硬件
 - 8个NCR 5251节点
 - 32C @733MHZ
 - 32GB 内存
 - 4TB 数据库空间
- 软件
 - OS: MP-RAS 4
 - DB: TD V2R5

- 硬件
 - 6个TD 5450H节点
 - 12C @3.0GHZ
 - 24GB 内存
 - 16TB数据库空间
- 软件
 - OS: MP-RAS 4
 - DB: TD V2R5

- 硬件
 - 18(+1)个TD 5500H节点
 - 36C @2.66GHZ 双核
 - 144GB 内存
 - 100TB 数据库空间
- 软件
 - OS: Suse Linux 9
 - DB: TD V2R6.2

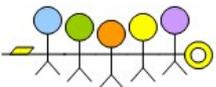


2006年DW上线

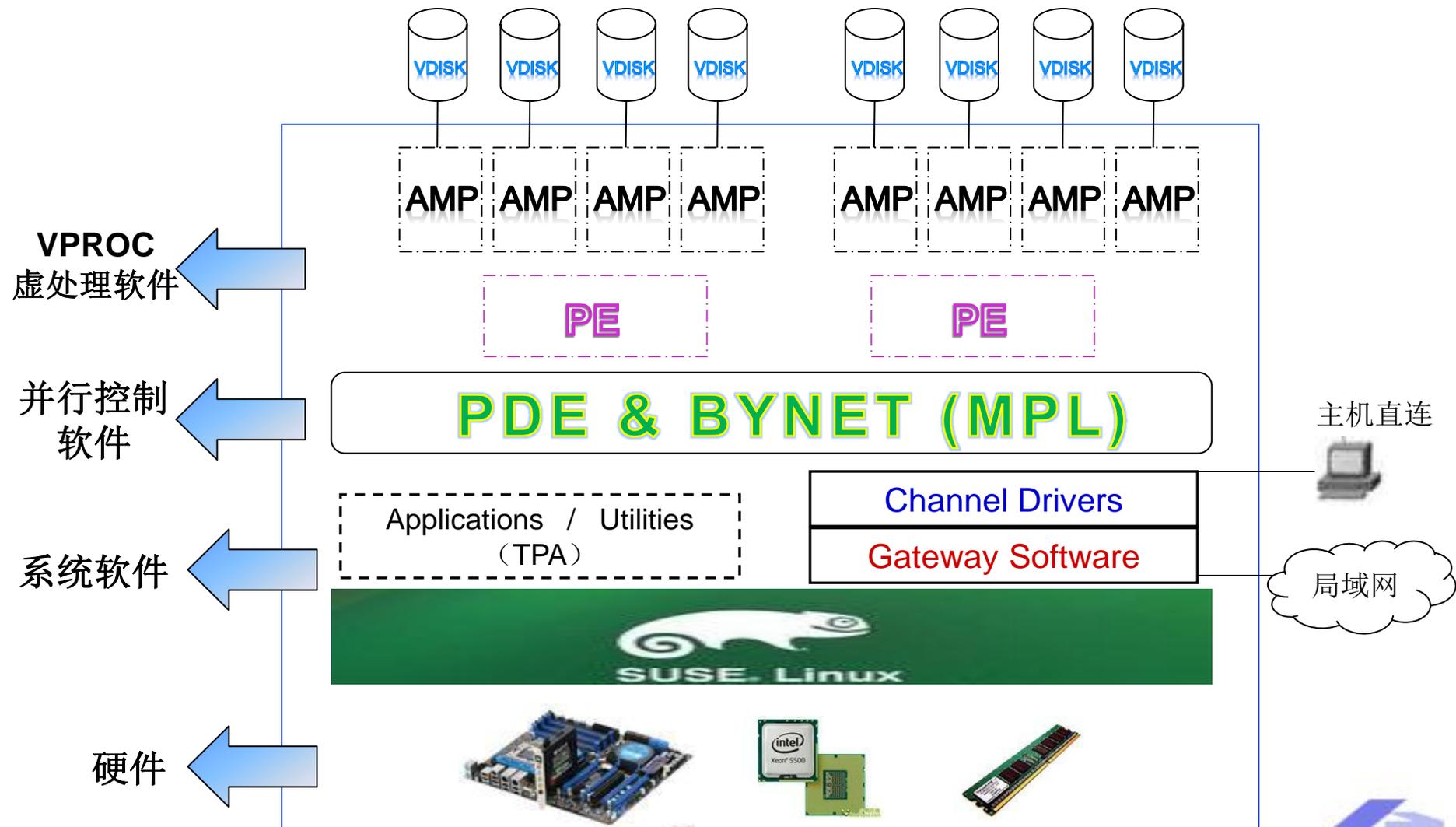
2007年设备更新

2008年设备更新

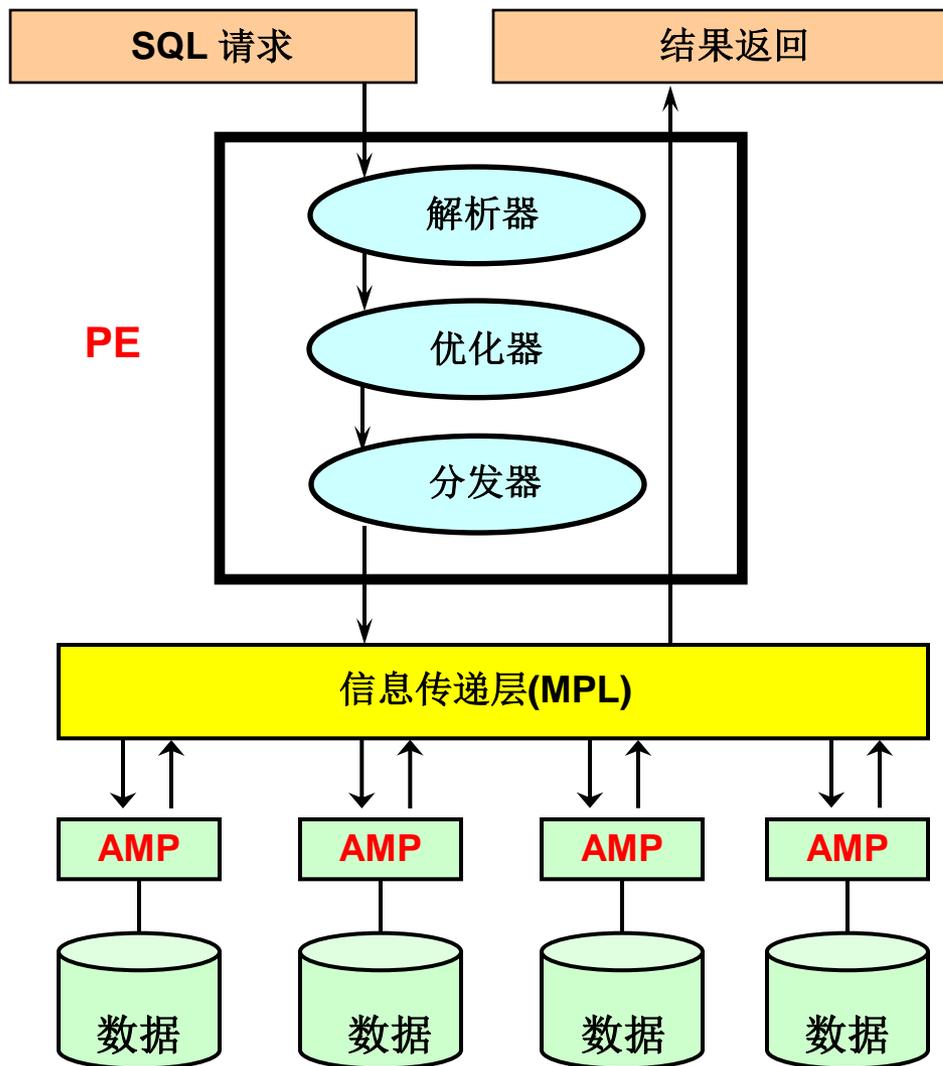




Teradata数据库底层结构



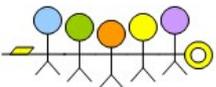
Teradata数据库工作原理



● 基本读写过程

- 解析引擎PE将SQL请求拆分成各AMP的请求以便并行处理
- 解析器分解接收到的SQL交易请求，验证语法、权限等
- 优化器产生最优的查询方案
- 分发所优化的方案到AMP
- 数据通过表PI的HASH值均匀分布到各AMP管理的磁盘（写）
- 信息传递层可汇总各AMP数据，将最终结果返回客户端（读）





