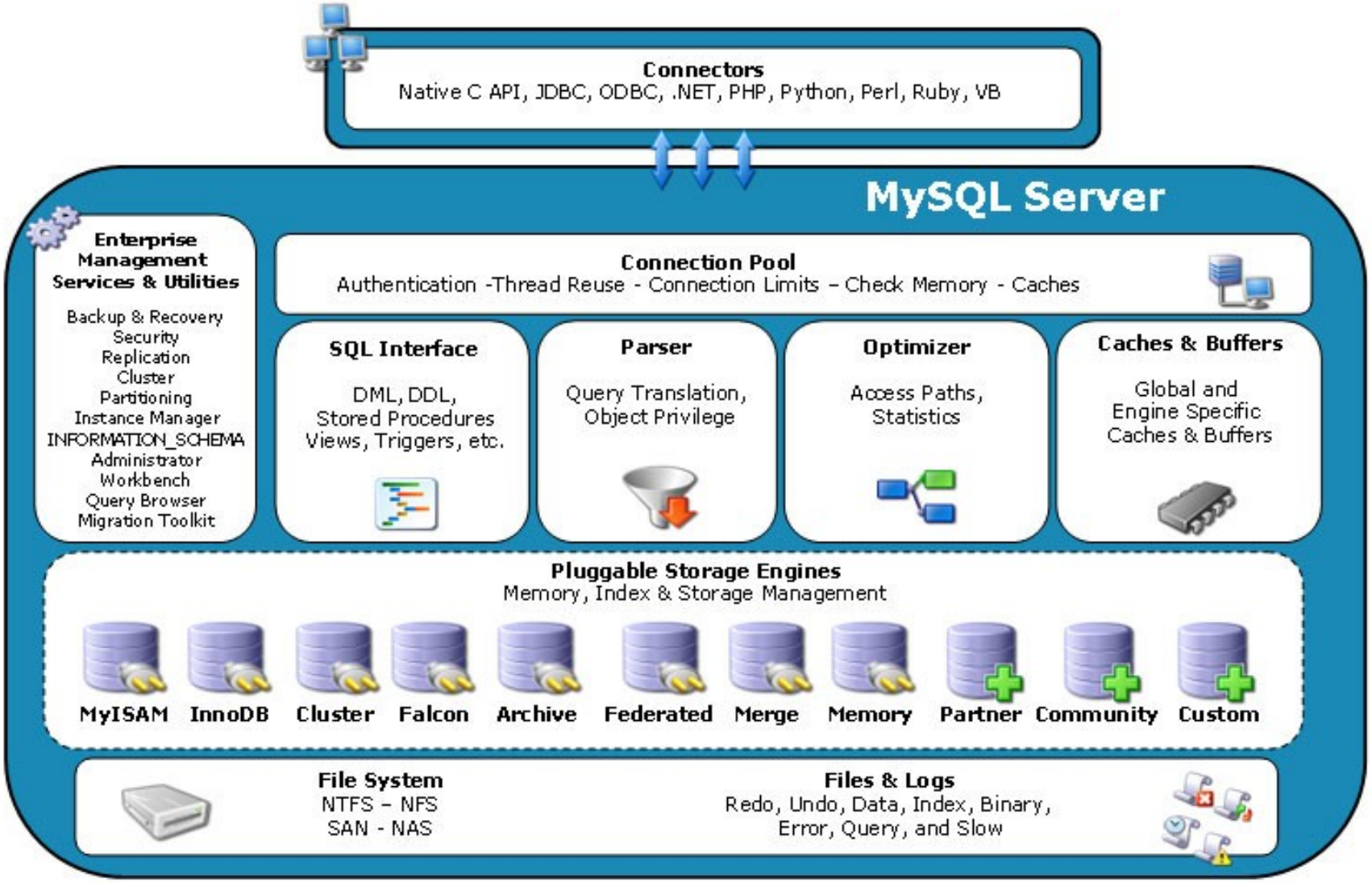
MySQL
Bests practices on Linux

Serge Frezefond
serge.frezefond@mysql.com
Dir. Technique
SUN / MySQL France
Solution Linux 2009 Paris, 02-04-2009

Agenda

- Architecture MySQL
- DRBD Heartbeat
- IO / File systems / ZFS
- LVM for backups
- Secure replication : checksum through SSL
- Virtualization
- Monitoring / Tuning
- Linux memory : locking , swappiness
- Dev MySQL on Linux.
- Conclusion / Q&A

MySQL Architecture



MySQL : Storage Engines

Partenaires

InnoDB

SolidBD for MySQL

InfoBright – Brighthouse DWH

NitroEDB

En cours :

