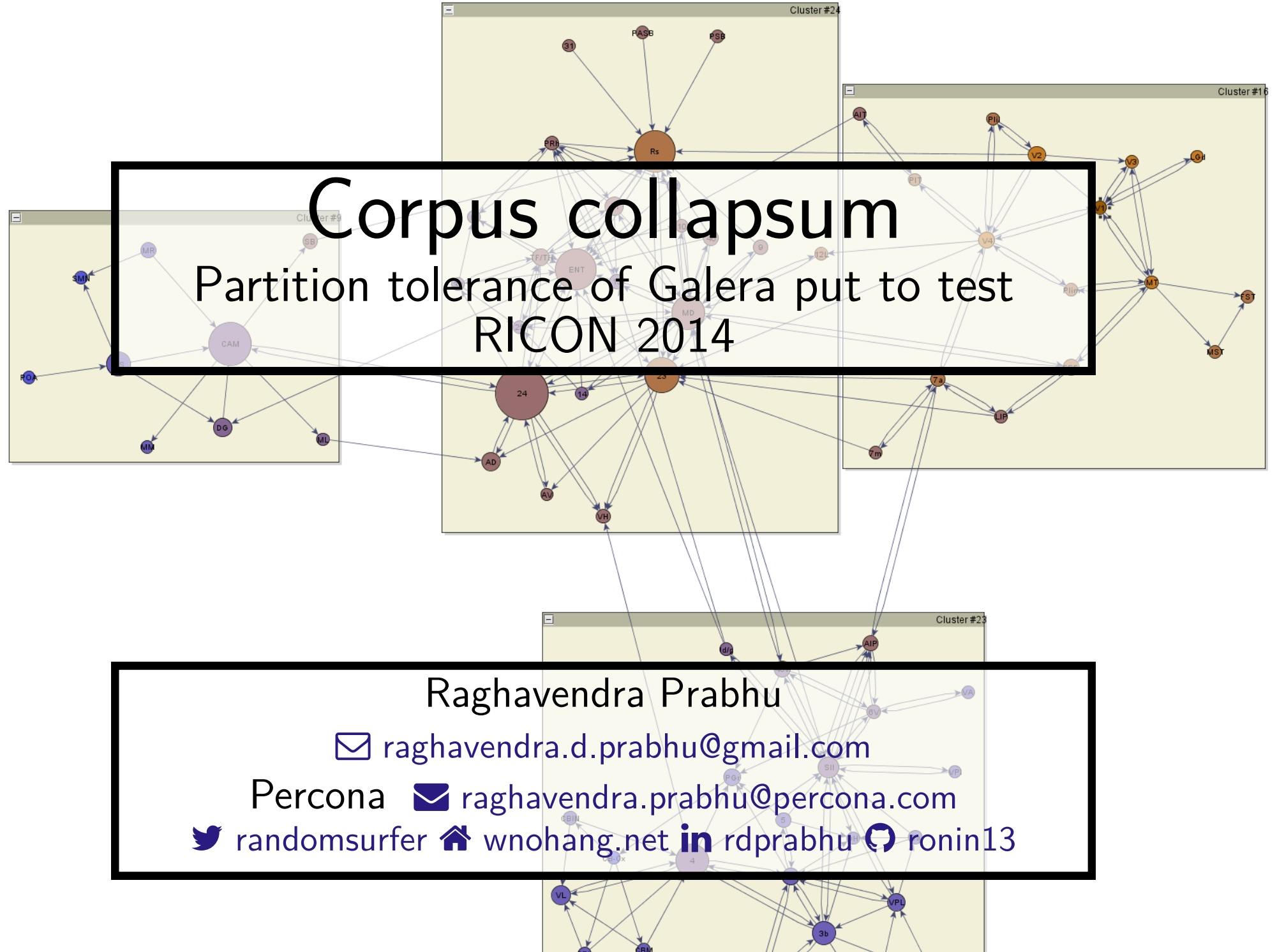


Corpus collapsum

Partition tolerance of Galera put to test

RICON 2014

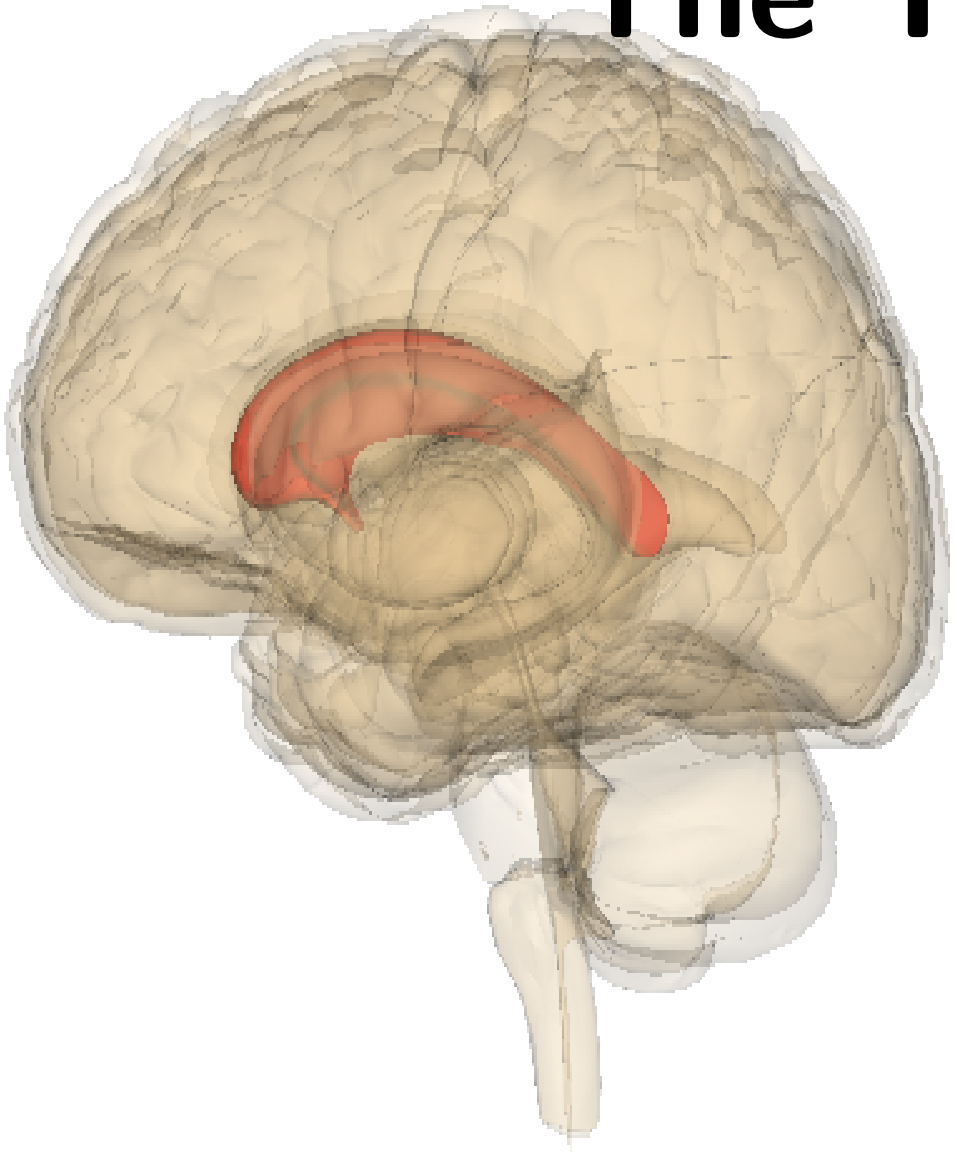
Raghavendra Prabhu

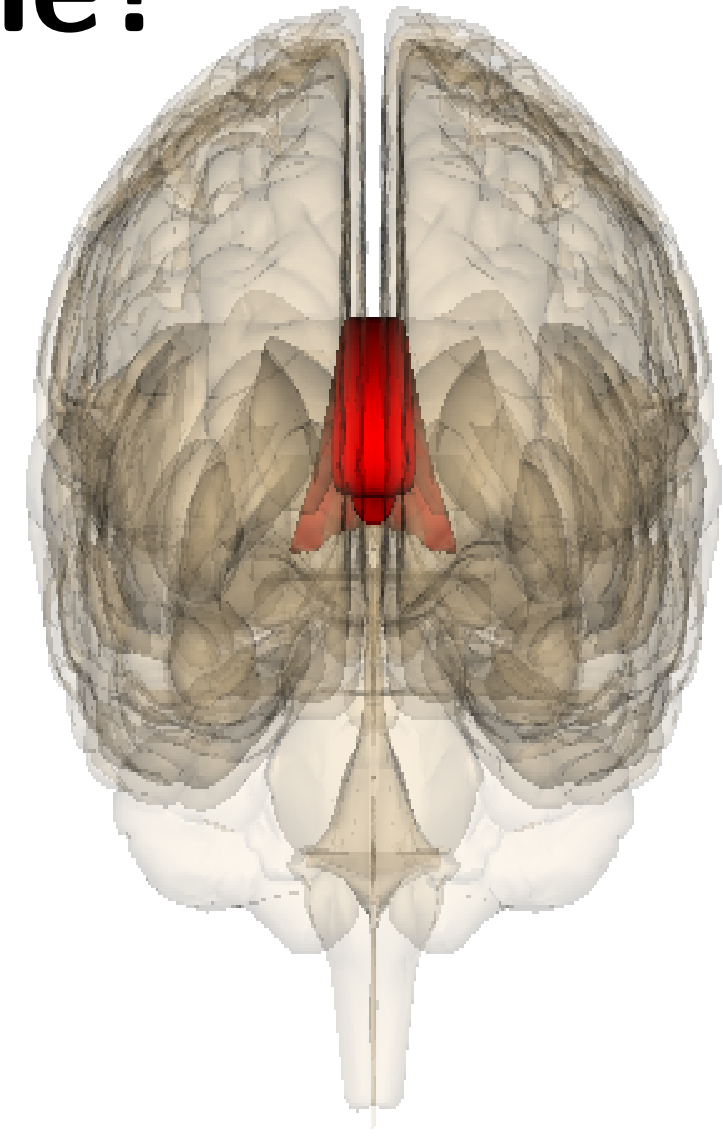
raghavendra.d.prabhu@gmail.com

Percona raghavendra.prabhu@percona.com

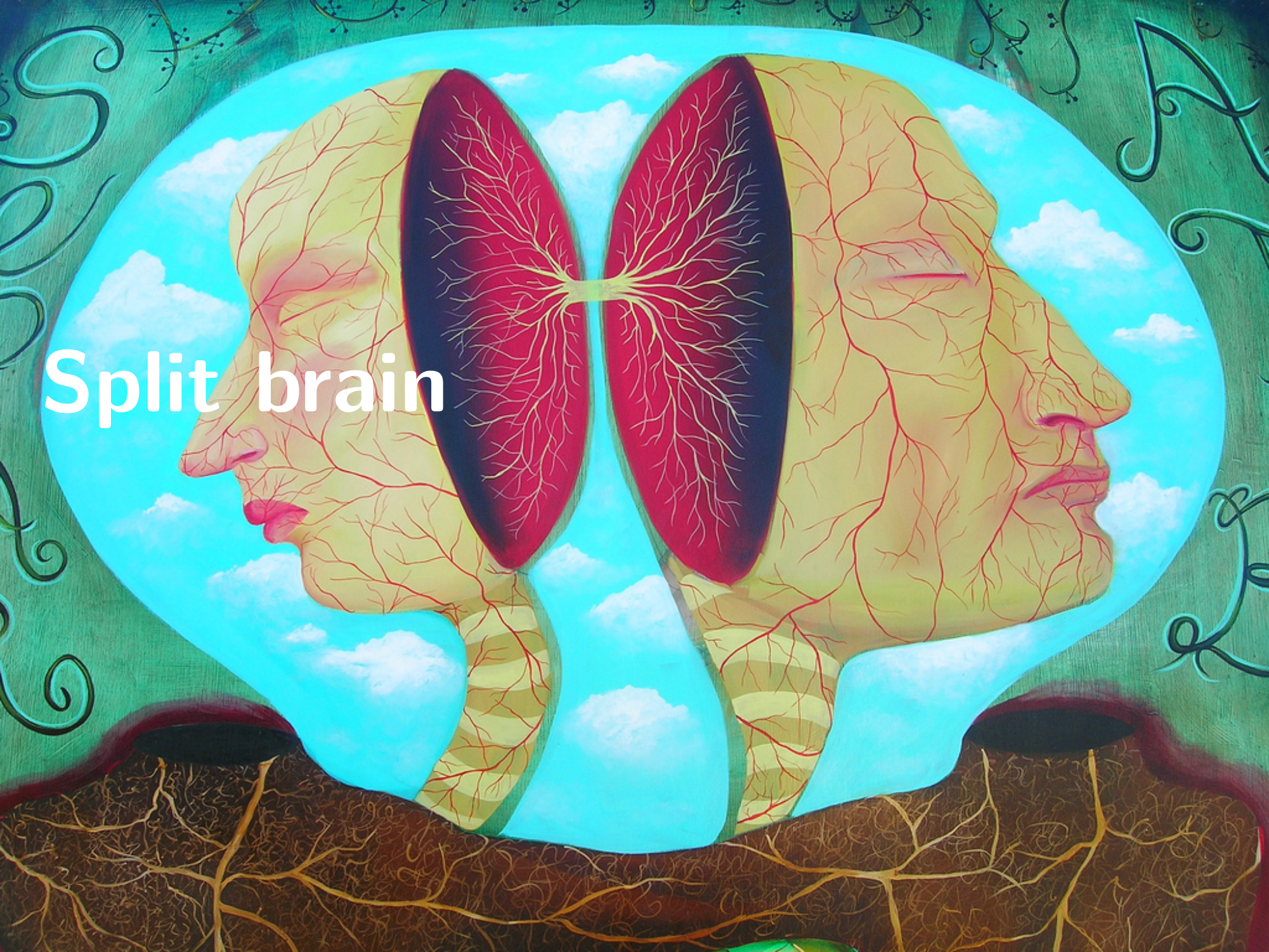
randomsurfer wnohang.net rdprabhu ronin13

The Title?






