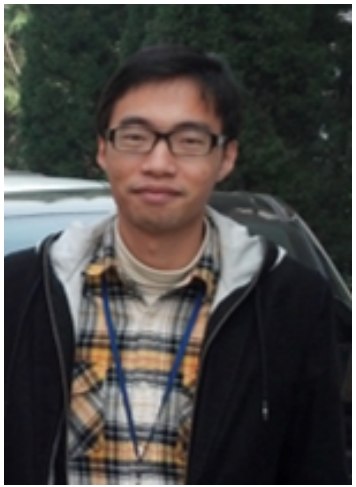

ParnassusData™

建立高可用MySQL数据库

by Biot Wang, May 2015

汪伟华



- 8年Oracle相关开发及数据库运维经验
  (Oracle DB, MySQL, Oracle Apps)
- 11g OCM
- MySQL OCP
- 上海Oracle用户组核心成员

ORACLE
Certified Professional
MySQL 5.6 Database
Administrator

SH'OUG
SHANGHAI ORACLE USERS GROUP
上海ORACLE用户组

Parnassus
诗 檀

- E-mail：biot.wang@parnassusdata.com

公司介绍

- 诗檀软件专注于数据服务
  - 对Oracle, MySQL, Oracle EBS提供远程数据库管理服务及咨询
  - 提供解决方案并进行安装、升级、迁移及运维
  - 数据救援(PRM工具)及优化

ORACLE Silver Partner

- 专业团队
  - 全天候的DBA专家服务
  - 24/7/365 DBA远程支持咨询，系统管理及特定项目紧急回复

- 服务客户
  - 现服务客户主要有在线大型电商，金融机构，政府部门及企事业单位。

议程

- Oracle -> MySQL数据迁移
- 可选MySQL HA架构
- 备份恢复

Oracle -> MySQL数据迁移

- MySQL数据迁移
  - 自动迁移：MySQL Migration Toolkit，Navicat等
  - 手工迁移：
  - 通过oracle相关对象查询了解情况(表，视图，主外键索引等)，并在mysql中建立表结构
  - 可使用sqldeveloper导出csv文件并修改处理后使用MySQL LOAD DATA …INFILE命令导入
  - 对oracle和mysql不兼容的列结构使用拼接后的insert语句进行批量插入。

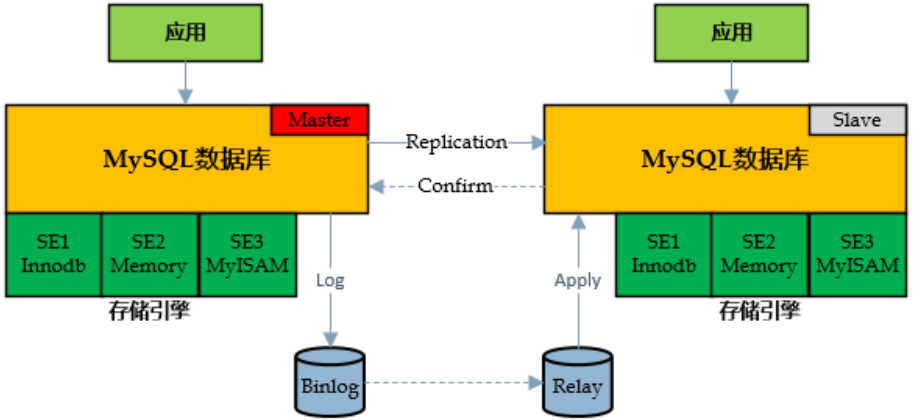
- 注意对应转换
  - 字段类型的变化，如：
    - Integer, Number => smallint, mediumint, decimal(10,2)
    - VARCHAR2 => VARCHAR
    - DATE, TIMESTAMP => DATETIME等
  - 字符集 / VARCHAR 表现 (空格处理不同)
  - Sequence => AUTO INCREMENT
  - InnoDB存储=> 每张表一个文件
  - 存储过程及视图，触发器等需要修改重建

