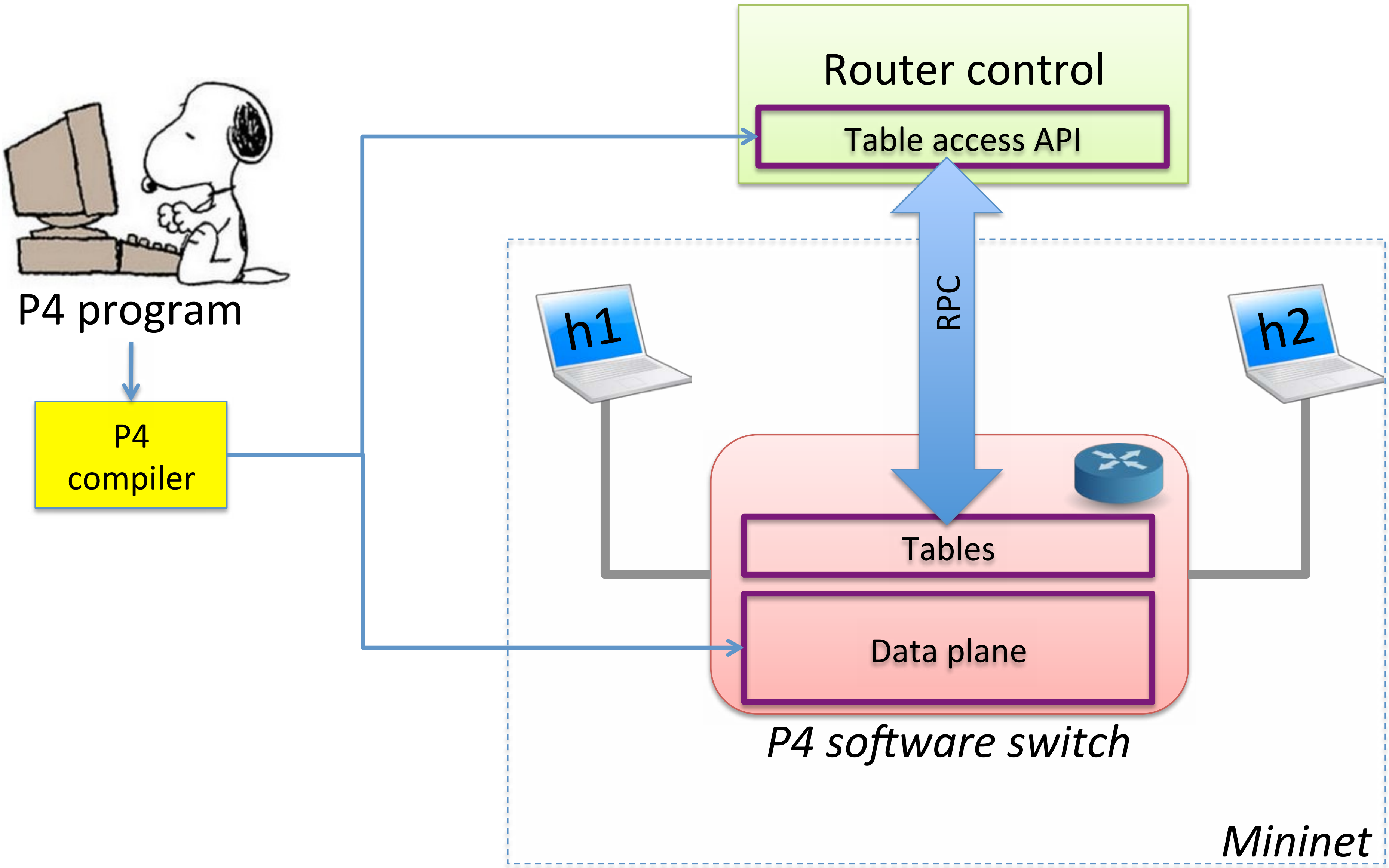
P4 demo:
a basic L2/L3 switch in 170 LOC