Our Cluster


Split brain

Seed quotes..

" 'Network is reliable' - a fallacy of the distributed system. "

" A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable. " - Leslie Lamport

" Never attribute to malice that which is adequately explained by stupidity. " - Hanlon's Razor

" Never attribute to Byzantine failure which can be explained by an ill node(s) " - Me

Seed quotes..

" 'Network is reliable' - a fallacy of the distributed system. "

" A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable. " - Leslie Lamport

" Never attribute to malice that which is adequately explained by stupidity. " - Hanlon's Razor

" Never attribute to Byzantine failure which can be explained by an ill node(s) " - Me

Seed quotes..

" 'Network is reliable' - a fallacy of the distributed system. "

" A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable. " - Leslie Lamport

" Never attribute to malice that which is adequately explained by stupidity. " - Hanlon's Razor

" Never attribute to Byzantine failure which can be explained by an ill node(s) " - Me

RICON 2014

Seed quotes..

“ 'Network is reliable' - a fallacy of the distributed system. ”

“ A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable. ” - Leslie Lamport

“ Never attribute to malice that which is adequately explained by stupidity. ” - Hanlon's Razor

“ Never attribute to Byzantine failure which can be explained by an ill node(s) ” - Me


20000 feet view

Actors

- Database - WSREP/PXC
- Plugin - Galera
- Traffic control
  - Traffic Control - tc
  - NetEm

RICON 2014

Actors

- Database - WSREP/PXC
- Plugin - Galera
- Traffic control
  - Traffic Control - tc
  - NetEm

Actors

- Database - WSREP/PXC
- Plugin - Galera
- Traffic control
  - Traffic Control - tc
  - NetEm

Actors

- Containers - Docker
- Load
  - Generators - Sysbench, RQG
- Network
  - Dnsmasq
  - nsenter

RICON 2014

Actors

- Containers - Docker
- Load
  - Generators - Sysbench, RQG
- Network
  - Dnsmasq
  - nsenter

Actors

- Jenkins
  - Build flow and CI
- Storage
  - Why

But why

- The ‘P’ in CAP
- WAN scalability
- Real Reason - fun!
- Tolerance to latency variance

But why

- The ‘P’ in CAP
- WAN scalability
- Real Reason - fun!
- Tolerance to latency variance

But why

- The ‘P’ in CAP
- WAN scalability
- Real Reason - fun!
- Tolerance to latency variance

But why

- The ‘P’ in CAP
- WAN scalability
- Real Reason - fun!
- Tolerance to latency variance

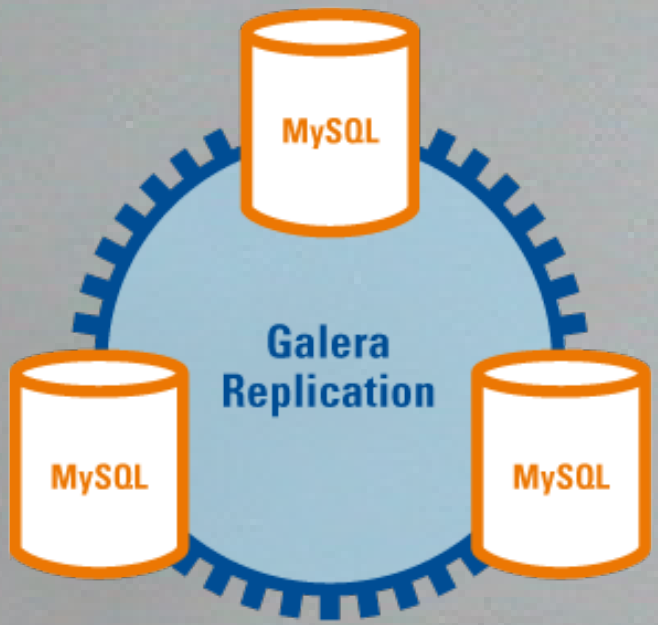
But why

- Failures in warehouses.
- Not quorum, but consensus.
- Real world networks and synchronous replication
  - Delay
  - Partition

