PERCONA
Performance Consulting Experts

Keep your MySQL backend online – no matter what

Date, time, place:	Froscon, 2010, Cologne, Germany August 21-22
Reporter:	Istvan Podor, Percona

who I am?

- Lead engineer at a payment processor for a site in alexa top100 (livejasmin.com)
- Performance engineer at ustream.tv (alexa top500)
- Consultant at Percona
- Some less interesting jobs ..

Main area of experience:

- web / database over clocking :)

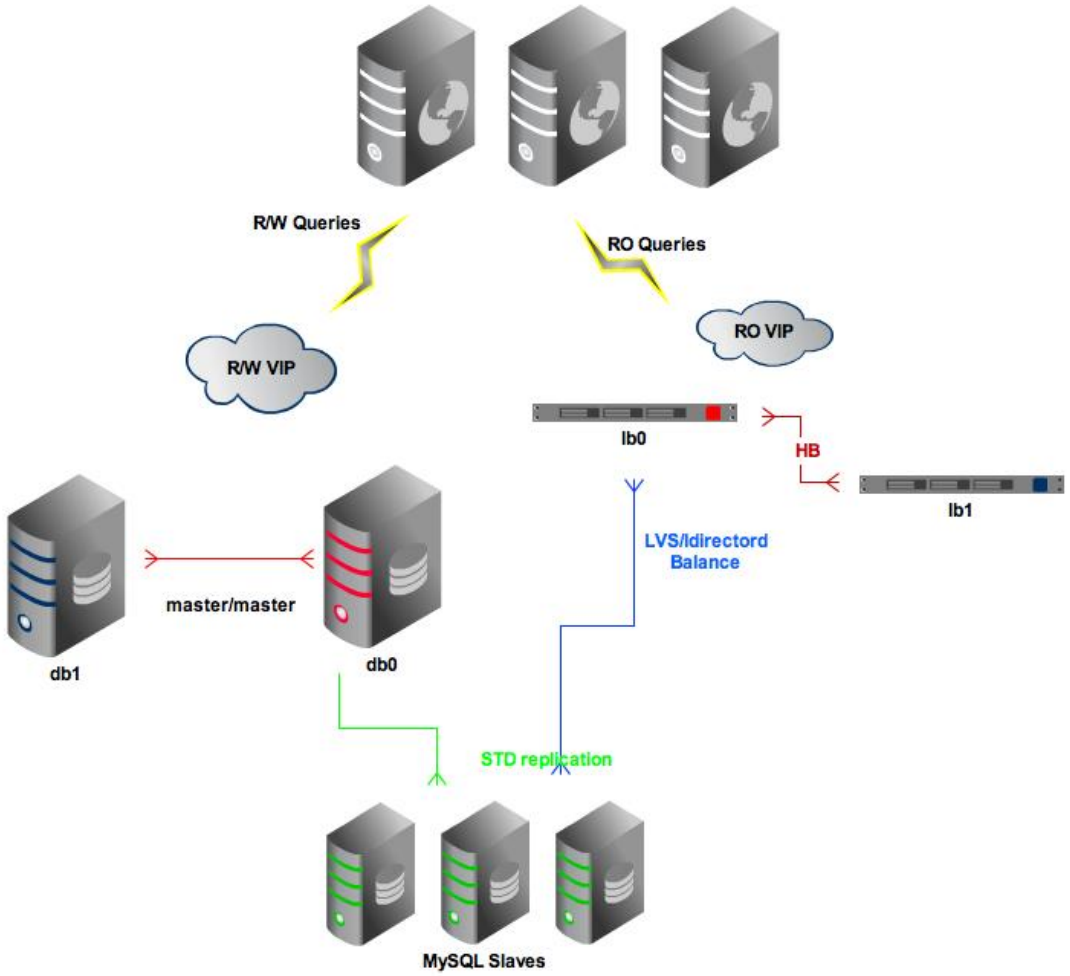
where the experience come from?

- Ustream.tv
  - Visitors online ::
    - Min: 9-10k
    - Peak: 60-200k
    - Max: 700k+
  - Web FE access / sec ::
    - min: ~1k, peak: ~10-15k
  - MySQL QPS ::
    - Master : min: ~1-2k/s, peak: ~6-12k
    - Slaves : min: ~2-3k/s, peak: ~6-10k
- livejasmin.com
  - I can confirm nothing. Very strict NDA.
  - Payment processor with very high traffic. No downtime was tolerated neither as any single bit of data loss.
  - Downtime was between 4 and 5 nines (<20 minutes) / year

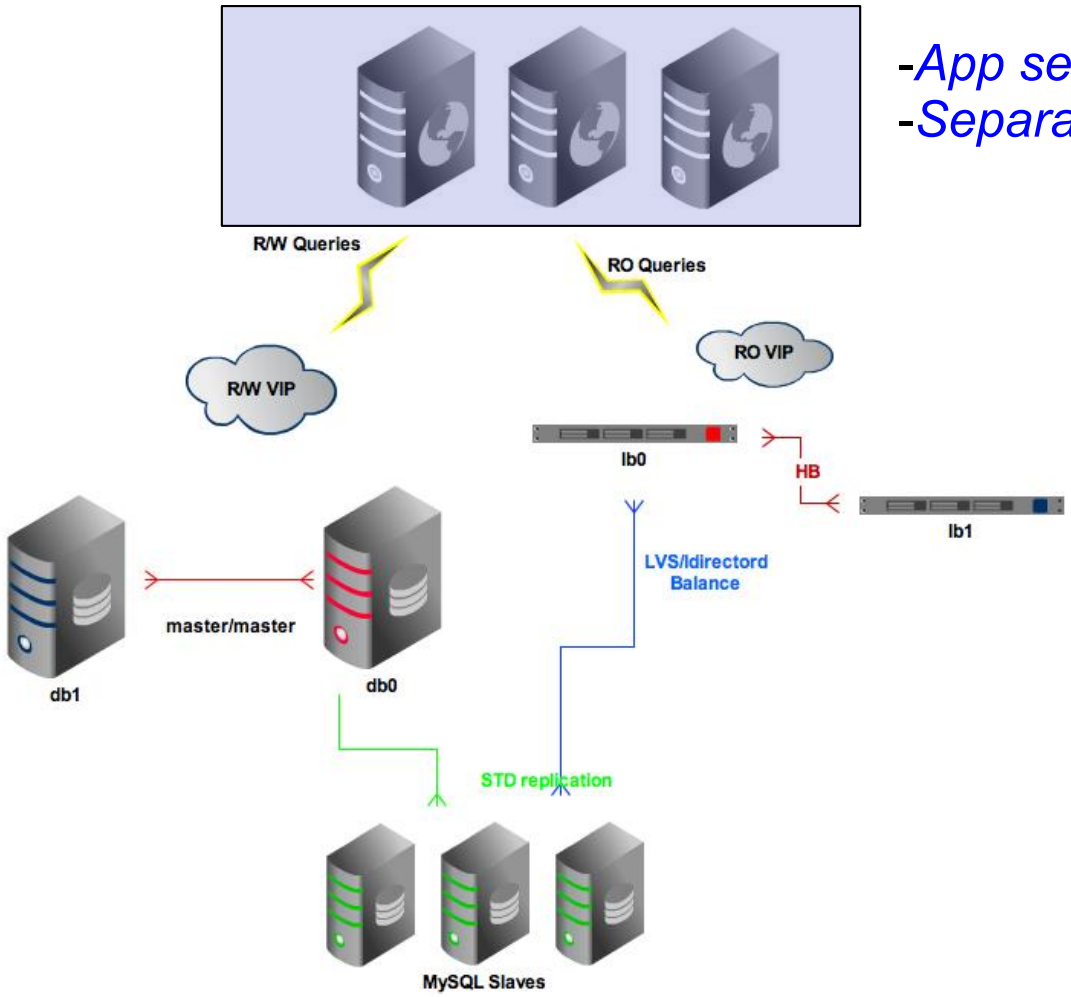
goals and requirements

- in general
  - Never go down (obviously :/)
  - Website must show up no matter what
  - Let the engineer sleep over the nights
- @livejasmin on payment (trust)
  - Never confuse any data in the DB (charge 2times or miss charging)
  - The site MUST show up!
  - Never loose any tracking records of the customer/visitor (fraud)
- @ustream (startup)
  - Stay online, show up the streams
  - As much performance as possible from the less resource as possible
  - Show must go on. (Michael Jackson's funeral, live event, you can't break the streaming)

how does it work?



how does it work?