netdev0.1

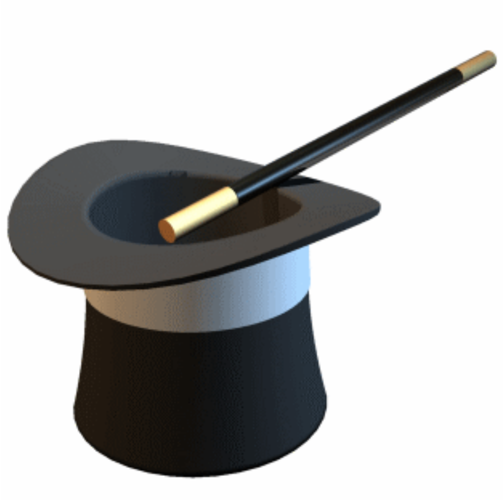
Ottawa, February 15, 2015

Mihai Budiu

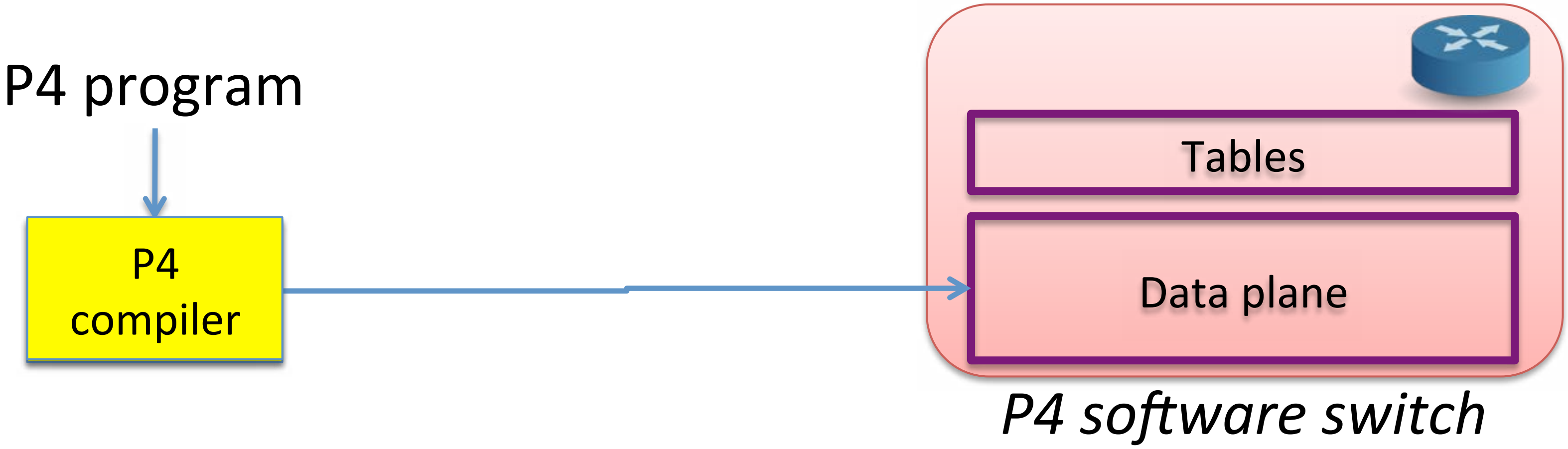
Setup


Router control
Table access API
P4 program
P4
compiler
RPC
h1
h2
Tables
Data plane
P4 software switch
Mininet

