PostgreSQL Clustering with Red Hat Cluster Suite

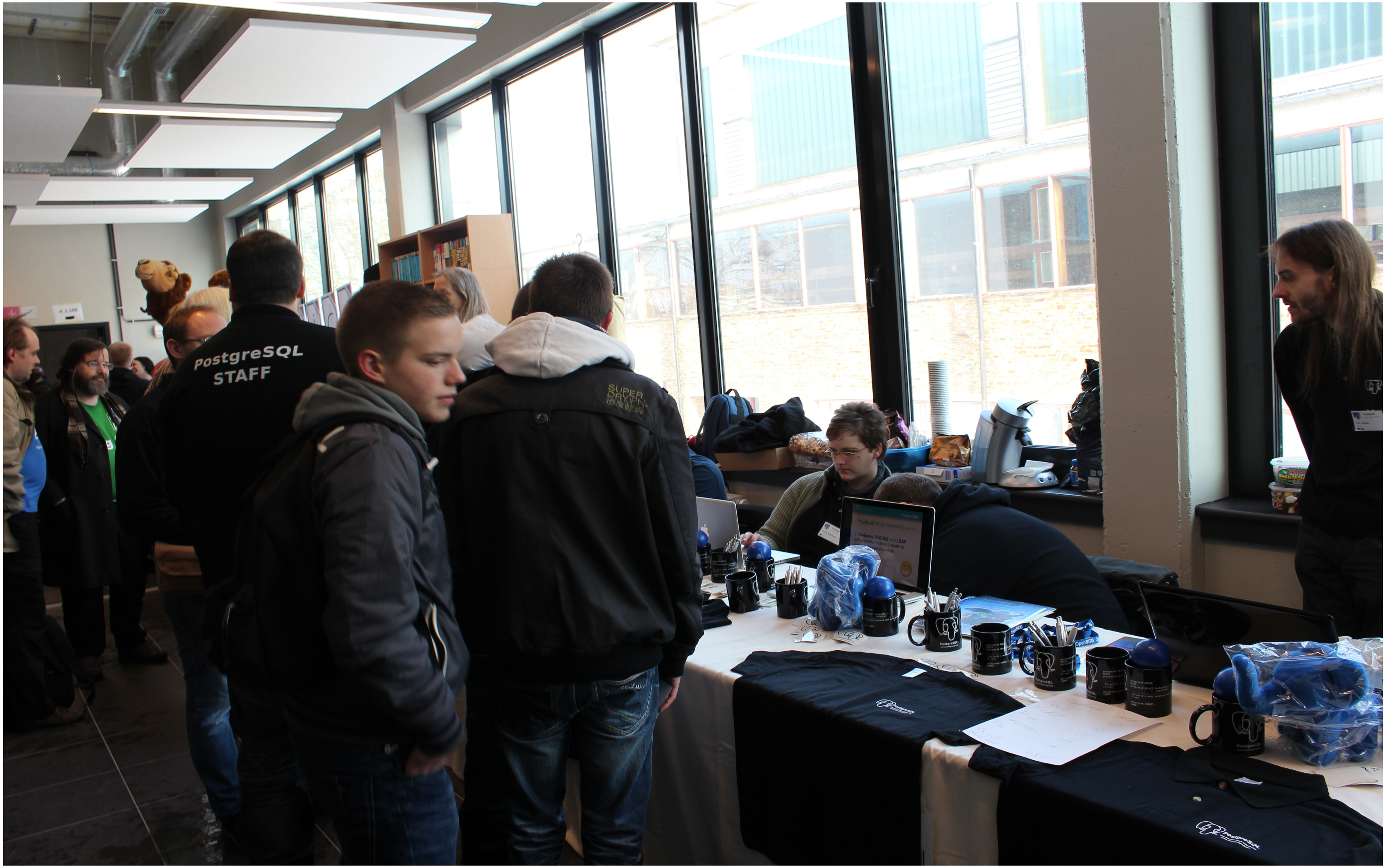
Devrim GÜNDÜZ
Principal Systems Engineer
EnterpriseDB
devrim.gunduz@EnterpriseDB.com

Use Red Hat Cluster Suite
for PostgreSQL Clustering

This guy...