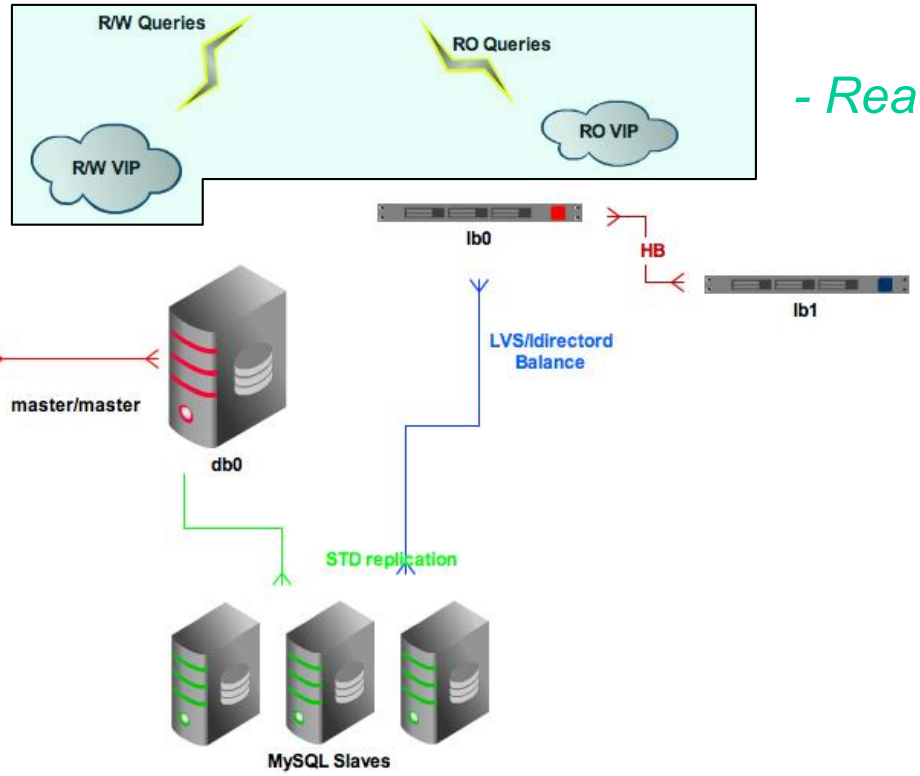


-*App servers*
-*Separate R/W queries*

how does it work?

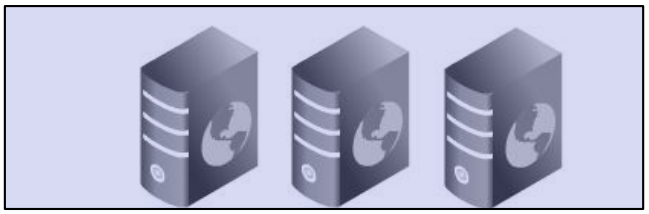
-*App servers*
-*Separate R/W queries*

- Writes go to RW VIP



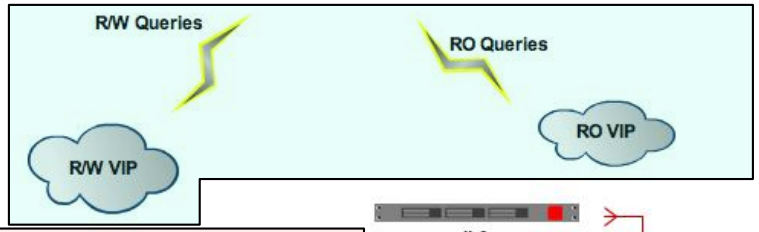
- *Reads go to RO VIP*

how does it work?

