Gdevops

全球敏捷运维峰会

京东弹性云技术

演讲人：京东南京研发中心 鲍永成

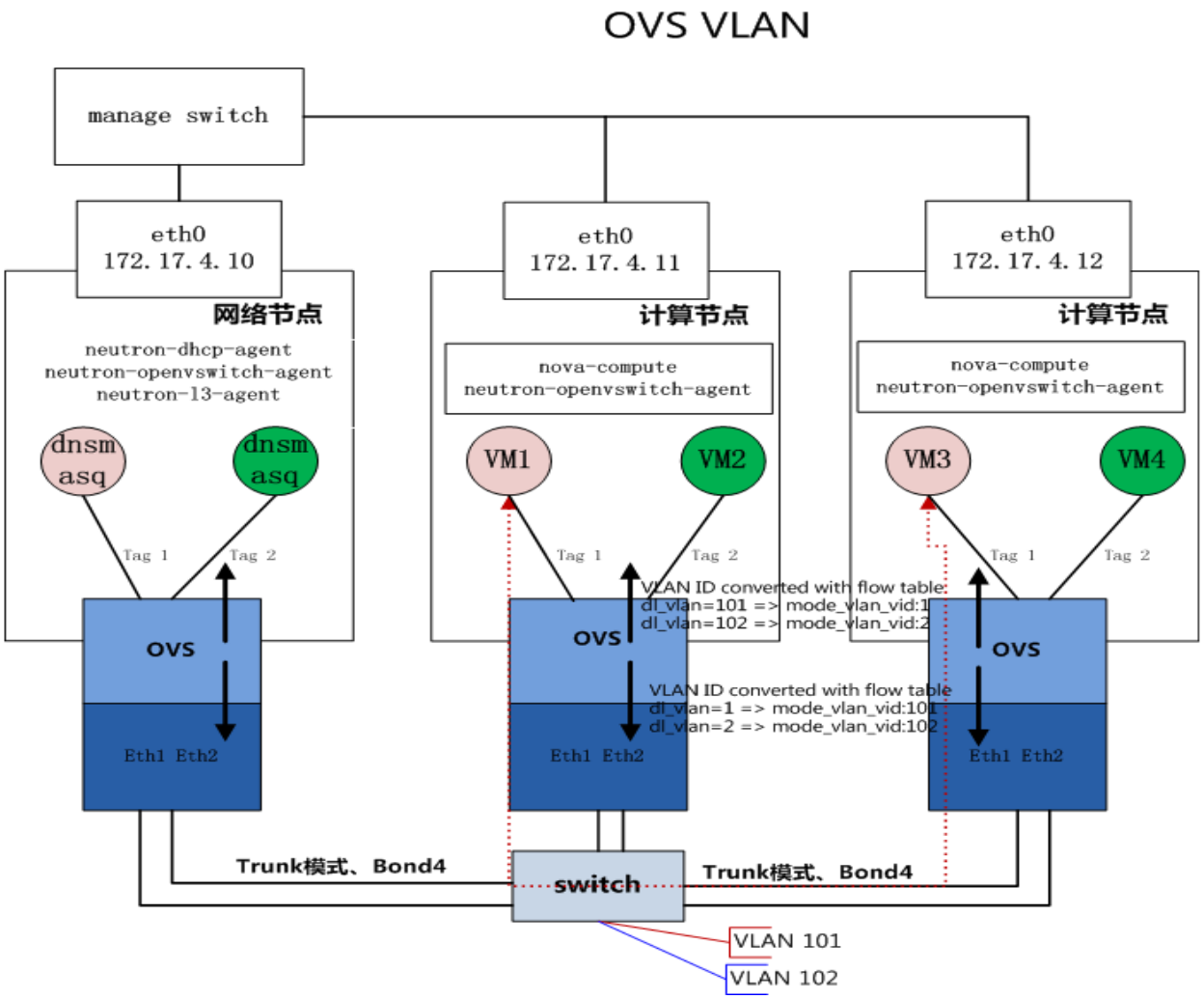
规模

京东全部业务run在弹性云平台

架构

Openstack+OVS+Docker

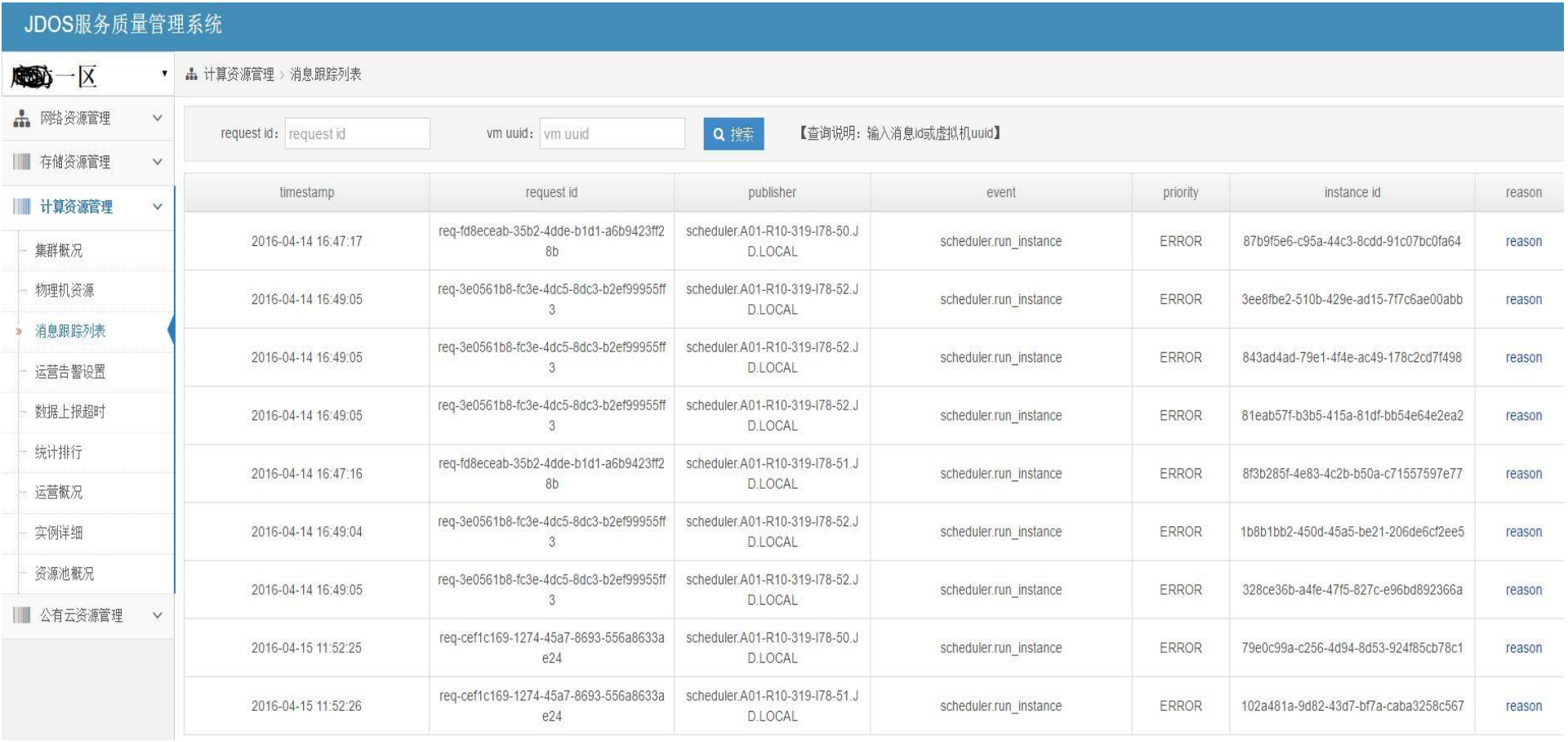
京东弹性云-网络


OVS VLAN
manage switch
eth0
172.17.4.10
网络节点
neutron-dhcp-agent
neutron-openvswitch-agent
neutron-l3-agent
dnsm asq
dnsm asq
Tag 1
Tag 2
OVS
Eth1 Eth2
eth0
172.17.4.11
计算节点
nova-compute
neutron-openvswitch-agent
VM1
VM2
Tag 1
Tag 2
OVS
Eth1 Eth2
VLAN ID converted with flow table:
dl_vlan=101 => mode_vlan_vid:1
dl_vlan=102 => mode_vlan_vid:2
VLAN ID converted with flow table
dl_vlan=1 => mode_vlan_vid:101
dl_vlan=2 => mode_vlan_vid:102
eth0
172.17.4.12
计算节点
nova-compute
neutron-openvswitch-agent
VM3
VM4
Tag 1
Tag 2
OVS
Eth1 Eth2
Trunk模式、Bond4
switch
Trunk模式、Bond4
VLAN 101
VLAN 102

