binlogic DataOps

" A journey between **infrastructure** and **data** "

JUNE 20-21 2019 WTCBCN

Stephane VAROQUI

Signal18

Wordpress Around the Clock

Agenda

- **Old School LAMP Stack**

- **LAMP Stack 2019.0**
  - Zone / Geo Dns
  - Docker Orchestration
    - Cluster
    - Services
    - Share Storage

- **Database Infrastructure**
  - Routing
    HA / Proxying / Critical writes
  - Geo queries

Mixr an old School LAMP Stack

- VMWare
- Single data center in Paris
- Nginx http servers - 2 MariaDB databases
- NFS store for images
- Database replication: master slave but slave as a backup
- HaProxy LB to fronts

Mixr Workload

- Wordpress is mostly serving dynamic content
  - Geo content
  - Customized via wordpress plugins

- Wordpress Transient:
  - Global store for all plugins
  - Transient storage default to database

- Wordpress database is generic
  - Highly normalized objects/attributes for everything
  - Plugins generic metadata key/value (wp_options)

What was wrong

- Wordpress is heavy database consumer
  - 2K avg queries per page
  - Mb transfert from database to load plugins
  - Plugins grrrr JSON is not a database

- Bad caching policy
  - No opcode cache
  - No CDN for statics

- NFS: Perf bottleneck & SPOF

- Response time > 10s from US

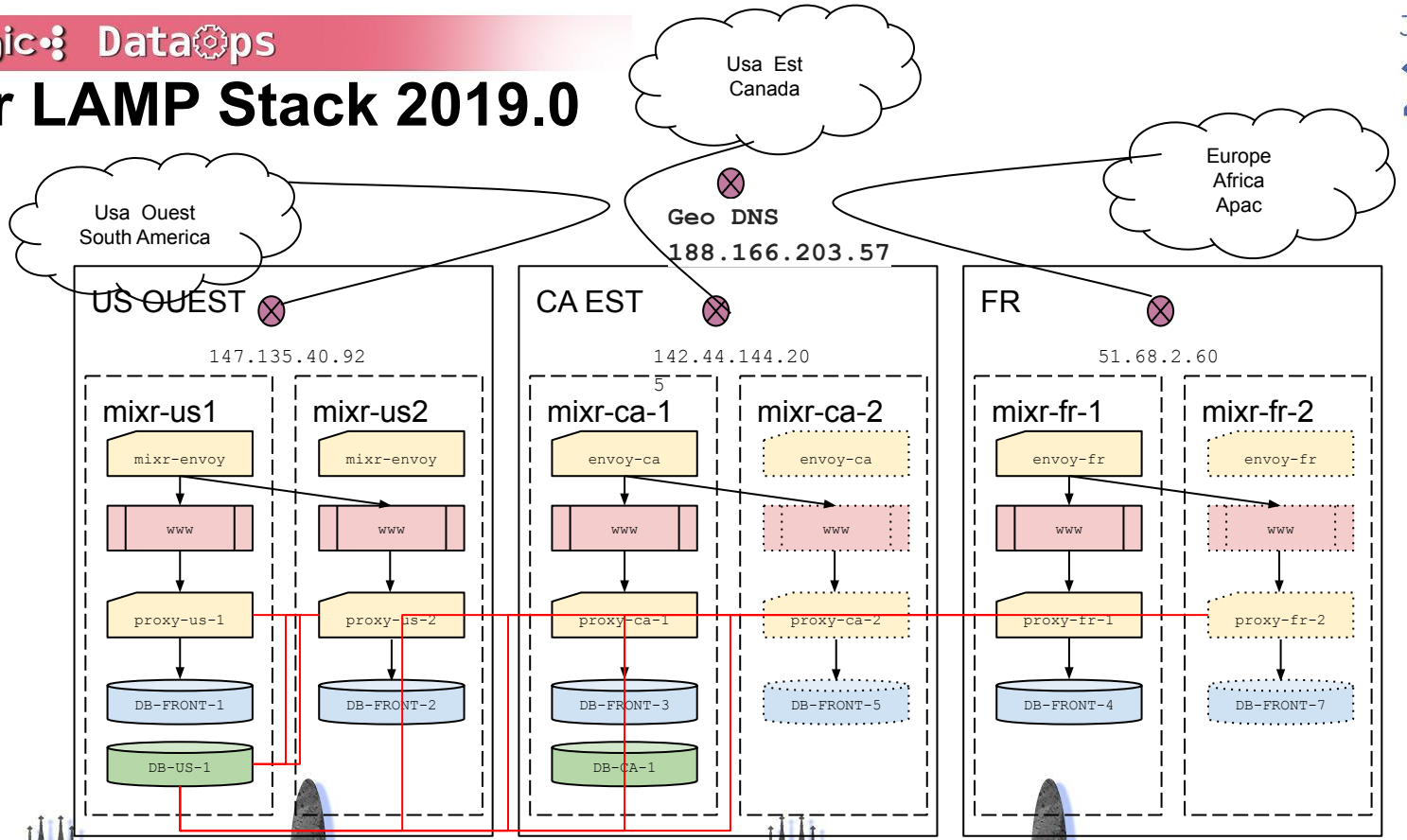
Too many external API calls

- Mostly synchronous calls

- API from bug players can fail
  - Google: 2s for analytics
  - Amazon toolings API like Geoloc can take 2-10s response time
  - Facebook GAuth JS can take 2s to load

- Doubling API, doubling 404 risk exposure
  - Internal API for completion
  - Internal API for Geo IP
  - Internal API for keyword suggestion

Mixr LAMP Stack 2019.0



OpenSVC OVH Cluster

om mon

```
Threads                                                  mixr-ca-1    mixr-ca-2    mixr-fr-1    mixr-fr-2    mixr-us-1    mixr-us-2    s18-fr-1     s18-fr-2
 daemon              running                           | O
 collector           running                           | O
 dns                 running
 hb#1.rx             running     0.0.0.0:10000         | O            O            O            O            O            /            O            O
 hb#1.tx             running                           | O            O            O            O            O            /            O            O
 hb#2.rx             running relay.opensvc.com         | O            O            O            O            O            /            O            O