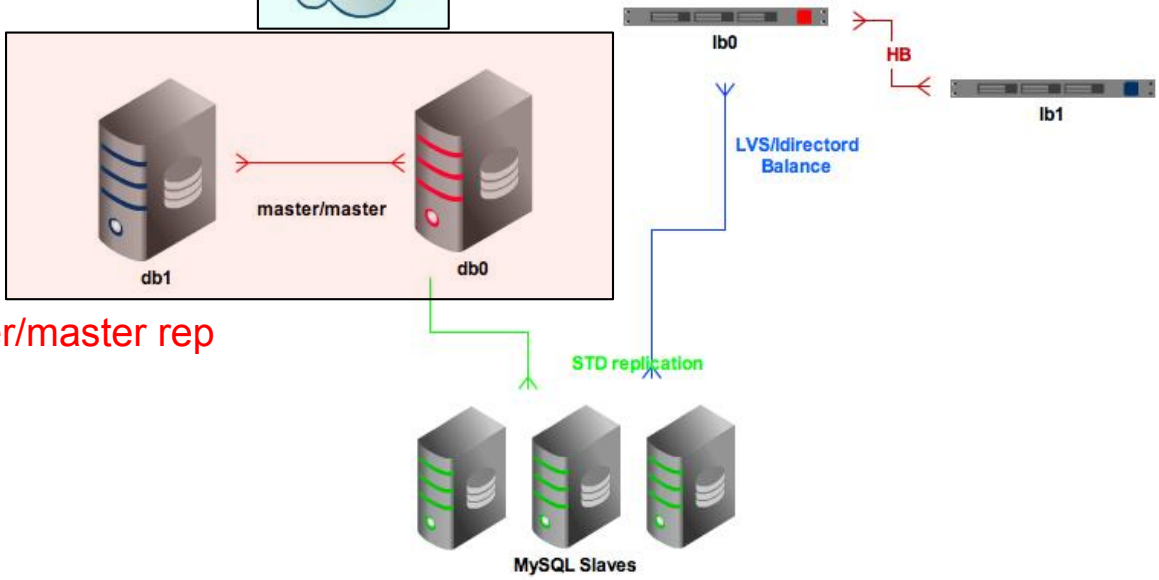


-*App servers*
-*Separate R/W queries*

- Writes go to RW VIP

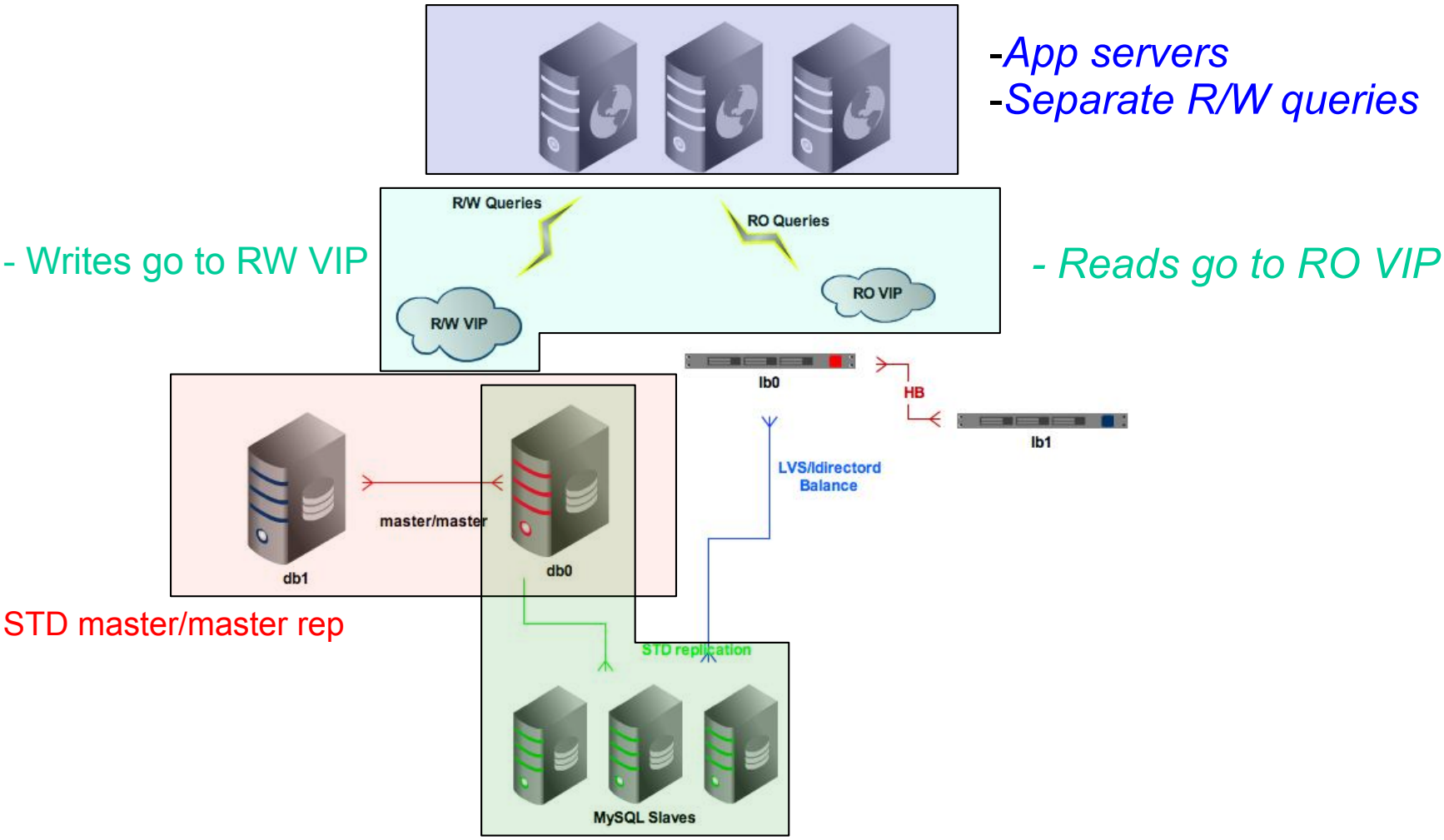


- Reads go to RO VIP

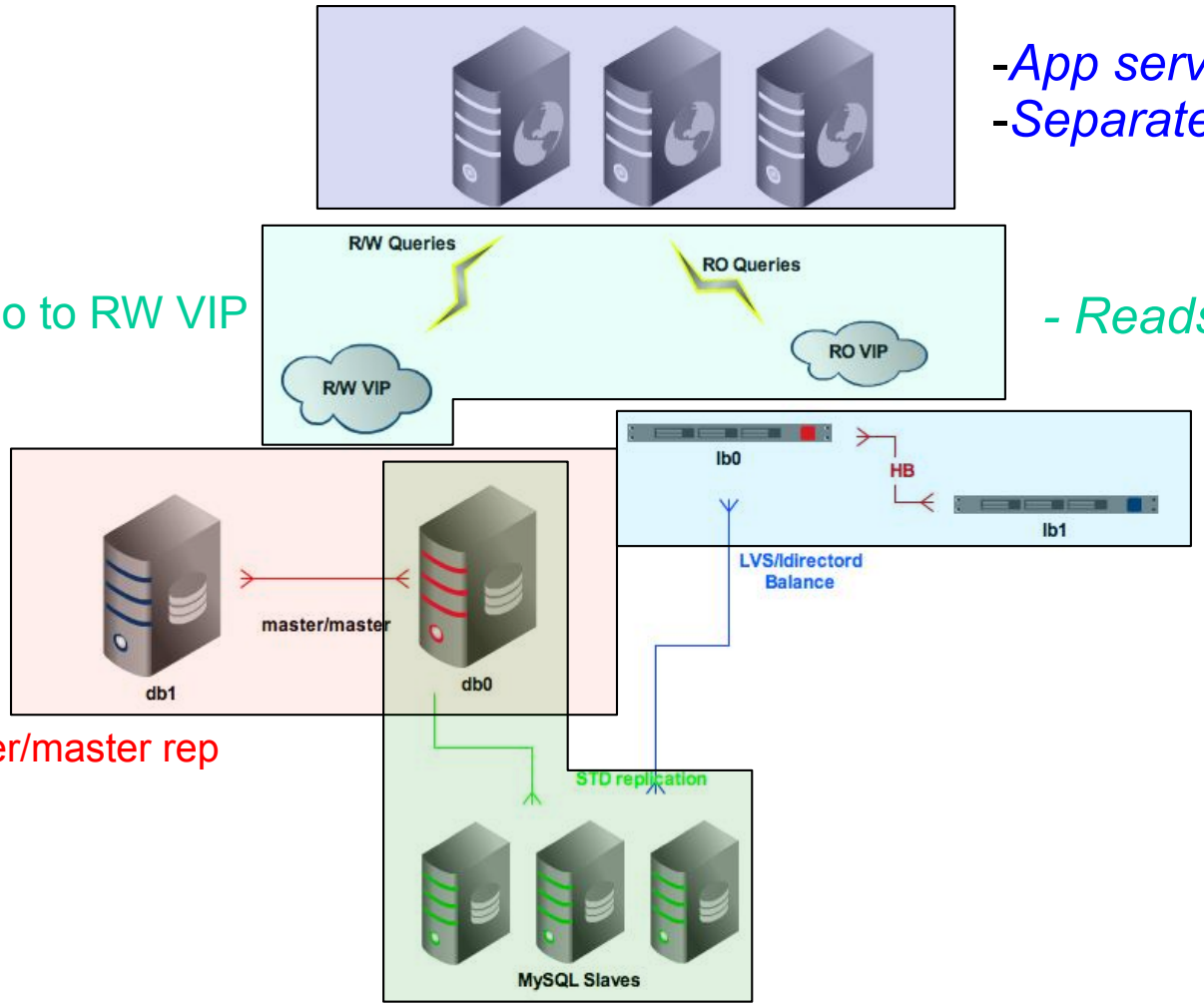


STD master/master rep

how does it work?



how does it work?