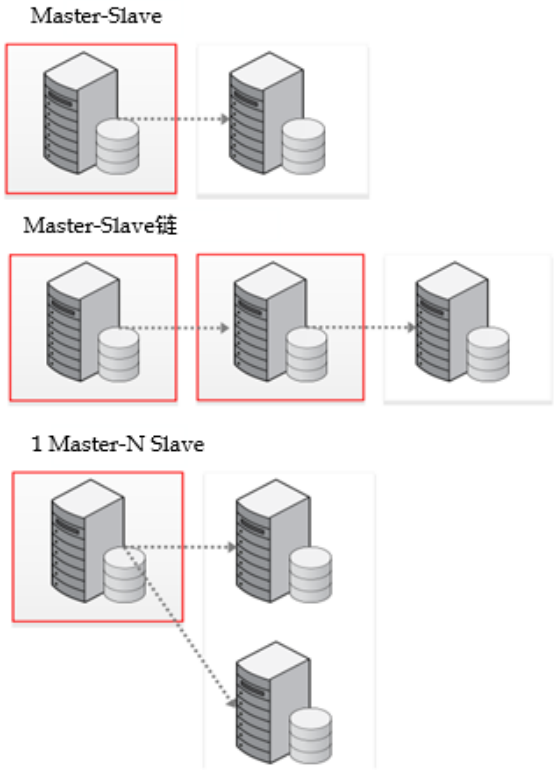
- 数据导出导入
  - ARCHIVE部分（低成本存储）：可之后分别导入
  - LIVE部分
  - 导入加速 （禁用binlogs/建立index及constraint等）

可选MySQL HA架构

- MySQL replication
  - 可能会丢失部分数据 (几秒), 可靠性一般
  - 需要两倍存储空间
  + 可扩展
- DRBD/Pacemaker/Corosync/Linux(过去的DRBD/heartbeat/Linux架构)
  - 受到SYNC模式的性能影响
  - 需要两倍存储空间（且备机不可用）
  - 不可扩展 (仅主 + 镜像)
  - 可扩大LVM，不过需要主备同时扩大，同时需要设置用以识别扩大的空间
  + 高可靠性
  + 此解决方案也同时被Oracle企业版支持并采纳
- LVS+MySQL Cluster
- 相关架构特点对比:
  http://dev.mysql.com/doc/mysql-ha-scalability/en/ha-overview.html
- 其他架构(MySQL Fabric, Heatbeat+共享存储, MHA架构)等