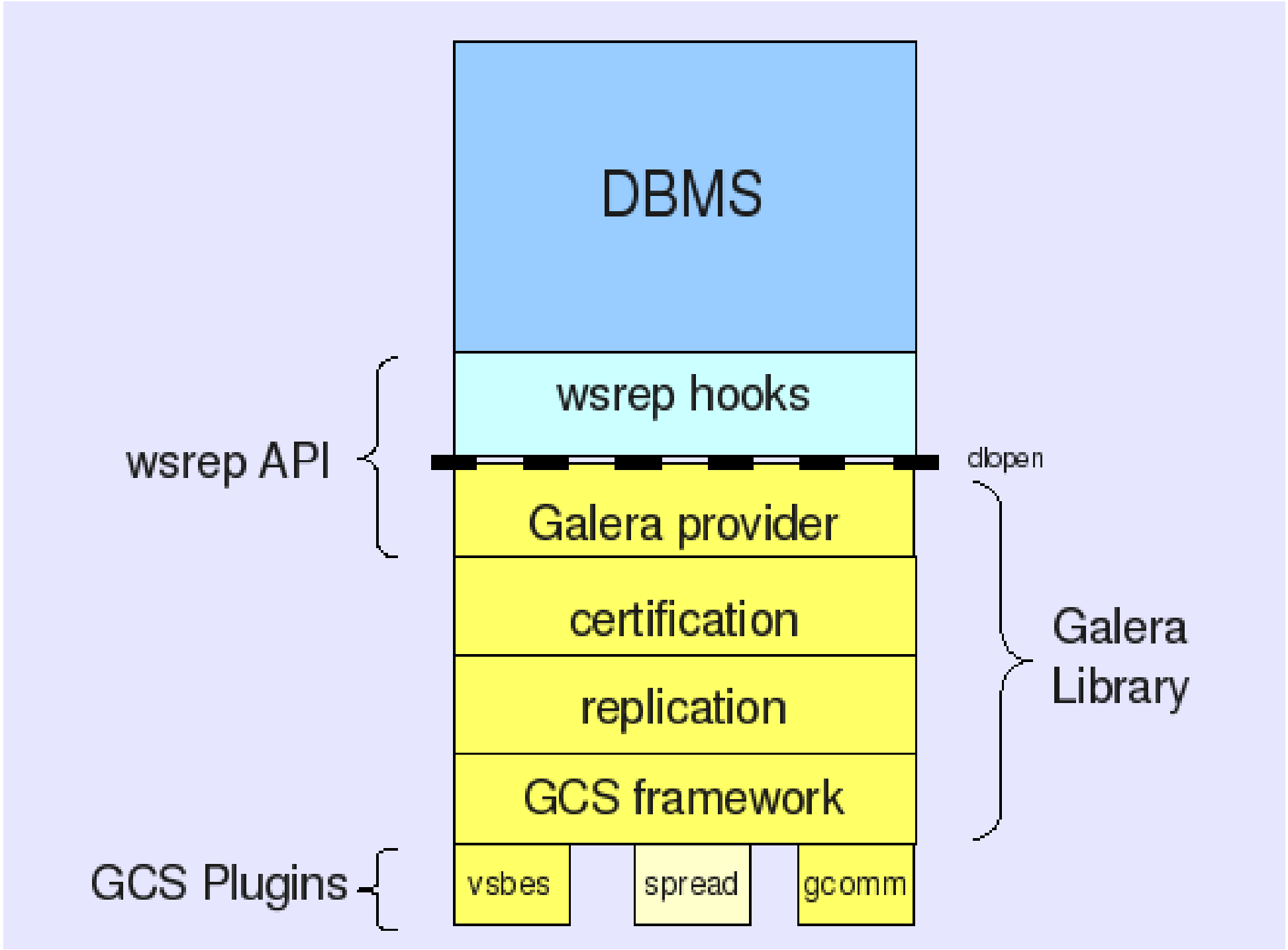
RICON2014

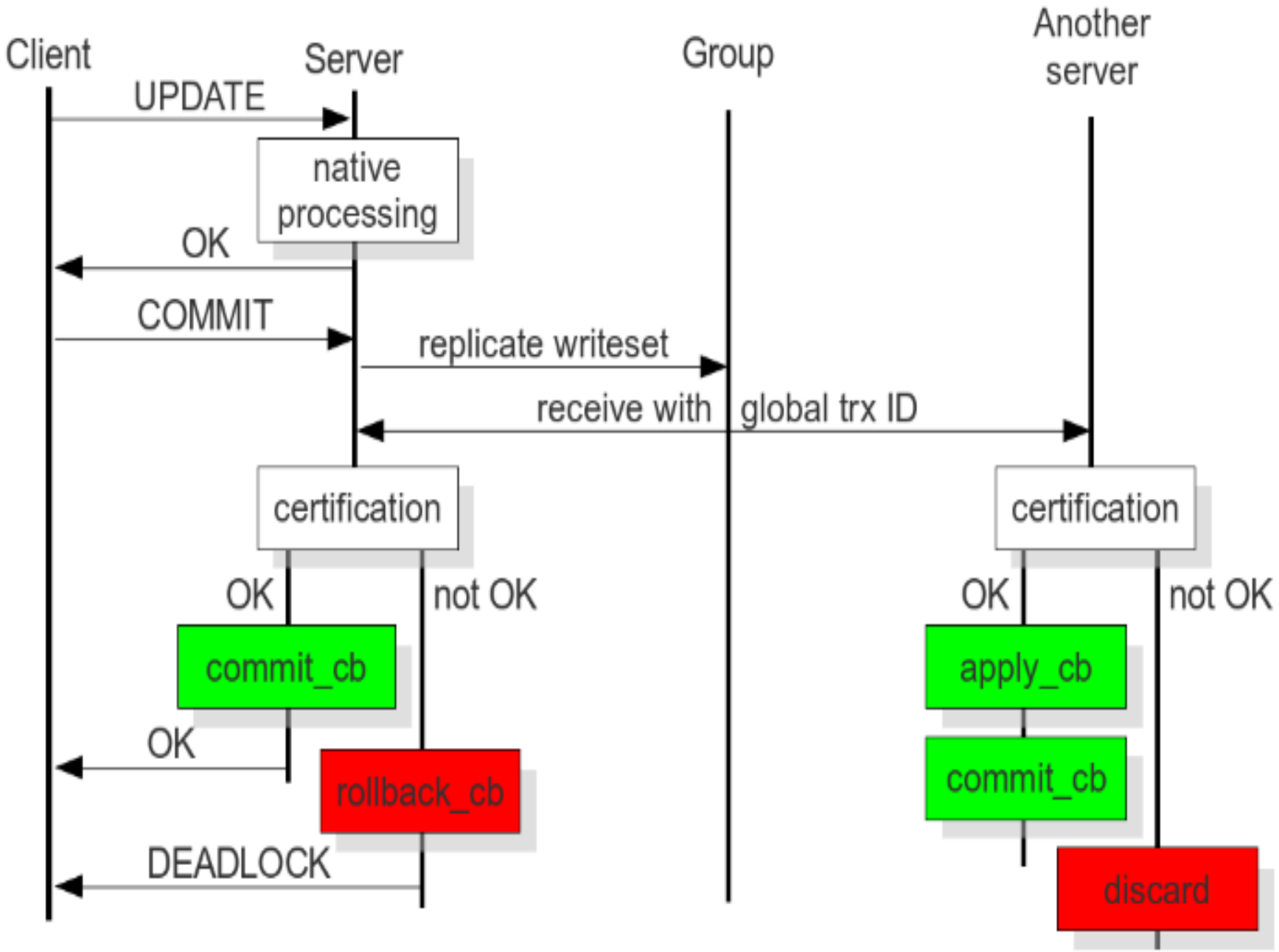

Galera

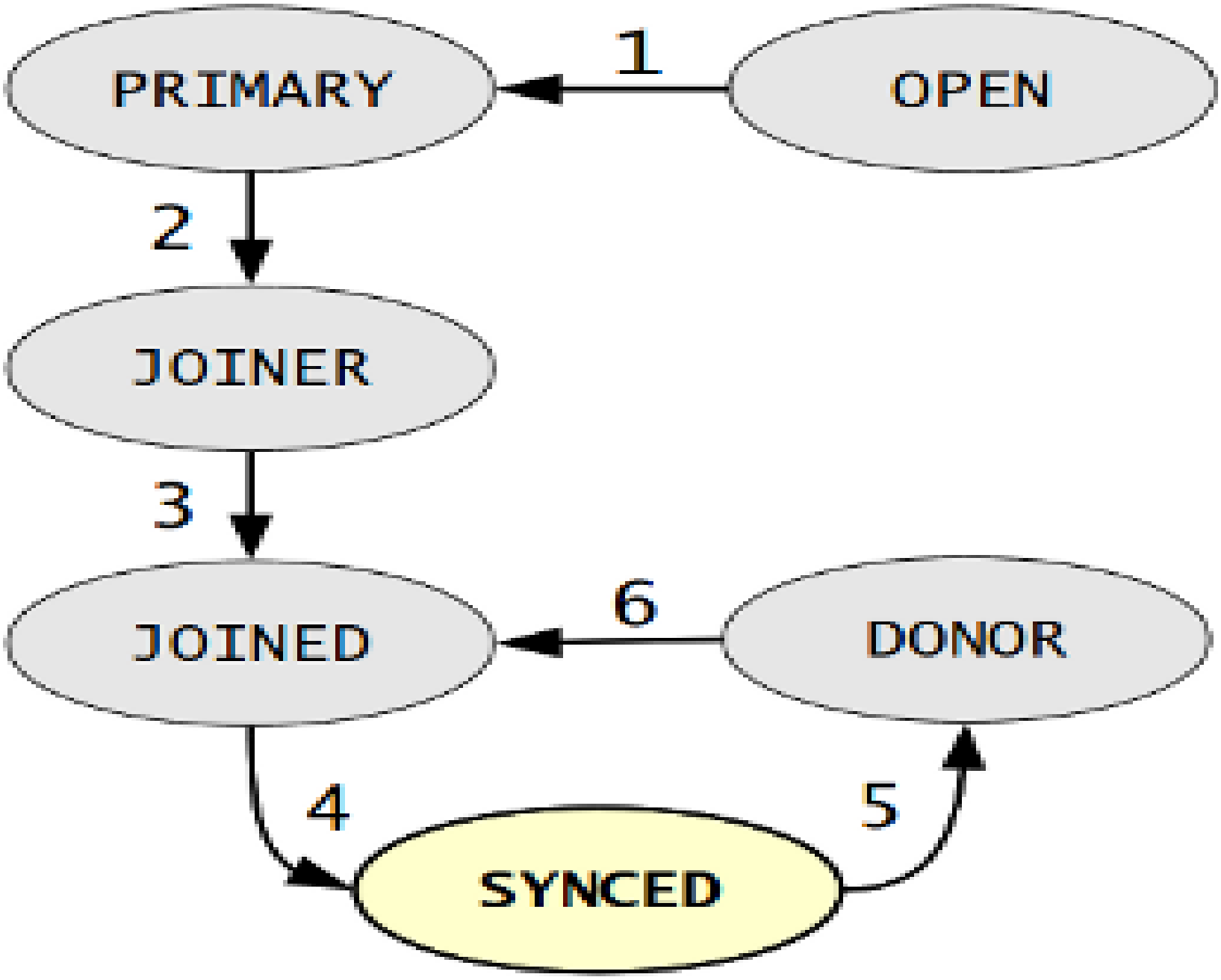
Galera



- Data-centric approach
- EVS
- Causality and Synchronous
- Latency


DBMS
wsrep hooks
wsrep API
dlopen
Galera provider
certification
replication
GCS framework
Galera
Library
GCS Plugins
vsbes
spread
gcomm


Client
Server
Group
Another server
UPDATE
native processing
OK
COMMIT
replicate writeset
receive with
global trx ID
certification
certification
OK
not OK
OK
not OK
commit_cb
apply_cb
OK
commit_cb
rollback_cb
DEADLOCK
discard


PRIMARY
1
OPEN
2
JOINER
3
JOINED
6
DONOR
4
5
SYNCED

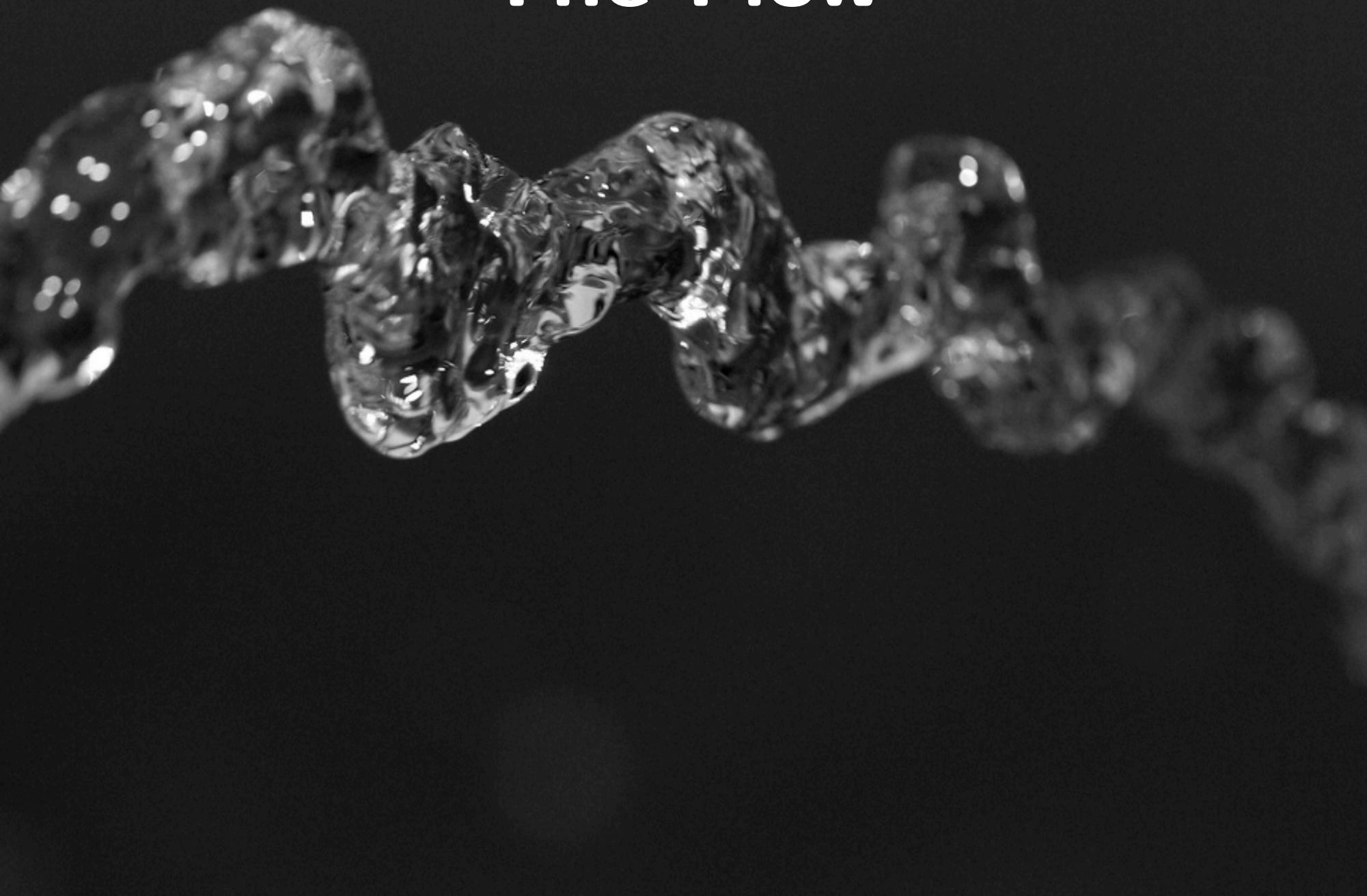

Where did it start

Where did it start

- Bug! https://bugs.launchpad.net/galera/+bug/1274192
- Loss of PC
- Crash
- HA goal

