From Dev to Prod with Sharded Mongo DB Clusters

May 2013

Johan Andersson

Severalnines AB

johan@severalnines.com

severalnines

Topics

- Best Practice Deployment
- Keeping it running
- Scaling with Shards

Note: Most features mentioned in this presentation are applicable only on MongoDB 2.2 or later.

severalnines

"An OS check-up a day keeps downtime away"

Ancient Chinese wisdom

Distributed Computing 101

severalnines

- Network latency is not 0
- Network is unreliable
- Network bandwidth is not infinite
- CPUs are not infinitely fast
- Disks are not infinitely fast
- Free RAM decreases fast
- Virtualization does not make the underlying hardware more powerful

MongoDB Concepts

severalnines

- Mongos
  - Query routers, routes queries to shards
  - Caches data from Config Servers
- Config Servers
  - Records which shard holds what chunks
- Shard Servers
  - Stores the actual data

Developer Setup

severalnines

- A good test setup for a developer:
  - One or three config Servers
  - Thee shard servers
  - One or more Mongos
- But do set it up as a shard server so you can validate queries (explain) to make sure they behave as intended.
  - Using correct indexes
- Vagrant image - http://alexyu.se/content/2013/03/vagrant-box-severalnines-database-cluster-deployment-scripts

Test Setup

severalnines

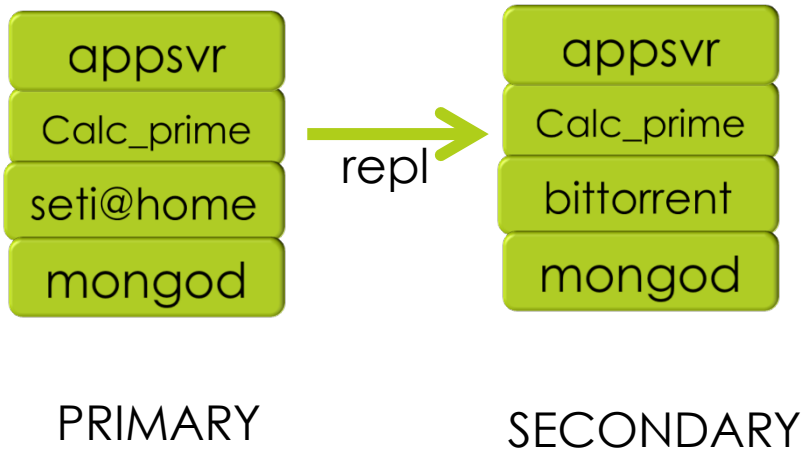
- Recommendations
  - One or three config Servers
  - >1 shard - thee shard servers
  - One or more Mongos
  - Stay close to your production system.
- Two shards (even if prod use one only) allows you to:
  - Verify queries use the correct indexes
  - Check if one or all shards are hit
    - Extremely useful when adding shards to Prod to avoid performance regression.