ScaleDB

AmazonS3

MemCache

PrimeBase XT

Moteur de stockage de Thinking Networks

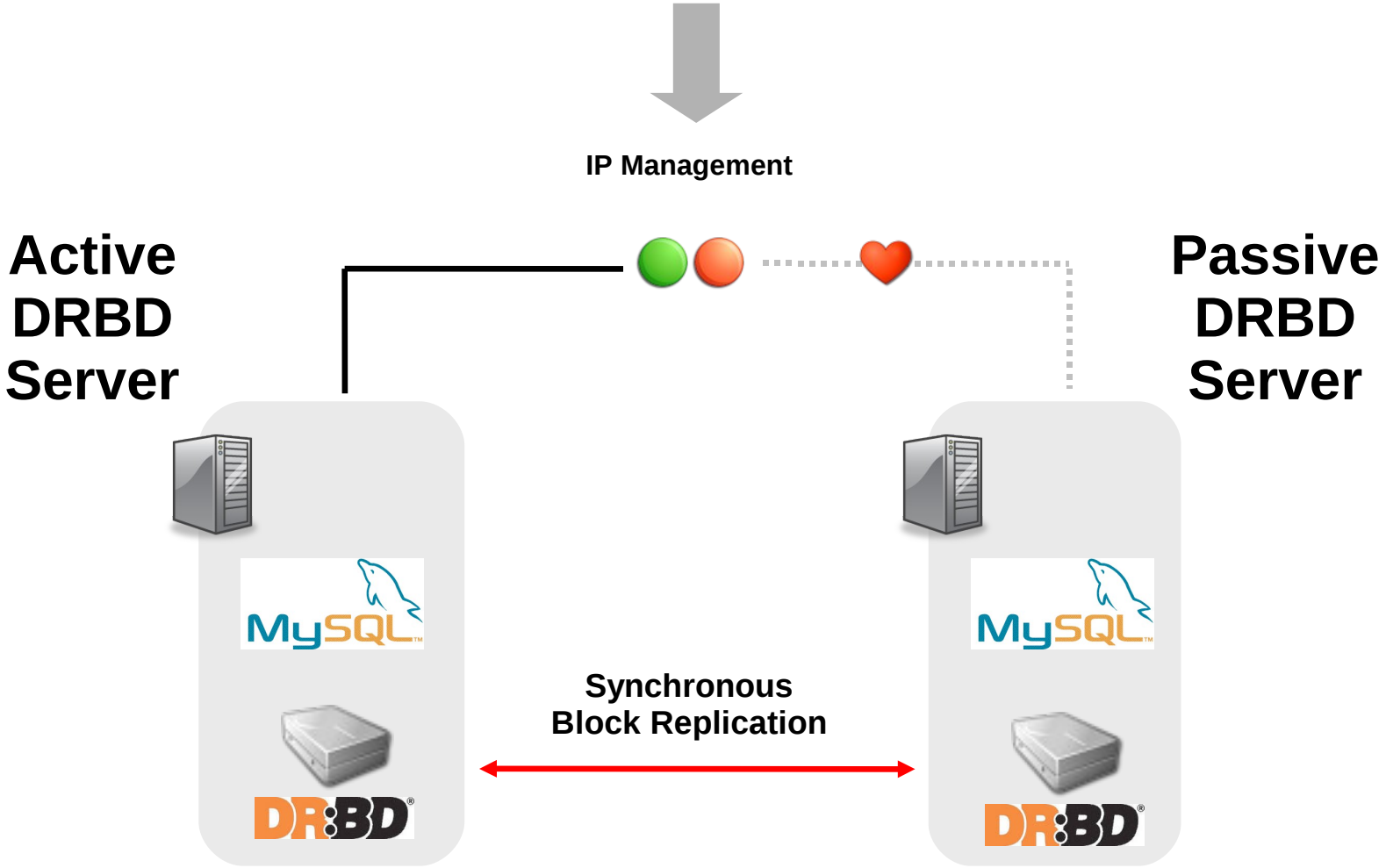
OpenOLAP

Communautaire

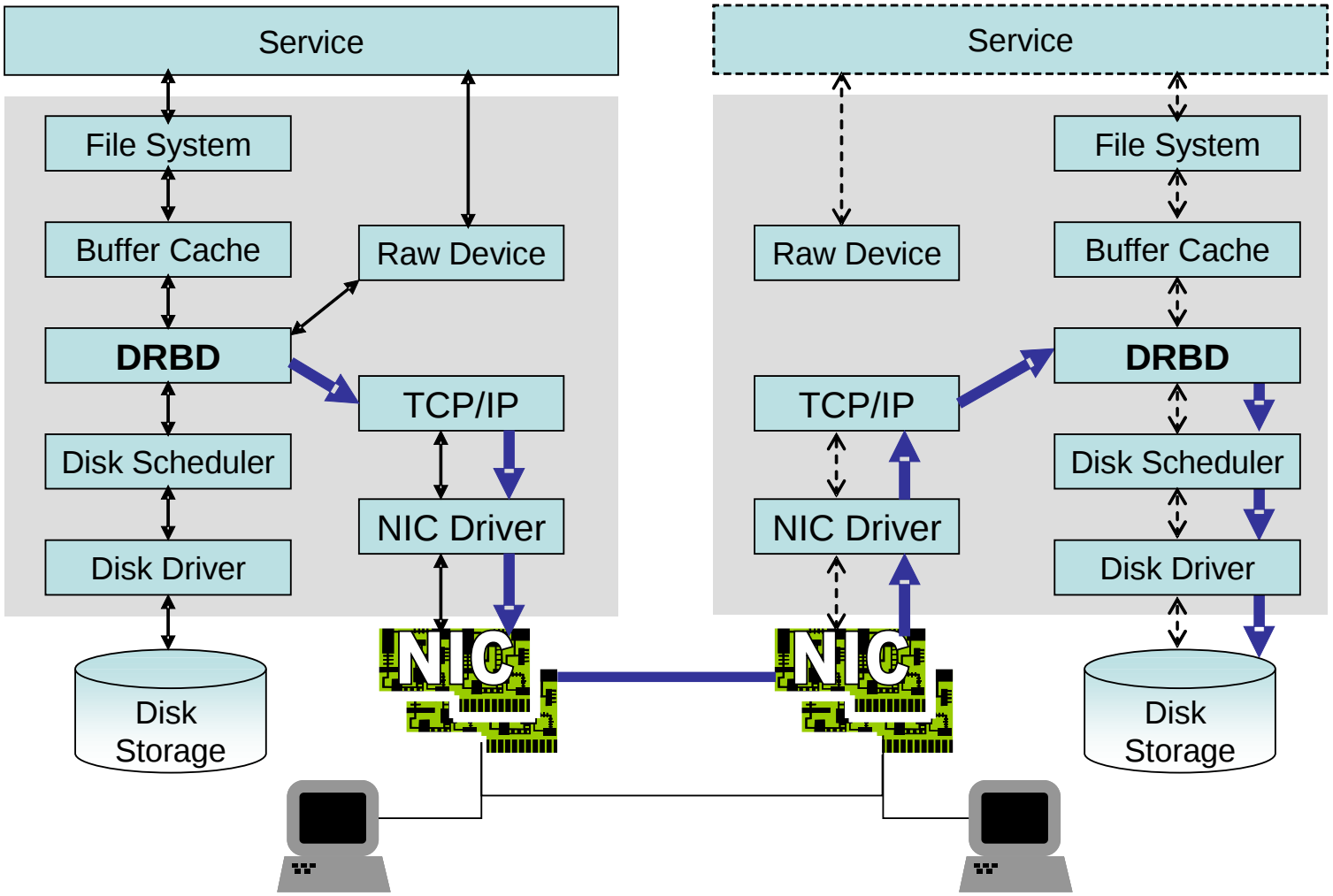
MySQL easy install on linux

- Rpm + repo
- Yum install
- Apt-get install
- Or install from a tar.gz
  - allow any case of installation : multiple base dir / multiple instances per base dir
- Automatic startup :
  - /etc/init.d/ + chkconfig (gestion des niveaux de rcx.d)