- Who is this guy?
  - I have been contributing to PostgreSQL over the last 8 years.
  - I'm not a hacker, I work on PostgreSQL Community RPMs and website.
  - I rarely break RPMs, but break website more often.
  - Work at EnterpriseDB right now.
  - Live in Istanbul, Turkey.
  - Have a son.

PostgreSQL booth in FOSDEM



Agenda

- **Why Red Hat Cluster Suite? (RHCS)**
- **Goals**
- **Before initializing setup...**
- **Choosing the right hardware**
- **Setting up RHCS**
- **Setting up PostgreSQL**
- **Failover, switchover**
- **Postgres-XC**
- **Tips & Tricks**
- **Questions**

Agenda

Why Red Hat Cluster Suite?

Why Red Hat Cluster Suite?

- Open Source Clustering Solution
- Developed by Red Hat, **with the community**
- Available through (Red Hat Network) RHN, but also available via the CentOS repositories (*unsupported by Red Hat, or supported by 3rd party support companies*)
- RHEL 5 and RHEL 6 provides RHCS (High Availability Addon) and GFS (Resilient Storage).
- It is the only open source clustering solution that has decent support.
- Use at least RHEL 5.4. All versions prior to that are broken in various ways. 6.2+ is the best.
- Minimizes downtime

Why Red Hat Cluster Suite?

- Support wide range of hardware
- Application/Service Failover - Create n-node server clusters for failover of key applications and services
- Load Balancing - Load balance incoming IP network requests across a farm of servers
- TGIOS! (Thanks God It is Open Source)

RHCS overview

- Supports up to 16 nodes (RHCS 5 and RHCS 6).
- All PostgreSQL nodes can access to the same storage, but they don't use it at the same time.
- Automatic failover
- http://www.redhat.com/cluster_suite/
- http://sources.redhat.com/cluster/wiki/ (Development site)

What else for RHCS?

- RHCS avoids cancer.
- It helps peace in the world.
- RHCS cannot be used as a replica. If you want to hear about replicas, this is not the right talk.
- RHCS does not run on Windows.
- It does not do “multimaster” clustering.
- Postgres-XC? We will talk about it later.

Agenda

Goals

Goals

- Clustering goals
  - Active/passive clustering
  - Having a redundant system
    - Data redundancy
    - Network redundancy
    - Server and power redundancy
  - Maximum uptime
  - Service failover (=PostgreSQL)
  - Data integrity

Agenda

Before initializing setup...

Before initalizing setup...

- Make sure that you have at least a RHCE or similar around.
- Make sure that the sysadmin, network and DBAs can work closely.

Agenda

Choosing the right hardware

Choosing the right hardware

- Database servers
  - Minimum hardware: Any hardware that Red Hat Enterprise Linux can run.
  - Typical hardware : Depends on your needs. *See related threads in pgsql-performance mailing list.*
  - SAN : Storage is the most important part – Use RAID arrays.
  - At least 2 NICs -- 4 would be much better for bonding.
  - Don't forget the fencing device -- I got nice results with HP, DELL and IBM servers.

Choosing right hardware

- Each node needs to have 1GB ram (not for PostgreSQL, it is for RHCS)
- Decent fiber channel switch to storage, decent switches for internal and external communications.
- Multicast is the key word. All switches must have multicast support.

Software requirements

- RHCS is built on GFS, which is the “Resilient Storage” addon in Red Hat Enterprise Linux.
- GFS is built on LVM.
- PostgreSQL :-)
- Feel free to use PostgreSQL 9.1.
- yum install perl-Crypt-SSLeay.x86_64 <-- For ILOs to work.