AMP (Access Module Processor)

- 一种VPROC，拥有内存和CPU资源，与一个VDISK连接，管理数据库/表的部分数据
- 每节点根据需求可划分多个AMP
- 控制所有磁盘交互及部分数据库的操作，如读、写、转换、格式化等
- 一个请求可以分发到所有AMP一起共同工作，每个AMP也可以同步工作于多个请求
- 各个AMP并行处理，互不干扰，交易处理结果在信息传递层汇总后，直接返回给应用程序



Teradata数据库特点

- 专为海量数据仓库等OLAP应用设计
- 多节点的单一数据库系统
- 跨多代设备线性扩展
- 自动数据分配机制
- 可实现多维并行
- 内嵌分析决策功能
- 采用SPOOL技术
- 易于管理

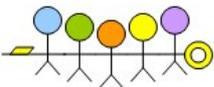


专为数据仓库等OLAP系统设计

OLAP数据库 VS OLTP数据库

	OLAP数据库 (Teradata)	OLTP数据库 (Oracle)
数据来源	本身不产生数据，来源于生产系统中的操作数据	数据在系统中产生
典型业务	基于查询的分析系统	基于交易的处理系统
数据量	复杂查询，经常使用多表连结、全表扫描等，涉及的数据量庞大	每次交易涉及的数据量小
响应速度	响应时间与具体查询有很大关系	对响应时间要求非常高
用户数量	用户数量相对较小，其用户主要是业务人员与管理人员	用户数量非常庞大，主要是操作人员
操作特性	由于业务问题的不固定，数据库的各种操作不能完全基于索引进行	数据库的各种操作主要基于索引进行



