Opendaylight scalability issues in super scale data center

Yi Yang@Inspur Cloud

inspur 浪潮云

Content

- Controller Clustering Stability, Reliability, Scalability
- South bound plugin scalability
- VXLAN scalability
- Other misc. issues: startup time, memory consumption, too many threads, …

Controller Clustering Stability Issues

- Not reliable, https://jira.opendaylight.org/browse/CONTROLLER-1892

https://git.opendaylight.org/gerrit/p/integration/test.git ./csit/suites/openstack/clustering/ha_l2.robot can reproduce this very easily

https://jira.opendaylight.org/browse/NETVIRT-1318 MDSAL best practice

https://jira.opendaylight.org/browse/NETVIRT-1384: Umbrella: Numerous new transaction leaks
examples: https://git.opendaylight.org/gerrit/#/c/62640/

https://git.opendaylight.org/gerrit/#/c/62886/

https://git.opendaylight.org/gerrit/#/q/topic:transaction-helper

https://git.opendaylight.org/gerrit/#/c/63372/

https://git.opendaylight.org/gerrit/#/c/63402/

- To Be Done:https://jira.opendaylight.org/browse/NETVIRT-1320,
- An example using managed transaction: https://git.opendaylight.org/gerrit/#/c/75005/

Other Issues of Controller Clustering

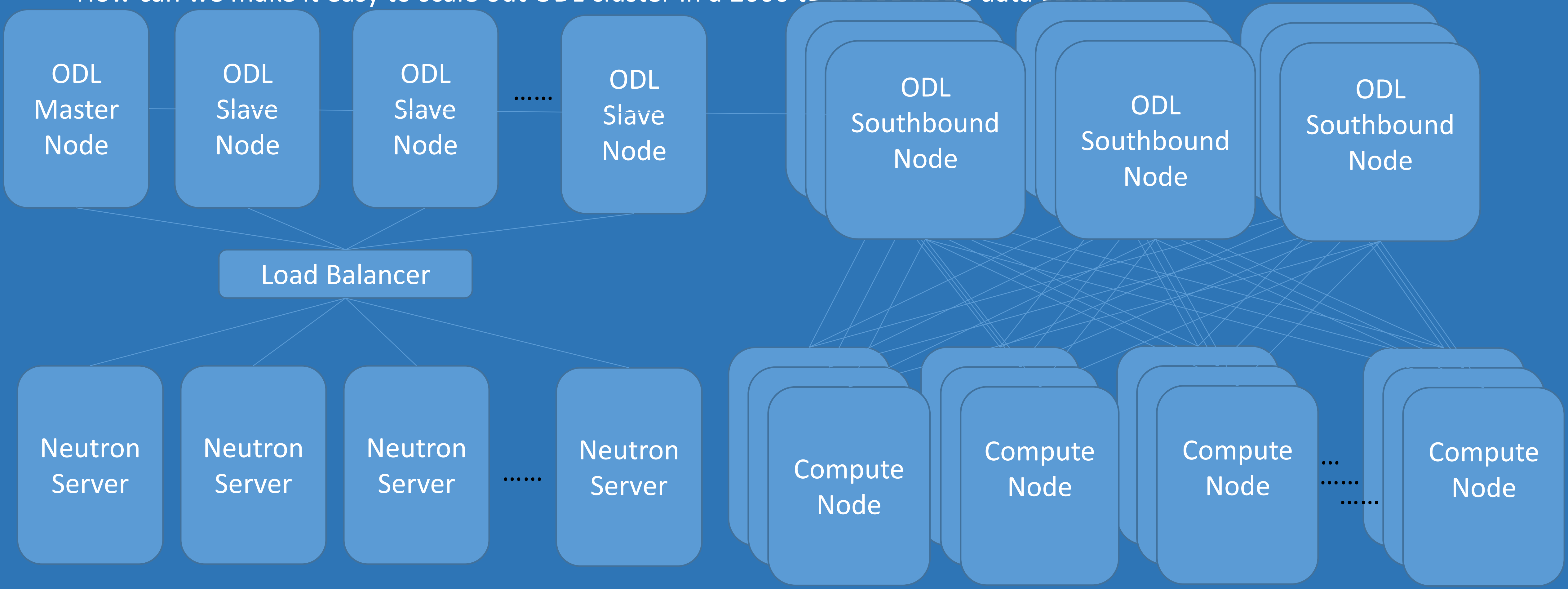
- How can it work with 127 cluster nodes?
- Replication to other 126?
- More granular shard: e.g. per openvswitch group for topology and inventory
- Cluster leader, shard leader and openvswitch master, it will be better if shard leader is same as openflowplugin master for openvswitch.
- Is read possible in any follower shard?
- Is asymmetric clustering possible? Nodes for neutron server and nodes for southbound device/openvswitch.
- Does Database backend help on these issues? https://wiki.opendaylight.org/view/Project_Proposals:Alt-datastores