MySQL replication

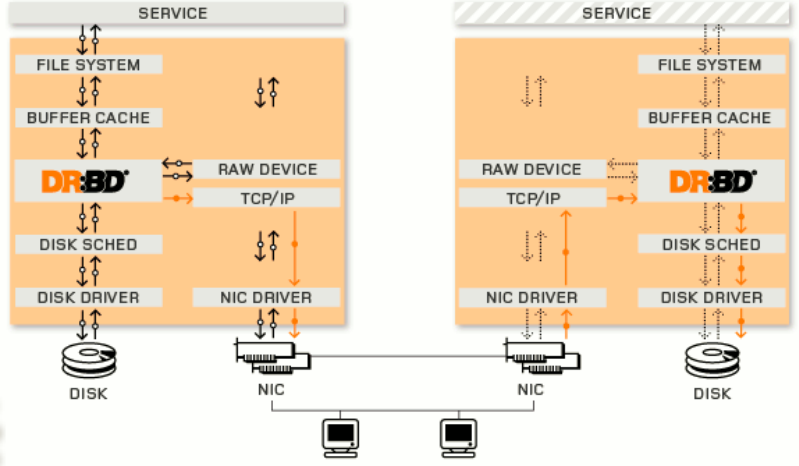


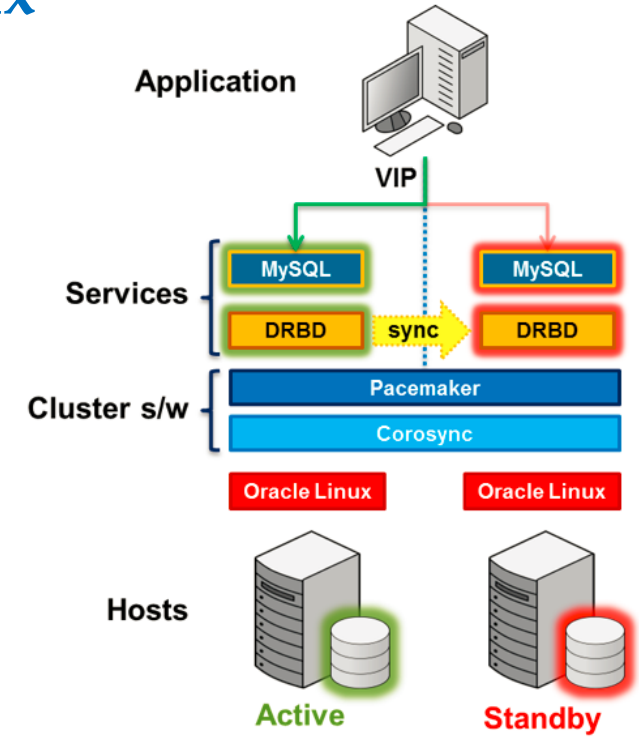


- 全局事务ID (GTID, 同步模式为Binary, 非Row)
- Crash-Safe多线程Slaves
- Group Commit
- Replication Checksums
- Binlog API