Production Setup

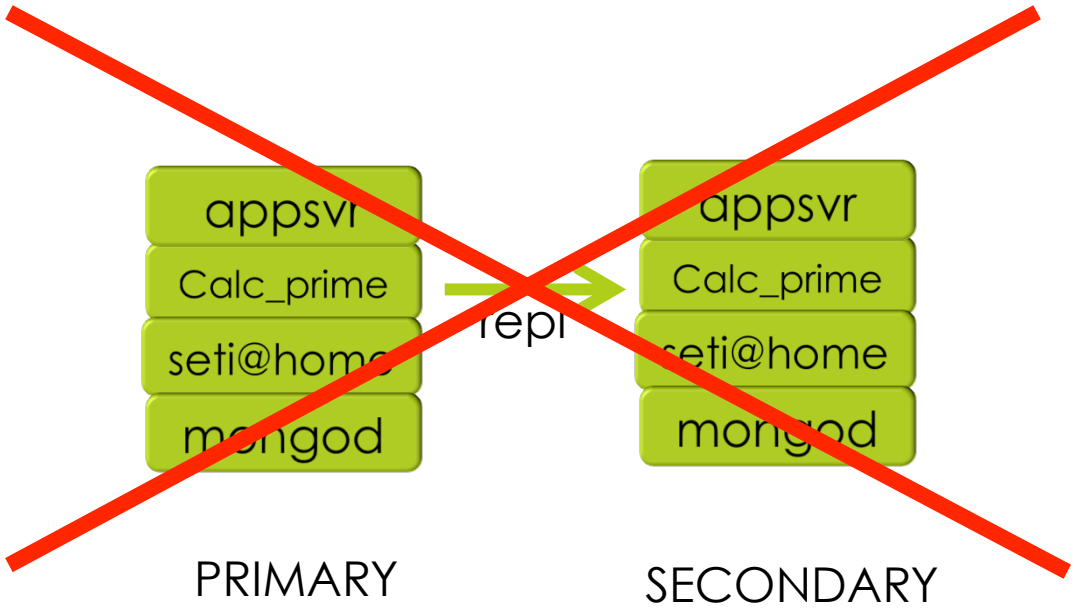
severalnines

- Start with a good and sound architecture from day one!
- A sharded architecture will allow you to scale easily
- Avoid un-necessary craziness and risk to switch a non-sharded setup to a sharded
- You do want three copies of your data