Southbound plugin scalability

- Inventory and network topology data store are big
- Openflowplugin clustering just uses 3 controller nodes (one master, two slaves), master can do read, write, flow statistics and async messages handling, slave only can read.
- A small lightweight southbound 3 node cluster is preferred for a group of compute node/network node.
- The same solution is applied to ovsdb

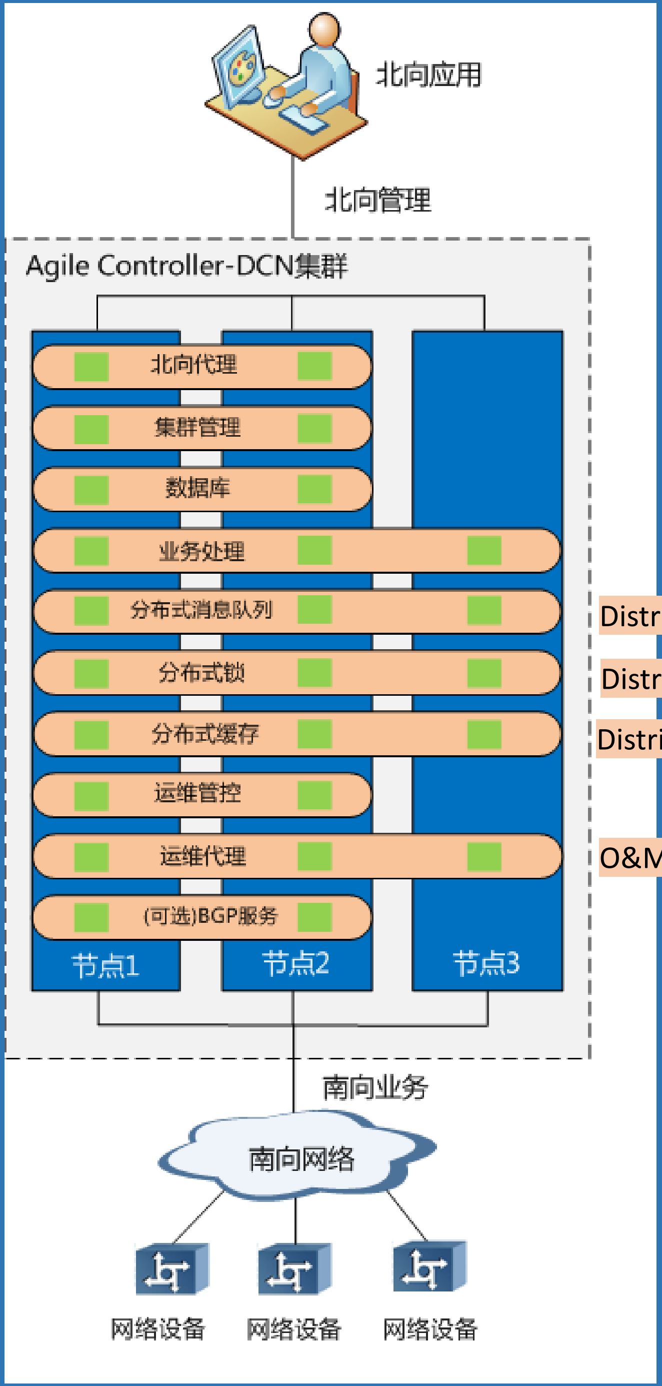
Is ODL Controller Cluster ready for super scale data center?

How can we make it easy to scale out ODL cluster in a 2000 to 10000 node data center?