Docker

- cgroup 需要被认真对待：
  - CPU默认调度cfs算法 cpu set & cpu share
  - blkio默认cfq vs deadline调度算法
- JD胖容器：/ + /export　　/=readonly, /export=LVM(JD工具链+应用系统)
- Disable docker network，use neutron（OVS-vlan）
- 监控加强: 加入process，open fd
- 自研镜像build增强：compress/tag/clear/include

京东弹性云-运维情况



timestamp	request id	publisher	event	priority	instance id	reason
2016-04-14 16:47:17	req-fd8eceab-35b2-4dde-b1d1-a6b9423ff28b	scheduler.A01-R10-319-I78-50.JD.LOCAL	scheduler.run_instance	ERROR	87b9f5e6-c95a-44c3-8cdd-91c07bc0fa64	reason
2016-04-14 16:49:05	req-3e0561b8-fc3e-4dc5-8dc3-b2ef99955ff3	scheduler.A01-R10-319-I78-52.JD.LOCAL	scheduler.run_instance	ERROR	3ee8fbe2-510b-429e-ad15-7f7c6ae00abb	reason
2016-04-14 16:49:05	req-3e0561b8-fc3e-4dc5-8dc3-b2ef99955ff3	scheduler.A01-R10-319-I78-52.JD.LOCAL	scheduler.run_instance	ERROR	843ad4ad-79e1-4f4e-ac49-178c2cd7f498	reason
2016-04-14 16:49:05	req-3e0561b8-fc3e-4dc5-8dc3-b2ef99955ff3	scheduler.A01-R10-319-I78-52.JD.LOCAL	scheduler.run_instance	ERROR	81eab57f-b3b5-415a-81df-bb54e64e2ea2	reason
2016-04-14 16:47:16	req-fd8eceab-35b2-4dde-b1d1-a6b9423ff28b	scheduler.A01-R10-319-I78-51.JD.LOCAL	scheduler.run_instance	ERROR	8f3b285f-4e83-4c2b-b50a-c71557597e77	reason
2016-04-14 16:49:04	req-3e0561b8-fc3e-4dc5-8dc3-b2ef99955ff3	scheduler.A01-R10-319-I78-52.JD.LOCAL	scheduler.run_instance	ERROR	1b8b1bb2-450d-45a5-be21-206de6cf2ee5	reason
2016-04-14 16:49:05	req-3e0561b8-fc3e-4dc5-8dc3-b2ef99955ff3	scheduler.A01-R10-319-I78-52.JD.LOCAL	scheduler.run_instance	ERROR	328ce36b-a4fe-47f5-827c-e96bd892366a	reason
2016-04-15 11:52:25	req-cef1c169-1274-45a7-8693-556a8633ae24	scheduler.A01-R10-319-I78-50.JD.LOCAL	scheduler.run_instance	ERROR	79e0c99a-c256-4d94-8d53-924f85cb78c1	reason
2016-04-15 11:52:26	req-cef1c169-1274-45a7-8693-556a8633ae24	scheduler.A01-R10-319-I78-51.JD.LOCAL	scheduler.run_instance	ERROR	102a481a-9d82-43d7-bf7a-caba3258c567	reason