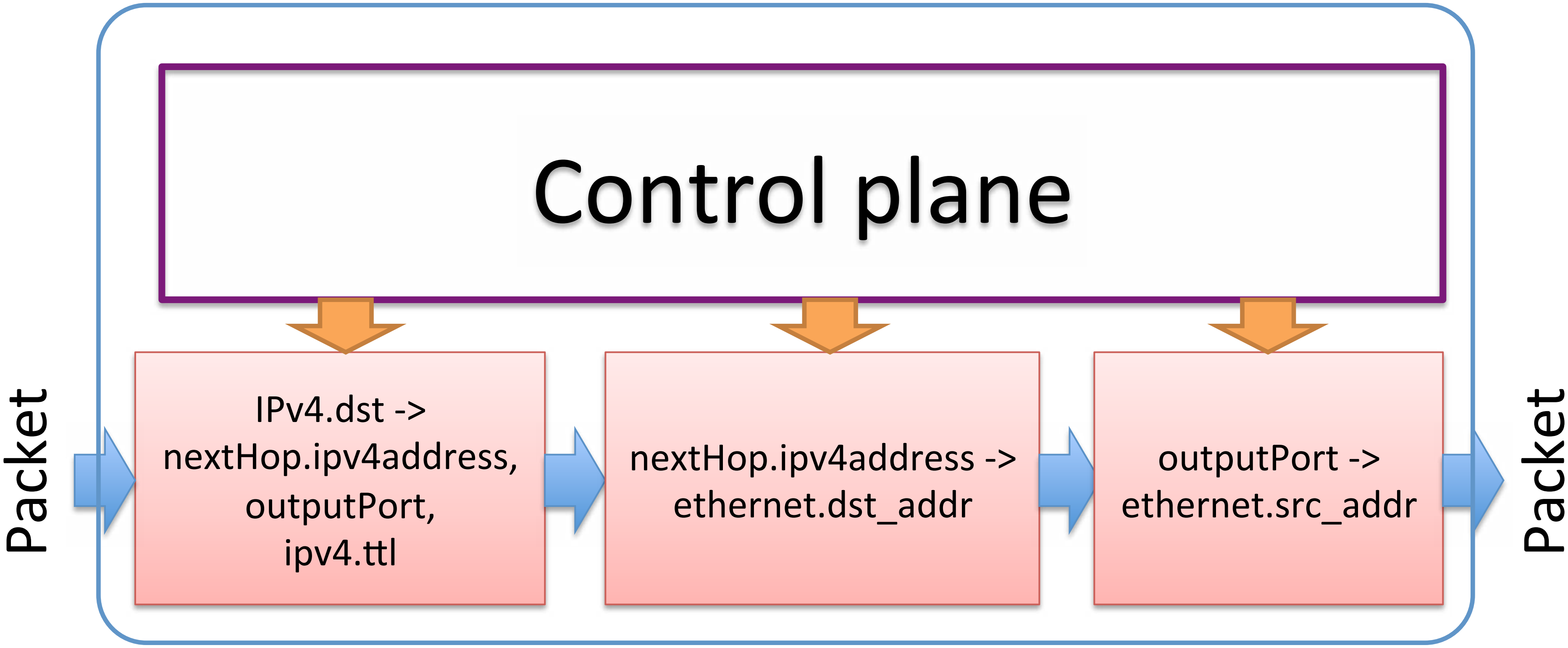
DEMO PART 1



Creating a basic ethernet+IPv4 switch from a P4 program



Layer 3 packet forwarding


Control plane
Packet
IPv4.dst -> nextHop.ipv4address, outputPort, ipv4.ttl
nextHop.ipv4address -> ethernet.dst_addr
outputPort -> ethernet.src_addr
Packet

Headers

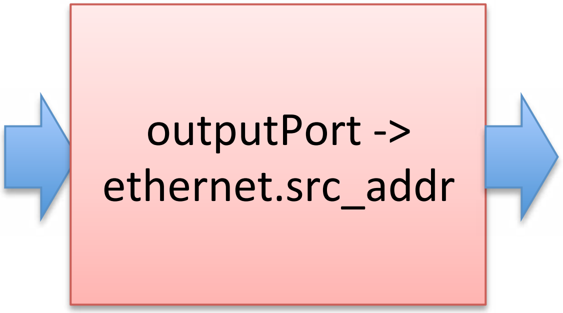
```
header_type ethernet_t {
   fields {
      dstAddr: 48;
      srcAddr: 48;
      etherType: 16;
   }
}

header_type ipv4_t  { … }
// no options
```

Parser

```
parser parse_ethernet {
  extract(ethernet);
  return select(latest.etherType) {
    0x0800 : parse_ipv4;
    default: ingress;
}}
…
parser parse_ipv4 { … }

calculated_field ipv4.hdrChecksum  {
  verify ipv4_checksum;
  update ipv4_checksum;
}
```

