Open vSwitch

Romain Lenglet

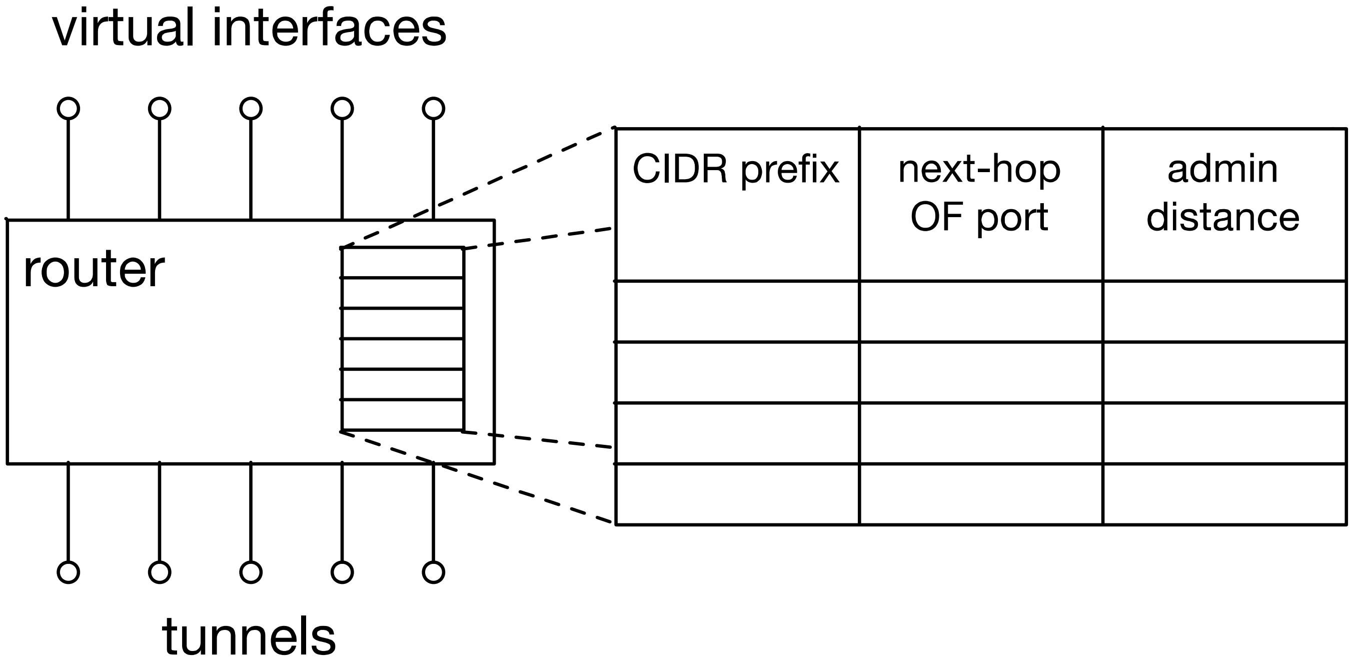
Oracle Public Cloud

IP FORWARDING WITH OVS

The Flow

```
table=7,priority=216,reg5=0x10000/0x10000,ip,nw_dst=
10.3.0.0/16, actions=load:0x0-
>NXM_NX_REG1[31..31],load:0x0->NXM_NX_REG2,load:0x0-
>NXM_NX_REG3,load:0x0->NXM_NX_REG4,load:0xffff-
>NXM_NX_REG5,load:0xffff-
>NXM_NX_REG6[0..15],bundle_load(symmetric_l3l4+udp,0
x95768,hrw,ofport,NXM_NX_REG6[16..31],slaves:12,43),
bundle_load(symmetric_l3l4+udp,0x95768,hrw,ofport,NX
M_NX_REG7[0..15],slaves:75),goto_table:8
```

model



CIDR prefix	next-hop OF port	admin distance

example routing table

CIDR prefix	next-hop OpenFlow port	administrative distance
192.168.1.2/32	A	1
10.3.0.0/16	B	1
10.3.0.0/16	C	1
10.3.0.0/16	A	2
0.0.0.0/0	B	1