京东弹性云-微笑运维尝试

亲，您好：

根据我们的走访调查，您申请的容器目前很多还处于沉睡状态，唤醒它们来为您效劳吧！脏活累活都交给它们搞定，别怕它们累着。

如它们不听使唤，请联系我们：http://jdos.jd.com/contact_us.html

项目【danpinye】容器排行

活跃容器TOP10

	容器IP	活跃系数
1	172.22.[redacted] cpu:2.79%;men:58.39%;disk:6.0%;tcpconns:687.0;net-in:0.01MB/s net-out:0.0MB/s	18
2	172.22.[redacted] cpu:2.59%;men:60.16%;disk:6.0%;tcpconns:693.0;net-in:0.01MB/s net-out:0.0MB/s	18
3	172.22.[redacted] cpu:2.58%;men:58.15%;disk:6.0%;tcpconns:688.0;net-in:0.01MB/s net-out:0.0MB/s	18
4	172.22.[redacted] cpu:2.68%;men:57.91%;disk:6.0%;tcpconns:699.0;net-in:0.01MB/s net-out:0.0MB/s	18
5	172.22.[redacted] cpu:2.77%;men:59.07%;disk:6.0%;tcpconns:688.0;net-in:0.01MB/s net-out:0.0MB/s	18
6	172.22.[redacted] cpu:9.32%;men:4.35%;disk:6.0%;tcpconns:150.0;net-in:1.59MB/s net-out:5.27MB/s	17
7	172.22.[redacted] cpu:2.53%;men:57.7%;disk:6.0%;tcpconns:699.0;net-in:0.01MB/s net-out:0.0MB/s	17
8	172.22.[redacted] cpu:9.16%;men:4.34%;disk:6.0%;tcpconns:150.0;net-in:1.59MB/s net-out:5.27MB/s	17
9	172.22.[redacted] cpu:9.38%;men:4.24%;disk:6.0%;tcpconns:148.0;net-in:1.59MB/s net-out:5.28MB/s	17
10	172.22.[redacted] cpu:5.4%;men:31.45%;disk:6.0%;tcpconns:466.0;net-in:0.7MB/s net-out:0.48MB/s	15

空闲容器TOP10

	容器IP	活跃系数
1	172.22.[redacted] cpu:0.59%;men:3.22%;disk:6.0%;tcpconns:None;net-in:0.0MB/s net-out:0.0MB/s	0
2	172.22.[redacted] cpu:0.92%;men:4.83%;disk:6.0%;tcpconns:None;net-in:0.0MB/s net-out:0.0MB/s	0
3	172.22.[redacted] cpu:0.73%;men:3.21%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
4	172.22.[redacted] cpu:0.56%;men:3.42%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
5	172.22.[redacted] cpu:0.53%;men:3.35%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
6	172.22.[redacted] cpu:0.87%;men:4.68%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
7	172.22.[redacted] cpu:0.61%;men:3.42%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
8	172.22.[redacted] cpu:0.84%;men:4.66%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
9	172.22.[redacted] cpu:0.9%;men:4.93%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1
10	172.22.[redacted] cpu:0.57%;men:3.44%;disk:6.0%;tcpconns:1.0;net-in:0.0MB/s net-out:0.0MB/s	1

Openstack集群规模-How

- 从F版开始使用自研brooder（default have redis） like nova-conductor
- 自研python RPC (eventlet+msgpack) 内部叫yaRPC，收录到 http://msgpack.org/
- neutron-openvswitch-agent update升级 （JD从I版开始支持）
- Openstack广泛使用的文件锁，JD重新写该函数
- 消息追踪，自研yaRPC自带消息tracking，并记录到DB
- 大量调度特性功能比如：应用根据app-id夸交换机/夸机架,根据业务系统级别，资源特性调度到不同的nova zone，不同隔离特性调度到不同的网络zone
- 开发JD IDC network zone功能，支持IDC不同物理pod
- OVS状态监控，内核信息追踪
- 巡检系统：

吐槽

- Kernel Bug：
  - https://bugs.centos.org/view.php?id=8703

```
2  kernel/futex.c

@@ -343,6 +343,8 @@ static void get_futex_key_refs(union futex_key *key)
343  343          case FUT_OFF_MMSHARED:
344  344                  futex_get_mm(key); /* implies MB (B) */
345  345                  break;
     346  +       default:
     347  +               smp_mb(); /* explicit MB (B) */
346  348          }
347  349  }
```