HUAWEI

Agile Controller - DCN



Database

Distributed Message Queue

Distributed Lock

Distributed Cache

O&M Proxy

ZTE

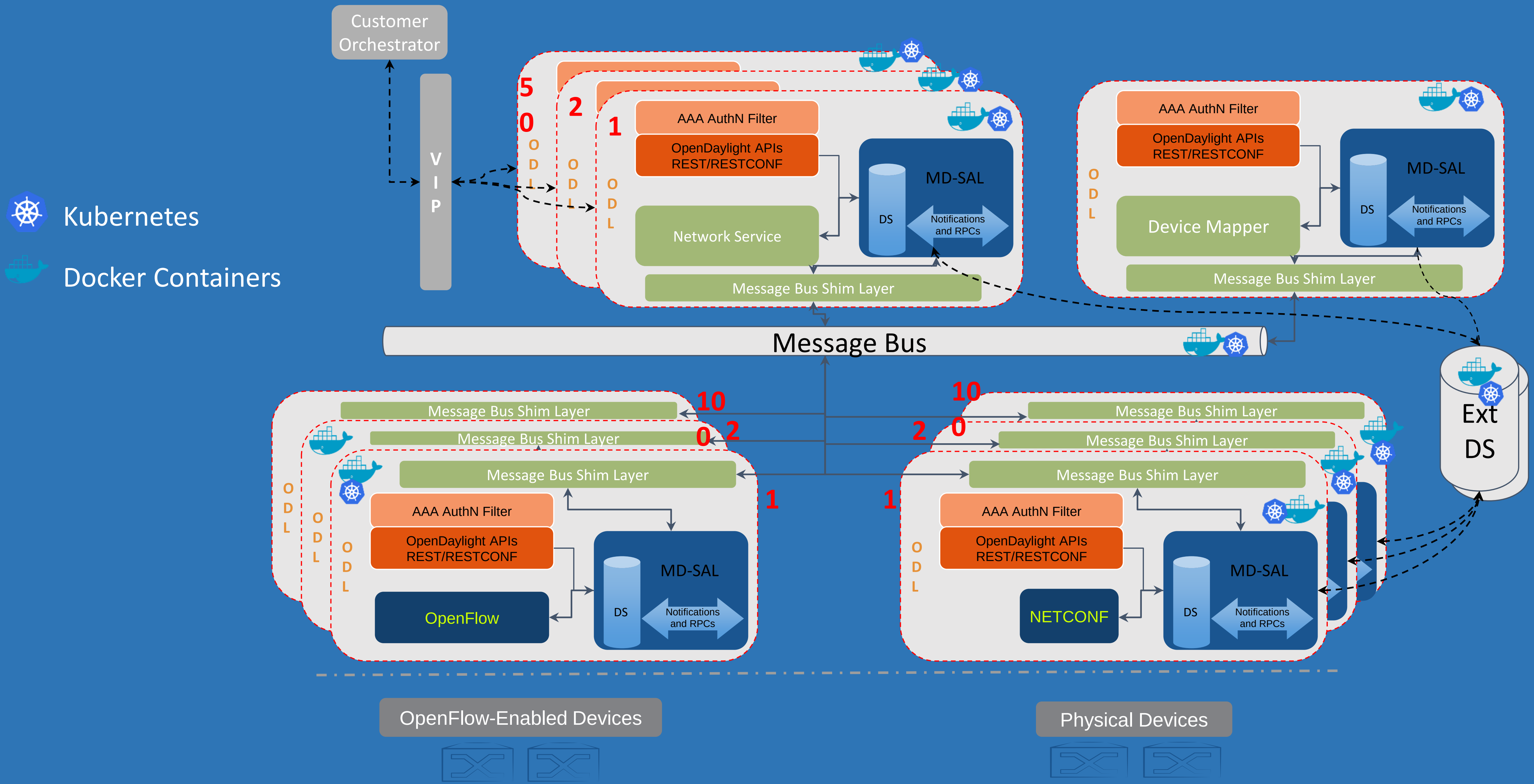
ZENIC vDC Controller

ZTE vDC ZENIC SDN Controller V2.00.10

Distributed Controller Cluster

2+N nodes： 2 master controllers（Active/Passive）， N（1-128）southbound controllers

inspur 浪潮云


Customer Orchestrator
VIP
50
2
1
ODL
ODL
ODL
AAA AuthN Filter
OpenDaylight APIs REST/RESTCONF
Network Service
MD-SAL
DS
Notifications and RPCs
Message Bus Shim Layer
ODL
AAA AuthN Filter
OpenDaylight APIs REST/RESTCONF
Device Mapper
MD-SAL
DS
Notifications and RPCs
Message Bus Shim Layer
Kubernetes
Docker Containers
Message Bus
Ext DS
10
20
1
Message Bus Shim Layer