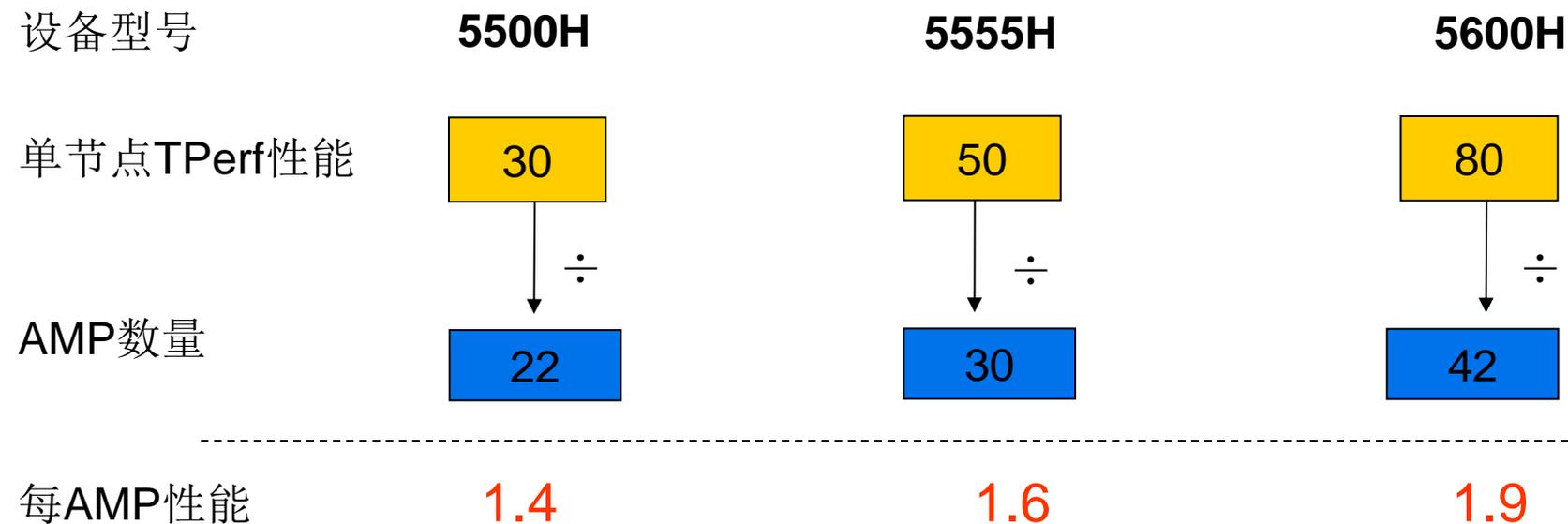


多节点的单一数据库系统

- 可运行于单个或多个节点
- 多个节点组成一个整体的数据库系统，每个结点有单独的IP地址，都连入系统网络
- 各结点之间自动进行负载平衡并提供结点互为备份的高可靠性
- 客户端可以从不同渠道以不同方式连接，连接时可自动实现负载均衡
- 客户端访问的不是某个具体结点，而是整个数据库
- 数据库资源无法从物理上实现完全的分割



不同代设备的线性扩展

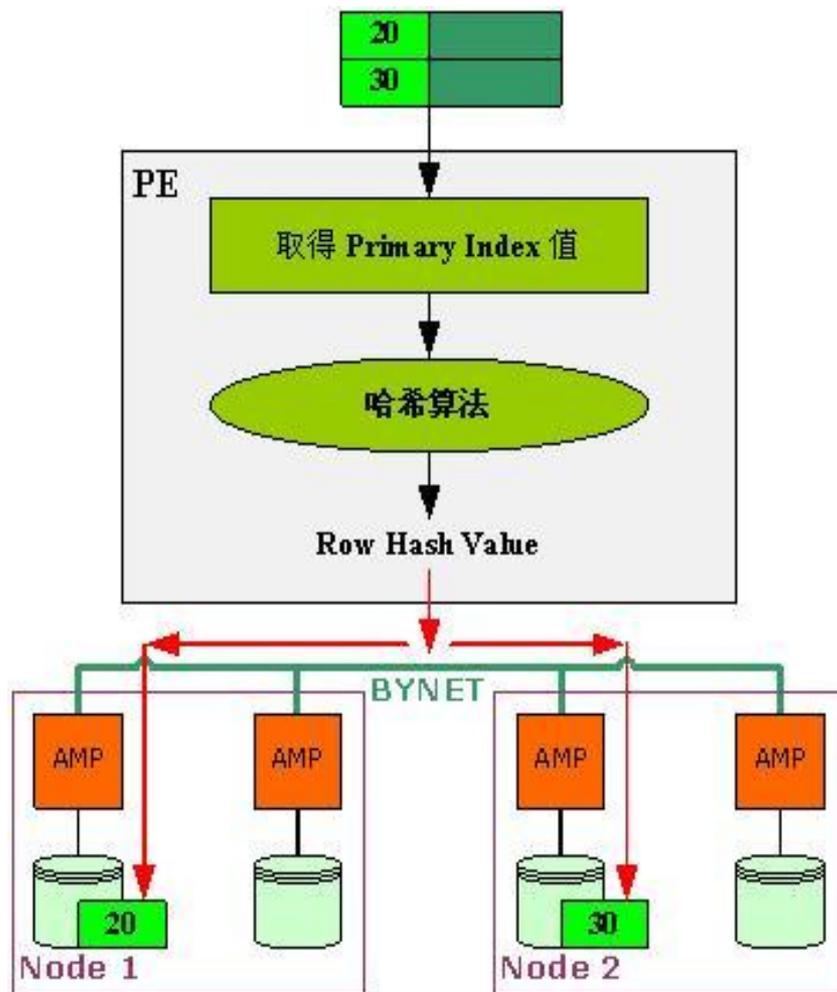


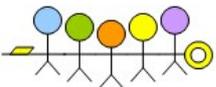
- TPerf值是衡量Teradata设备性能的指标，以第一代型号设备5100性能为基准1，后续型号Tperf是与5100的性能比值
- AMP数量可以根据要求进行增减，但受到磁盘数、背板带宽、接口数量、CPU、内存等限制
- 多代混存会产生资源浪费，一般最多4-5代共存