Network setup

- Preparing network
  - Multicast traffic must be supported / enabled in network switches.
  - Testing: ping -t 1 -c 2 224.0.0.1
  - Cluster services will not work if they don't respond to ICMP echo requests.
    - On RHEL 6:

```
echo "net.ipv4.icmp_echo_ignore_broadcasts = 0" \
          >> /etc/sysctl.conf
sysctl -p
```

Fencing device

- Fencing: Disconnection of a node from the cluster's shared storage (RHCS docs)
- It cuts off I/O from share storage to ensure data integrity.
- System **must** have a supported fencing device.

Fencing device

- Power fencing : Uses a power controller to power off an inoperable node.
- Fibre Channel switch fencing : Disables the Fibre Channel port that connects storage to an inoperable node.
- GNBD fencing :Disables an inoperable node's access to a GNBD server. **Not supported as of RHEL 6.**
- Other fencing :Several other fencing methods that disable I/O or power of an inoperable node, including IBM Bladecenters, PAP, DRAC/MC, HP ILO, IPMI, IBM RSA II, and others.

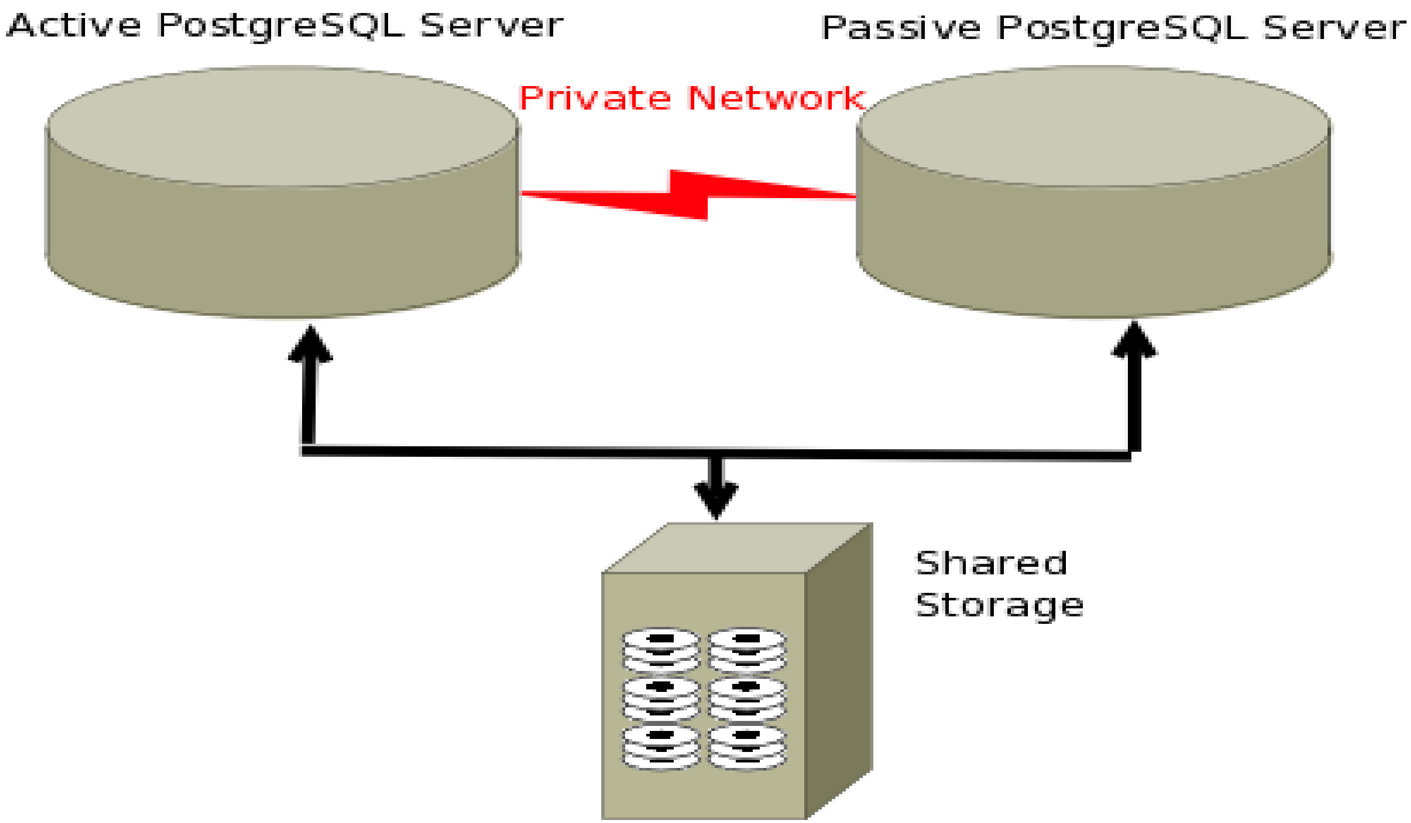
Design and howtos

- We need two servers that has been setup identically.
- Only OS and PostgreSQL will run
- Same PostgreSQL versions.
- Using GFS, all data will be mounted from the storage. GFS is not a very strict requirement, but we would better be safe.
- When node1 goes down, node2 will act as "active" server by announcing specified cluster ip.
- When node1 comes back, the process may be reverted, depending on the setup.