Message Bus Shim Layer
Message Bus Shim Layer
ODL
ODL
ODL
AAA AuthN Filter
OpenDaylight APIs REST/RESTCONF
OpenFlow
MD-SAL
DS
Notifications and RPCs
10
20
1
Message Bus Shim Layer
Message Bus Shim Layer
Message Bus Shim Layer
ODL
AAA AuthN Filter
OpenDaylight APIs REST/RESTCONF
NETCONF
MD-SAL
DS
Notifications and RPCs
OpenFlow-Enabled Devices
Physical Devices

inspur 浪潮云

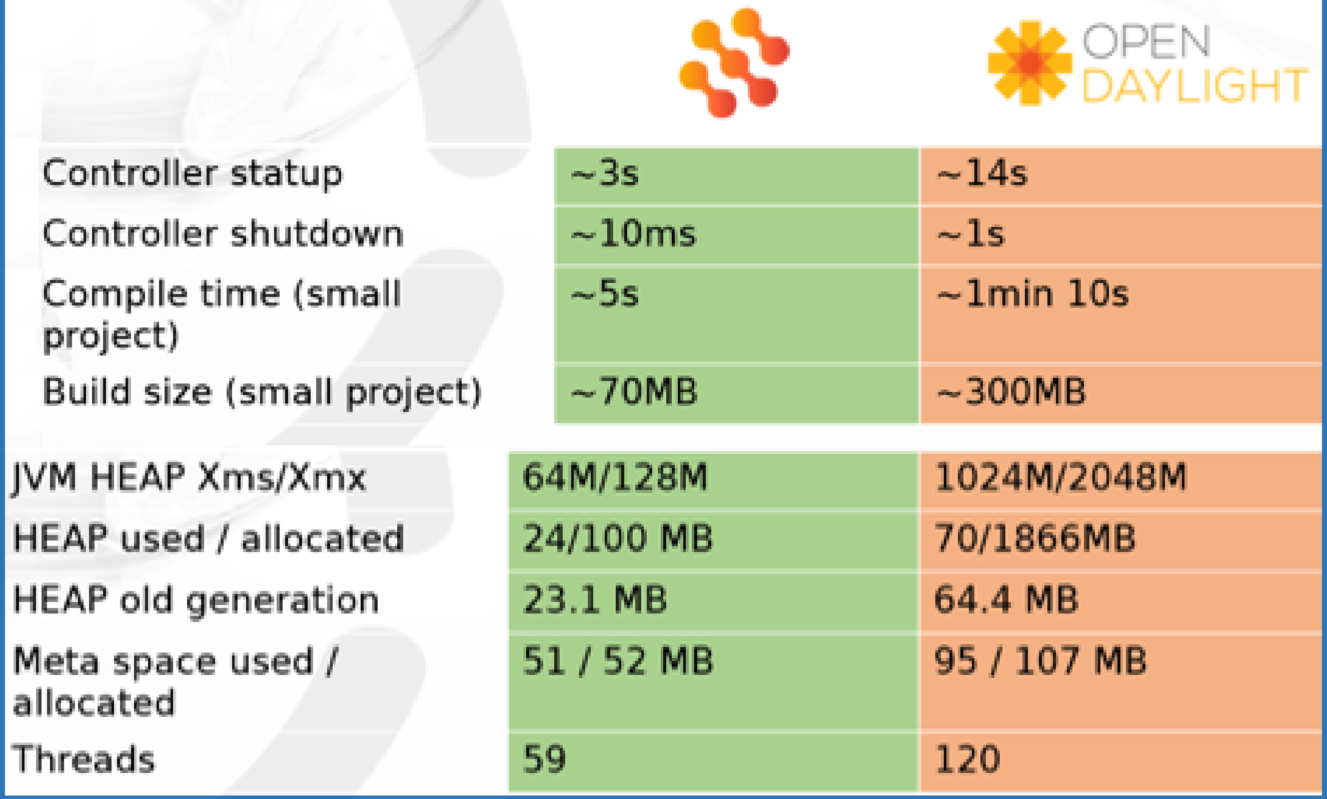
VXLAN scalability

- VxLAN tunnels are full-meshed between all the nodes, it is non-scalable
- ODL doesn’t support l2population
- L2population is also non-scalable although it is a big leap forward
- Ericsson folks are working on of-tunnel in itm-direct-tunnel in genius
- It is almost ready for merge
- Demo