自动数据分配机制

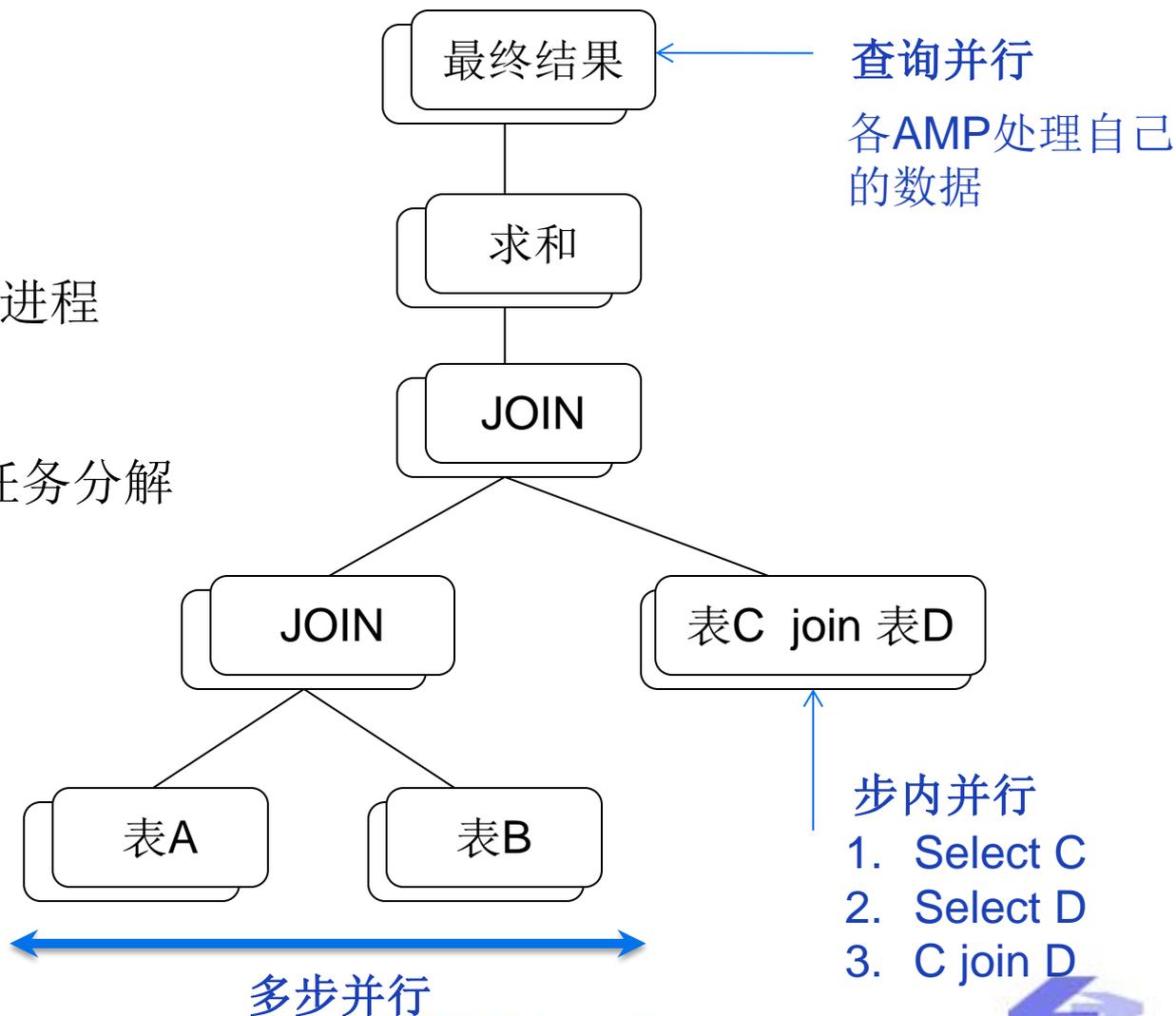
- 通过对PI的哈希运算将数据记录均匀分布到各AMP;
- 记录RowID由行哈希值和一个32位的UV组成;
- AMP根据数据记录的RowID确定物理存储位置;
- 最新TD R13提供了Non-PI表
- 解决了传统数据库的“数据重组”问题