-*App servers*
-*Separate R/W queries*

- Writes go to RW VIP

- Reads go to RO VIP

- Two HB node

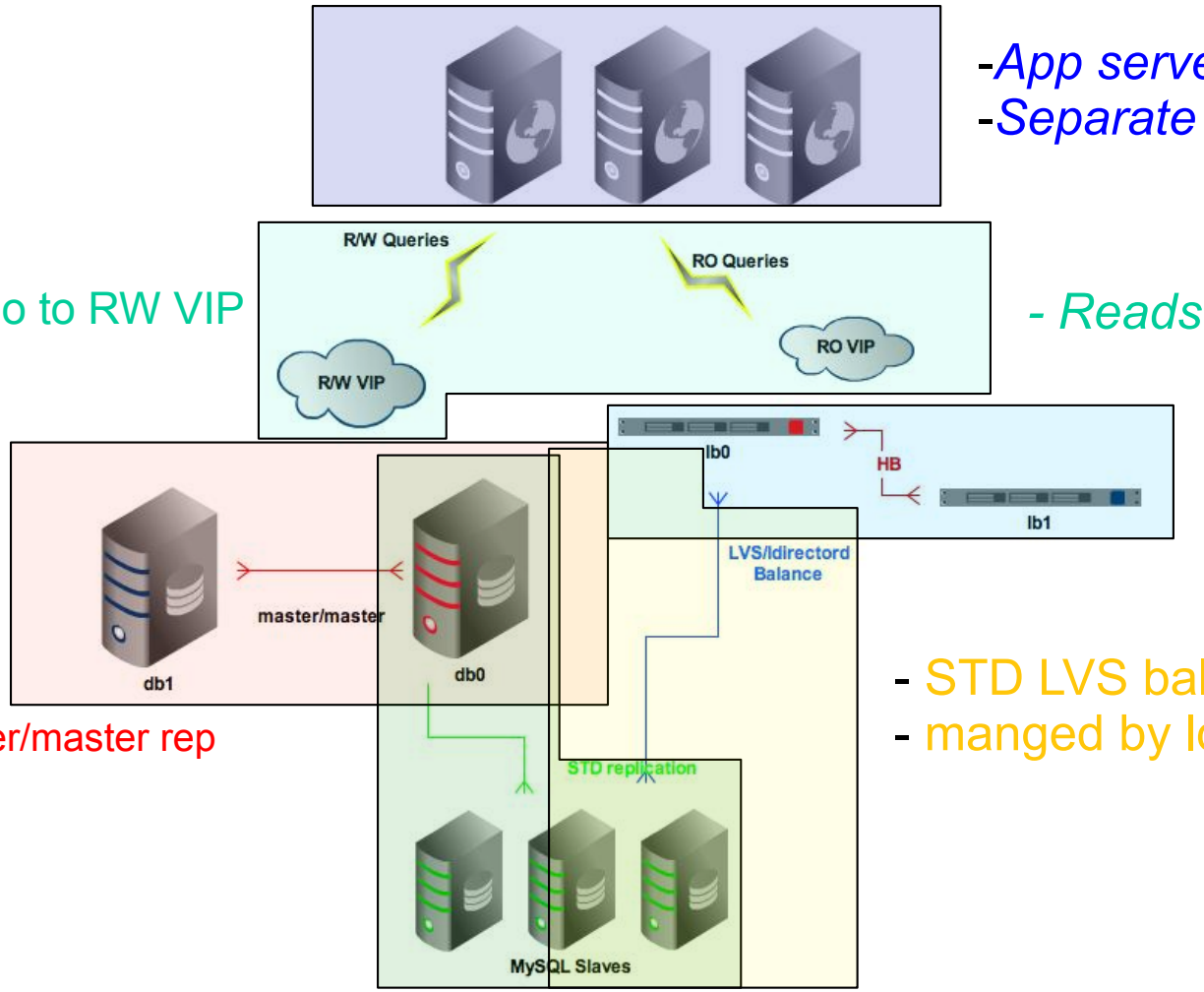
STD master/master rep

STD replication

gliffy

PERCONA
Performance Consulting Experts

how does it work?