DEMO

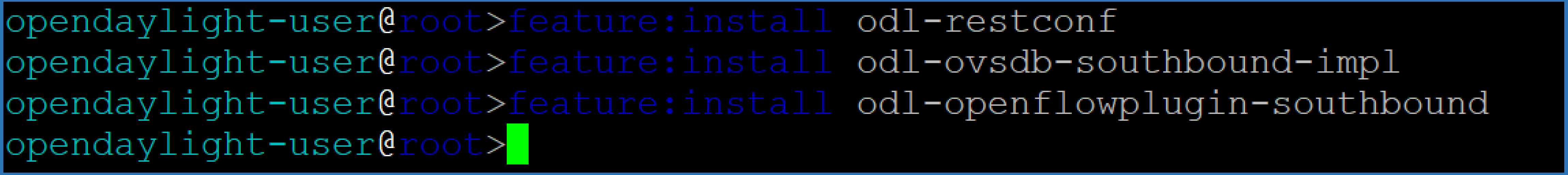
Other misc. issues

- Use too much memory
- Slow startup
- Too many threads

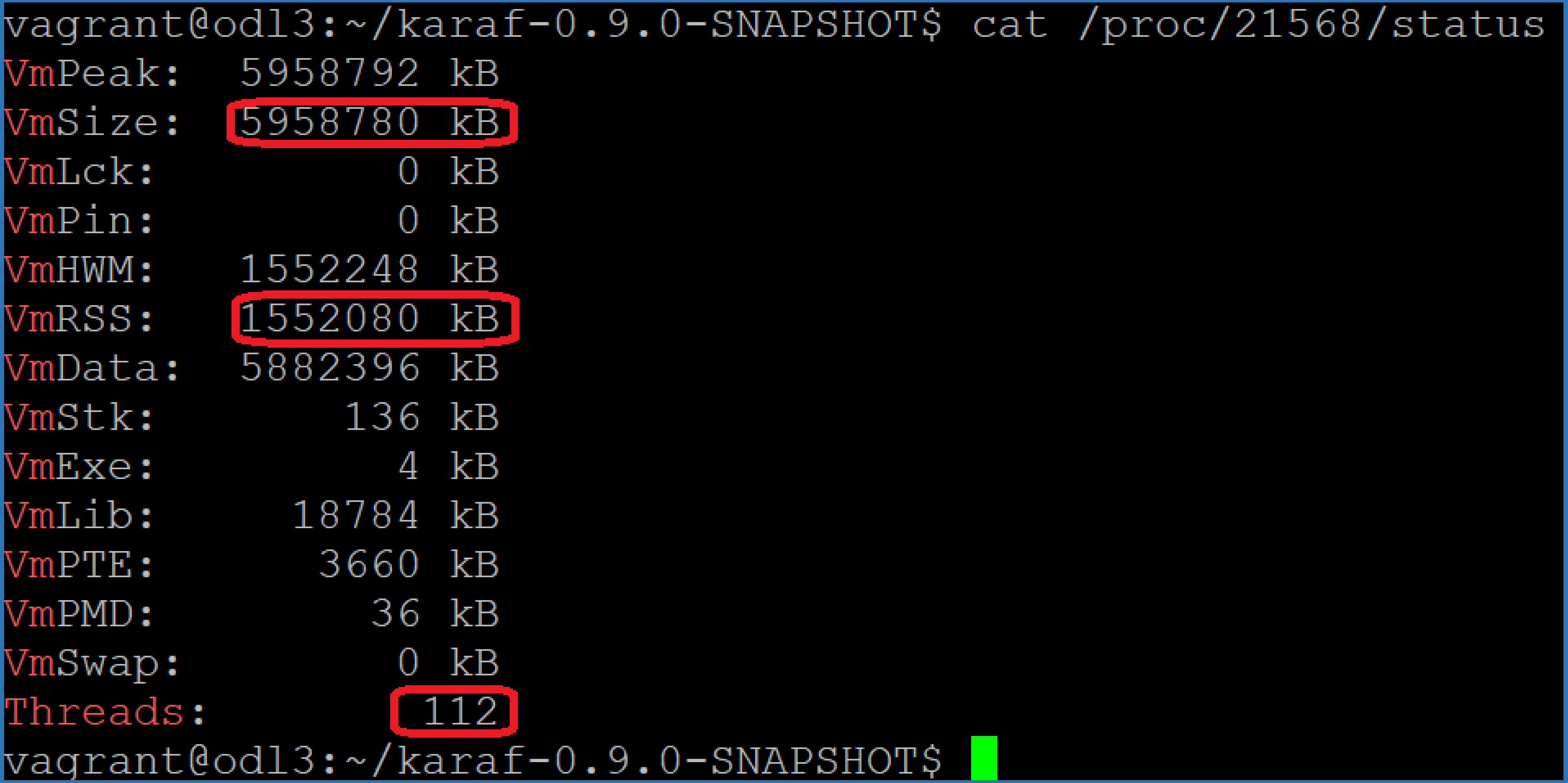


		OPEN DAYLIGHT
Controller statup	~3s	~14s
Controller shutdown	~10ms	~1s
Compile time (small project)	~5s	~1min 10s
Build size (small project)	~70MB	~300MB
JVM HEAP Xms/Xmx	64M/128M	1024M/2048M
HEAP used / allocated	24/100 MB	70/1866MB
HEAP old generation	23.1 MB	64.4 MB
Meta space used / allocated	51 / 52 MB	95 / 107 MB
Threads	59	120

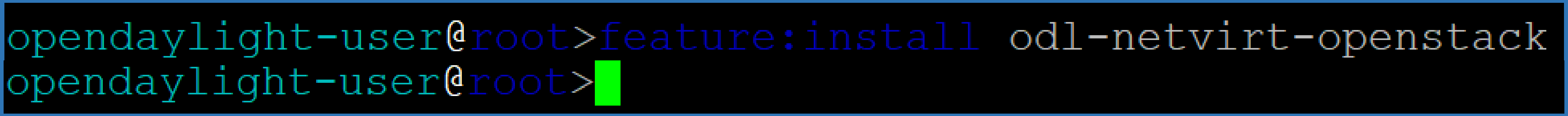
- Optimization:
  - ➢lighty.io (https://lighty.io/ remove karaf, faster with better memory efficiency)
  - ➢opendaylight-simple (https://github.com/vorburger/opendaylight-simple), use guice (pronounced '**juice**‘, a lightweight dependency injection framework) instead of karaf