Teradata的多维并行技术

- 查询并行
 - ▶ 多个VPROC并行
- 步内并行
 - ▶ 每个VPROC中多进程
- 多步并行
 - ▶ SQL语句的并行任务分解



- 提供多种OLAP函数

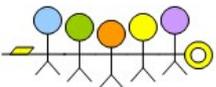
- ▶ 累计和 CSUM、移动平均 MAVG、移动和 MSUM、移动差分 MDIFF、采样 SAMPLE、限定 QUALIFY等
- ▶ 所有函数在Teradata内部以并行方式来工作

- 可以自定义函数UDF

- 可嵌入外部厂商的产品功能

- ▶ SAS、MicroStrategy等BI功能
- ▶ SilkRoute 、 SAP等企业管理功能

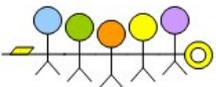




SPOOL技术

- SPOOL是未使用的且连续的数据库空间（类似虚拟内存），与Perm、Temp空间一起以AMP为单位分配，且使用不同Cylinder
- 适合大数据量、并行处理的特点（与传统数据库在内存中处理相比）
- 在工作量适中、无Fallback的系统中，SPOOL最少占总数据库空间的25%—30%
- 好的调优策略可减少对SPOOL空间的占用
- 每个用户的SPOOL的在建立时设置
- SPOOL的类型
 - ▶ Volatile Spool
 - ▶ Intermediate Spool
 - ▶ Output Spool

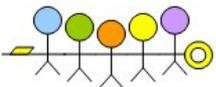




与传统数据库管理复杂度对比

数据库管理任务	OLTP RDBMS	Teradata
物理数据建模	高	低
数据分区定义	高	低
数据分布定义	高	自动
自由空间管理	高	低
数据重组	高	无
索引重组	高	无
工作空间管理	高	自动
查询调整	高	自动
负载管理	高	自动
变换管理	高	低



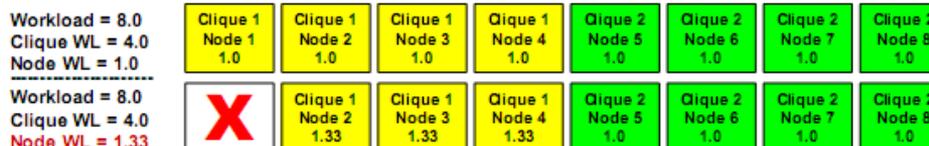


- **Teradata**软硬件体系结构
- **Teradata**数据库原理及特点
- **Teradata**数据保护机制
- **Teradata**系统访问配置及连接方式
- **Teradata**使用中的一些问题及案例分析

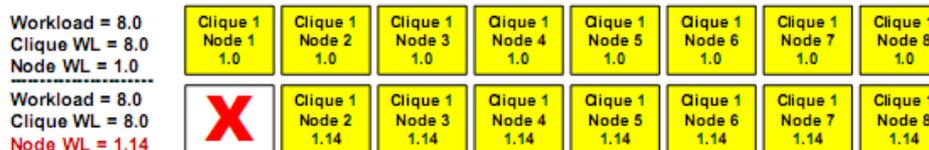


- 定义：共享磁盘阵列的两个或更多的数据库节点
- 一个Clique包含2到8个节点，数量有减少的趋势，热备节点（HSN）逐渐成为标准配置
- Clique中如果一个节点宕机，该节点上的AMP会迁移到其他节点，系统仍然可以运行。
- 迁移过程中数据库会重启，且性能有损失

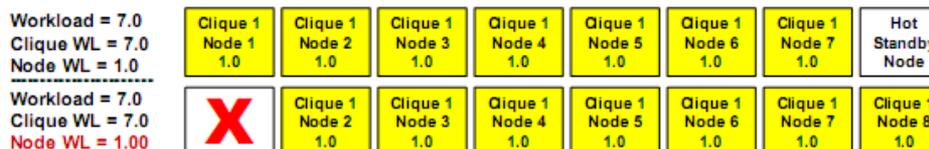
2 High Availability Cliques (4 nodes) – performance degradation of 33% with node failure.