One can bring the whole down

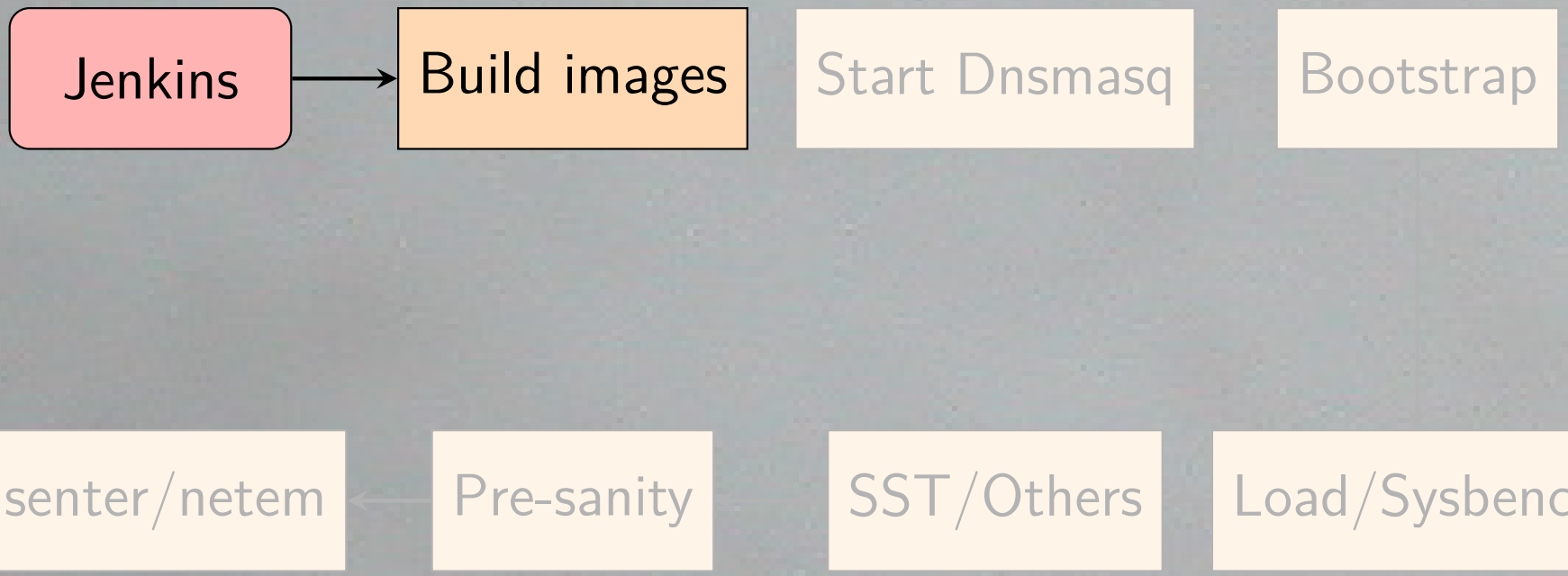
The Flow



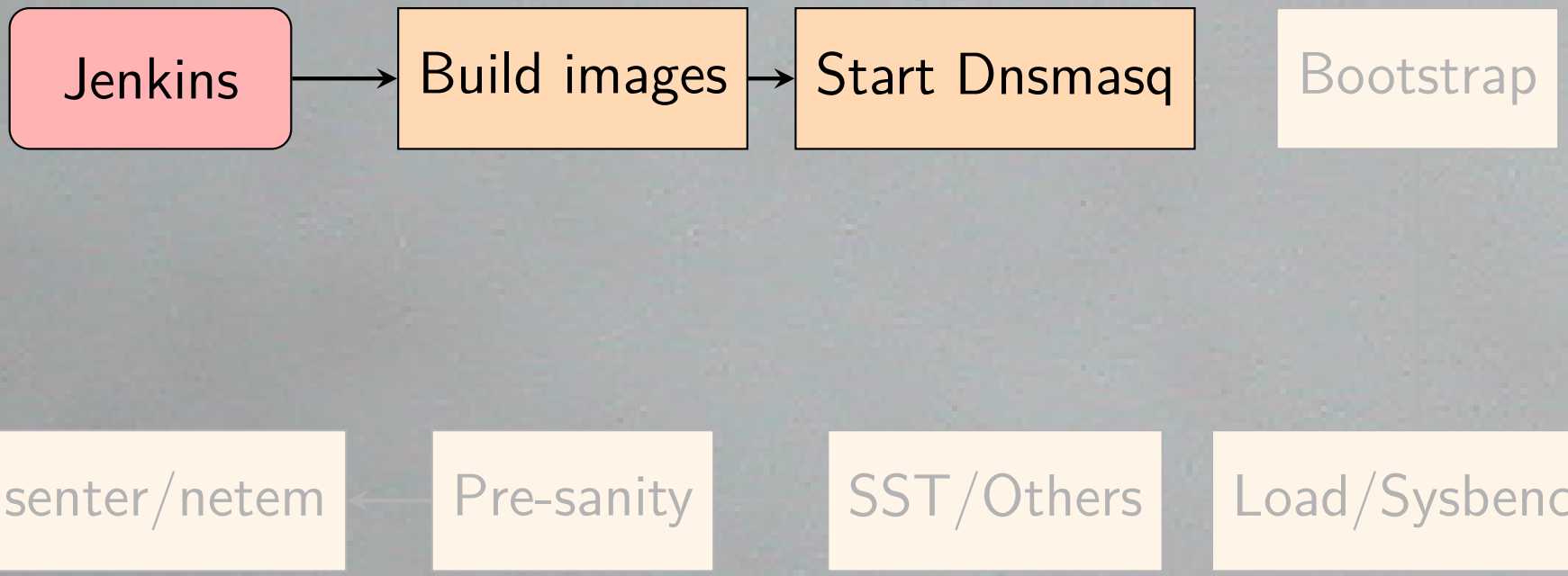
Basic Flow

RICON 2014

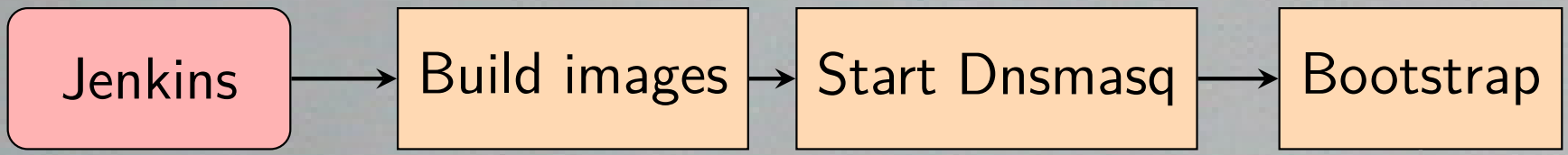
Basic Flow

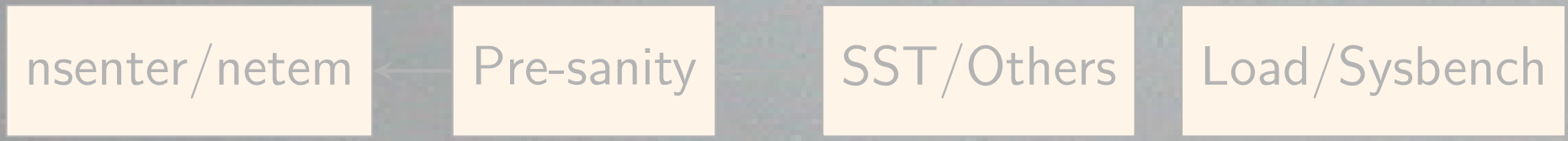

Jenkins
Build images
Start Dnsmasq
Bootstrap
nsenter/netem
Pre-sanity
SST/Others
Load/Sysbench

Basic Flow


Jenkins
Build images
Start Dnsmasq
Bootstrap
nsenter/netem
Pre-sanity
SST/Others
Load/Sysbench

Basic Flow





RICON 2014

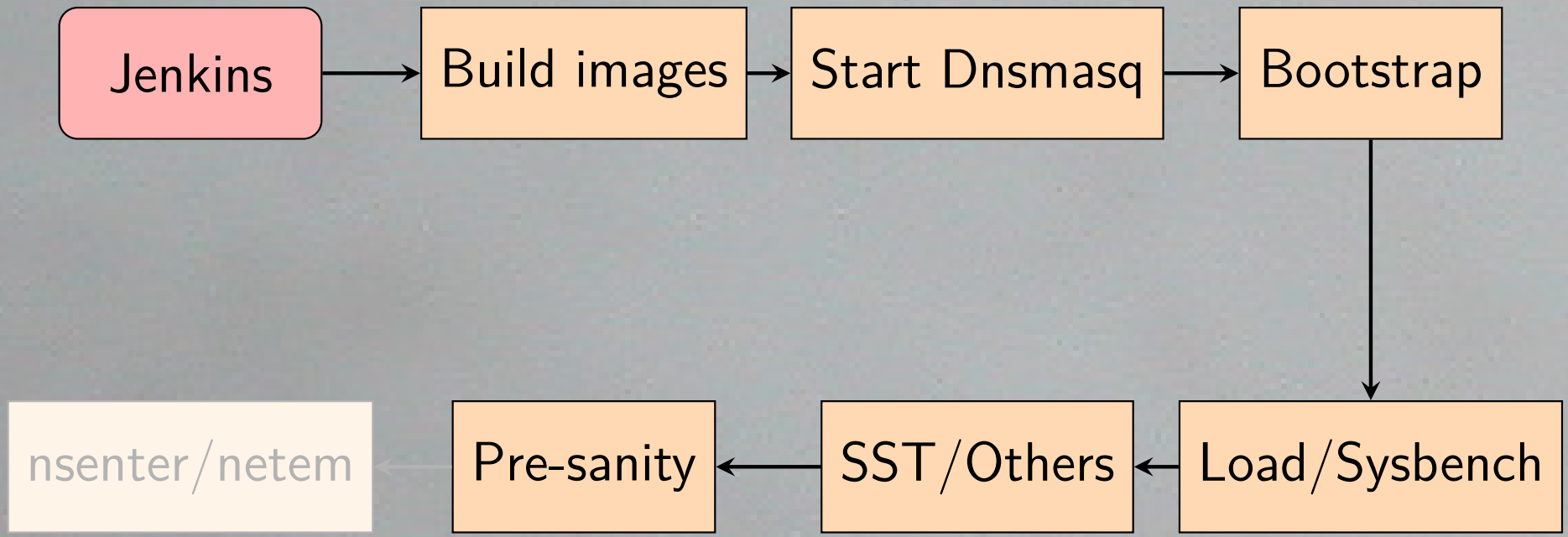
Basic Flow


Jenkins
Build images
Start Dnsmasq
Bootstrap
Load/Sysbench
SST/Others
Pre-sanity
nsenter/netem

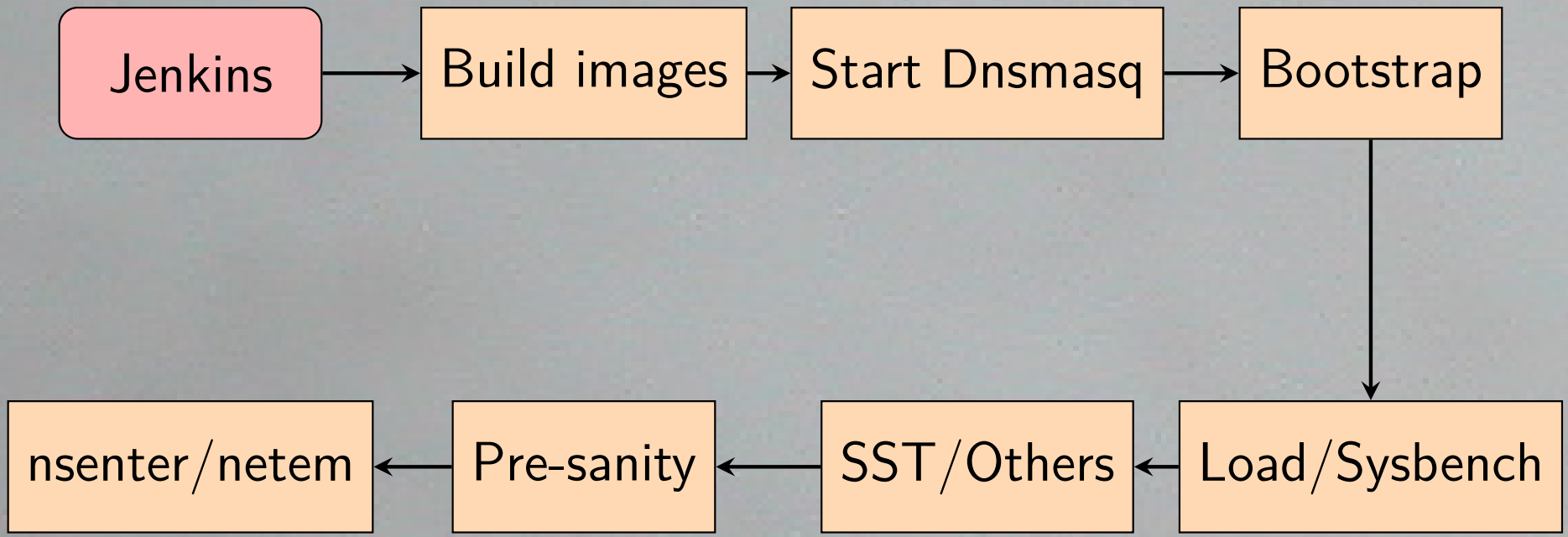
Basic Flow

RICON 2014

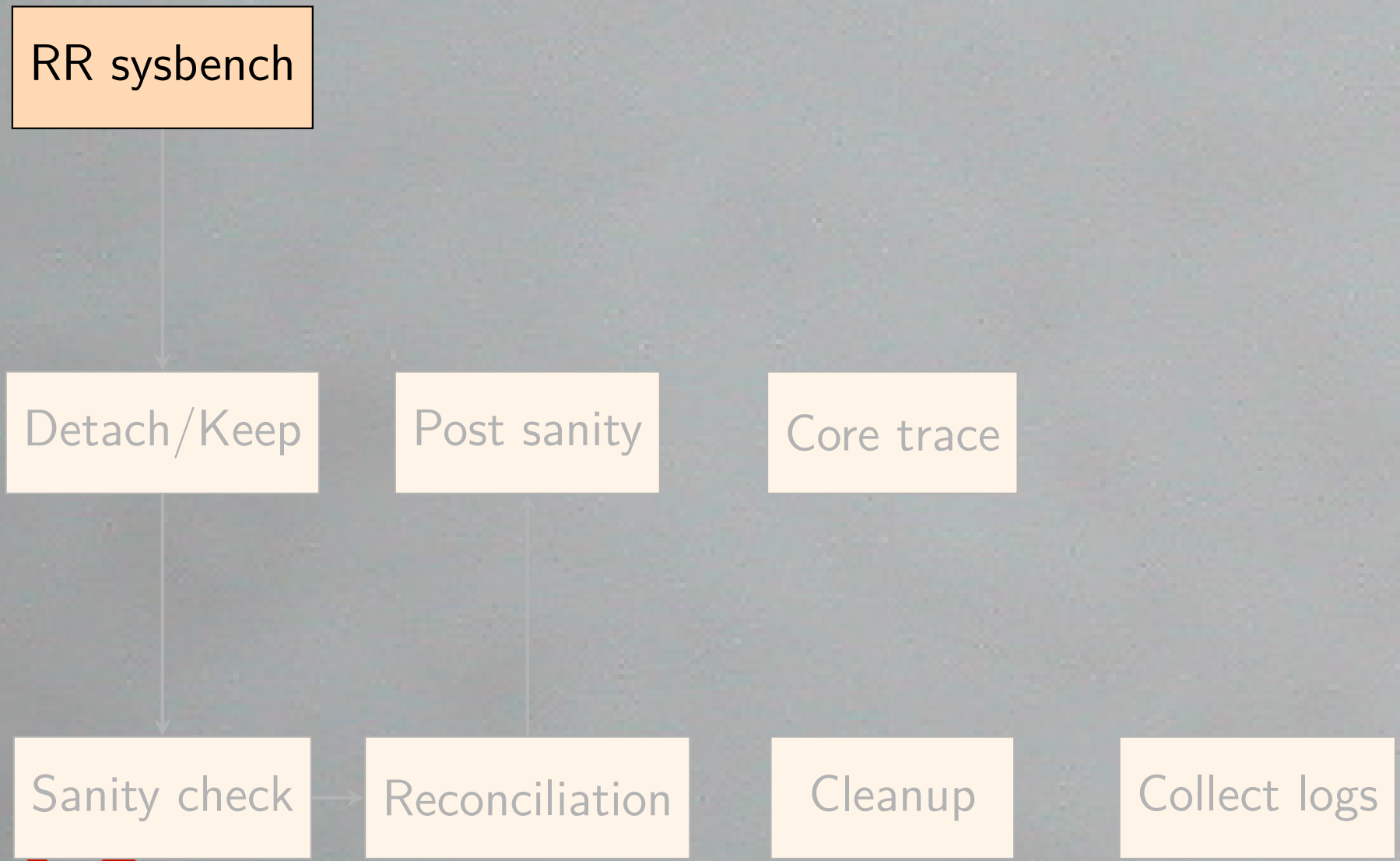
Basic Flow


Jenkins
Build images
Start Dnsmasq
Bootstrap
Load/Sysbench
SST/Others
Pre-sanity
nsenter/netem