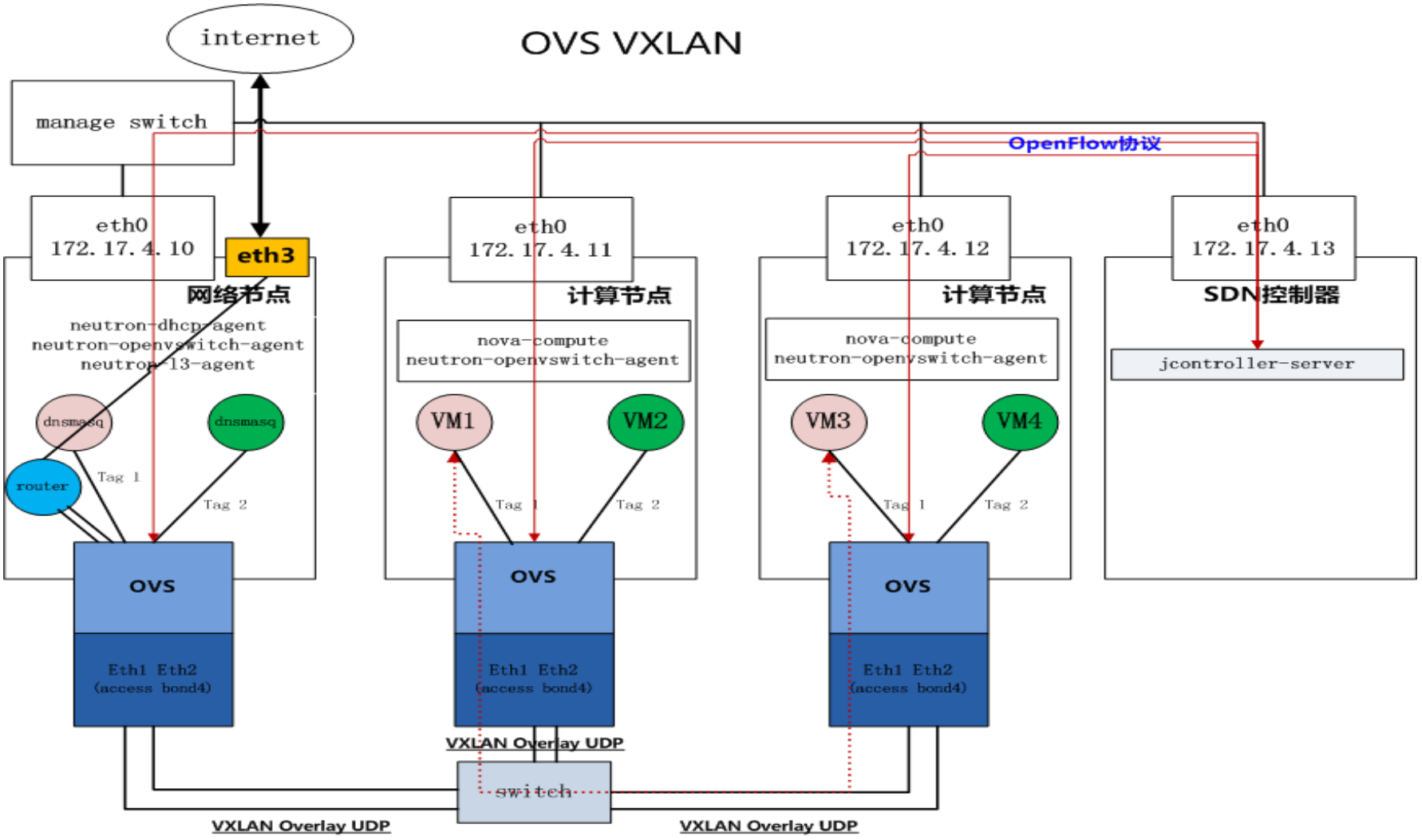
- 64CPU
  - /proc/sys/vm/min_free_kbytes
  - ulimit -v
  - MALLOC_ARENA_MAX
  - 服务器功率
- 10Gbps：SMP IRQ affinit设置更多CPU处理网卡中断
- CPU使用率xx%话题