1 Large Clique (8 nodes) – performance degradation of 14% with node failure.



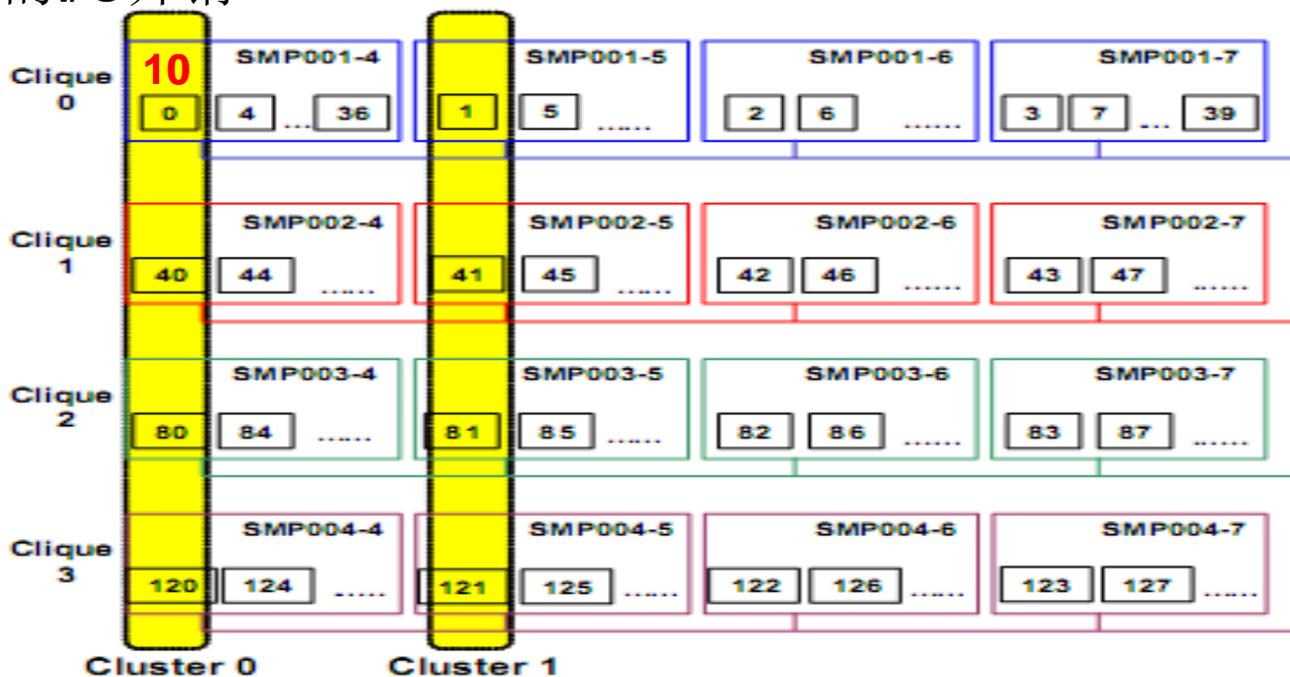
1 Large Clique with Hot Standby Node – performance degradation of 0% with node failure.



10010010100001010010010101001010



- 定义：在同一Cluster中其他的AMP上保存一份相同的记录来达到保护数据的目的
- 如果一个AMP失效，AMP上的数据仍然可用。用户可以继续使用Fallback的表，而不会丢失任何数据
- 有额外的开销，包括两倍的磁盘空间、两倍的Insert、Update、Delete的I/O开销



- 防止多用户在同一时刻修改同一数据而影响数据的完整性
- 请求操作时自动加载，请求完成后自动释放
- 用户是可以改变锁的类型(*LOCKING TABLE TableA FOR Access;*)
- 锁的作用对象：数据库、表、记录
- 锁的类型及优先级

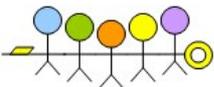
申请的锁	所持有的锁				
	无	ACCESS	READ	WRITE	EXCLUSIVE
ACCESS	接受	接受	接受	接受	等候
READ	接受	接受	接受	等候	等候
WRITE	接受	接受	等候	等候	等候
EXCLUSIVE	接受	等候	等候	等候	等候