Production Setup

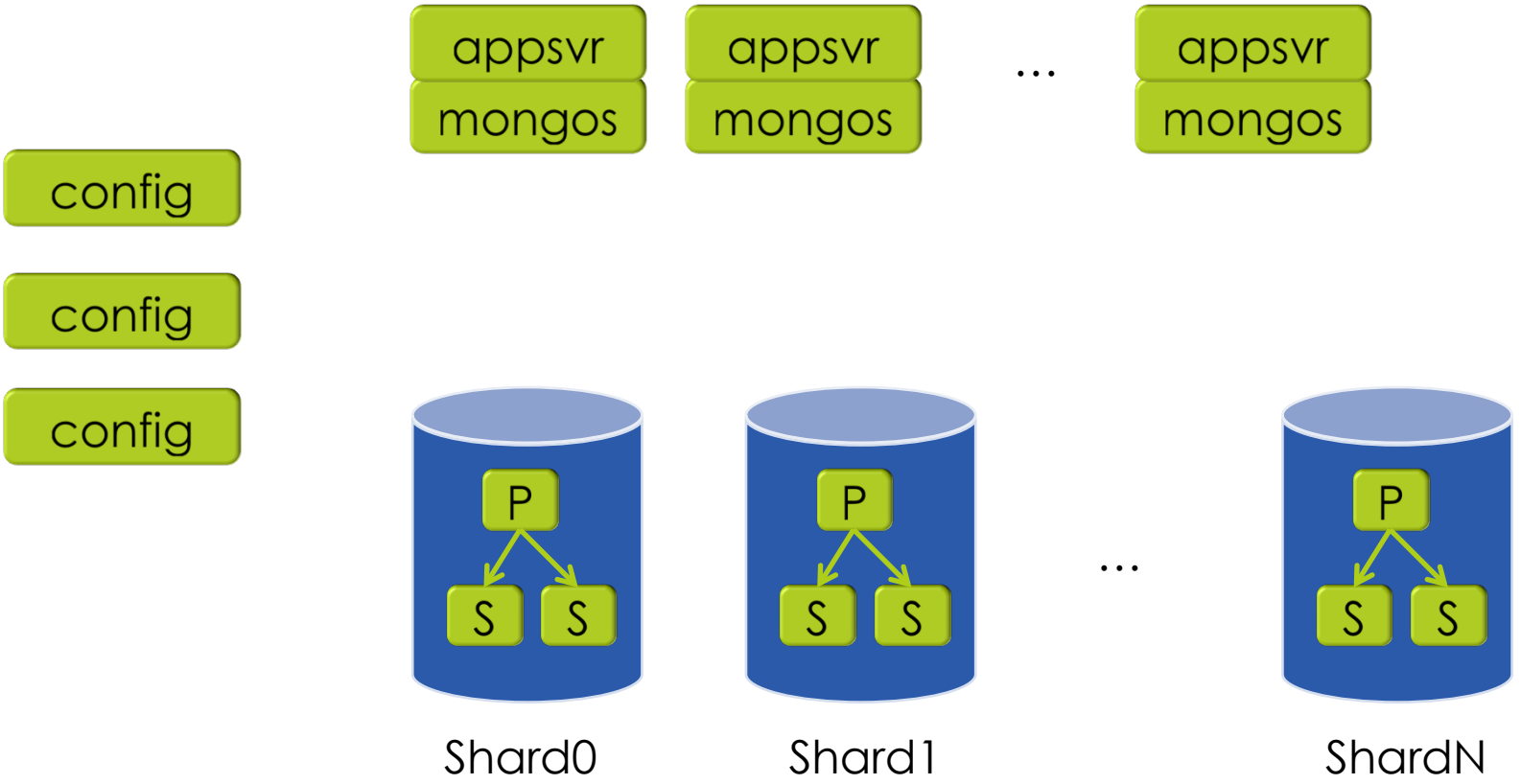


Production Setup



Hard to scale, too crowded
Go for Best Practice instead

Production Setup



Production Setup

severalnines

- Allows you to scale - Add more shards and mongos easily
- It is proven – used by many large organizations
- Deploy…
  - Mongos
  - Config servers
  - Shard servers (mongod --shardsvr)
- … on individual servers

When processes fails

severalnines

- Config Server process fails
  - If one config server fails, no chunk migration between shards
  - No chunk splitting
  - Shards are still read/writeable
- Mongos fails
  - You will have many
- Shard server fails
  - If PRIMARY fails → One SECONDARY will become PRIMARY
  - If SECONDADY fails → Restart and it will resync with PRIMARY
- Restart the failed process!
  - Troll the logs to see it starts without errors

Keeping it running

- RAM is a scarce resource – keep as much as possible in it!
- Use short field names in the Documents to reduce size:
  - { first_name:'not good'}
  - {fn:'better'}
  - Important when you have many Documents
- Compact collections:
  - Run on SECONDARY
    - db.mycollection.runCommand("compact")
  - Stop PRIMARY, SECONDARY becomes new PRIMARY.

Keeping it running

severalnines

- Replication Lag
  - Producer / Consumer problem
  - Temporary or Permanent Problem?
  - Permanent problem
    - Expunge the oplog?
    - Magic wand?
  - Disk, network, or load-dependent problem?
- Remedy:
  - Faster disks: Improve disk subsystem -> IOPS or SSD, RAID configuration
  - Faster CPUs
  - Can you reduce writes to the shard(s)?
  - Scale with shards