-*App servers*
-*Separate R/W queries*

- Writes go to RW VIP

- Reads go to RO VIP

- Two heartbeat node

STD master/master rep

- STD LVS balacing
- manged by ldirectord

STD replication

gliffy

PERCONA
Performance Consulting Experts

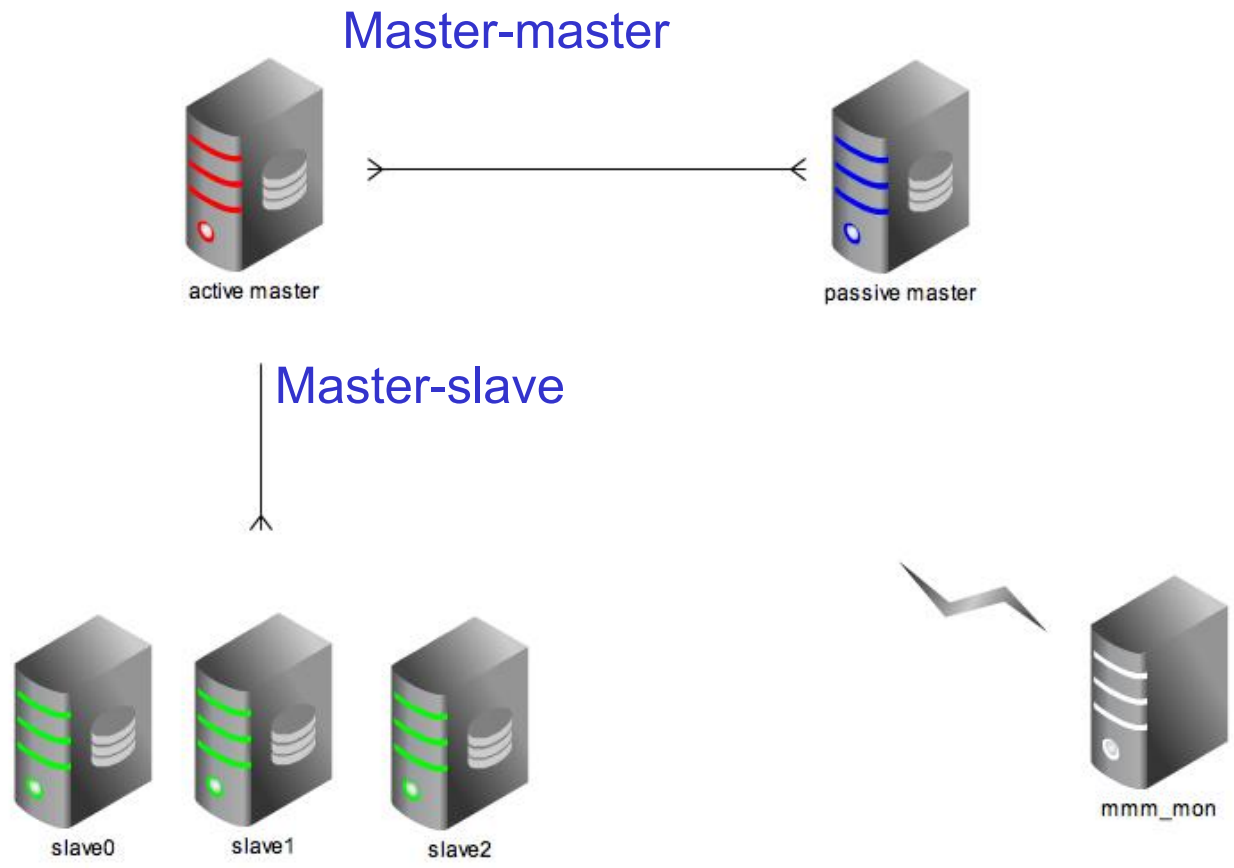
We need to …

- Manage master-master + slaves with MMM

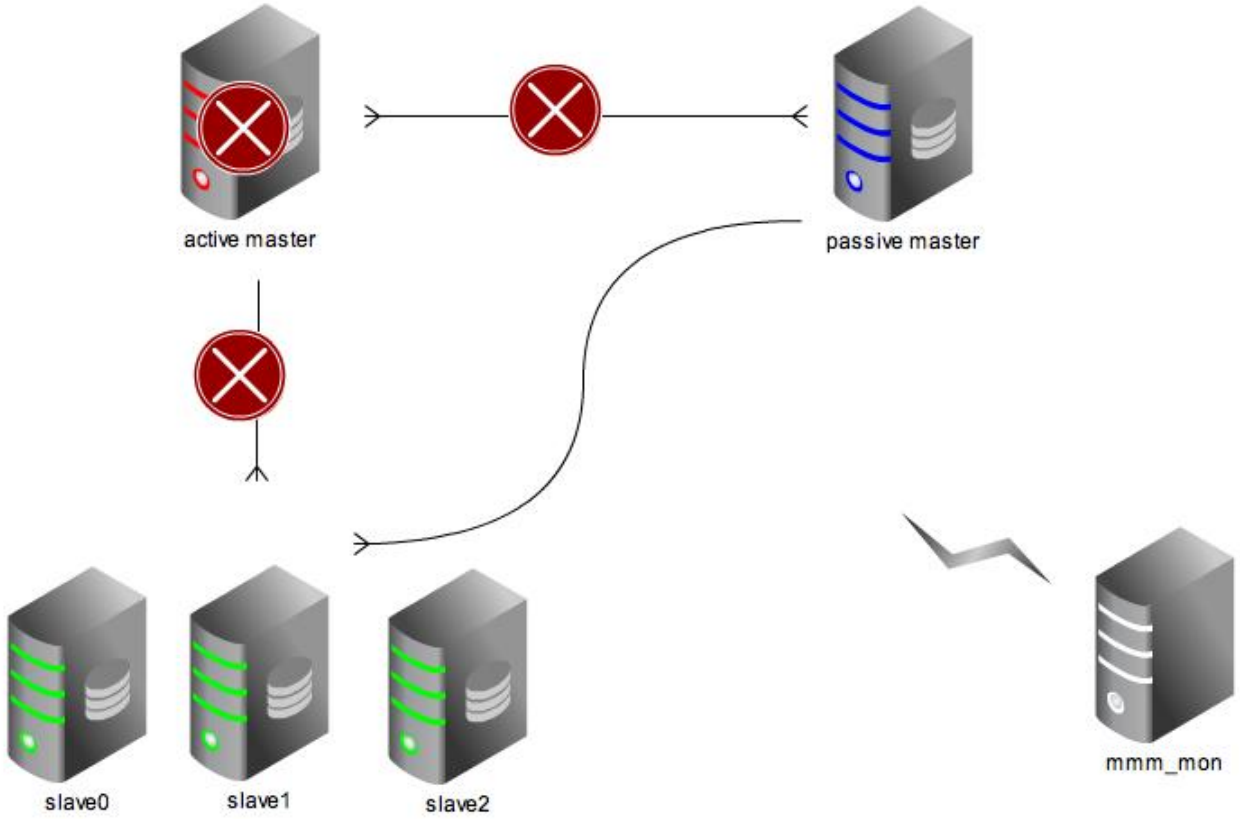
Manage master-master + slaves with MMM

- This is not a real master-master.
- MMM is a simple perl script collection
- It checks for the health of replication
- Can handle active master failure
- Weakness: it's based on the checking of the replication status

Manage master-master + slaves with MMM



Manage master-master + slaves with MMM



We need to …

- Manage master-master + slaves with MMM
- Set up HB+LVS/ldirectord to do the balancing

HB+LVS/ldirectord

- HB is only responsible, to keep one of the two balancer online all the time (www.linux-ha.org)

haresources: lb1 192.168.1.10 ldirectord

- LVS is kernel level implementation of software based loadbalancing
- ldirectord is an application what manage the balancing

```
# MySQL
virtual = 192.168.2.1:3306
real=192.168.2.1:3306 gate 10
real=192.168.2.2:3306 gate 10
#passive master
fallback=192.168.1.10:3306 gate
request = "select * from mydb.loadbalance”
```

PERCONA
Performance Consulting Experts

We need to …

- Manage master-master + slaves with MMM
- Set up LVS/ldirectord to do the balancing
- Extend capabilities

Extend capabilities

- Split reads/writes from your application
- Prepare for replication delay

```php
<?php

$my_writer = '192.168.1.1';
$my_reader = '192.168.2.1';

//build connection for insert
$writer_connection = mysql_connect($my_writer,"user","pass")
mysql_select_db("database", $writer_connection);

//build connection for select
$reader_connection = mysql_connect($my_reader,"user","pass")
mysql_select_db("database", $reader_connection);

//execute insert
$result_of_insert = mysql_query("insert into users (name, email, password) values ('Istvan', 'istvan.podor@percona.com', '******'", $writer_connection);

//execute select
$result_of_read = mysql_query("select name from users where `email` = 'istvan.podor@percona.com'", $reader_connection);

?>
```

PERCONA
Performance Consulting Experts

Extend capabilities

- Split read/writes from application
- Set fallback node to your passive master
- Set up ldirectord to check for a table if exists

```
# MySQL
virtual = 192.168.2.1:3306
real=192.168.2.1:3306 gate 10
real=192.168.2.2:3306 gate 10
#passive master
fallback=192.168.1.10:3306 gate
request = "select * from mydb.loadbalance"   //Interacting with the balancer
```

Extend capabilities

- Split read/writes from application
- Set fallback node to your passive master
- Set up ldirectord to check for a table if exists
- Manage read-only pool by yourself

Balancer mgmt script

- Simple (bash/php/perl) script on your slaves
- Check for slave health
- Manipulate balancing by renaming the `loadbalance' table