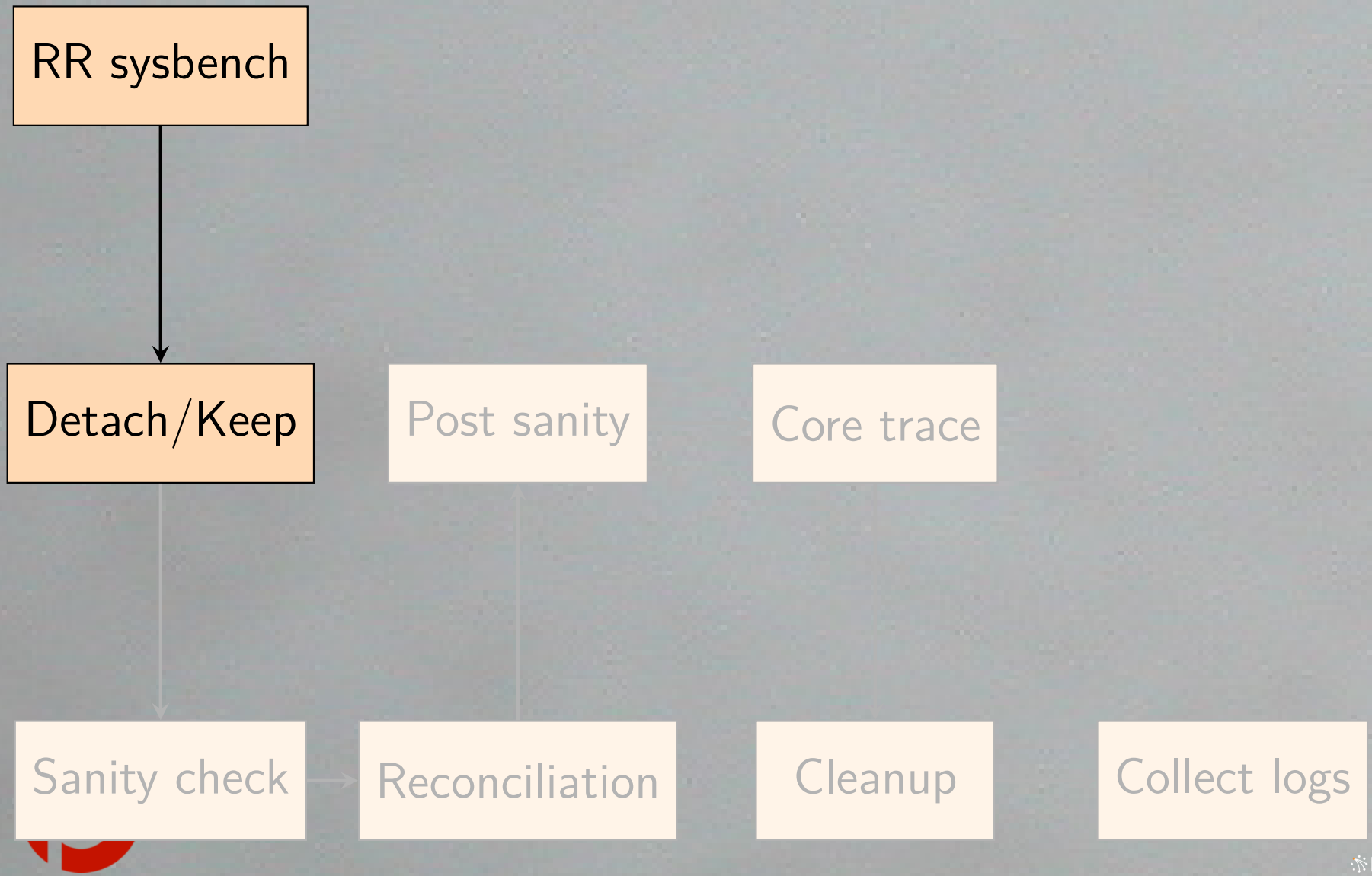
Basic Flow


Jenkins
Build images
Start Dnsmasq
Bootstrap
Load/Sysbench
SST/Others
Pre-sanity
nsenter/netem

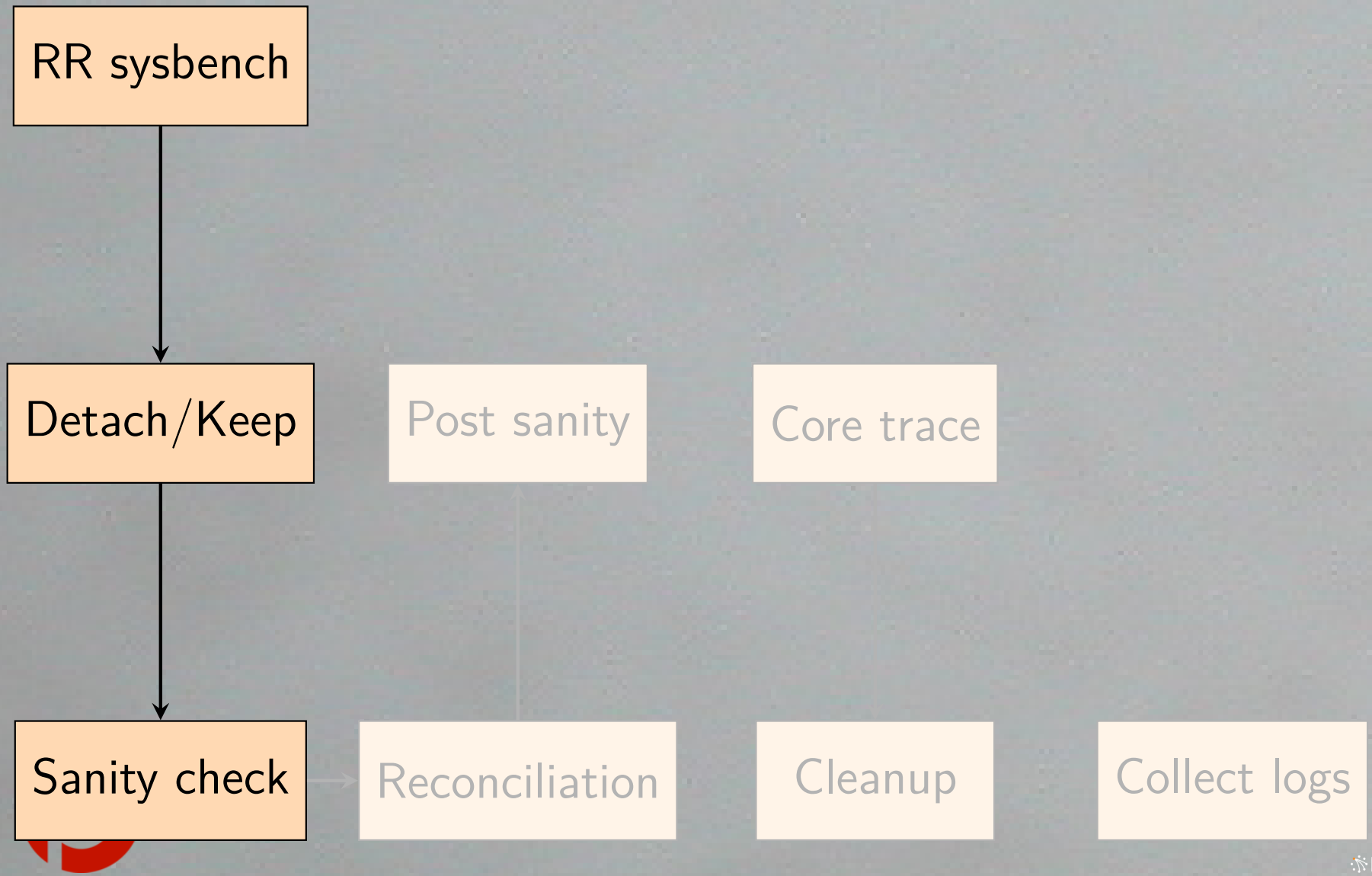
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

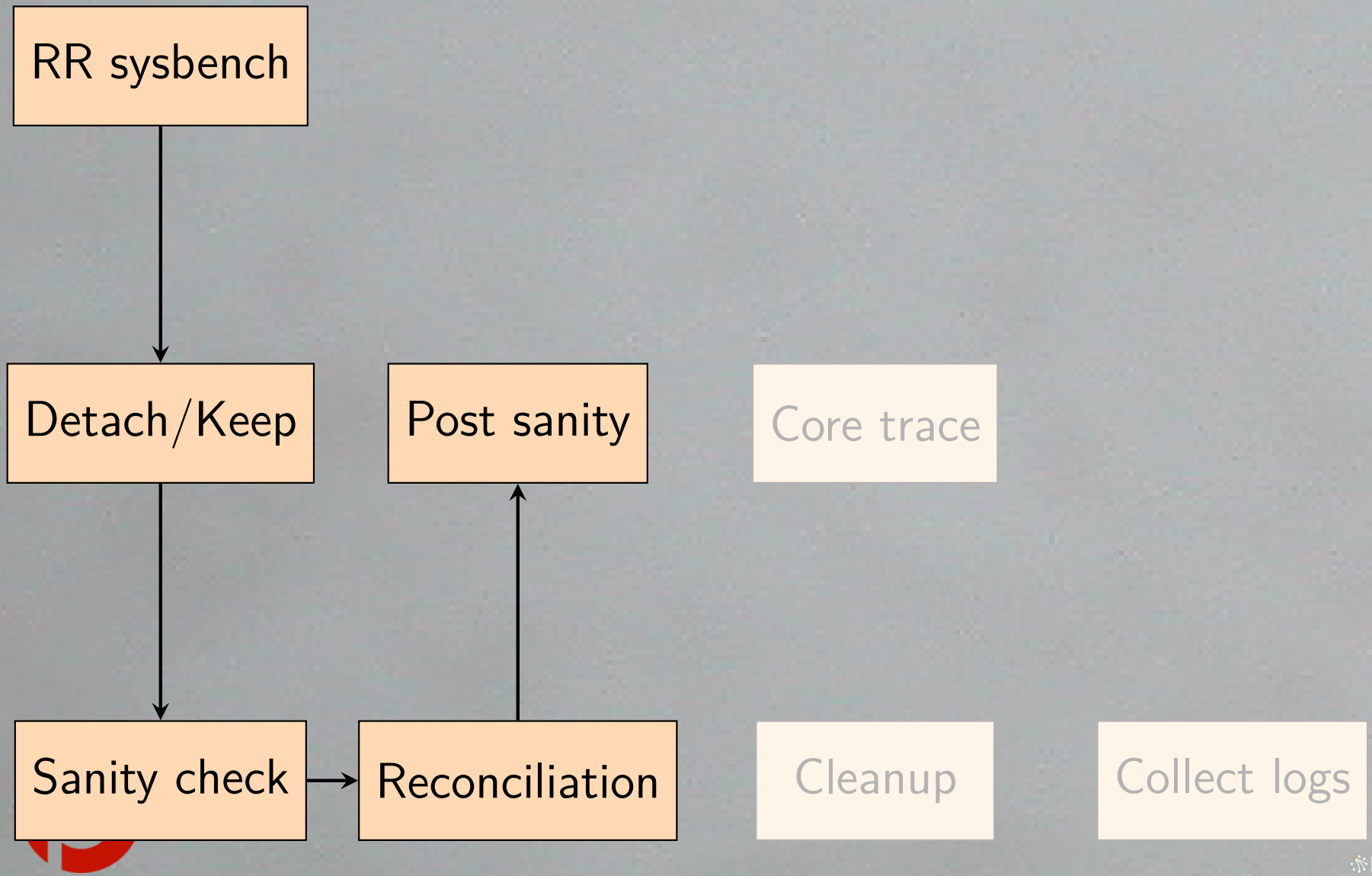
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