 hb#2.tx             running                           | O            O            O            O            O            /            O            O
 listener            running     0.0.0.0:1214
 monitor             running
 scheduler           running

Nodes                                                    mixr-ca-1    mixr-ca-2    mixr-fr-1    mixr-fr-2    mixr-us-1    mixr-us-2    s18-fr-1     s18-fr-2
 score                                                 | 86           99           63           71           85           89           86           87
  load 15m                                             | 0.3          0.0          0.8          0.2          0.4          0.2          1.0          0.7
  mem                                                  | 94/98%:31.1g 6/98%:15.6g  79/98%:15.6g 3/98%:62.8g  95/98%:31.1g 72/98%:31.1g 80/98%:31.3g 78/98%:31.3g
  swap                                                 | 2/90%:10.00g 0/90%:510m   89/90%:3.50g -            4/90%:10.00g 1/90%:10.00g 8/100%:1.50g 4/99%:1.50g
 version             warn                              | 1.9-3335     1.9-3349     1.9-3335     1.9-3335     1.9-3335     1.9-3335     1.9-3335     1.9-3335
```

OpenSVC Orchestration Ressources

om net status

```
name          type            network         size      used   free      pct
|- default    bridge          10.22.0.0/16    65536     0      65536     0.00%
|- lo         loopback        127.0.0.1/32    1         0      1         0.00%
|- mixrnet    routed_bridge   10.48.0.0/16    65536     45     65491     0.07%
`- repman     weave           10.32.0.0/12    1048576   29     1048547   0.00%
```

om pool status

```
name          type        caps                        head                              vols   size    used    free
|- tank       zpool       rox,rwx,roo,rwo,snap,blk    data                              0      676g    75.0g   601g
|- default    directory   rox,rwx,roo,rwo,blk         /var/lib/opensvc/pool/directory   0      76.8g   23.9g   49.0g
`- shm        shm         rox,rwx,roo,rwo,blk         /dev/shm                          1      15.5g   2.16m   15.5g
```

- IPIP routed bridge best on latency / response time / robustness
- Weavenet a real SPOF in case of corruption
- OVH VRack routing is pure marketing US/EU not possible
- OVH VRack in same DC can bring no IP failover that is unpredictable in time
- OpenSVC have so many more pool options DRBD, SAN multipath, NFS ……. ……