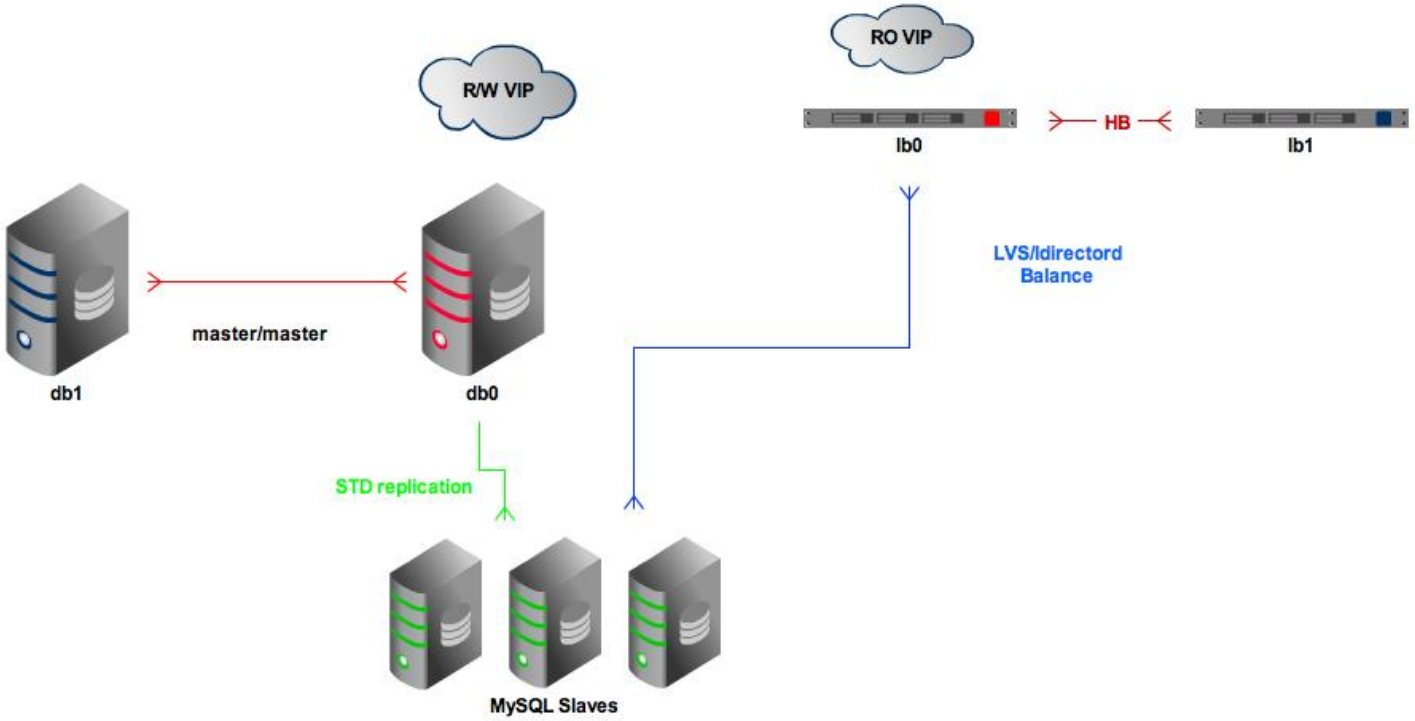
- http://istvanpodor.ath.cx/froscon/slave_mgmt.sh

PERCONA
Performance Consulting Experts

We need to …

- Manage master-master + slaves with MMM
- Set up LVS/ldirectord to do the balancing
- Extend capabilities
- At this point, we should know what happens if something fail :)

Lets brake things ..



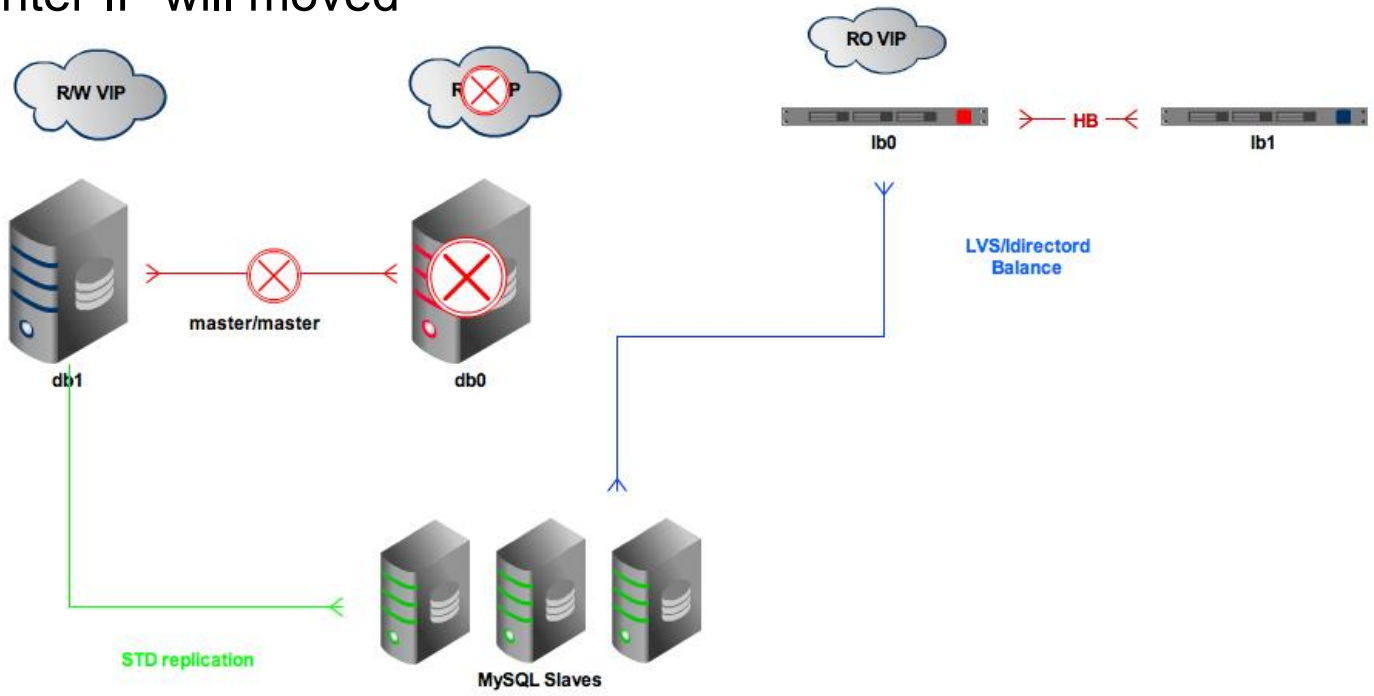
create and share your own diagrams at gliffy.com

gliffy

PERCONA
Performance Consulting Experts

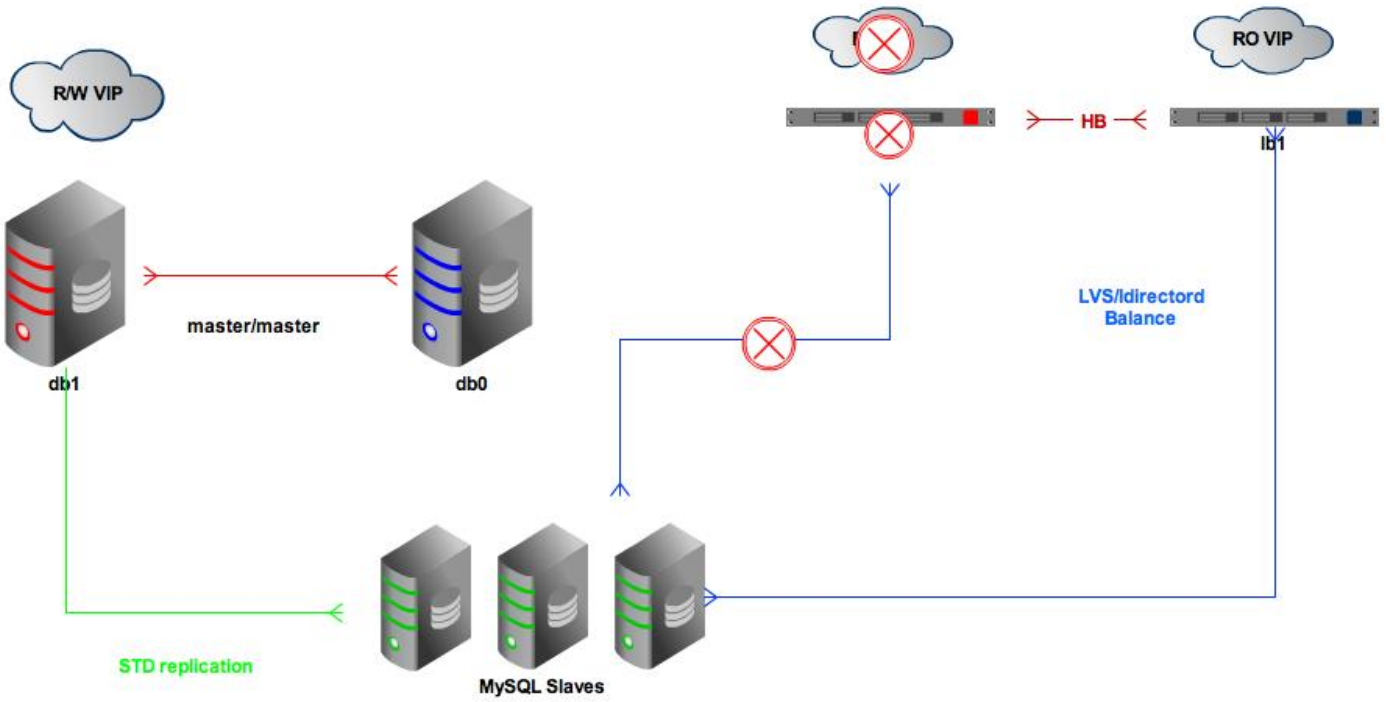
Lets brake things ..