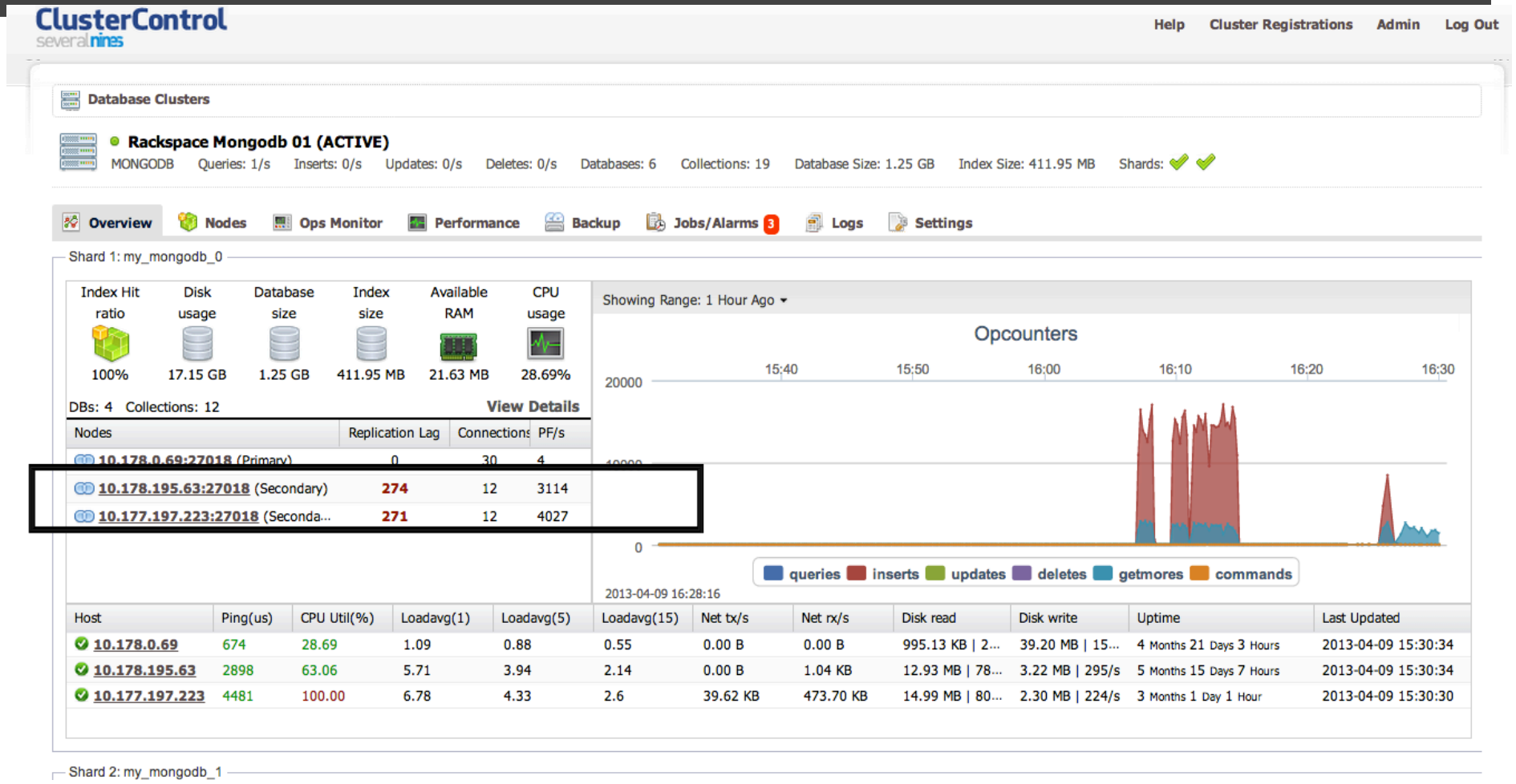
Keeping it running

severalnines

- Page Faults
  - MongoDB uses mmap:ed files, allocated from the VMM address space. Mmapped files does not need to fit in RAM
    - OS manages what pages are mapped in RAM and not
  - Accessing data on a page not in RAM cause a page fault
  - Data will be loaded from DISK
  - DISK is slow
  - Performance will degrade
  - If the number of PFs is constant, no worry, else if increasing, problem in the horizon.
- Remedy:
  - Increase RAM in server (Active working set should fit in RAM)
  - Scale with shards
  - Add SSD

Not cool

severalnines



Not cool

severalnines

```
root@server06:~# vmstat 1
procs -----------memory---------- ---swap-- -----io---- -system-- ----cpu----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa
 0  1 956880   9860   1076 148984    0    0     4    20    2    3  0  0 99  0
 0  1 956880  10108   1068 148572    0    0 24496    16 1406 1175  0  4  0 94
 0  1 956844  44064   1100 152156  112    0  6632   476  722 15556 30 28  0 41
 0  1 956844  27572   1100 167252    0    0 15088     0 1212  762  2  3  0 94
 0  1 956844  13684   1100 179900    0    0 12664     0 1038  666  3  2  0 94
 0  1 956844  10152   1096 182352    0    0 12324  4096 1449  661  6  7  0 86
 0  1 956844  10072   1096 181412    0    0 12536  4096 1459  642  4  4  0 90
 0  2 956796   8884   1100 180688  112    0 11536     0  899  601  4  3  0 91
 0  1 956796  10088   1100 178540    0    0 12212  1096 1196  657  3  3  0 93
 0  1 956796   9828   1100 177872    0    0 13608  1328 1343  694  2  3  0 93
 0  2 956796  10152   1076 176560    0    0  7676 12780 2180  644  4  6  0 88
 0  2 956796   9780   1076 176820    0    0  3020 14184 1955  541  1  6  0 92
 1  0 956796  10156   1076 175336    0    0  6240  1192  856  497  1  2  0 96
 0  1 956796   9524   1076 176168    0    0  4076     0  460  451  3  1  0 96
 0  1 956796   9532   1076 175340    0    0  9480     0  809  562  5  2  0 92
```