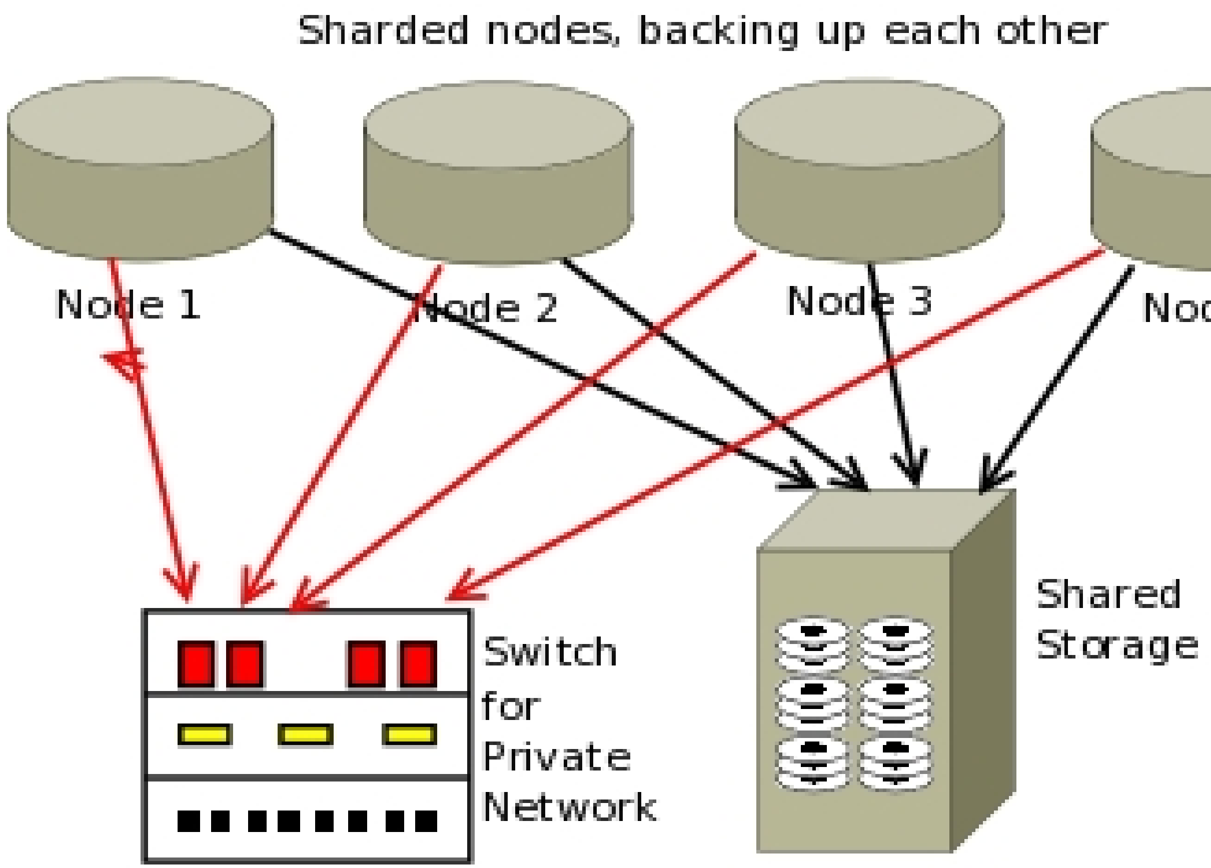
General recommendations

- https://access.redhat.com/kb/docs/DOC-30004 (formerly http://www.redhat.com/cluster_suite/hardware)
- Check this list **before** you purchase the hardware.
- HP, DELL and IBM servers have been proved to be working well with RHCS. Recommended.
- Make sure that you have updated firmware.