- 避免死锁的发生



- 将数据备份到Teradata数据库外部，可以是文件、磁带等形式
- 可选的基本ARC方式
 - ▶ Fastload/FastExport
 - ▶ Arcmain
- 备份工具：Netvault、NetBackup、ASF2等
- 备份方式
 - ▶ 增量备份:以分析日志为基础，需开启Permanent Journal
 - ▶ 全量备份：最安全的方式，但多进程（带机）下无法均分备份任务
 - ▶ Cluster备份：效率最高，但多进程（带机）下风险较大
- 备份文件的分析
- 数据恢复方式





- **Teradata**软硬件体系结构
- **Teradata**数据库原理及特点
- **Teradata**数据保护机制
- **Teradata**系统访问配置及连接方式
- **Teradata**使用中的一些问题及案例分析



- 配置HOST文件，例：

128.64.96.21 *xm*cop1

128.32.101.221 *bj*cop1

128.32.101.222 *bj*cop2

- 配置ODBC数据源（略）

- 配置过程中注意的问题

- ▶ 可为不同部门的用户配置不同的Clique组主机
- ▶ HOST中的主机所有地址务必准确且连通
- ▶ ODBC配置中应填入主机名而非IP地址

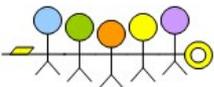


Teradata Server Info

Name[s]
or IP address(es) 128.64.96.227 [dbccop1]

Do not resolve alias name to IP address





连接Teradata的方式

● CLI (Call Level Interface)接口

- ▶ Teradata专用接口，速度快，可并行
- ▶ Teradata代表工具：bteq,fastload,fastexport
- ▶ 可使用接口进行二次开发

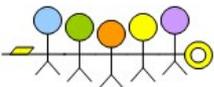
● ODBC/JDBC接口

- ▶ 通用接口，可供外部工具连接
- ▶ Teradata代表工具：Teradata Manager, Administrator...
- ▶ 可进行通用的程序开发

● 嵌入式预处理接口

- ▶ 在高级语言中嵌入SQL语句





- **Teradata**软硬件体系结构
- **Teradata**数据库原理及特点
- **Teradata**数据保护机制
- **Teradata**系统访问配置及连接方式
- **Teradata**使用中的一些问题及案例分析



案例1——数据倾斜

● T05事件主题表倾斜导致整个系统运行效率低

- ▶ 时间：2008年
- ▶ 现象：T05所有脚本运行效率低
- ▶ 原因：T05主题表物理化时直接以主键作PI，虽然从业务角度看是合理的，但忽视了源系统数据库与Teradata技术特性上的差异，导致PI设置不合理而发生数据倾斜。

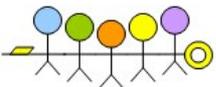
● Accessright系统表倾斜导致整个系统运行效率低

- ▶ 时间：2007年
- ▶ 现象：ODS数据加载效率低
- ▶ 原因：由于dbc.accessrights 系统表的PI为(userid,databaseid)，ODS采用一个用户dwjob2向dwsdata2加数，导致大量权限记录分布在1个AMP上

	UserName	DatabaseName	count	HashValue	HashedAmp
1	dwbpsuser	dwEPMart	191455	0000F8090000F609	229
2	dwjob10	dwSDATA2	24650	00007E0700007C04	191
3	dwjob11	dwSDATA2	22780	00008B0700007C04	244

1001001010000101001010010101001010





附：Teradata的一些限制

注:V2R6及以前版本

限制项	值
SQL长度	1 MB
活动的事务数量	2,048
数据包最大长度	65,104
每个消息的数据包数量	256
字符串常量的长度	255
数据格式描述长度	30
每个PE处理的最大会话数	120
每个AMP最多控制的磁盘	64
每个AMP能并发执行的最大任务数	80
并发FLOAD/FEXP/MLOAD任务数量	15
系统中Vprocs最大数量	16,384
每个节点中Vprocs最大数量	128
每AMP可管理的最大数据量	1.3 TB
表中的字段最大数量	2,048
表中的大对象LOB最大数量	32