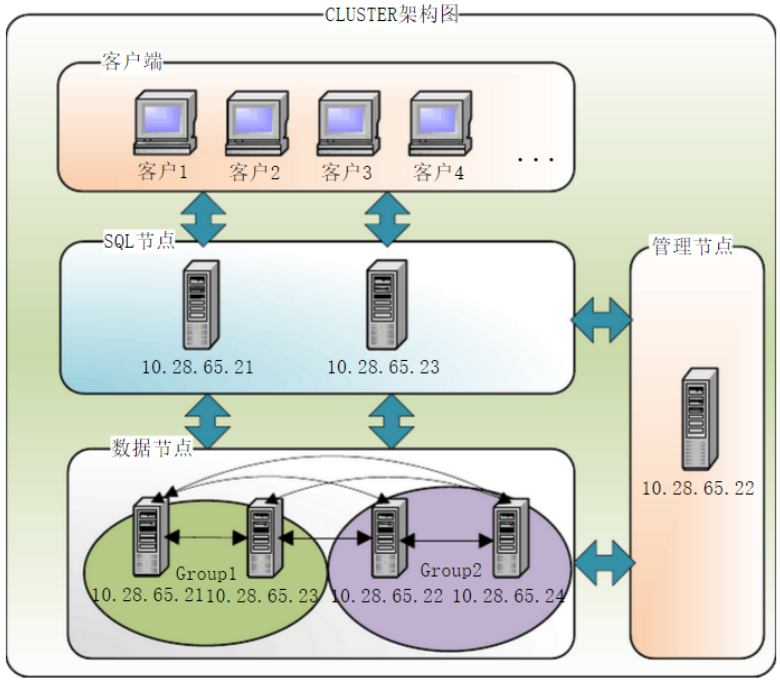
DRBD/Pacemaker/Corosync/Linux

- Corosync控制主备资源切换
- DRBD类似主机热备
- 主备库不共享
- Virtual IP / VIP

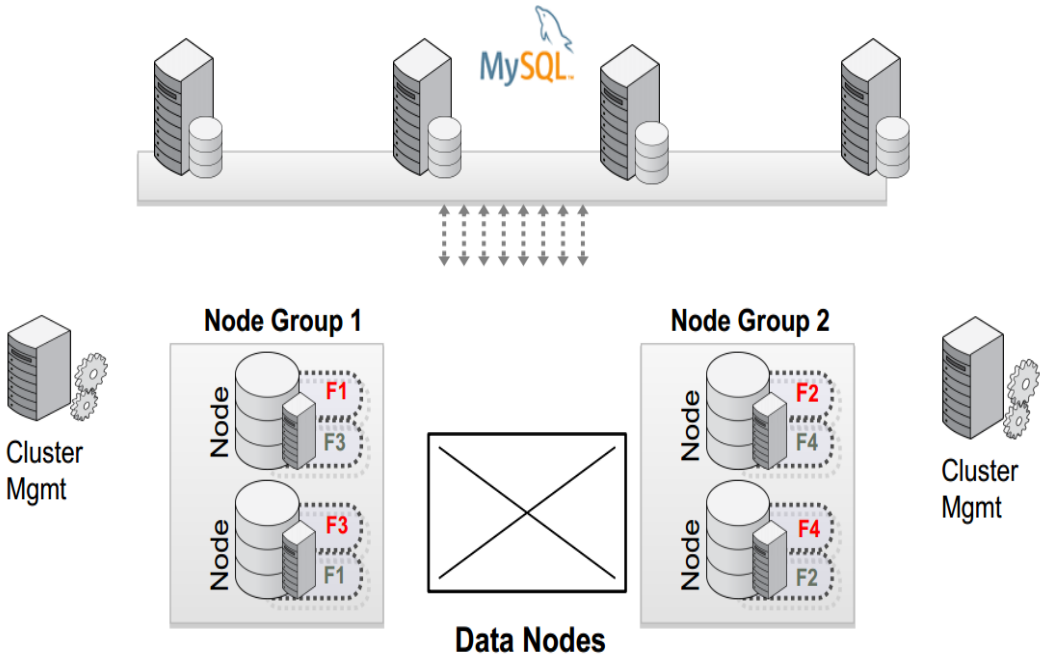




LVS + MySQL Cluster架构



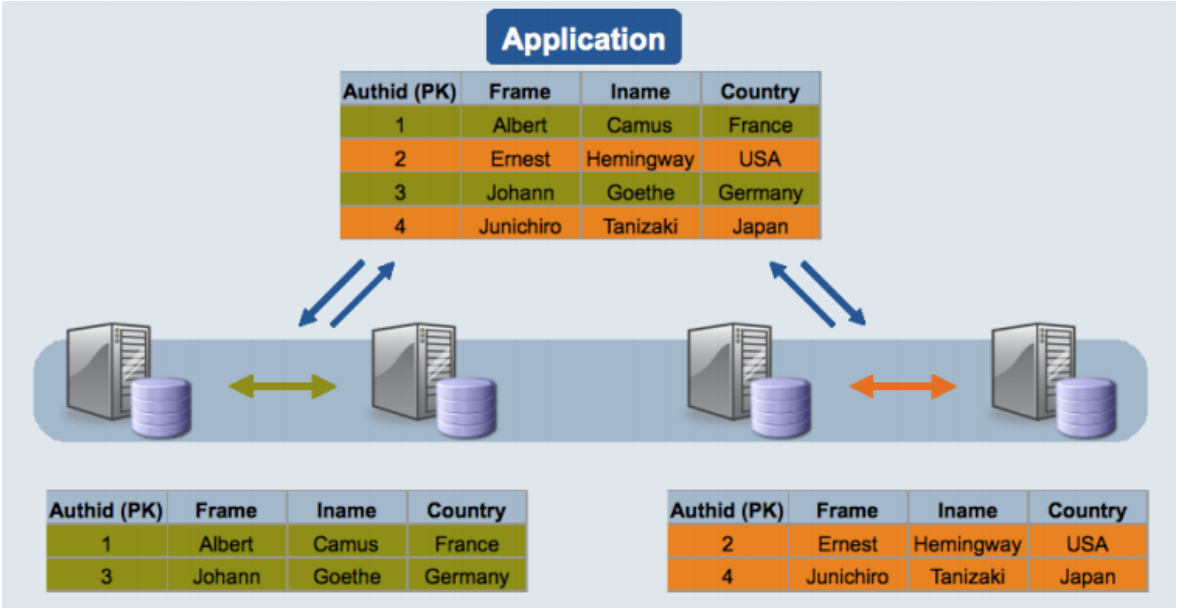
MySQL NDB Cluster – Shared Nothing



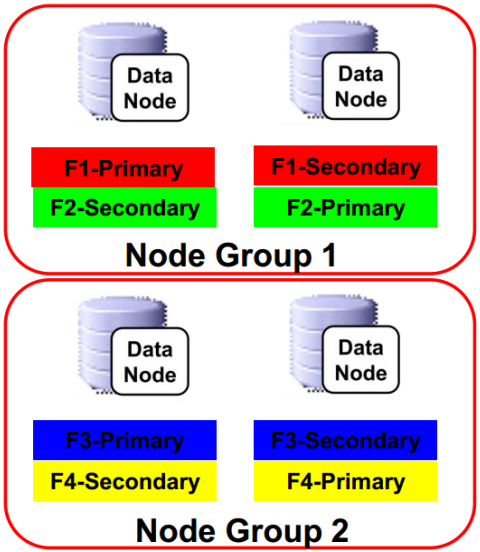
优点:

- 分布式、无共享架构:
  集群中的每个节点都是冗余的，可以放在单独的主机上，从而确保在发生进程、硬件或网络故障时的持续可用性。
- 无单点故障
- 同步复制
- 自动故障切换
- 多站点集群

MySQL NDB Cluster – 分片



ID	Capital	Country	UTC	
1	Copenhagen	Denmark	2	Partition 1
2	Berlin	Germany	2	
3	New York City	USA	-5	Partition 2
4	Tokyo	Japan	9	
5	Athens	Greece	2	Partition 3
6	Moscow	Russia	4	
7	Oslo	Norway	2	Partition 4
8	Beijing	China	8	



MySQL NDB Cluster

缺点：

- 对需要进行分片的表需要修改引擎Innodb为NDB，不需要分片的可以不修改。
- NDB的事务隔离级别只支持Read Committed，即一个事务在提交前，查询不到在事务内所做的修改；而Innodb支持所有的事务隔离级别，默认使用Repeatable Read，不存在这个问题。
- 外键支持：外键性能有问题（因为外键所关联的记录可能在别的分片节点中），所以建议去掉所有外键。
- Data Node节点数据会被尽量放在内存中，对内存要求大，如果内存不够用会导致性能大幅下降。

优点：

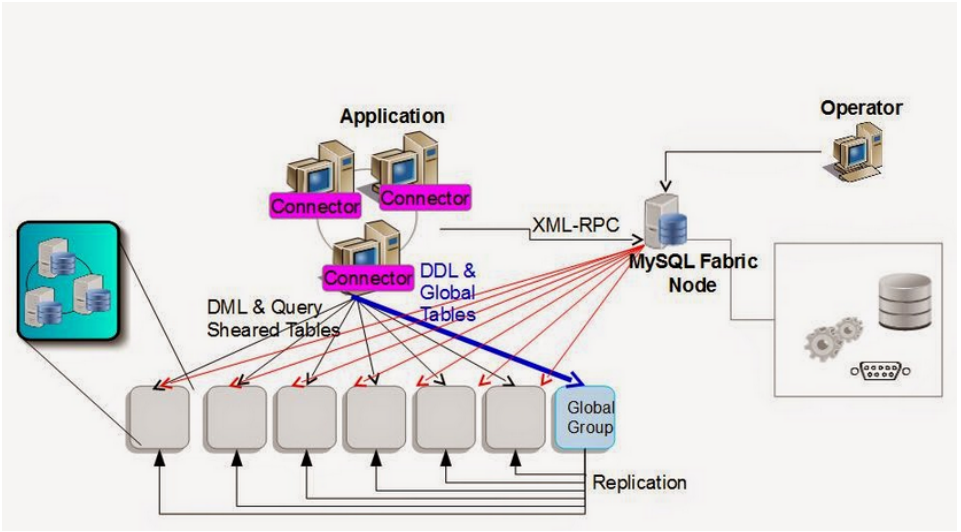
- 可将数据分布于多地
  – 在多地进行同步复制(Synchronous replication)和自动故障切换(auto-failover)
- 是一个无冲突处理的Active-Active双活方案

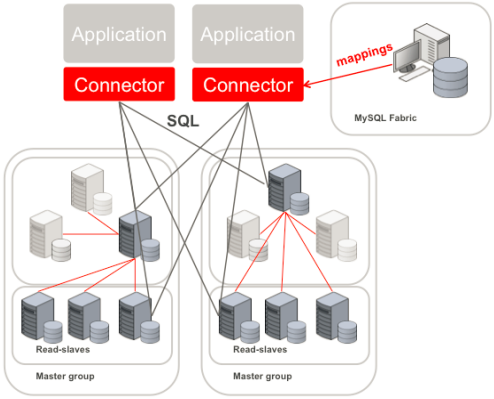
其他架构

- 1. MySQL Fabric
  - 2014年年中发布的新解决方案
  - 需要使用新的Connector API应用接口来访问
  - MySQL Fabric Node管理整个MySQL Farm
  - 开源并基于MySQL Replication
  - 可自动分片和主备切换