Last table

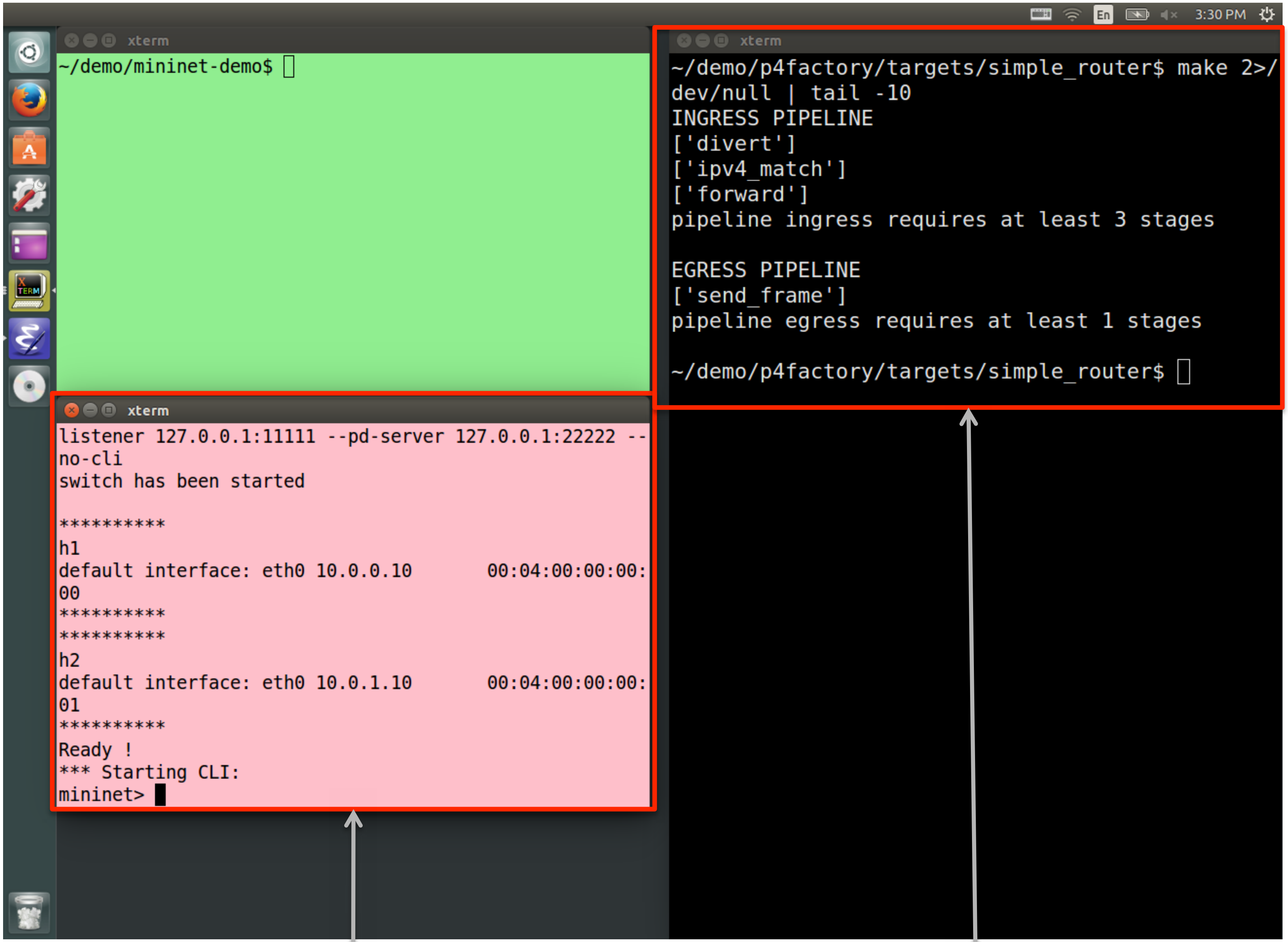
```
action rewrite_mac(smac) {
   modify_field(ethernet.srcAddr, smac);
}

table send_frame {
   reads { std_metadata.egress_port: exact; }
   actions {
      rewrite_mac;
      drop;
   }
   size: 256;
}
```



outputPort -> ethernet.src_addr

Complete pipeline

```
control ingress {
  apply(ipv4_match);
  apply(forward);
}

control egress {
  apply(send_frame);
}
```