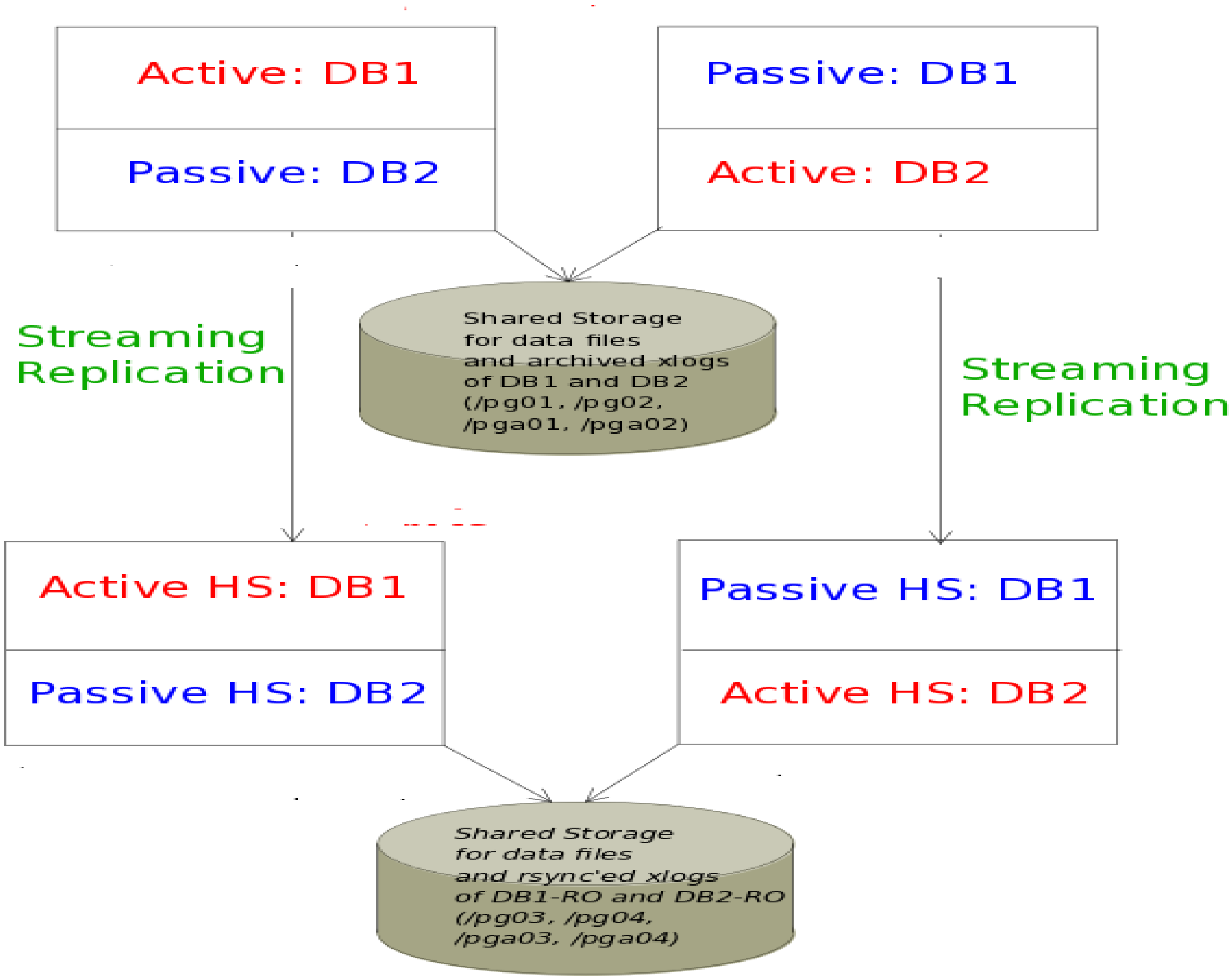
Active-Passive Cluster Overview



Sharding, multiple nodes


Sharded nodes, backing up each other
Node 1
Node 2
Node 3
Node 4
Switch for Private Network
Shared Storage

Another setup, with Streaming Replication on 2 separate clusters


Active: DB1
Passive: DB2
Passive: DB1
Active: DB2
Streaming Replication
Shared Storage for data files and archived xlogs of DB1 and DB2 (/pg01, /pg02, /pga01, /pga02)
Streaming Replication
Active HS: DB1
Passive HS: DB2
Passive HS: DB1
Active HS: DB2
Shared Storage for data files and rsync'ed xlogs of DB1-RO and DB2-RO (/pg03, /pg04, /pga03, /pga04)

Agenda

Setting up RHCS

Before we start

- Do **NOT** edit contents of cluster.conf manually, if you don't know what you are doing.
- If you choose to edit cluster.conf manually, make sure that xml version numbers are identical on each node.
- If you think that you will screw up things, ask someone else.
- Be patient. This is not a plug-and-play solution.

Services that needs to run on boot

- clvmd
- cman
- gfs2
- rgmanager

chkconfig is your friend (and it is going away, it seems...)

Don't start PostgreSQL on boot. It is RHCS' responsibility!

Packages

- yum groupinstall “High Availability” “Resilient Storage”
- perl-Crypt-SSLeay package is essential for HP iLO fencing mechanism to function properly.

Setting up the cluster

- RHEL 5 provides system-config-cluster (scc), which is not supported in RHEL 6 (Thanks!)
- If you have to stick to RHEL 5, use only very recent versions of scc, otherwise you may screw up your cluster.
- scc helps you versioning your cluster configuration. Make sure that it is the same in all nodes.
- clusterssh will be your best friend during setup.