MySQL, Heartbeat & DRBD Cluster

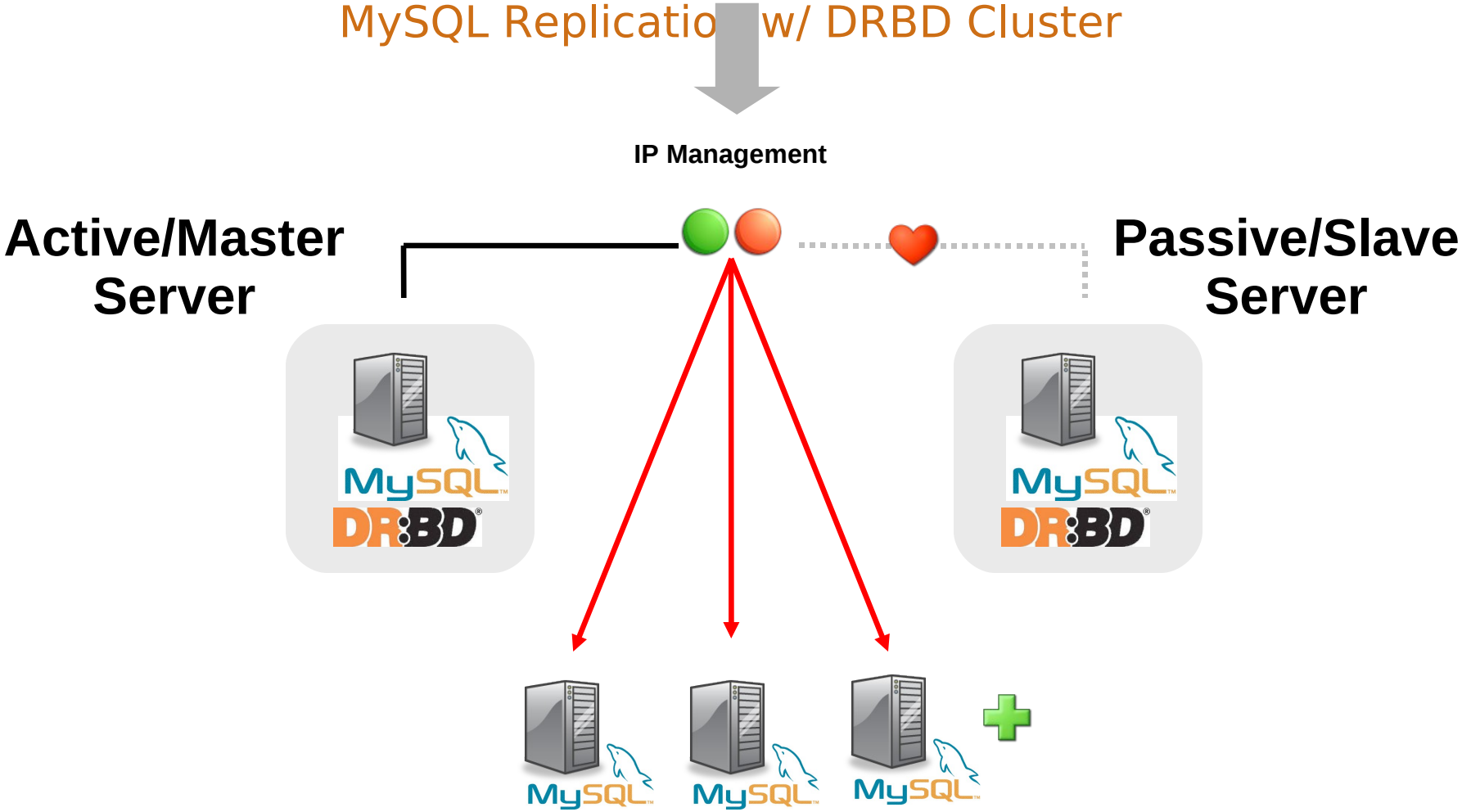


DRBD


Service
File System
Buffer Cache
DRBD
Disk Scheduler
Disk Driver
Raw Device
TCP/IP
NIC Driver
NIC
Disk Storage
Service
Raw Device
TCP/IP
NIC Driver
NIC
File System
Buffer Cache
DRBD
Disk Scheduler
Disk Driver
Disk Storage