OpenSVC Secrets

om */sec/* mon

/sec/			mixr-ca-1	mixr-ca-2	mixr-fr-1	mixr-fr-2	mixr-us-1	mixr-us-2	s18-fr-1	s18-fr-2
mixr-dev1/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-dev2/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-dev3/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-dev4/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-dev5/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-pre/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr-preprod/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
mixr/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
monitor/sec/cert-s18	n/a	-	/	/	/	/	/	/	/	/
s18/sec/cert-mixr	n/a	-	/	/	/	/	/	/	/	/
s18/sec/cert-s18	n/a	-	/	/	/	/	/	/	/	/

- Instant key/value change push to all consumers
- Exposed to containers via volumes
- Crypto secret auto delivered from the cluster

```
[volume#1]
type = shm
name = cert-s18
size = 1m
secrets = cert-s18/*:/
```

LAMP Stack 2019.0 - Services on Docker

om mixr/svc/* mon

```
mixr/svc/*                                    mixr-ca-1 mixr-ca-2 mixr-fr-1 mixr-fr-2 mixr-us-1 mixr-us-2
 mixr/svc/blackfire      up  ha    1/1    |                                            O^         X*
 mixr/svc/ci-front-1     up  ha    1/1    |                                       O^
 mixr/svc/ci-front-2     up  ha    1/1    |                                                  O^
 mixr/svc/cron           up  start 1/1    |                                            X         O^
 mixr/svc/db-ca-1        up  ha    1/1    | O^
 mixr/svc/db-front-1     up  ha    1/1    |                                       O^
 mixr/svc/db-front-2     up  ha    1/1    |                                                  O^
 mixr/svc/db-front-3     up  ha    1/1    | O^
 mixr/svc/db-front-4     up  ha    1/1    |                     O^
 mixr/svc/db-us-1        up  ha    1/1    |                                       O^
 mixr/svc/envoy-ca       up  ha    1/1    | O^
 mixr/svc/envoy-fr       up  -     1/1    |                               O^
 mixr/svc/envoy-us       up^ ha    1/1    |                                            X^         O
 mixr/svc/gitlab         up  start 1/1    |  O^                                                    X
 mixr/svc/img-us         up  start 1/1    |                                            O^         X
 mixr/svc/img-ca         up  start 1/1    |           O
 mixr/svc/img-fr         up  start 1/1    |                     O^
 mixr/svc/proxy-ca-1     up  ha    1/1    | O^
 mixr/svc/proxy-fr-1     up  ha    1/1    |                     O^
 mixr/svc/proxy-front-1  up  ha    1/1    |                                       O^
 mixr/svc/proxy-front-2  up  ha    1/1    |                                                  O^
 mixr/svc/repman         up  start 1/1    |                               O^
 mixr/svc/www-us         up  ha    2/2-2  |                                            O*         O^
 mixr/svc/www-ca         up  ha    1/1-1  | O*
 mixr/svc/www-fr         up  ha    1/1-1  |                     O^
```

- HA Failover
- HA Flex placement
- Logical Replication

LAMP Stack 2019.0 - Welcome Docker - OpenSVC

om mixr-*/svc/* mon

```
mixr-*/svc/*                                mixr-ca-1 mixr-ca-2 mixr-fr-1 mixr-fr-2 mixr-us-1 mixr-us-2 s18-fr-1 s18-fr-2
 mixr-dev/svc/backup            up  start 1/1   |                 O^
 mixr-dev/svc/db-fr-1           up  start 1/1-1 |                 O*
 mixr-dev/svc/db1               up  start 1/1-1 |                                                   O^
 mixr-dev/svc/db2               up  start 1/1-1 |                                                             O^
 mixr-dev/svc/proxysql1         up  start 1/1-1 |                                                   O^
 mixr-dev/svc/s3                up  start 1/1   |                 O^
 mixr-dev1/svc/www              up  start 1/1   |                 O^
 mixr-dev2/svc/www              up  start 1/1   |                 O^
 mixr-dev3/svc/www              up  start 1/1   |                 O^
 mixr-dev4/svc/www              up  start 1/1   |                 O^
 mixr-dev5/svc/www              up  start 1/1   |                 O^
 mixr-preprod/svc/db-ca-1       up  ha    1/1   | O^
 mixr-preprod/svc/db-front-1    up  ha    1/1   |                                   O^
 mixr-preprod/svc/db-front-2    up  ha    1/1   |                                             O^
 mixr-preprod/svc/proxy-ca-1    up  ha    1/1   | O^
 mixr-preprod/svc/proxy-front-1 up  ha    1/1   |                                   O^
 mixr-preprod/svc/proxy-front-2 up  ha    1/1   |                                             O^
 mixr-preprod/svc/www           up^ ha    2/1-2 |                                   O^        O
```