优缺点：

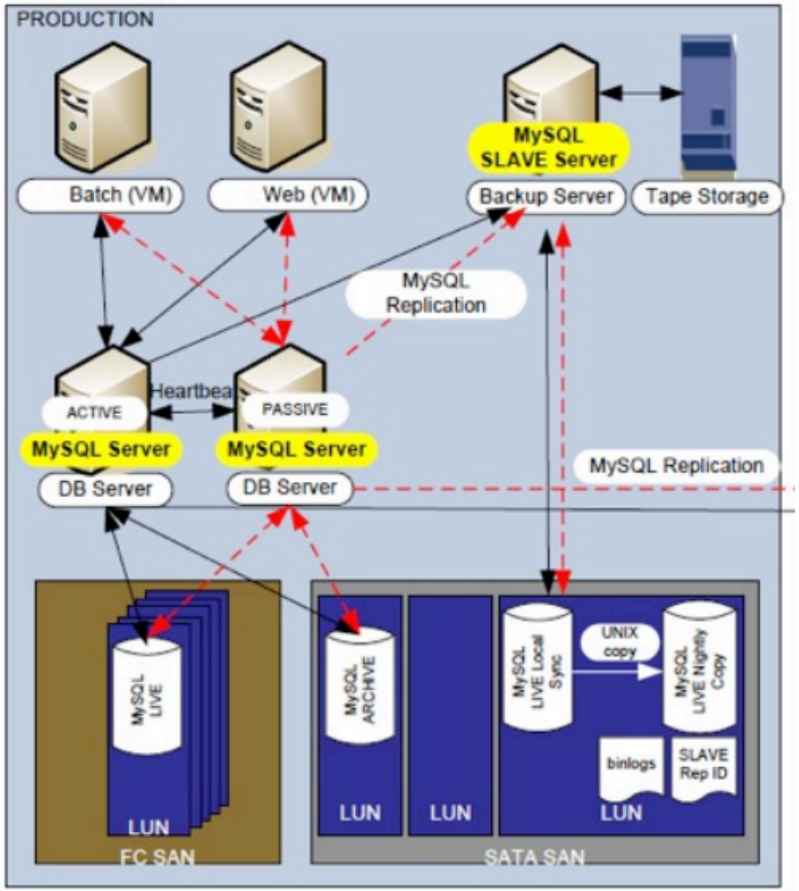
- 自增长键不能作为分片的键；
- 事务及查询只支持在同一个分片内，事务中更新的数据不能跨分片，查询语句返回的数据也不能跨分片。
- 当前为止还没有成熟的实际实施案例研究。





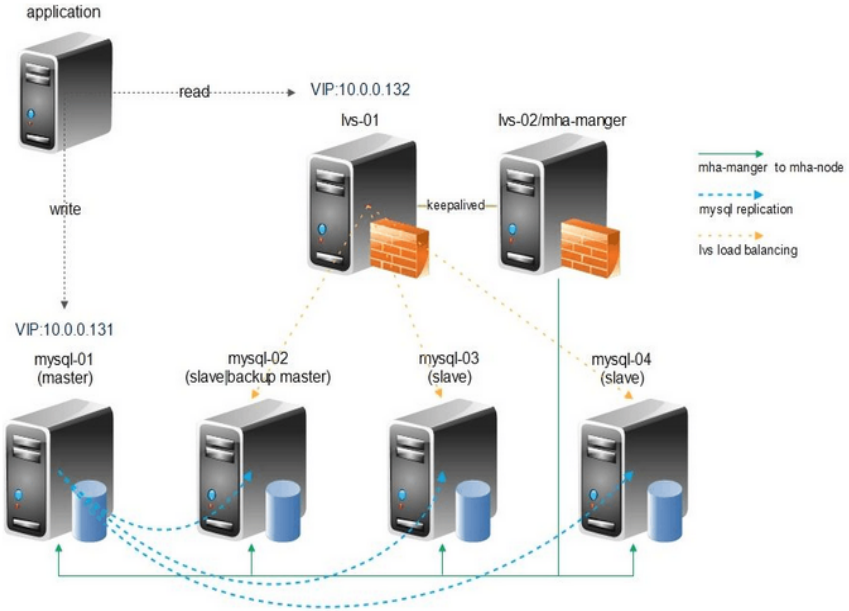
其他架构

- 2. Heartbeat+共享存储HA架构
  - Heartbeat控制资源
  - 共享存储
    - LUN's accessible from two servers
    - ext3 – 仅mount到活动的节点
    - no LVM - LVM is not clustered
  - Virtual IP / VIP
  - MySQL 5.7实例运行在一个节点上
    - read-write数据必须为InnoDB
    - read-only数据可以是MyISAM
  - 优缺点
    - 共享一个数据文件
    - 一旦主库down, 切换备库恢复时间长(10min多分钟+)