Scale Out

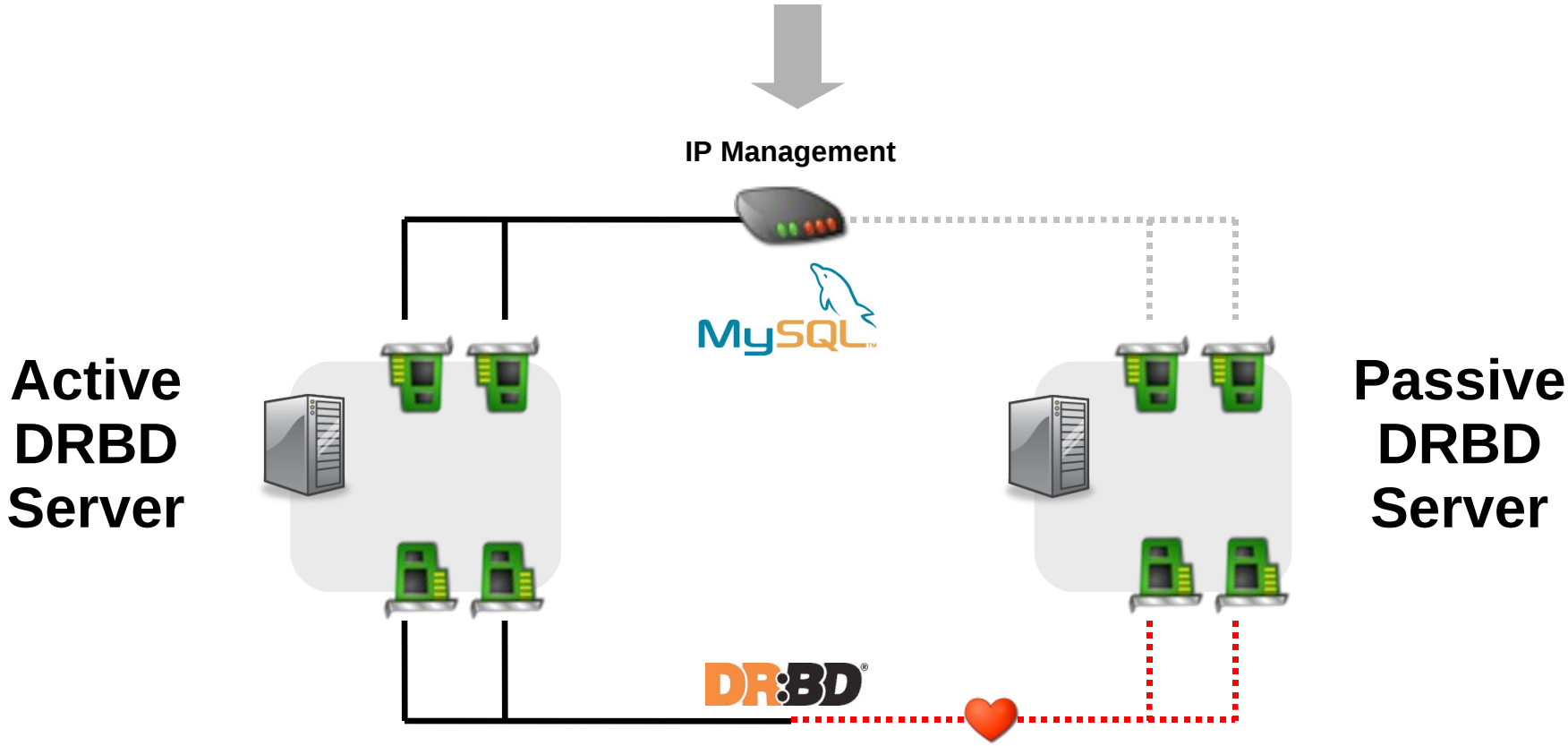
MySQL Replication w/ DRBD Cluster



MySQL Replication Slaves – Read Scalability - Asynchronous

DRBD : Setup & Configuration

- Multiple NICs will increase redundancy
- MySQL traffic switched over public LAN
- DRBD & Heartbeat over private LAN



Setup & Configuration - DRBD

- Install DRBD package
- Edit and copy configuration file (*drbd.conf*) to nodes
- Choose primary node
- Synchronize the underlying devices
- Device is now ready
  - Create a file system if one does not exist
- Mount DRBD on primary
- Heartbeat handles
  - Changing of DRBD primary or secondary status
  - Mounting and unmounting volumes

```
global {
minor-count 1;
}
resource mysql {
protocol C;
on node0.ka6wke.net {
device /dev/drbd0; # The name of our drbd device.
disk /dev/sdb1;    # Partition we wish drbd to use.
address 192.168.12.21:7788; # node0 IP address and port number.
meta-disk internal; # Stores meta-data in lower portion of sdb1.
}
on node1.ka6wke.net {
device /dev/drbd0; # Our drbd device, must match node0.
disk /dev/sdb1;    # Partition drbd should use.
address 192.168.12.22:7788; # IP address of node1, and port number.
meta-disk internal; #Stores meta-data in lower portion of sdb1.
}
```

Admin DRBD

Sur node 1

drbdadm create-md mysql

drbdadm -- --overwrite-data-of-peer primary mysql

mkfs.ext3 -L mysql /dev/drbd0

drbdadm secondary mysql

Sur node 2

drbdadm primary mysql

[root@node1 ~]# mount /dev/drbd0 /mnt/mysql

Heartbeat

Fichier haresource por DRBD :

IPaddress::192.168.12.30/24 - Runs /etc/ha.d/resources.d/IPaddress 192.168.12.30/24 {start,stop}

drbddsk::mysql - Runs /etc/ha.d/resources.d/drbddsk mysql {start,stop}

Filesystem::/dev/drbd0::/mnt/mysql::ext3::defaults - Runs /etc/ha.d/resources.d/Filesystem /dev/drbd0 /mnt/mysql ext3 defaults {start,stop}

mysqld - Runs mysqld {start,stop}

DRBD : Setup & Configuration - MySQL

- Ensure all MySQL files are installed on the DRBD volume
- Create ‘mysql’ group and user
- Create MySQL directory
- Install MySQL
- Shutdown MySQL
- Unmount the DRBD volume

Clustering solutions under Linux

- Heartbeat
- Red Hat Cluster
- HP Service Guard
  - Monitor Mysql
- Can be complemented by a cluster filesystem

OCFS , GFS

Heartbeat + mon

```
 cat > /etc/ha.d/haresources
master 192.168.0.2 mysqld mon

cat > /etc/ha.d/ha.cf
logfile /var/log/ha-log
keepalive 2
deadtime 10
initdead 20
bcast eth0
node master.mydomain.com
node slave.mydomain.com
# Mettez cette valeur à "on" seulement si vous êtes dans un setup multi-
    master
auto_failback off
# Nous allons pinger la passerelle pour vérifier la connectivité réseau
ping 192.168.0.1
respawn hacluster /usr/lib64/heartbeat/ipfail
```

Heartbeat + Mon

```
 cat > mon.cf

# IP virtuelle
hostgroup mysql_servers 192.168.0.2
watch mysql_servers
        mysql
                interval 1m
                monitor mysql.monitor
                period wd {Mon-Sun}
                        alert bring-ha-down.alert
                        alert mail.alert -S "Host1 MYSQL est tombé"
    admin@example.com
                        upalert mail.alert -S "Host1 MYSQL le serveur est en ligne"
    admin@example.com

cat > /usr/local/mon/alert.d/bring-ha-down.alert
/etc/rc.d/init.d/heartbeat stop
```