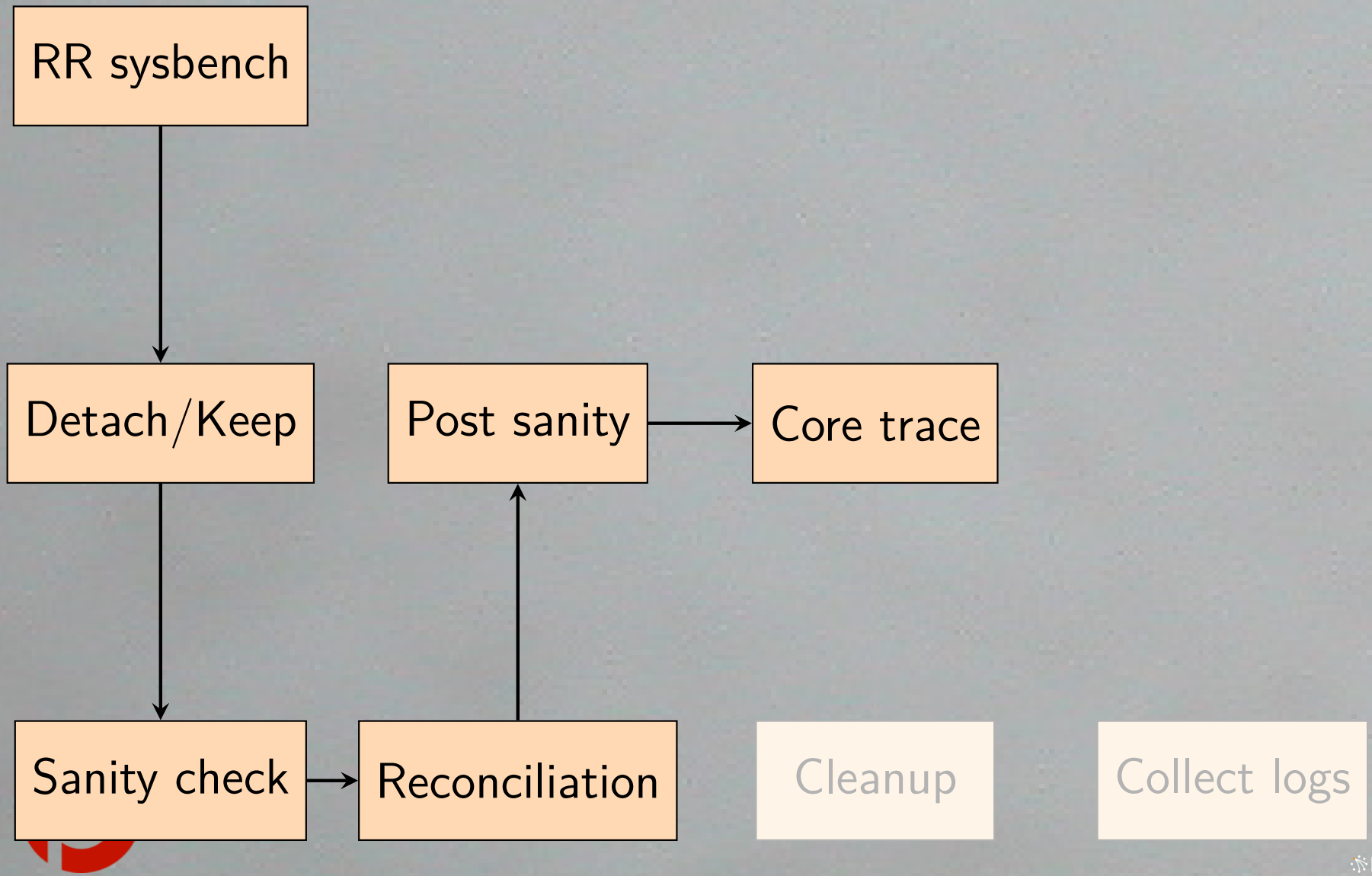
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

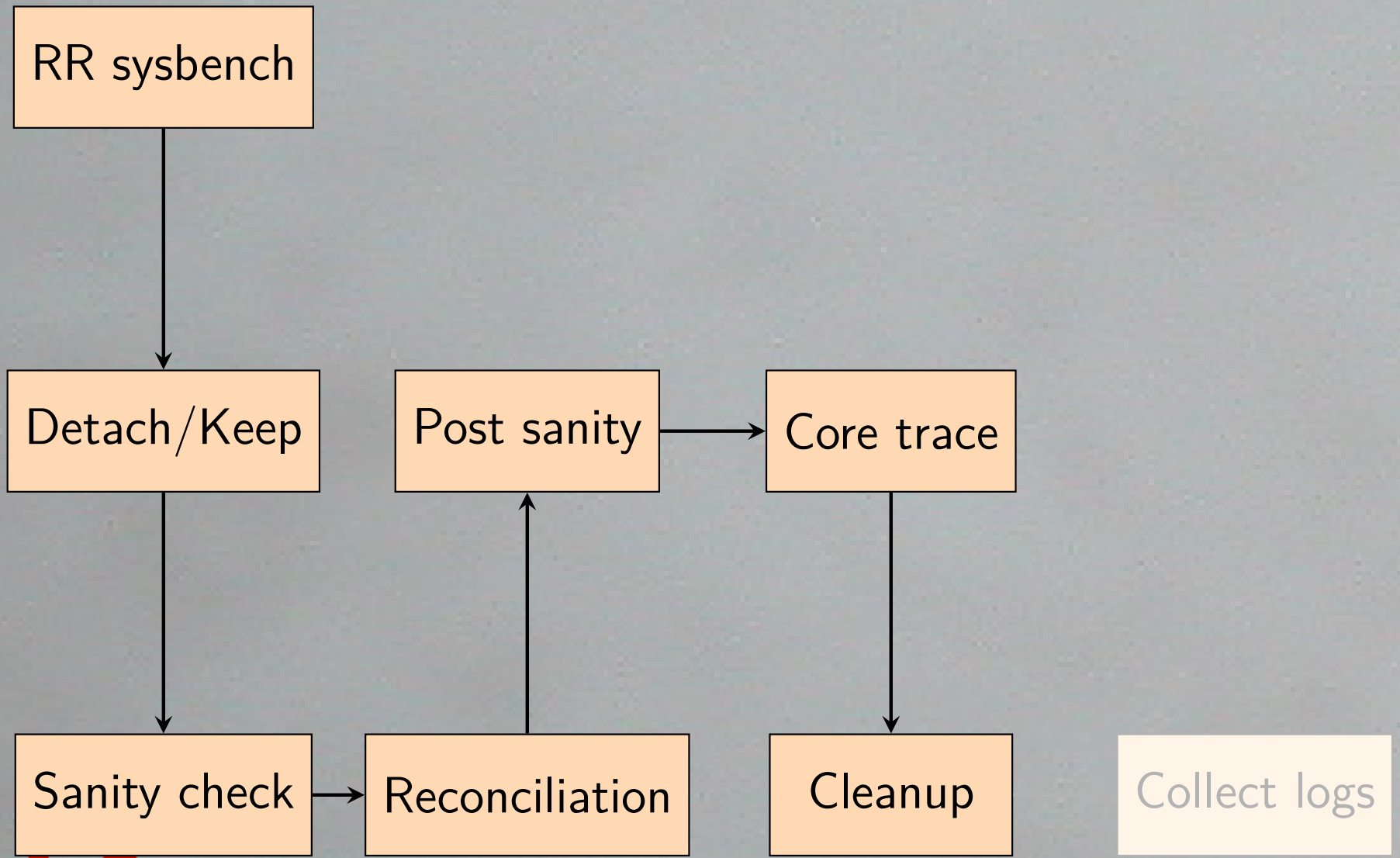
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

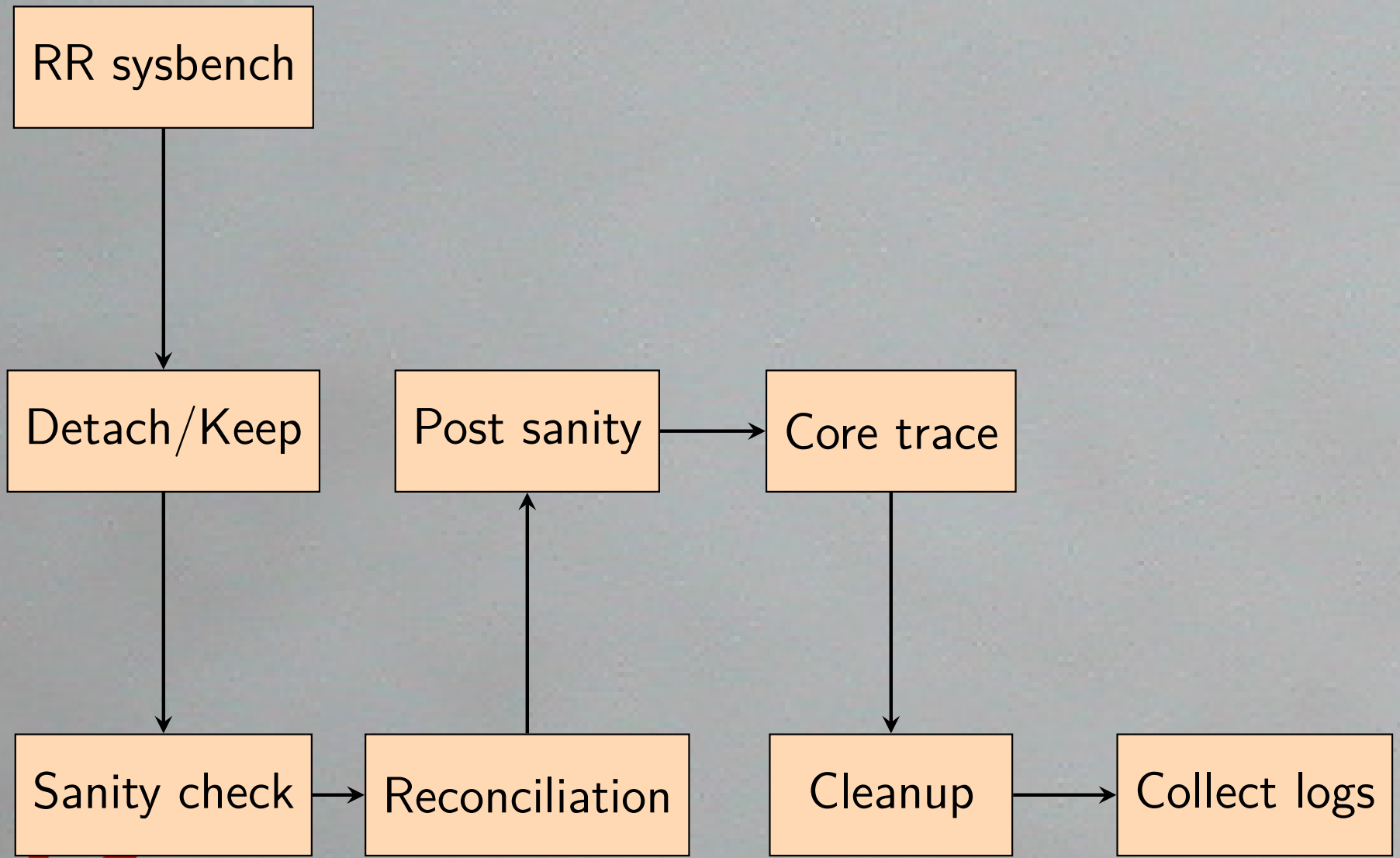
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

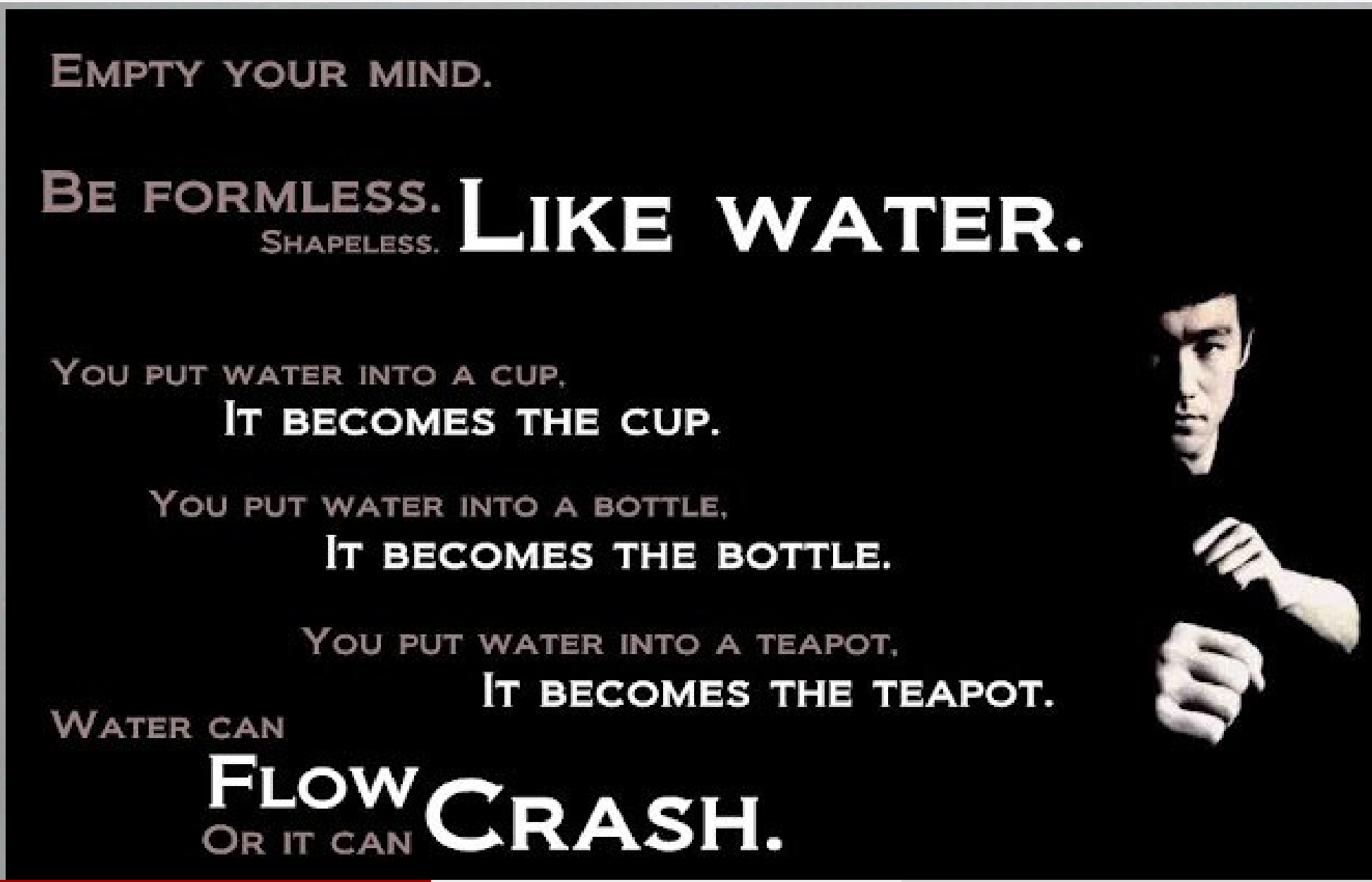
Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