om mixr/svc/www-us print status

```
mixr/svc/www-us                  up
`- instances
   |- mixr-us-1                  up         frozen, idle, started
   `- mixr-us-2                  up         idle, started
      |- ip#0           ........ up         cni mixrnet 10.48.48.10/20 eth12 expose#1 expose#2
expose#3 expose#4 expose#5
      |- fs#1           ........ up         zfs data/mixr-www@/srv/mixr-www
      |- container#0    ........ up         docker google/pause
      |- container#1    ........ up         docker signal18/nfsmixr:latest
      |- container#2    ........ up         docker memcached:1.5.10-alpine
      |- container#3    ........ up         docker signal18/php-fpm-72:latest
      |- container#4    ........ up         docker adminer:standalone
      |- container#5    ........ up         docker nginx/nginx-prometheus-exporter:0.2.0
      |- container#6    ........ up         docker bakins/php-fpm-exporter:v0.5.0
      |- sync#i0        ...O./.. up         rsync svc config to nodes
      |- task#01        ...O.... n/a        task
      |- certificate#0  ...../.. n/a        tls certificate
      |- expose#1       ...../.. n/a        envoy expose 80/tcp via 0.0.0.0:443
      |- expose#2       ...../.. n/a        envoy expose 443/tcp via 0.0.0.0:80
      |- expose#3       ...../.. n/a        envoy expose 8081/tcp via 0.0.0.0:443
      |- expose#4       ...../.. n/a        envoy expose 8082/tcp via 0.0.0.0:443
      |- expose#5       ...../.. n/a        envoy expose 80/tcp via 0.0.0.0:80
      |- hash_policy#1  ...../.. n/a        envoy hash policy
      |- route#0        ...../.. n/a        envoy route
      |- route#1        ...../.. n/a        envoy route
      |- route#5        ...../.. n/a        envoy route
      |- vhost#0        ...../.. n/a        envoy vhost us.mixr.net, www.mixr.net, mobile.mixr.net,
beta.mixr.net
      |- vhost#1        ...../.. n/a        envoy vhost us.mixr.net, www.mixr.net, mobile.mixr.net,
beta.mixr.net
      |- vhost#3        ...../.. n/a        envoy vhost cdn-us.mixr.net
      |- vhost#4        ...../.. n/a        envoy vhost landing.mixr.net
      `- vhost#5        ..
```