Q&A

大胆假设，小心求证

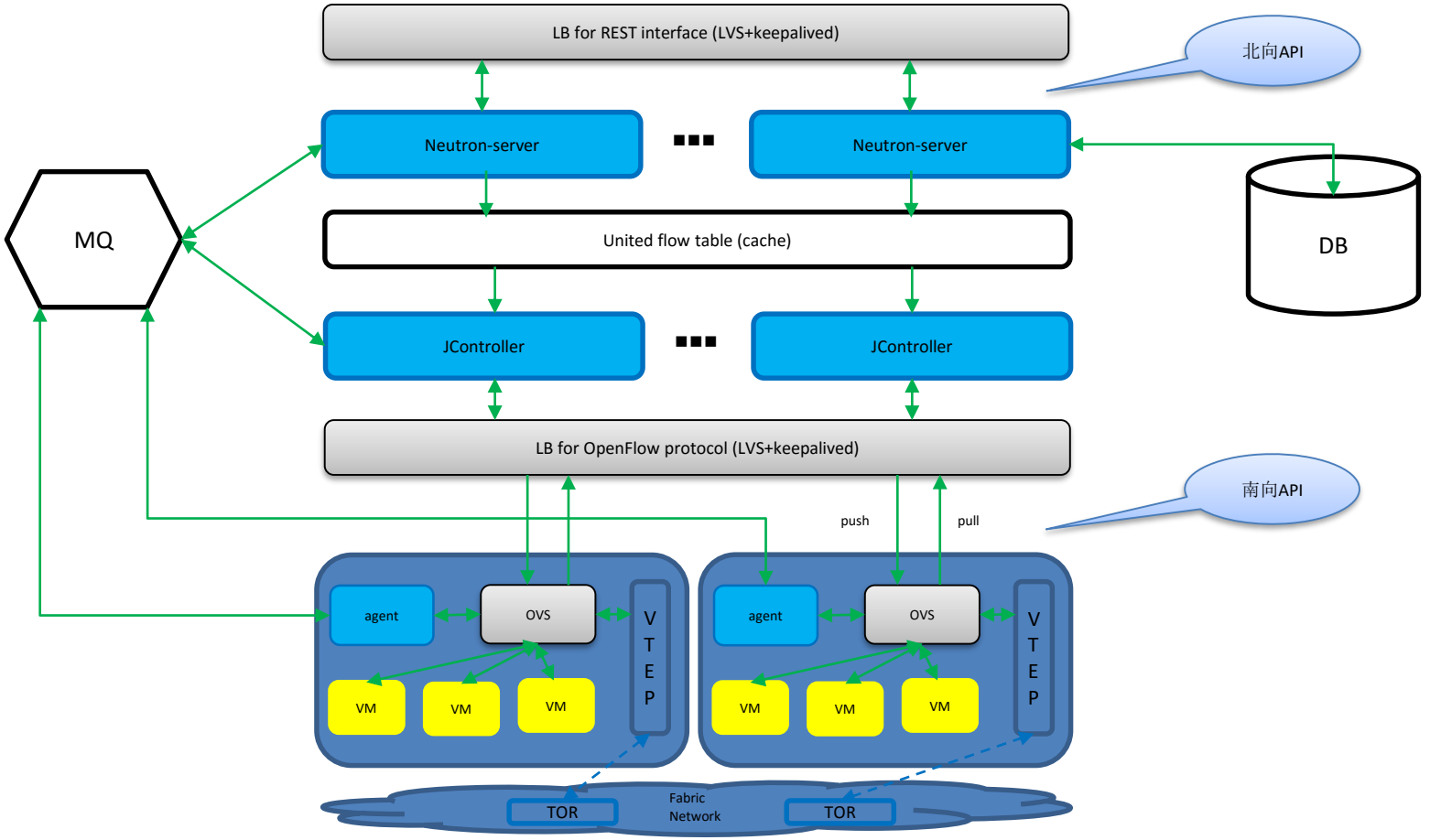
规模驱动技术持续演进

公有云使用的SDN


internet
OVS VXLAN
manage switch
OpenFlow协议
eth0
172.17.4.10
eth3
网络节点
neutron-dhcp-agent
neutron-openvswitch-agent
neutron-l3-agent
dnsmasq
dnsmasq
router
Tag 1
Tag 2
OVS
Eth1 Eth2
(access bond4)
eth0
172.17.4.11
计算节点
nova-compute
neutron-openvswitch-agent
VM1
VM2
Tag 1
Tag 2
OVS
Eth1 Eth2
(access bond4)
eth0
172.17.4.12
计算节点
nova-compute
neutron-openvswitch-agent
VM3
VM4
Tag 1
Tag 2
OVS
Eth1 Eth2
(access bond4)
eth0
172.17.4.13
SDN控制器
jcontroller-server
VXLAN Overlay UDP
switch
VXLAN Overlay UDP
VXLAN Overlay UDP