Basic Flow


RR sysbench
Detach/Keep
Post sanity
Core trace
Sanity check
Reconciliation
Cleanup
Collect logs

Cluster Resilience



Parameters

- Sysbench
- Segment
- Reconciliation period
- Loss nodes

Parameters

- Sysbench
- Segment
- Reconciliation period
- Loss nodes

Parameters

- Sysbench
- Segment
- Reconciliation period
- Loss nodes

Parameters

- Sysbench
- Segment
- Reconciliation period
- Loss nodes

Parameters

- NetEm
- Detach loss
- Fsync
- Shutdown

RICON 2014

Parameters

- NetEm
- Detach loss
- Fsync
- Shutdown

RICON 2014

Parameters

- NetEm
- Detach loss
- Fsync
- Shutdown

Parameters

- NetEm
- Detach loss
- Fsync
- Shutdown


Containers!
APL
tex
WESTFALIA EXPRESS
LONDON
IMO 9233832

Docker

- Why not virtualize
  - Occam
  - Namespaces
- Simplicity
  - Network
  - One application per node

RICON 2014

Docker

- Portability
  - See same qualitative behavior that I do.
- Reproducibility
  - Makes it determinstic
- Configurable and CI
  - Byproducts

Docker

- QEMU and Docker
- Scalability
  - Performance
  - Feature
- Abstraction of channels

Container Networking

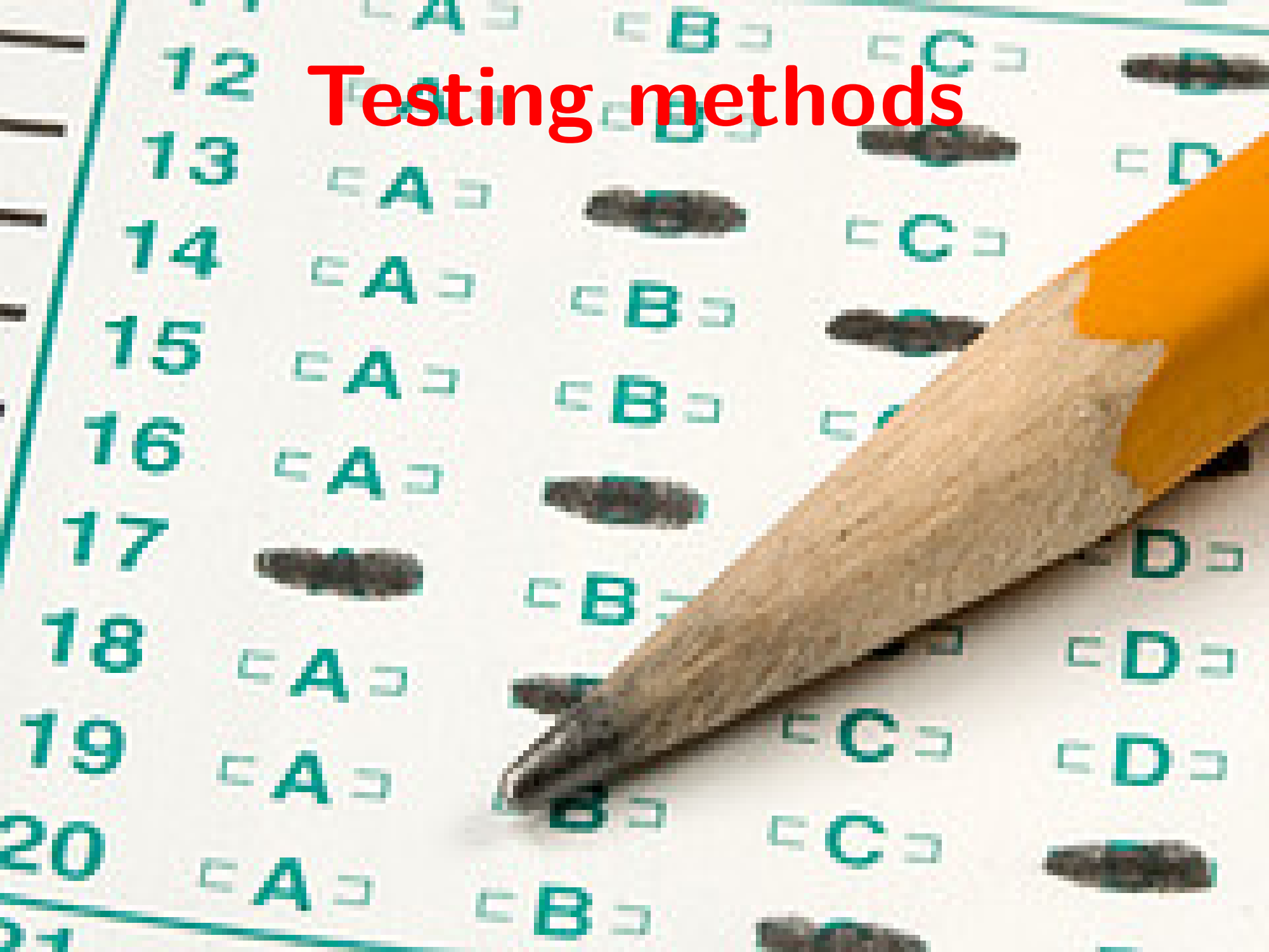
- Linking didn't help
- Dnsmasq to rescue!
  - Hosts file and volumes
  - SIGHUP and refresh
- More elegant methods
  - Swarm

RICON2014

Noise

- Initial setup
  - Bridge
  - Egress only
  - IFB
- Present state
- NetEm
  - tc qdisc buckets
  - packet loss, delay, corruption, duplication, reordering
  - nsenter
- Future
  - Docker exec

Testing methods



Method I

- Qdisc is detached after load
- Objective
  - Time to recover of full cluster
- Done with a larger subset

Method II

- Qdisc is kept till the end
- Objective
  - Formation of primary component
- Comparatively smaller set

Observations

- Post sanity types
  - Why
- Which method is more pertinent
- State transfer issues
  - Beginning
  - During re-emergence

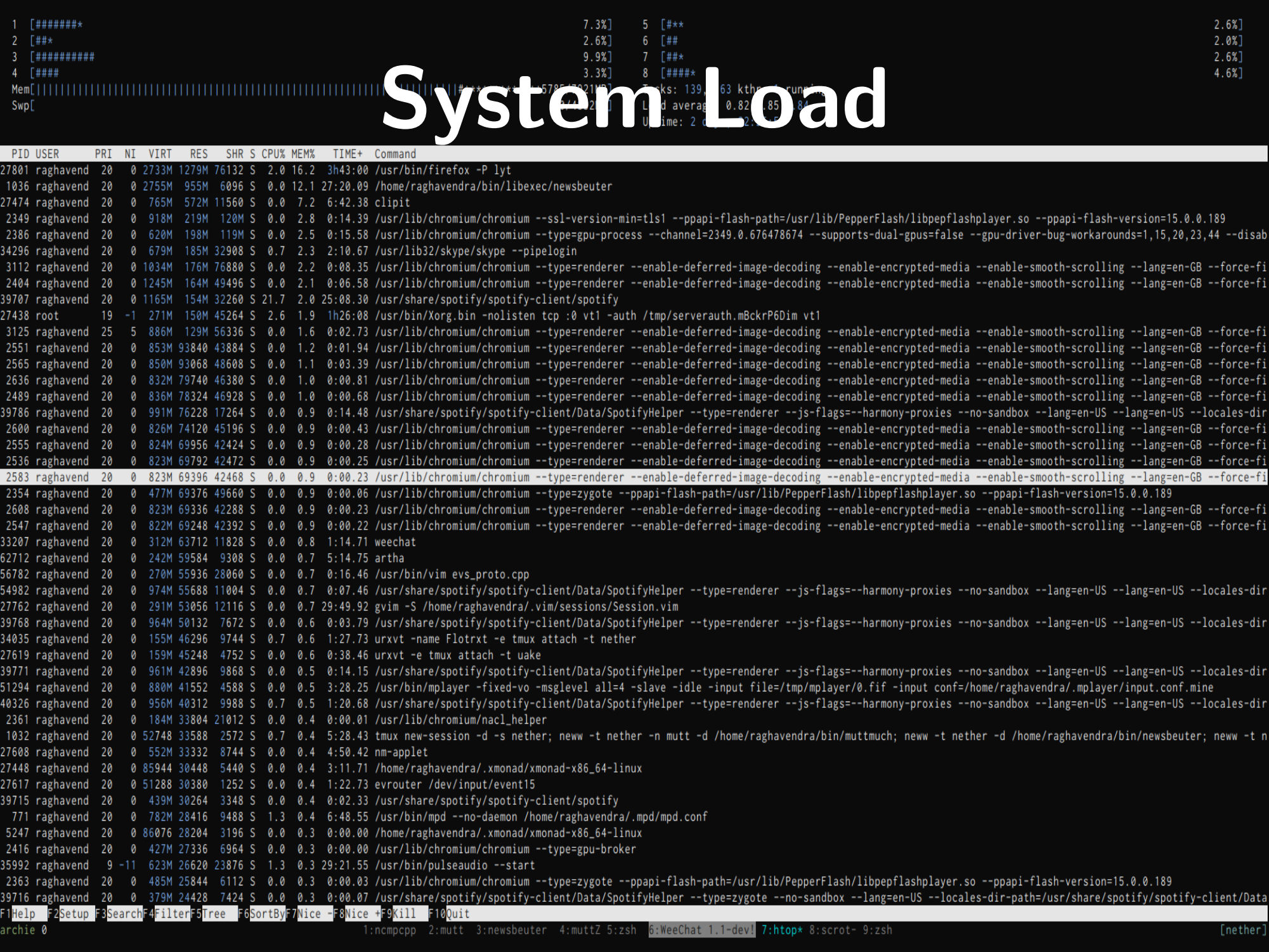
Observations

- Direct load to affected nodes
- Logs
  - journalctl
  - Streaming?

Other noises

- Aim
- Fsync
  - libeatmydata
  - Variance
- Correlation with network
- How with Docker
  - LD_PRELOAD