Partitionement des données

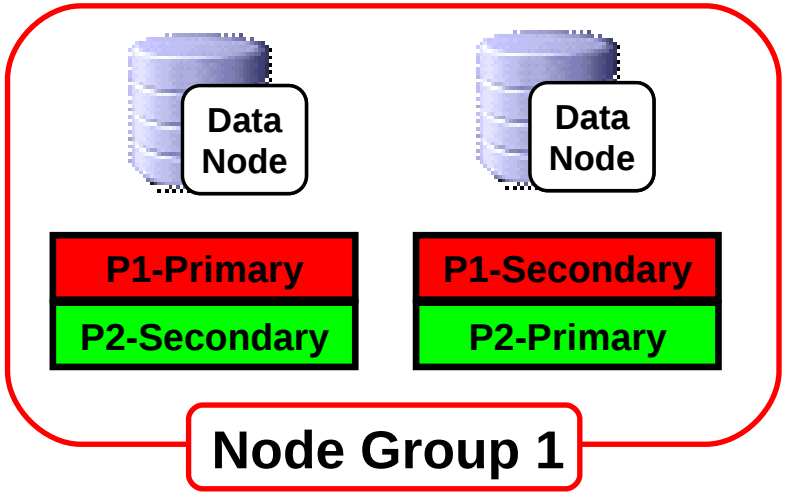
ID	Capital	Country	UTC
1	Copenhagen	Denmark	2
2	Berlin	Germany	2
3	New York City	USA	-5
4	Tokyo	Japan	9
5	Athens	Greece	2
6	Moscow	Russia	4
7	Oslo	Norway	2
8	Beijing	China	8

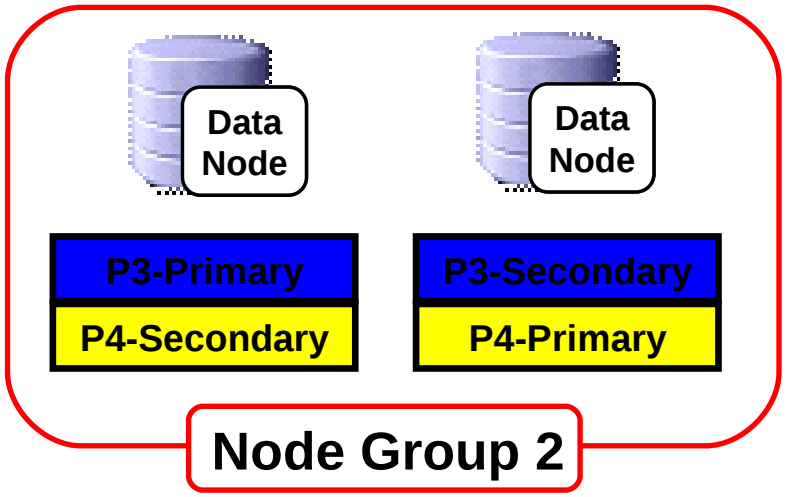
Partition 1

Partition 2

Partition 3

Partition 4





- **Fonctionement real time**
- **Attachement à un processeur**

Files system & linux

- Ext3 Journaled filesystem
- Recovering the filesystem does not mean a database recovery !!
- LVM as underlying layer : Offer RAID / stripping / resizing
- IO scheduler can be important
- Ext4 , btrfs coming
- mount -o noatime,remount,rw /dev/hda3

What about RAMDISk and TMPFS or SSD

Good for temporary table (in /tmp)

It does not transform MySQL in a in memory database !

In that case think about MySQL Cluster)

Think also about memory storage engine

```
/bin/mount -t tmpfs -o
   size=1G,nr_inodes=10k,mode=0775,noatime,nodiratime
   tmpfs /tmp
to dynamically increase its size
/bin/mount -t tmpfs -o
   size=2G,nr_inodes=20k,mode=0775,noatime,nodiratime,re
   mount tmpfs /tmp
Tuning the swappiness
/bin/echo “1" > /proc/sys/vm/swappiness
```

Ramdisk on Linux

$ ls -lh /dev/ram*.

To know default size of RAM disk use

```
$ dmesg | grep RAMDISK
$ mkdir -p /ramdisk
$ mkfs -b 1024 -o Linux -L RAMDisk -T ext3 /dev/ram0 65536
$ mount /dev/ram0 /ramdisk
$ df -k /ramdisk
$ time `dd if=/dev/zero of=/ramdisk bs=1M count=100`
$ sync; time `dd if=/dev/zero of=/ramdisk bs=1M count=100 &&
   sync`
```

config GRUB : pour augmenter la taille du ramdisk

```
 kernel /boot/vmlinuz-2.6.11-1.1369_FC4 ro root=LABEL=/1
 ramdisk_size=512000 quiet
```

Storage / LVM & Linux

```
$ pvcreate /dev/sdb
$ vgcreate dbvolgrp /dev/sdb
$ lvcreate -L50M -ndbvol dbvolgrp
$ mkfs.ext3 /dev/dbvolgrp/dbvol
$ mount /dev/dbvolgrp/dbvol /mnt/data
```