longest prefix match – priority & flow per distinct prefix

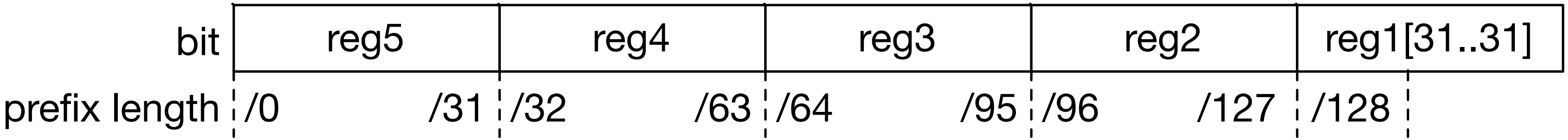
```
table=N,priority=232,…,ip,nw_dst=192.168.1.2/32,
actions=…,goto_table:N+1

table=N,priority=216,…,ip,nw_dst=10.3.0.0/16,
actions=…,goto_table:N+1

table=N,priority=200,…,ip,
actions=…,goto_table:N+1

table=N,priority=100,actions=drop
```

iterative lookup – prefix length bitmap in registers



bit	reg5	reg4	reg3	reg2	reg1[31..31]
prefix length	/0 /31	/32 /63	/64 /95	/96 /127	/128

longest prefix match + iterative lookup – init flow

```
table=N-1,…,actions=load:0x1-
>NXM_NX_REG1[31..31],load:0xffffffff-
>NXM_NX_REG2,load:0xffffffff-
>NXM_NX_REG3,load:0xffffffff-
>NXM_NX_REG4,load:0xffffffff-
>NXM_NX_REG5,goto_table:N
```

longest prefix match + iterative lookup – route flow

```
table=N,priority=216,
reg5=0x10000/0x10000,ip,nw_dst=10.3.0.0/16,
actions=load:0x0->NXM_NX_REG1[31..31],load:0x0-
>NXM_NX_REG2,load:0x0->NXM_NX_REG3,load:0x0-
>NXM_NX_REG4,load:0xffff-
>NXM_NX_REG5,…,goto_table:N+1
```

match on the bit for /16

enable all prefix lengths < 16, from /0 to /15 → set 16 bits

OVS / OpenFlow limitations

- too few registers
  - ½ of registers used for prefix length bitmap

- resubmit limit too small
  - hardcoded constant: 64
  - should really be > 2x129, maybe 300?

port status check and selection (ECMP)

```
table=N,…,nw_dst=10.3.0.0/16,
actions=…,
load:0xffff->NXM_NX_REG6[0..15],
bundle_load(…,NXM_NX_REG6[16..31],slaves:B,C),
bundle_load(…,NXM_NX_REG7[0..15],slaves:A),
goto_table:N+1
```

eliminate routes to ports down:

one ½ register with output port per admin distance (may be OFPP_NONE)

bundle load for every administrative distance [0..2] → **status check & ECMP**

or load OFPP_NONE if no route with given distance (e.g. 0)

ordering by admin distance in Table N+1

```
if distance 0 port is not NONE, output to distance 0 port
else if distance 1 port is not NONE, output to distance 1 port
else if distance 2 port is not NONE, output to distance 2 port
else resubmit to lookup shorter prefixes
```

problem: no “is not equal” predicate in OpenFlow

ordering by admin distance in Table N+1

```
table=N+1,priority=203,
reg6=0xffffffff,reg7=0xffff/0xffff, actions=resubmit(,N)

table=N+1,priority=202,
reg6=0xffffffff, actions=output to NXM_NX_REG7[0..15]

table=N+1,priority=201,
reg6=0xffff/0xffff, actions=output to NXM_NX_REG6[16..31]

table=N+1,priority=200,
actions=output to NXM_NX_REG6[0..15]
```

The Flow

```
table=N,priority=216,
reg5=0x10000/0x10000,ip,nw_dst=10.3.0.0/16, actions=
load:0x0->NXM_NX_REG1[31..31],load:0x0-
>NXM_NX_REG2,load:0x0->NXM_NX_REG3,load:0x0-
>NXM_NX_REG4,load:0xffff->NXM_NX_REG5,
load:0xffff->NXM_NX_REG6[0..15],
bundle_load(…,NXM_NX_REG6[16..31],slaves:B,C),
bundle_load(…,NXM_NX_REG7[0..15],slaves:A),
goto_table:N+1
```

Romain Lenglet

romain.lenglet@oracle.com