om mixr-dev/svc/db-fr-1 print config

```
[DEFAULT]
nodes = {env.nodes}
flex_primary = {env.nodes[0]}
topology = flex
rollback = false
app = mixr
docker_daemon_private = false
docker_data_dir = {env.base_dir}/docker
docker_daemon_args =  --storage-driver=zfs
orchestrate = start

[ip#01]
pod01
type = cni
container_rid = container#0001
network = mixrnet

[task#00]
schedule = @1
command =
{env.base_dir}/pod01/init/trigger-dbjobs
user = root
run_requires = fs#01(up,stdby up)
container#0001(up,stdby up)

[env]
nodes = mixr-fr-1
size = 100g
db_img = mariadb:10.3
ip_pod01 = db-fr-1.mixr-dev.svc.cloud18
port_pod01 = 3306
mysql_root_password = xxxx
mysql_root_user = root
base_dir = /srv/{namespace}-{svcname}
```

```
[fs#00]
type = zfs
dev = data/{namespace}-{svcname}_docker
mnt = {env.base_dir}/docker
size = 2g

[fs#01]
type = zfs
dev =data/{namespace}-{svcname}_pod01
size = {env.size}
mkfs_opt = -o recordsize=16K -o
primarycache=metadata -o atime=off -o
compression=gzip -o mountpoint=legacy
mnt = {env.base_dir}/pod01
standby = true

[fs#02]
type = zfs
dev =
data/{namespace}-{svcname}_pod01-system
size = 20G
mkfs_opt = -o recordsize=4K -o
primarycache=metadata -o atime=off -o
mountpoint=legacy
mnt = {env.base_dir}/pod01/data/.system

[fs#03]
type = tmpfs
mnt = {env.base_dir}/tmp
dev = none
```

```
[container#0001]
type = docker
hostname = {svcname}.{namespace}.svc.{clustername}
image = google/pause
rm = true
```

```
[container#0002]
detach = false
type = docker
image = busybox
netns = container#0001
rm = true
volume_mounts = /etc/localtime:/etc/localtime:ro
{env.base_dir}/pod01:/data
command = sh -c 'wget -qO-
http://{env.mrm_api_addr}/api/clusters/{env.mrm_clust
er_name}/servers/{env.ip_pod01}/{env.port_pod01}/conf
ig|tar xzvf - -C /data'
```

INIT CONTAINER

```
[container#2001]
tags = pod01
type = docker
run_image = {env.db_img}
netns = container#0001
rm = true
run_args = -e
MYSQL_ROOT_PASSWORD={env.mysql_root_password}
 -e MYSQL_INITDB_SKIP_TZINFO=yes
volume_mounts = /etc/localtime:/etc/localtime:ro
 {env.base_dir}/pod01/data:/var/lib/mysql:rw
 env.base_dir}/pod01/etc/mysql:/etc/mysql:rw
```

Envoy Routing

om mixr/svc/www-us print config

```
[DEFAULT]
docker_data_dir = /srv/{namespace}-{svcname}/do
docker daemon args = --storage-driver=
nodes = mixr-us-1 mixr-us-2
rollback = false
docker daemon private = true
topology = flex
orchestrate = ha
app = mixr
flex_min_nodes = 2

[certificate#0]
certificate_secret = cert-mixr
type = tls
```

Load balancing to each DC front

HTTPS redirect

```
[expose#1]
type = envoy
listener_addr = 0.0.0.0
listener_port = 443

[expose#2]
type = envoy
listener_addr = 0.0.0.0
listener_port = 80
```

cdn-us.mixr.net to serve static to Cloudflare

```
[expose#3]
type = envoy
listener_port = 443
port = 8081
listener_certificates = certificate#0
vhosts = vhost#3
gateway = {namespace}-envoy-us
lb_policy = ring_hash
```

```
[vhost#0]
domains = us.mixr.net www.mixr.net
mobile.mixr.net beta.mixr.net
routes = route#0

[vhost#1]
domains = us.mixr.net www.mixr.net
mobile.mixr.net beta.mixr.net
routes = route#1

[vhost#3]
domains = cdn-us.mixr.net
routes = route#0

[vhost#4]
domains = landing.mixr.net
routes = route#0

[route#0]
match_prefix = /
hash_policies = hash_policy#1
route_timeout = 300s

[route#1]
                    true
```

Session host affinity

```
[hash_policy#1]
cookie_name = envoy_hash
cookie_ttl = 0
```

Before: Mixr's around me


```
SELECT DISTINCT
               w1.post_id,

FROM           wp_posts p
               INNER JOIN wp_postmeta w1 ON p.ID = w1.post_id
               INNER JOIN wp_postmeta w2 ON w1.post_id = w2.post_id
               INNER JOIN wp_postmeta w4 ON w1.post_id = w4.post_id
               INNER JOIN wp_term_relationships tr ON w1.post_id = tr.object_id
               INNER JOIN wp_term_taxonomy tt ON tr.term_taxonomy_id = tt.term_taxonomy_id
               INNER JOIN wp_terms t ON t.term_id = tt.term_id
               LEFT JOIN wp_em_events e ON e.post_id = p.ID
               LEFT JOIN wp_secret_group_invitations sgi ON p.ID = sgi.post_id
               LEFT JOIN wp_em_attending ea ON p.ID = ea.id_post
               $custom_join
 WHERE          (w1.meta_key = 'groupe_latitude' OR w1.meta_key = 'event_latitude')
               AND (w2.meta_key = 'groupe_longitude' OR w2.meta_key = 'event_longitude')
               AND ((w1.meta_key = 'groupe_latitude' AND w4.meta_key != '_end_ts')
               OR (w4.meta_key = '_end_ts' AND '$current_date' <= w4.meta_value))
               $custom_where_post_status
               $custom_where_post_type
               $custom_where
               $custom_where_terms_id
ORDER BY       $base_orderby
               $custom_orderby
LIMIT          $sql_limit
```