Lvextend / lvreduce /resize2fs possible

LVM snapshot for backup on Linux

FLUSH TABLES WITH READ LOCK

lvcreate -L16G -s -n dbbackup /dev/Main/Data

SHOW MASTER STATUS

UNLOCK TABLES

mount /dev/Main/dbbackup /mnt/backup

Copy

umount /mnt/backup

lvremove -f /dev/Main/dbbackup

CHANGE master TO master_host="master",
master_user="user", master_password="password",
master_log_file="host-bin.000335",
master_log_pos=401934686

Storage / iscsi & Linux

tgtadm --lld iscsi --mode target --op new --tid=1 –
    targetname=sharedstorage.serge.localdomain

dd if=/dev/zero of=/var/lib/libvirt/images/mysql-shared.img bs=1M count=256

tgtadm --lld iscsi --mode logicalunit --op new --tid=1 --lun=1
            --backing-store=/var/lib/libvirt/images/mysql-shared.img

tgtadm --lld iscsi --mode target --op bind --tid=1 --initiator-address=ALL

tgtadm --mode target --op show

iptables -I INPUT -p tcp --dport 3260 -j ACCEPT

Storage / nfs & Linux

Be carefull about NFS mounting options
Fsync : Synchronous write should be synchronous

No problem with NAS if mounted with the correct option (NetApp, SUN Open Storage)

Linux fuse and ZFS

ZFS is an very interesting filesystem : copy-on-write transactional semantics, fast snapshots, and optional compression. SSD cache. ZIL (Zfs intent log)

zpool create zp1 c2t0d0s2

zpool create zp1 ~/storage/myzfile

zfs create zp1/data

zfs create zp1/logs

innodb_flush_method = O_DIRECT

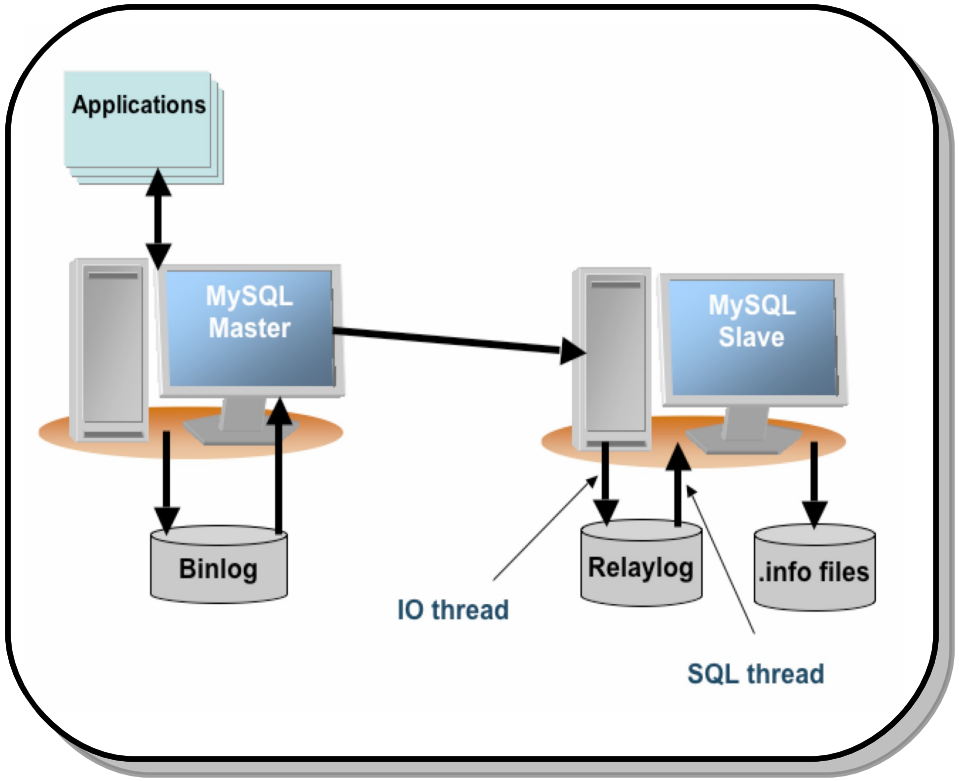
ZFS is smart about block size

zfs set recordsize=16K zp1/data

ZFS guarantees that partial writes never happen.

Réplication MySQL

- Le Maître stocke les requêtes DML au Binlog
- Threads de la réplication
  - IO Thread
  - SQL Thread
- Fichiers de la réplication
  - Binlog
  - Fichier de contrôle
  - Relay log
  - Fichier de contrôle relay log



MySQL Replication / checksum ?