JD SDN控制器架构

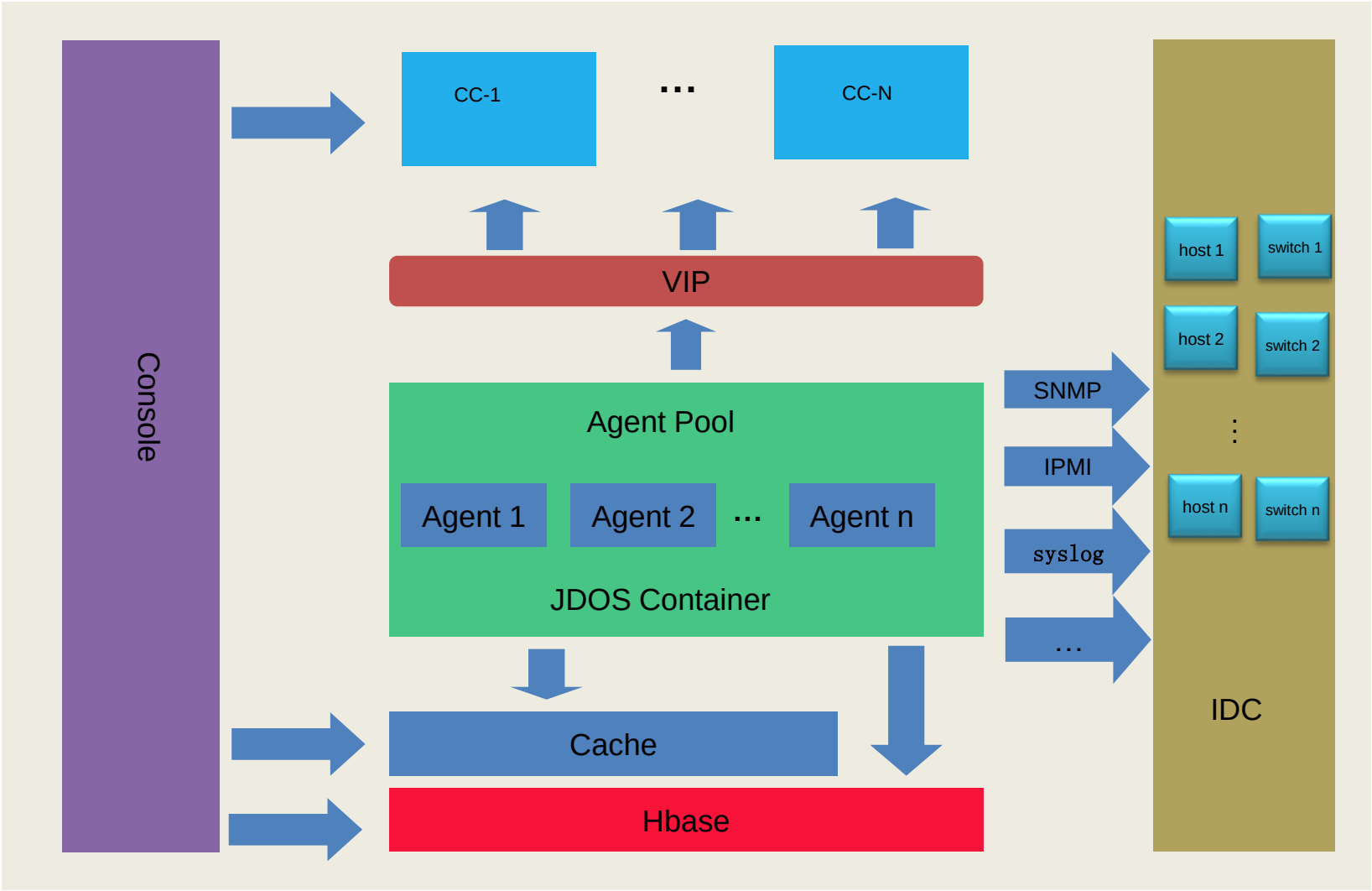
SDN(东西流向)



网络性能优化

OVS + DPDK

运维监控系统-整体架构


Console
CC-1
...
CC-N
VIP
Agent Pool
Agent 1
Agent 2
...
Agent n
JDOS Container
Cache
Hbase
SNMP
IPMI
syslog
...
host 1
switch 1
host 2
switch 2
host n
switch n
IDC

Gdevops

全球敏捷运维峰会

THANK YOU！