System Load


1 [******* 7.3%]
2 [##* 2.6%]
3 [########## 9.9%]
4 [#### 3.3%]
5 [#** 2.6%]
6 [## 2.0%]
7 [##* 2.6%]
8 [####* 4.6%]
Mem[
Swp[
Tasks: 139, 63 kthr
Load averag 0.82 .85
Uptime: 2 d
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
27801 raghavend 20 0 2733M 1279M 76132 S 2.0 16.2 3h43:00 /usr/bin/firefox -P lyt
1036 raghavend 20 0 2755M 955M 6096 S 0.0 12.1 27:20.09 /home/raghavendra/bin/libexec/newsbeuter
27474 raghavend 20 0 765M 572M 11560 S 0.0 7.2 6:42.38 clipit
2349 raghavend 20 0 918M 219M 120M S 0.0 2.8 0:14.39 /usr/lib/chromium/chromium --ssl-version-min=tls1 --ppapi-flash-path=/usr/lib/PepperFlash/libpepflashplayer.so --ppapi-flash-version=15.0.0.189
2386 raghavend 20 0 620M 198M 119M S 0.0 2.5 0:15.58 /usr/lib/chromium/chromium --type=gpu-process --channel=2349.0.676478674 --supports-dual-gpus=false --gpu-driver-bug-workarounds=1,15,20,23,44 --disab
34296 raghavend 20 0 679M 185M 32908 S 0.7 2.3 2:10.67 /usr/lib32/skype/skype --pipelogin
3112 raghavend 20 0 1034M 176M 76880 S 0.0 2.2 0:08.35 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2404 raghavend 20 0 1245M 164M 49496 S 0.0 2.1 0:06.58 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
39707 raghavend 20 0 1165M 154M 32260 S 21.7 2.0 25:08.30 /usr/share/spotify/spotify-client/spotify
27438 root 19 -1 271M 150M 45264 S 2.6 1.9 1h26:08 /usr/bin/Xorg.bin -nolisten tcp :0 vt1 -auth /tmp/serverauth.mBckrP6Dim vt1
3125 raghavend 25 5 886M 129M 56336 S 0.0 1.6 0:02.73 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2551 raghavend 20 0 853M 93840 43884 S 0.0 1.2 0:01.94 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2565 raghavend 20 0 850M 93068 48608 S 0.0 1.1 0:03.39 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2636 raghavend 20 0 832M 79740 46380 S 0.0 1.0 0:00.81 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2489 raghavend 20 0 836M 78324 46928 S 0.0 1.0 0:00.68 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
39786 raghavend 20 0 991M 76228 17264 S 0.0 0.9 0:14.48 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=renderer --js-flags=--harmony-proxies --no-sandbox --lang=en-US --lang=en-US --locales-dir
2600 raghavend 20 0 826M 74120 45196 S 0.0 0.9 0:00.43 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2555 raghavend 20 0 824M 69956 42424 S 0.0 0.9 0:00.28 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2536 raghavend 20 0 823M 69792 42472 S 0.0 0.9 0:00.25 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2583 raghavend 20 0 823M 69396 42468 S 0.0 0.9 0:00.23 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2354 raghavend 20 0 477M 69376 49660 S 0.0 0.9 0:00.06 /usr/lib/chromium/chromium --type=zygote --ppapi-flash-path=/usr/lib/PepperFlash/libpepflashplayer.so --ppapi-flash-version=15.0.0.189
2608 raghavend 20 0 823M 69336 42288 S 0.0 0.9 0:00.23 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
2547 raghavend 20 0 822M 69248 42392 S 0.0 0.9 0:00.22 /usr/lib/chromium/chromium --type=renderer --enable-deferred-image-decoding --enable-encrypted-media --enable-smooth-scrolling --lang=en-GB --force-fi
33207 raghavend 20 0 312M 63712 11828 S 0.0 0.8 1:14.71 weechat
62712 raghavend 20 0 242M 59584 9308 S 0.0 0.7 5:14.75 artha
56782 raghavend 20 0 270M 55936 28060 S 0.0 0.7 0:16.46 /usr/bin/vim evs_proto.cpp
54982 raghavend 20 0 974M 55688 11004 S 0.0 0.7 0:07.46 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=renderer --js-flags=--harmony-proxies --no-sandbox --lang=en-US --lang=en-US --locales-dir
27762 raghavend 20 0 291M 53056 12116 S 0.0 0.7 29:49.92 gvim -S /home/raghavendra/.vim/sessions/Session.vim
39768 raghavend 20 0 964M 50132 7672 S 0.0 0.6 0:03.79 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=renderer --js-flags=--harmony-proxies --no-sandbox --lang=en-US --lang=en-US --locales-dir
34035 raghavend 20 0 155M 46296 9744 S 0.7 0.6 1:27.73 urxvt -name Flotrxt -e tmux attach -t nether
27619 raghavend 20 0 159M 45248 4752 S 0.0 0.6 0:38.46 urxvt -e tmux attach -t uake
39771 raghavend 20 0 961M 42896 9868 S 0.0 0.5 0:14.15 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=renderer --js-flags=--harmony-proxies --no-sandbox --lang=en-US --lang=en-US --locales-dir
51294 raghavend 20 0 880M 41552 4588 S 0.0 0.5 3:28.25 /usr/bin/mplayer -fixed-vo -msglevel all=4 -slave -idle -input file=/tmp/mplayer/0.fif -input conf=/home/raghavendra/.mplayer/input.conf.mine
40326 raghavend 20 0 956M 40312 9988 S 0.7 0.5 1:20.68 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=renderer --js-flags=--harmony-proxies --no-sandbox --lang=en-US --lang=en-US --locales-dir
2361 raghavend 20 0 184M 33804 21012 S 0.0 0.4 0:00.01 /usr/lib/chromium/nacl_helper
1032 raghavend 20 0 52748 33588 2572 S 0.7 0.4 5:28.43 tmux new-session -d -s nether; neww -t nether -n mutt -d /home/raghavendra/bin/muttmuch; neww -t nether -d /home/raghavendra/bin/newsbeuter; neww -t n
27608 raghavend 20 0 552M 33332 8744 S 0.0 0.4 4:50.42 nm-applet
27448 raghavend 20 0 85944 30448 5440 S 0.0 0.4 3:11.71 /home/raghavendra/.xmonad/xmonad-x86_64-linux
27617 raghavend 20 0 51288 30380 1252 S 0.0 0.4 1:22.73 evrouter /dev/input/event15
39715 raghavend 20 0 439M 30264 3348 S 0.0 0.4 0:02.33 /usr/share/spotify/spotify-client/spotify
771 raghavend 20 0 782M 28416 9488 S 1.3 0.4 6:48.55 /usr/bin/mpd --no-daemon /home/raghavendra/.mpd/mpd.conf
5247 raghavend 20 0 86076 28204 3196 S 0.0 0.3 0:00.00 /home/raghavendra/.xmonad/xmonad-x86_64-linux
2416 raghavend 20 0 427M 27336 6964 S 0.0 0.3 0:00.00 /usr/lib/chromium/chromium --type=gpu-broker
35992 raghavend 9 -11 623M 26620 23876 S 1.3 0.3 29:21.55 /usr/bin/pulseaudio --start
2363 raghavend 20 0 485M 25844 6112 S 0.0 0.3 0:00.03 /usr/lib/chromium/chromium --type=zygote --ppapi-flash-path=/usr/lib/PepperFlash/libpepflashplayer.so --ppapi-flash-version=15.0.0.189
39716 raghavend 20 0 379M 24428 7424 S 0.0 0.3 0:00.07 /usr/share/spotify/spotify-client/Data/SpotifyHelper --type=zygote --no-sandbox --lang=en-US --locales-dir-path=/usr/share/spotify/spotify-client/Data
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit
archie 0
1:ncmpcpp 2:mutt 3:newsbeuter 4:muttZ 5:zsh 6:WeeChat 1.1-dev! 7:htop* 8:scrot- 9:zsh
[nether]

Load generation

- Sysbench
  - Generation
  - Reconnect on partition
- Sockets chosen
  - Load on affected nodes
- Distribution of Load
  - RR with socat
  - Native sysbench support
  - HAProxy?

Load generation

- Nature of data/load
  - DDL
- RQG in future
  - Fuzz testing

The Fix


Strike Out!
Los Angeles
66
LA
Los Angeles
17

Eviction

- STONITH
- Permanent eviction
- 'N' strikes & out!
  - Timers - evs parameters
  - wsrep_evs_delayed and wsrep_evs_evict_list

Eviction

- Aim
- Quorum required
  - Why? - Not shoot each other
  - Non-PC nodes also.

Eviction

- Aim
- Quorum required
  - Why? - Not shoot each other
  - Non-PC nodes also.

Eviction

- EVS version and upgrade
- TODO!
  - Ingress only
  - Follow here.
- Credits to Teemu Ollakka, Yan Zhang and Alex Yurchenko from codership.

Coredumps with Docker

- Breakdown of abstraction
- Lack of isolation
- What was done
  - Volumes
  - core_pattern & sysctl
  - suid and ulimit

WAN Segments

- How they work
- Random allocation
- Joiner starvation
- Simulates data center
- Donor selection

The code

- Github: https://github.com/percona/pxc-docker
- Jenkins: http://jenkins.percona.com/job/PXC-5.6-netem/
- Contributions/testing welcome!
- Dependencies
  - Sysbench

RICON 2014

Code: todo

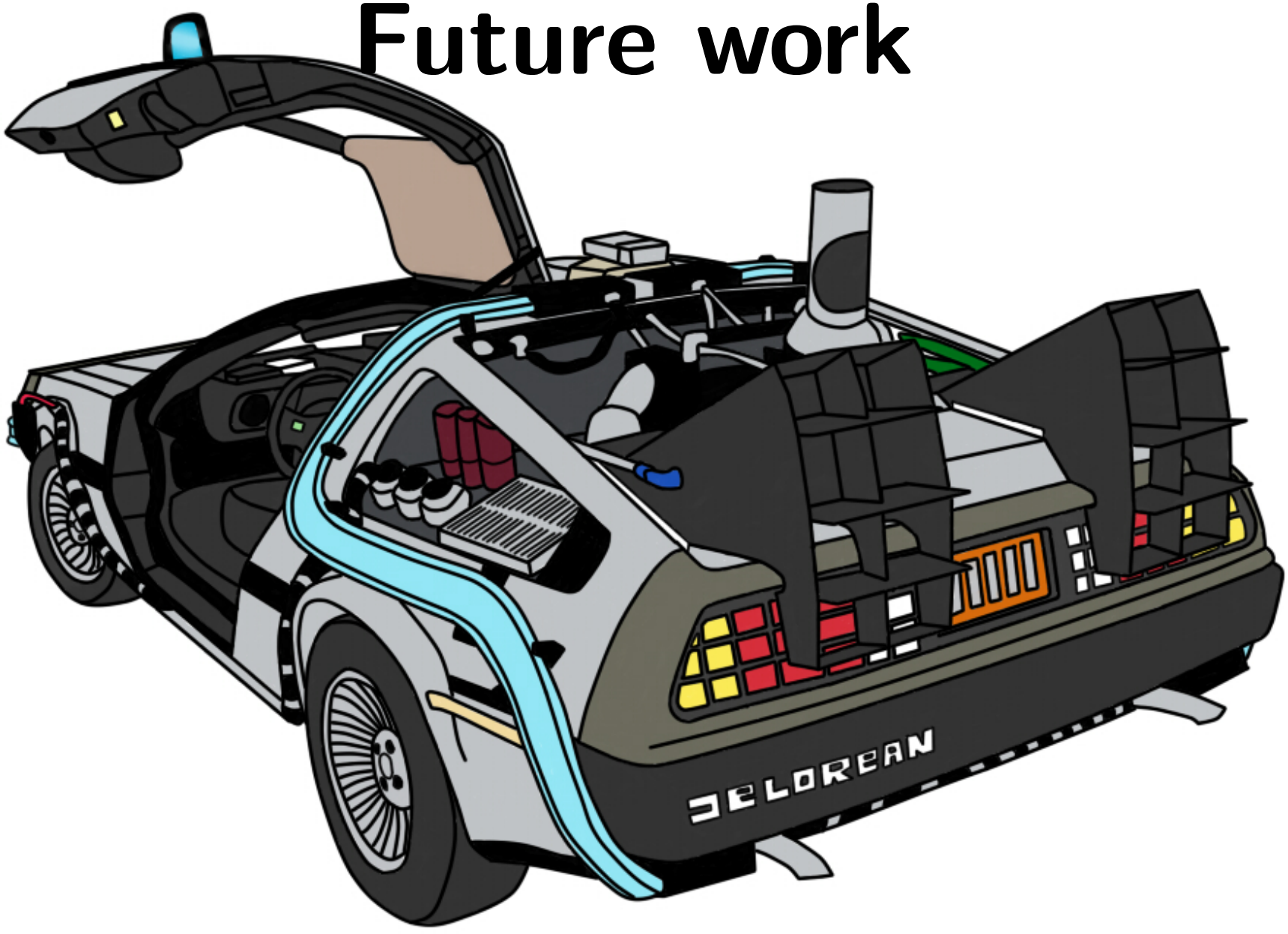
- Docker automated builds
- Orchestration
- Docker
  - Injection
  - Signal proxying

Code: todo

- Use Hoare's channels - Go!
- Run it bare - CoreOS
- Overlay with etcd/fleet/libswarm

RICON 2014

Future work



Future work

- Fault injection
  - Memory
    - Poisoned memory
  - Disk
    - libeatmydata
    - Opposite
    - ENOSPC

Fault injection

- CPU
  - NUMA?
  - Hotplug
- More network
  - corruption, duplication, reordering, rate-limit
  - Better distribution
  - Other shaping


More Chaos

Future work

- Disturb cluster more!
  - Membership changes
  * Manual eviction
  * Pull the cord!
  - Corrupt nodes
- Consistency voting

Further Reading

- Byzantine fault tolerance
  - Reaching agreement in presence of faults
- The Network is Reliable
- NetEm
- Latency: The New Web Performance Bottleneck
- Galera Cluster Documentation
- Auto eviction code
- Don't Settle for Eventual Consistency
- Extended Virtual Synchrony

About

- /me: Raghavendra Prabhu, Product Lead, Percona XtraDB Cluster, Percona.
- Slides will be at slideshare.net/slidunder and owncloud
- About.me: raghavendra.prabhu
- Keybase.io: rdprabhu
- Presentation under CC BY-SA 4.0

RICON2014

Image Credits

- http://galeracluster.com/documentation-webpages/
- http://www.thelastdragontribute.com/40th-anniversary-death-of-bruce-lee/
- https://upload.wikimedia.org/wikipedia/commons/6/60/Corpus_callosum.png
- http://www.thebarrow.org/Neurological_Services/Epilepsy/204354
- https://flic.kr/p/9J6GNu
- https://secure.flickr.com/photos/brewbooks/7780990192
- https://www.flickr.com/photos/kwerfeldein/2649294869
- https://secure.flickr.com/photos/mindmob/51951632
- https://secure.flickr.com/photos/arenamontanus/2227769907
- https://www.flickr.com/photos/markop/477199204
- https://www.flickr.com/photos/gcwest/281385801
- https://www.flickr.com/photos/29233640@N07/13466208953
- https://www.flickr.com/photos/bob_in_thailand/9782777742/