Writer IP will moved



Change master on slaves

Lets brake things ..



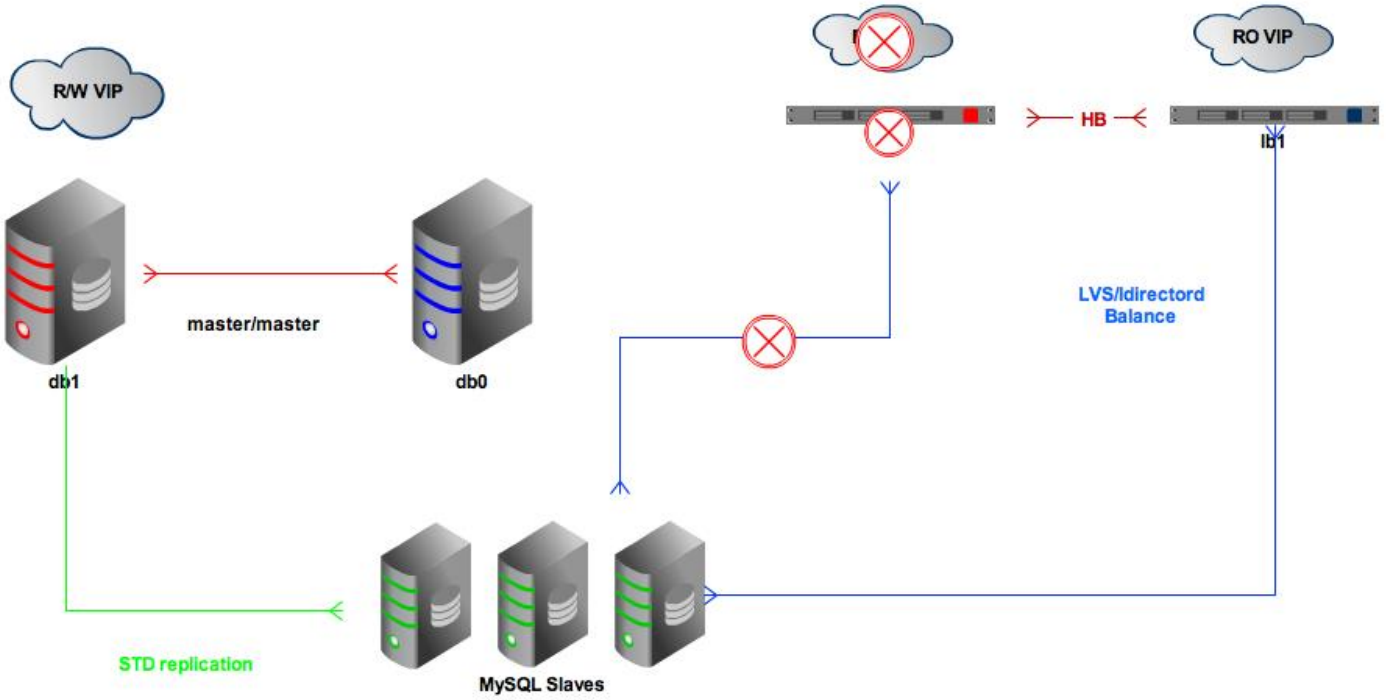
create and share your own diagrams at gliffy.com

gliffy

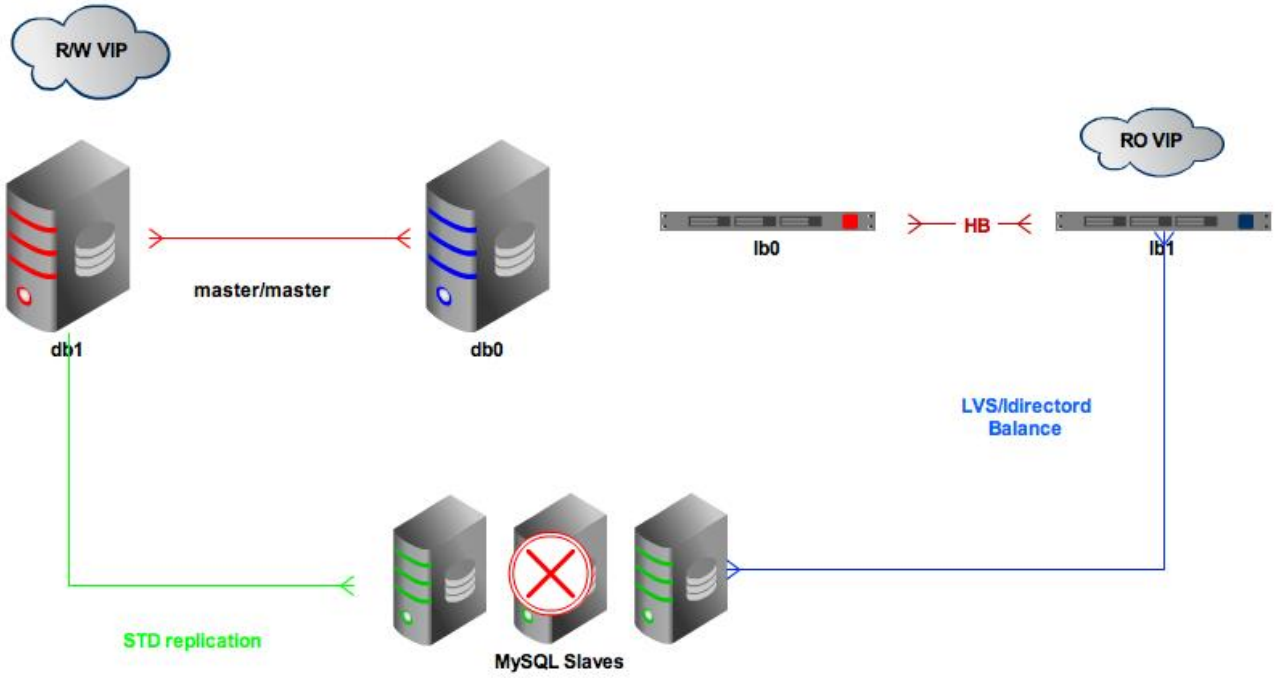
PERCONA
Performance Consulting Experts

Lets brake things ..