You are done for!

severalnines

```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
          6.89    0.00    6.36   86.11    0.64    0.00

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
xvdc              0.00     0.00    2.55    0.00    10.18     0.00     8.00     0.18   69.33   69.33    0.00   8.67   2.21
xvda              0.11   601.27  813.79  221.85 19820.57  3418.45    44.88    14.37   13.63    4.34   47.69   0.97 100.78
```

Avoid the situation

- Keep data set in RAM
- Use fast io-subsystem since OS is managing the paging
  - SSD
  - IOPS on EC2
- Make sure the Host OS (Virtualized env) does not swap.
- Lock mongod's to CPUs, check /etc/interrupts, bind to CPUs not handling ETH interrupts, avoids ctx switching.
- Add shards before its too late
  - More disks/CPUs,RAM to handle the load

Scaling with Shards

- Shard Keys
- When to scale?
- Add a shard

Database Size

- db.stats()
  - Size of one db = dbStats.dataSize + dbStats.indexSize
  - Does it fit in RAM?
- Plan ahead:
  - avgObjSize x expectedNoOfDocuments
  - Does it fit in RAM?
  - How many documents can you store?

Working set

- Working Set Analyzer in 2.4 helps:

```
"workingSet" : {
        "note" : "thisIsAnEstimate",
        "pagesInMemory" : 33688, //4KB pg_size
        "computationTimeMicros" : 15283,
        "overSeconds" : 1923 //dist pg_new to pg_oldest
    },
```

- Trend `overSeconds` and `pagesInMemory`

Shard Key

severalnines

- Collections must have a Shard Key
- The shard key must be indexed
- Hash-based or Range based Sharding
  - Hash-bashed great for exact match
  - Subscriber/user databases
- Queries must use the shard key
  - If not, all shards will be queried → slow
  - Check and verify with .explain()

Adding Shards

severalnines

- Adding a new Shard is easy:
  - Deploy a new Replica-set:
    - PRIMARY - 192.168.0.151
SECONDARY - 192.168.0.152
SECONDARY - 192.168.0.153
  - Register the Replica Set:
    - sh.addShard("myshard2/192.168.0.151:27017");
- Finding the time to add shards is harder
  - Rebalancing costs and query performance will suffer due to chunk migration.
  - Do it off-peak if possible
- MongoDB will automatically rebalance the shards.

SPHINX INFORMATION TECHNOLOGIES

severalnines

Q&A

Resources

severalnines

- MongoDB Configurator
  http://www.severalnines.com/mongodb-configurator/
- -ClusterControl for MongoDB
  http://www.severalnines.com/resources/cmon-mongodb-cluster-management-monitoring-tool
- Installing ClusterControl on top of existing MongoDB Cluster
  http://www.severalnines.com/blog/install-clustercontrol-top-existing-mongodb-sharded-cluster
- Vagrant image -
  http://alexyu.se/content/2013/03/vagrant-box-severalnines-database-cluster-deployment-scripts

Follow us on

- Twitter : @severalnines
- Facebook: www.facebook.com/severalnines
- Slideshare : www.slideshare.net/severalnines
- Linked-in: www.linkedin.com/company/severalnines

SPHINX INFORMATION TECHNOLOGIES

severalnines

Thank you!