After: Mixr's around me

- Geo position is the biggest filtering factor
- Search by increasing distance until enough candidates have been found

```
SELECT            DISTINCT p.guid,
FROM        geo_filter p $custom_join ci
WHERE       get_geolocalized_post($lat,$lon,$distance,$nb_rows)
                  $custom_where_post_futur
                  $custom_where_post_status
                  $custom_where_post_type
                  $custom_where_restriction_age
                  $custom_where
                  $custom_where_terms_id
                  $custom_filter_dates_range
GROUP BY    p.guid
ORDER BY    $custom_orderby
                  $base_orderby
                  $sql_limit
```

After: Mixr's around me

- To enable index merge, one need all criteria in a single table.
- Data have been "denormalized" in named geo_filter table.

```
SPATIAL KEY coord (coord),
FULLTEXT KEY ci (ci),
FULLTEXT KEY invitations (invitations),
FULLTEXT KEY attending (attending),
FULLTEXT KEY cit (cit)
```

coord
POINT(49.6196726 6.0721393)

```
MySQL [wpmixrgen]> select ci from geo_filter where id = 44205;
+-----------------------------------------------------------------------------------------------------------------+
| ci |
+-----------------------------------------------------------------------------------------------------------------+
| 0000002291 0000002295 0000002297 0000004623 0000005915 0000007119 0000007739 0000011916 0000012437 |
+-----------------------------------------------------------------------------------------------------------------+
```

API Call Replacement

- **get_current_geoloc_by_ip**
  http://api.eurekapi.com/
  100-400ms

- **get_geolocation_data_by_lat_lon**
  Elastic search on geonames
  200ms -300ms

- **get_current_geoloc_by_ip**
  GeoLite2-City.mmdd
  Native php module
  < 1ms

- **get_geolocation_data_by_lat_lon**
  geonames direct GIS SQL query
  = 1ms

MariaDB HA : replication-manager

SWITCHOVER

Monitor Status	
Manager Status	ACTIVE
Cluster Name	mixr
Failover Count	0 / 5
Last Failover Time	0

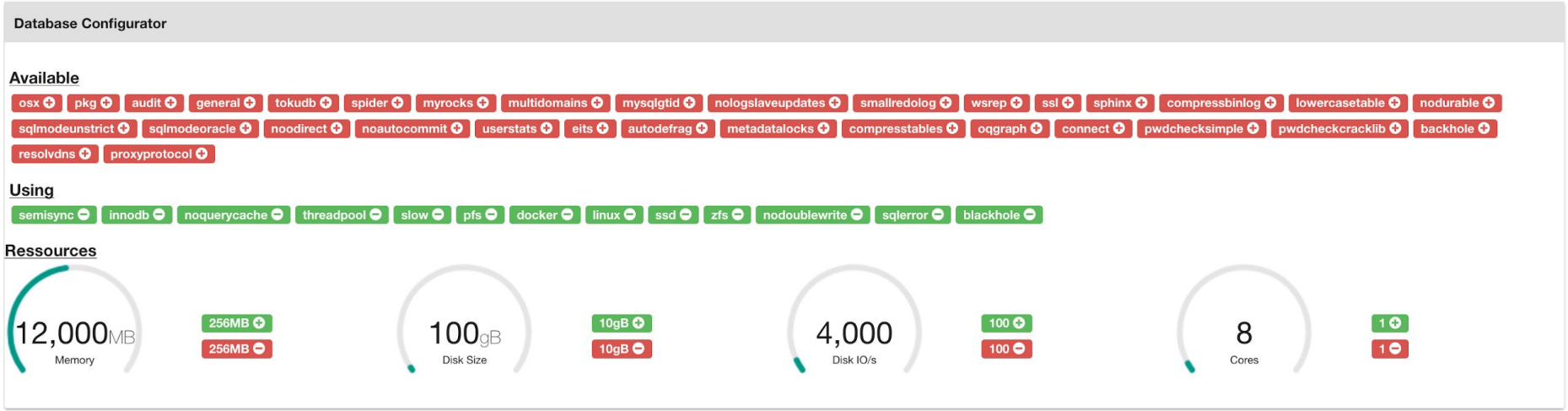
Cluster Availability	
Monitor Tickers	3788
Master Available	91.58606
Acceptable Data Loss	91.58601
No Data Loss	91.58606