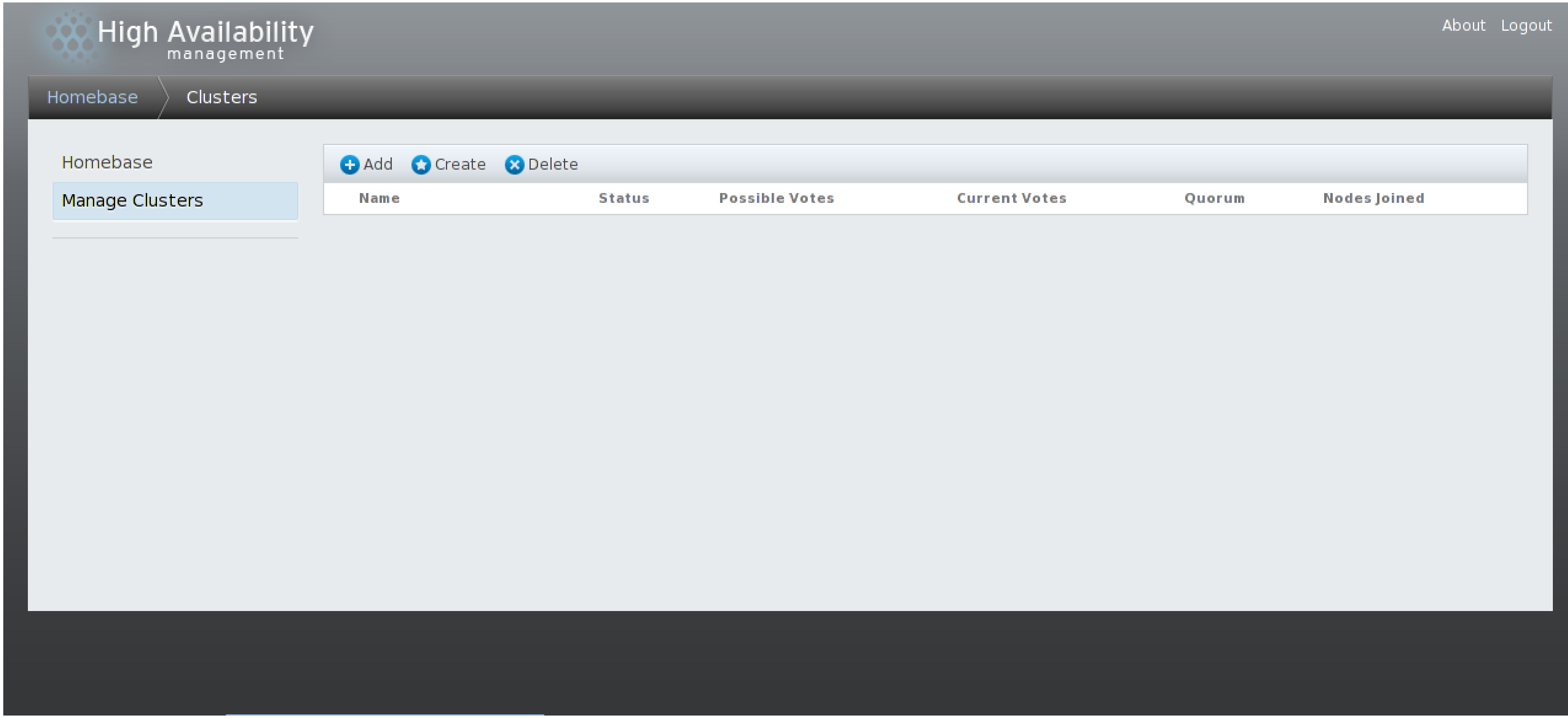
Features of Conga

- One Web interface for managing cluster and storage
- Automated Deployment of Cluster Data and Supporting Packages
- Easy Integration with Existing Clusters
- Integration of Cluster Status and Logs