其他架构

- 3. Lvs+KeepAlived+MHA+MySQL架构
  - 1）从宕机崩溃的Master保存二进制日志事件（binlogevent）；
  - 2）识别含有最新更新的Slave；
  - 3）应用差异的中继日志(relaylog)到其他Slave；
  - 4）应用从Master保存的二进制日志事件；
  - 5）提升一个Slave为新的Master；
  - 6）使其他的Slave连接新的Master进行复制；

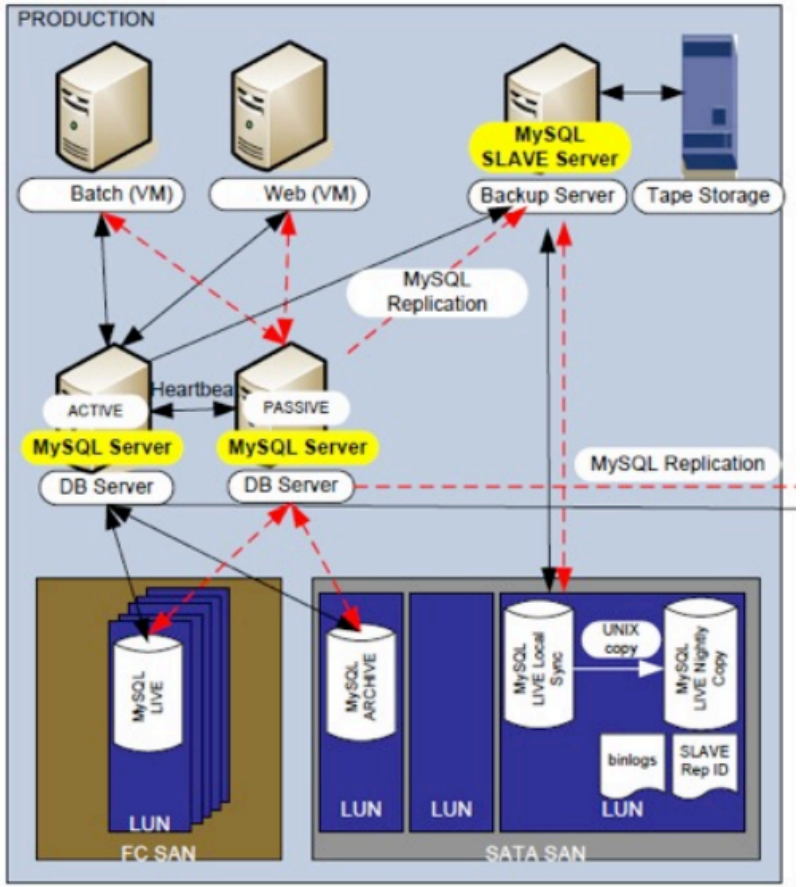


优缺点

- 需要至少半同步复制以降低数据丢失风险
- 故障切换速度快(0~30s)
- MHA能最大程度保证数据一致性
- 注意：MySQL服务挂了，但是可以从服务器拷贝二进制。但如果硬件宕机或者SSH不能连接，不能获取到最新的binlog日志，如果复制出现延迟，会丢失数据。

备份恢复

- 分为LIVE和ARCHIVE
  - LIVE - InnoDB 200-500GB
  - ARCHIVE - MyISAM 2 TB
- LIVE备份 - on slave
  - FLUSH ... WITH READ LOCK
  - 停止slave SQL thread
  - LVM snapshot 或RSYNC
  - 或者使用mysqldump，mysqlbinlog进行备份转储
- 恢复
  - LIVE first as a whole instance
  - ARCHIVE later - it's MyISAM



专注于数据

PARNASSUSDATA

软件，方案，服务供应商


ParnassusData™