Use crypto insted of checksum

ssh -f user@master.server -L 4306:master.server:3306 -N

```
mysql> STOP SLAVE;
mysql> CHANGE MASTER TO master_host='localhost',
    master_port=4306;
mysql> START SLAVE;
```

Innodb & linux

MySQL : memlock et swapiness

cat /proc/meminfo" will have lines like:
HugePages_Total: vvv
HugePages_Free: www
HugePages_Rsvd: xxx
HugePages_Surp: yyy
Hugepagesize: zzz kB

- Memlock
  - Use to lock mysqld in memory

- echo 0 > /proc/sys/vm/swappiness

MySQL & linux Network

- Mutipath / Chunking /bonding
- Netfilter / Iptables for security
- Network Loadbalancing through LVS
- virtual IP

Bonding NIC

```
# vi /etc/sysconfig/network-scripts/ifcfg-bond0
DEVICE=bond0
IPADDR=192.168.1.20
NETWORK=192.168.1.0
NETMASK=255.255.255.0
USERCTL=no
BOOTPROTO=none
ONBOOT=yes

# vi /etc/modprobe.conf
alias bond0 bonding
options bond0 mode=balance-alb miimon=100
# modprobe bonding
# service network restart
# less /proc/net/bonding/bond0
```

```
# vi /etc/sysconfig/network-scripts/ifcfg-eth0
DEVICE=eth0
USERCTL=no
ONBOOT=yes
MASTER=bond0
SLAVE=yes
BOOTPROTO=none
```

```
# ifconfig
bond0     Link encap:Ethernet  HWaddr 00:0C:29:C6:BE:5
eth1      Link encap:Ethernet  HWaddr 00:0C:29:C6:BE:59
 UP BROADCAST RUNNING SLAVE MULTICAST  MTU:1500  Metr
```

Virtual IP

```
ifconfig eth0:1 192.168.30.128 netmask
    255.255.255.0
# cat /etc/sysconfig/network-scripts/ifcfg-
    eth0:1
DEVICE=eth0:1
BOOTPROTO=static
IPADDR=192.168.30.128
NETMASK=255.255.255.0
NETWORK=192.168.30.0
BROADCAST=192.168.30.255
ONBOOT=yes
```

```
howie# route add -net 192.168.30.0 netmask 255.255.255.0

Kernel IP routing table
Destination   Gateway    Genmask         Flags Metric Ref    Use If
172.16.0.0    *          255.255.255.0   U     0      0        21 eth0
192.168.30.0  *          255.255.255.0   U     0      0         1 eth0:1
127.0.0.0     *          255.0.0.0       U     0      0        10 lo
default       bazooka    0.0.0.0         UG    0      0         4 eth0

ip addr add 192.168.0.14 dev eth0
```

MySQL & linux Security

- Basic : special user for mysqld daemon
- Running Mysql under Chroot(chroot option in my.cnf)
- Using virtualisation to have a black box approach
  - Main issue is separation of duties
- Iptables for security

Chrooter MySQL

```
mkdir -p /chroot/mysql ;cd /chroot/mysql
mkdir share ;# cp -R /opt/mysql/share/mysql ./share
mv /opt/mysql/data/ .
mkdir tmp ; chmod 1777 tmp
mkdir etc ;  grep "^mysql:" /etc/group > etc/group  ; grep "^mysql:" /etc/passwd
    > etc/passwd
cp /etc/hosts etc ; cp /etc/host* etc; cp /etc/resolv.conf etc; cp /etc/localtime
    etc
cd /chroot/mysql;# chown -R root:mysql .;# chown -R mysql data
mkdir dev ;# mknod dev/null c 2 2; # chown root:root dev/null;# chmod 666
    dev/null
cp /opt/mysql/support-files/my-medium.cnf /etc/my.cnf
/opt/mysql/bin/mysqld &
```

Verification:

```
 ps aux | grep ^mysql | awk {'print $2'}
5254
readlink /proc/5254/root
/chroot/mysql
```

```
[client]
socket = /chroot/mysql/tmp/mysql.sock
[mysqld]
chroot = /chroot/mysql
socket = /tmp/mysql.sock
basedir = /
datadir = /data
```

Encrypting the database

- Using triggers and crypting function
- Using Linux capabilities

Use a crypted file system

```
cryptsetup luksFormat /dev/vgsf/lv1
cryptsetup luksOpen /dev/vgsf/lv1 cryptlv1
mkfs.ext3 /dev/mapper/cryptlv1
mkdir /mnt/t
mount /dev/mapper/cryptlv1 /mnt/t
df
umount /mnt/t
cryptsetup luksClose cryptlv1
```

Virtualisation et linux

- Xen / Kvm
  - Libvirt
  - Permet d'avoir des poin de reprise grace au snapshot
- Vserver (not to be confused with Linux Virtual Server) or OpenVZ
  - Equivalent des zones/container de linux

Monitoring / Tunning

- top, iostat, dstat, vmstat...
- Nagios / Cacti

Development on Linux : Netbeans

Netbeans can be used to developp and debug MySQL on Linux

With C/ C++ API

For developping UDF

For libmysqld development

And what about SOLARIS

- Solaris is a very stable / scalable OS
- Adopted by Dell / HP / IBM on X86
- Very interesting on the new Intel highly threaded processor Intel® Microarchitecture (Nehalem)
- And Solaris is open source as all SUN software products
- Come with nice things like the zone/container for virtualization
  - Good canditate for consolidation
- ZFS in Solaris : Used in Sun Open Storage appliances

Questions ?

serge.frezefond@mysql.com