Features of Conga

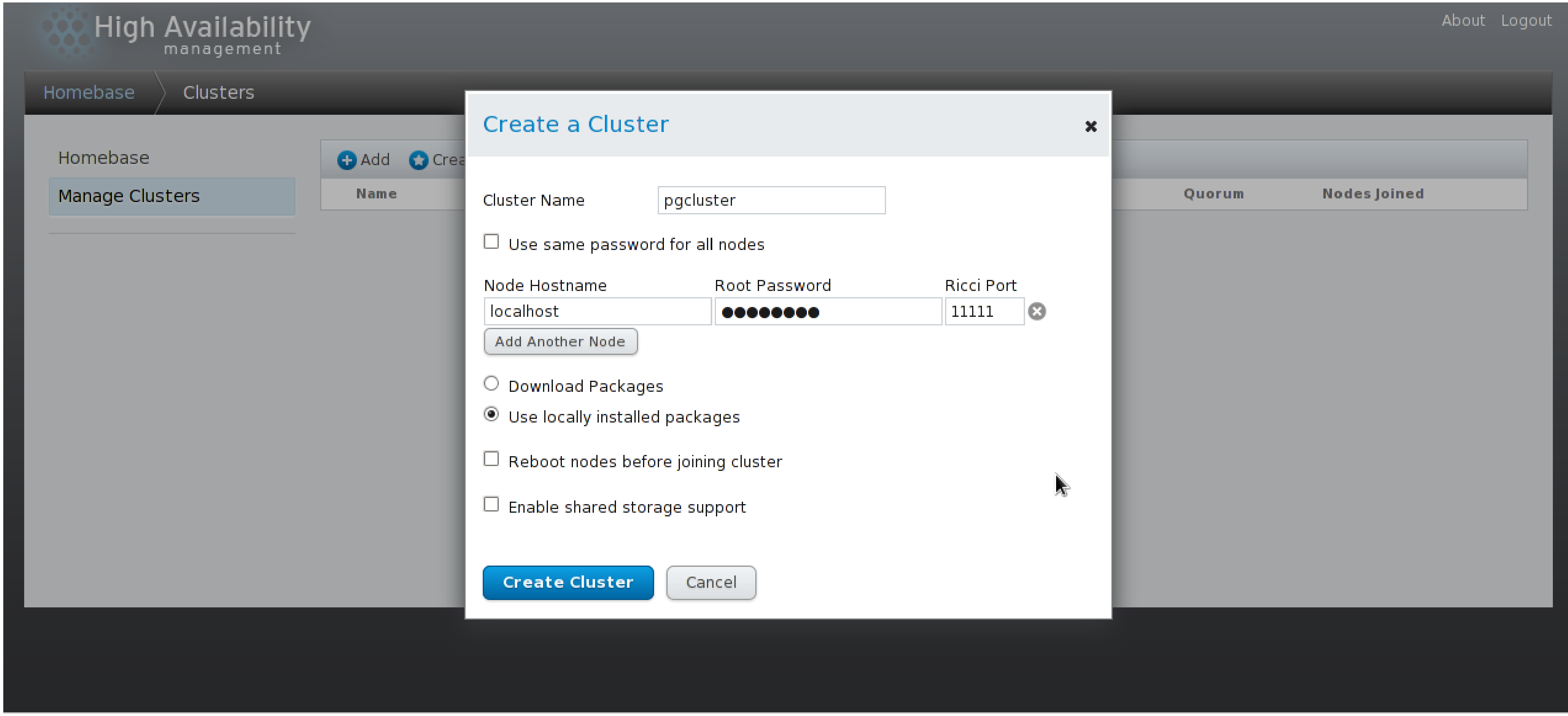
- 2 components: luci and ricci
- Luci: server side tool, communicates with ricci.
- Ricci: agent tool that runs on cluster members, and communicates with luci.
- TGIP (Thanks God It's Python!)

Features of Conga



Name	Status	Possible Votes	Current Votes	Quorum	Nodes Joined

Features of Conga



Create a Cluster

Cluster Name: pgcluster

□ Use same password for all nodes

Node Hostname	Root Password	Ricci Port
localhost	●●●●●●●●	11111

Add Another Node

○ Download Packages

● Use locally installed packages

□ Reboot nodes before joining cluster

□ Enable shared storage support

Create Cluster | Cancel

An example to cluster.conf

- Let me run an editor first :)

Agenda

Setting up PostgreSQL

Setting up PostgreSQL

- No specific **tuning** needed, except:
  - listen_addresses
  - unix_socket_directory
  - external_pid_file

- However, if you are using more than one node, you will want to be careful while sharing hardware resources.
- Many people use Streaming Replication and Hot Standby nowadays, along with RHCS, in order to be able to use the standby machine even for read-only queries.

Agenda

Failover

Failover

- RHCS handles failover properly.
- It detects dead node, and moves service to the next machine, as configured in cluster.conf
- Once the dead machine is up, service may or may not be transferred back to the “master” node.
- ~30 - 60 seconds of downtime during this operation.

Agenda

Postgres-XC!

Postgres-XC

- A new synchronous and transparent clustering solution for PostgreSQL, providing both read and write scalability
- 0.9.7
- http://postgres-xc.sourceforge.net
- Can be used with or without RHCS, and it will work more or less like Oracle RAC.
- Under heavy development

Agenda

Tips & Tricks

Tips and tricks

- Use quorum feature. It will incredibly increase HA.
- Try not to move to back to the “master” node, when it comes back. -> Increased availability

Agenda

Questions?

Questions

Questions please.

PostgreSQL Clustering with Red Hat Cluster Suite

Devrim GÜNDÜZ
Principal Systems Engineer
EnterpriseDB
devrim.gunduz@EnterpriseDB.com