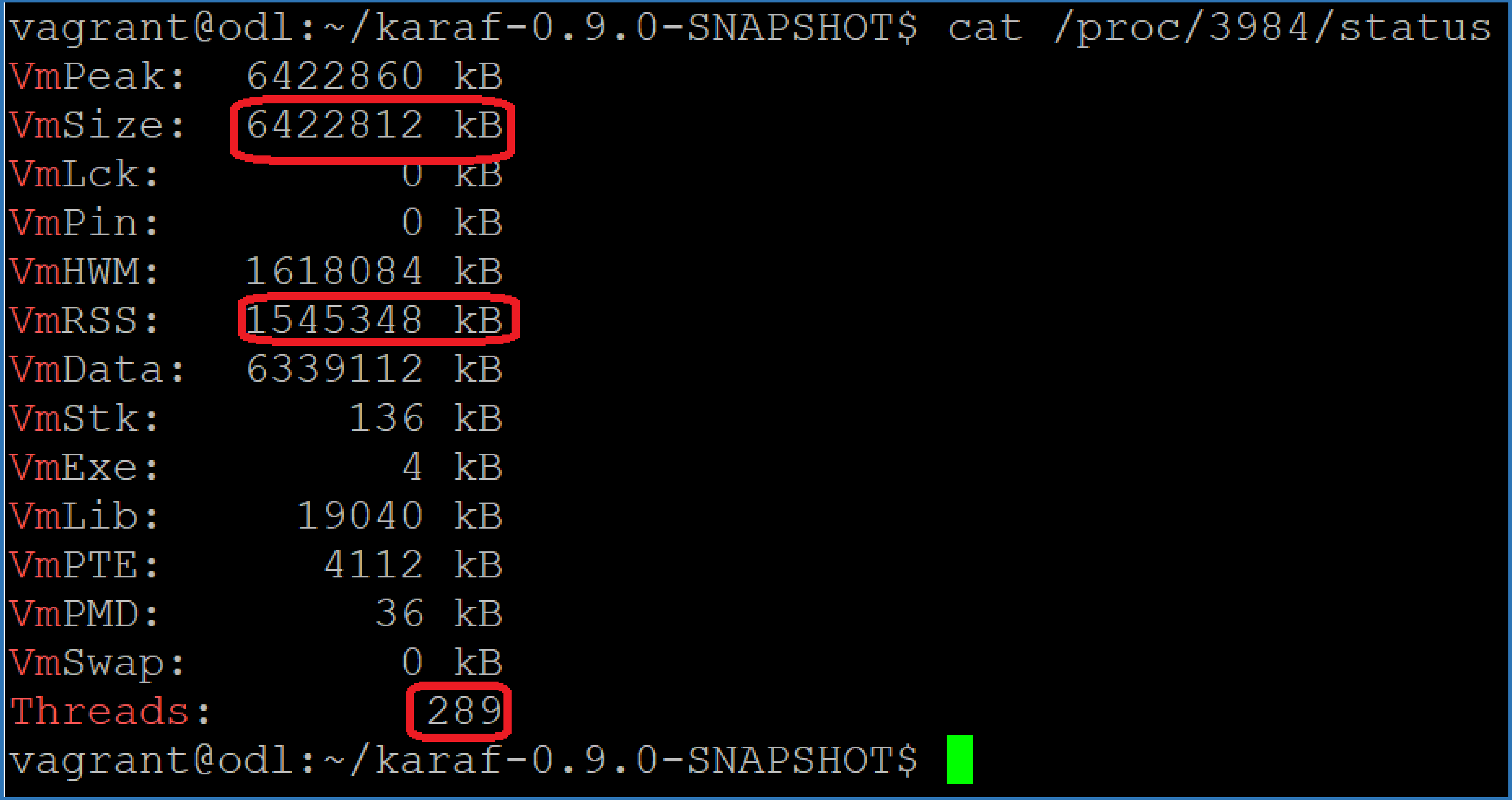
```
opendaylight-user@root>feature:install odl-restconf
opendaylight-user@root>feature:install odl-ovsdb-southbound-impl
opendaylight-user@root>feature:install odl-openflowplugin-southbound
opendaylight-user@root>
```



```
vagrant@odl3:~/karaf-0.9.0-SNAPSHOT$ cat /proc/21568/status
VmPeak:   5958792 kB
VmSize:   5958780 kB
VmLck:          0 kB
VmPin:          0 kB
VmHWM:    1552248 kB
VmRSS:    1552080 kB
VmData:   5882396 kB
VmStk:        136 kB
VmExe:          4 kB
VmLib:      18784 kB
VmPTE:       3660 kB
VmPMD:         36 kB
VmSwap:         0 kB
Threads:        112
vagrant@odl3:~/karaf-0.9.0-SNAPSHOT$
```



```
opendaylight-user@root>feature:install odl-netvirt-openstack
opendaylight-user@root>
```



```
vagrant@odl:~/karaf-0.9.0-SNAPSHOT$ cat /proc/3984/status
VmPeak:  6422860 kB
VmSize:  6422812 kB
VmLck:         0 kB
VmPin:         0 kB
VmHWM:   1618084 kB
VmRSS:   1545348 kB
VmData:  6339112 kB
VmStk:       136 kB
VmExe:         4 kB
VmLib:     19040 kB
VmPTE:      4112 kB
VmPMD:        36 kB
VmSwap:        0 kB
Threads:     289
vagrant@odl:~/karaf-0.9.0-SNAPSHOT$
```

Summary

- ODL Controller Clustering is NOT stable, NOT reliable and NOT scalable for super scale cloud data center
- ODL is NOT container friendly
- Southbound plugin scalability is BAD
- VXLAN is NOT scalable (in progress)
- Startup is SLOW
- Memory consumption is BIG
- Too MANY threads

Call for action: ODL community needs to take efforts on these directions, they are very important for ODL if we want to push ODL to cloud data center.