Passive HB node will take over
move the IP, startup ldirectord

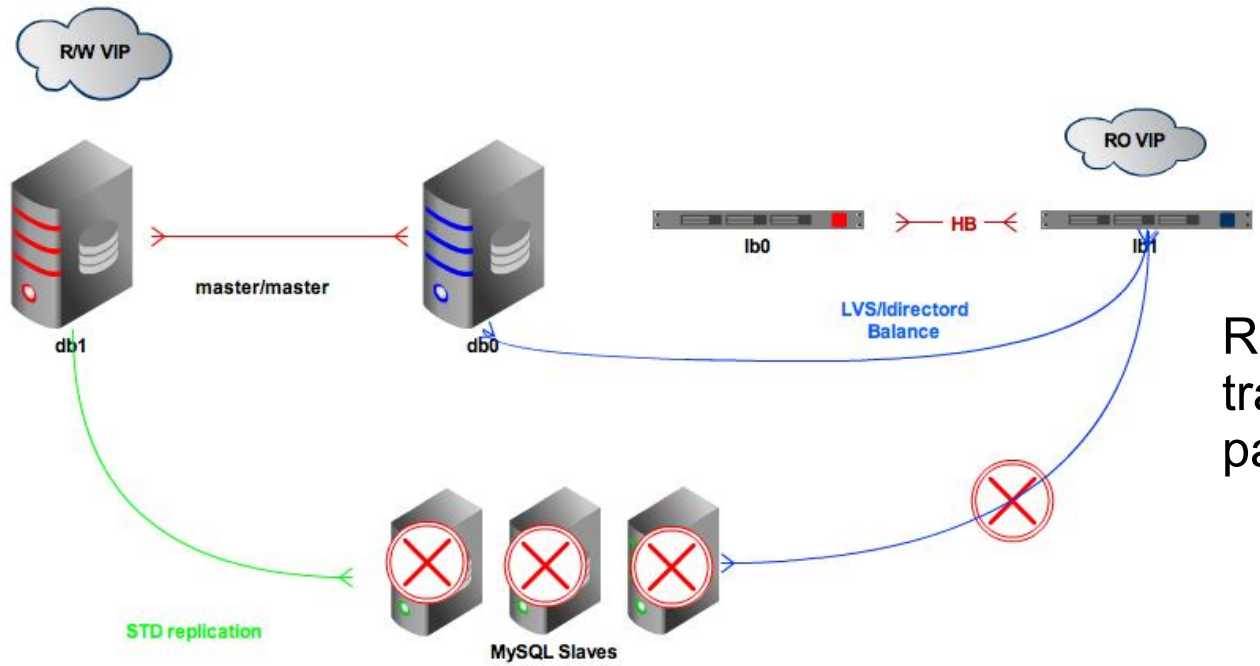


Lets brake things ..



Slave fail or fall behind

Lets brake things ..



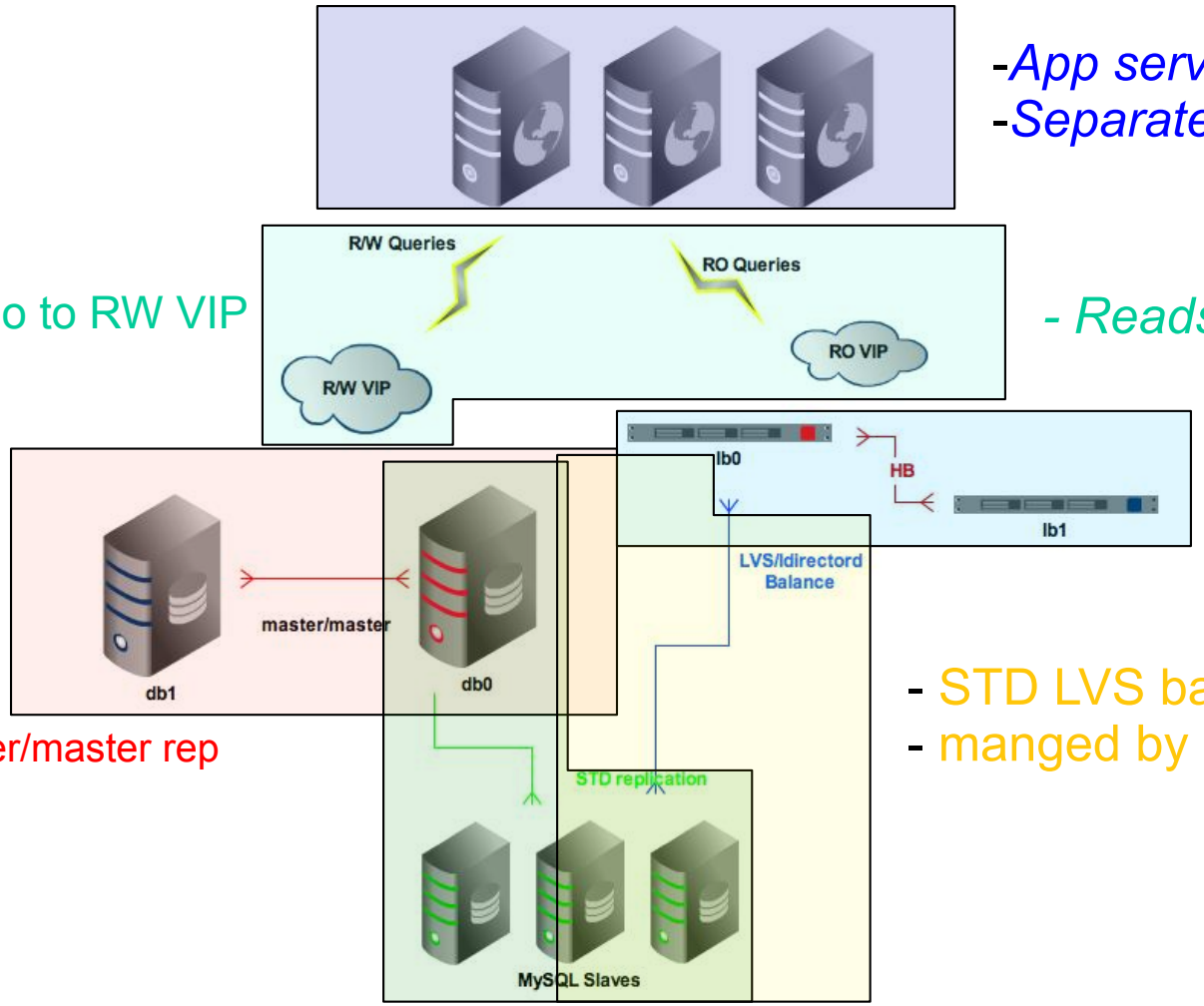
Read queries will transmitted to your passive master

All slave fail (wrong query)

gliffy

PERCONA
Performance Consulting Experts

Better?



-*App servers*
-*Separate R/W queries*

- Writes go to RW VIP

- Reads go to RO VIP

- Two HB node

- STD LVS balacing
- manged by ldirectord

STD master/master rep

STD replication

What could go wrong?

- One or more servers fail (no problem)
- Somebody/something broke replication
- MMM won't work (manual fail over)
- Load balancer goes down
- Data inconsistency
  - Wait, what do you mean, inconsistency?
  - mk-checksum daily
  - mk-sync-table on demand

So we won?

- In some cases, yes. If the show must go on.
- If we can have a relatively small backend for our website/mission critical application and store other data elsewhere (datamining, logging etc)
- Useful for altering large tables, making larger upgrades/updates

- Best win: once a netop pulled out the uplink from a whole rack with 24 servers, all the active nodes. We felt over to the other rack without anyone realizing it.

Any questions?

- Links:
  - www.percona.com
  - www.maatkit.org
  - www.linux-ha.org
  - www.mysql-mmm.org
  - http://www.linuxvirtualserver.org/docs/ha/heartbeat_ldirectord.html
  - Istvan.podor@percona.com

PERCONA
Performance Consulting Experts