Cluster Info	
Topology	master-slave
QPS	5
Table Size	1111369308
Index Size	214887424

Database Servers	Status	Using GTID	Current GTID	Slave GTID	Delay	Fail Cnt	Prf Ign	IO Thr	SQL Thr	Ro Sts	Evt Sch	Mst Syn	Rep Syn
db-front-1.mixr.svc.cloud18:3306	Slave	Slave_Pos	0-9-1211348571	0-9-1211348571	0	0/0		✓	✓	✓	✗	✗	✗
db-front-2.mixr.svc.cloud18:3306	Slave	Slave_Pos	0-9-1211348571	0-9-1211348571	0	0/0		✓	✓	✓	✗	✗	✗
db-front-3.mixr.svc.cloud18:3306	Slave	Slave_Pos	0-9-1211348571	0-9-1211348571	0	0/0		✓	✓	✓	✗	✗	✓
db-front-4.mixr.svc.cloud18:3306	Slave	Slave_Pos	0-9-1211348570	0-9-1211348570	0	0/0		✓	✓	✓	✗	✗	✓
db-ca-1.mixr.svc.cloud18:3306	Slave	Slave_Pos	0-9-1211348571	0-9-1211348571	0	0/0	👍	✓	✓	✓	✗	✗	✓
db-us-1.mixr.svc.cloud18:3306	Master		0-9-1211348571	0-3-1207946785		0/0	👍	✗	✗	✗	✗	✓	✗

MariaDB HA : replication-manager

- Hands on configurations
- Manual failover is our choice, alerts via slack and email
- Default GTID // conservative , strict, durable for data and binary logs
- Semisync used in single DC



MariaDB HA : proxysql

- Read write splitting SELECT goes to local slave
- SELECT /*hints*/ goes to master
- Weighted local slave
- No READ on master but READ can failover inside same DC
  - this enables transparent proxy restart

```
        ping_timeout_server=1000
        commands_stats=true
        sessions_sort=true
        connect_retries_on_failure=10
        monitor_writer_is_also_reader=0
}

# defines all the MySQL servers
mysql_servers =
(
{ address="db-ca-1.mixr.svc.cloud18" , port=3306 , hostgroup=1, max_connections=1024, weight=1  },
{ address="db-us-1.mixr.svc.cloud18" , port=3306 , hostgroup=1, max_connections=1024 , weight=1 },
{ address="db-front-3.mixr.svc.cloud18" , port=3306 , hostgroup=1, max_connections=1024 , weight=1000 }
)
```

MASTER_GTID_WAIT

- On Wordpress critical READS can happen just after a WRITE caused by object propagation to plugins

```
SELECT * FROM wp_comments WHERE comment_ID = ? LIMIT 1
```

- Master gtid wait function will timeout on slaves having replication delay

```
function wait_4_master_sync() {
  global $wpdb; $sql = "SELECT @@gtid_current_pos as id;";
  $id = $wpdb->get_row($sql);
  if ( is_object( $id ) ){
    $value = $id->id;
    $sql = "SELECT MASTER_GTID_WAIT('".$value."', ".GTID_WAIT.");";
    $wpdb->query($sql);
  }
}
```

Conclusion

- Fixed cost infrastructure
- No share servers with others
- GeoDNS to each zone
- Embedding best practice on MySQL/MariaDB
- Services = no loss of sysops knowledge
- Monitored and secured stack
- Easy resource reallocation
- No virtualisation extra cost
- Enjoying dynamic change via RPC routing and secrets