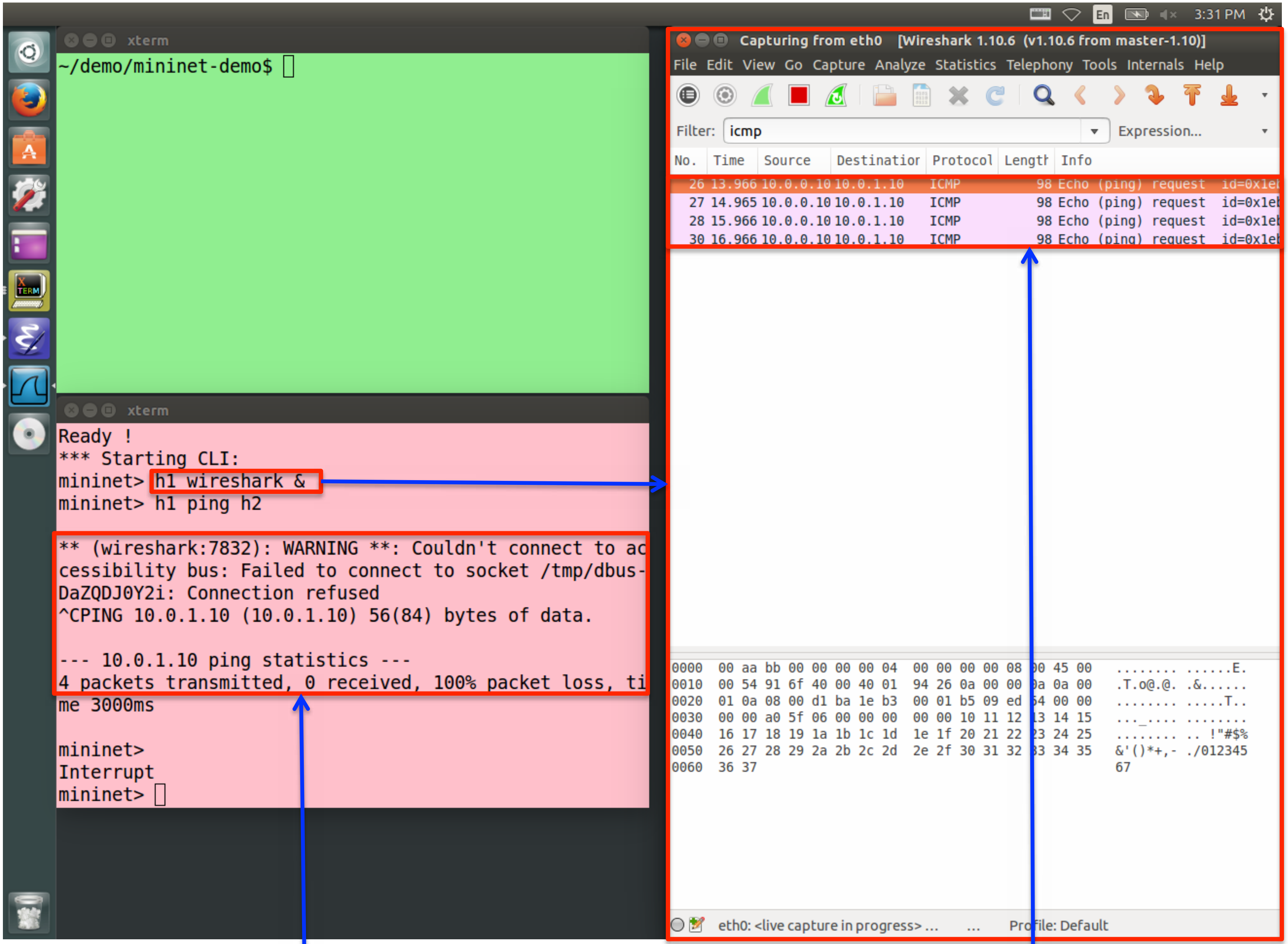
```
~/demo/mininet-demo$
```

```
~/demo/p4factory/targets/simple_router$ make 2>/
dev/null | tail -10
INGRESS PIPELINE
['divert']
['ipv4_match']
['forward']
pipeline ingress requires at least 3 stages

EGRESS PIPELINE
['send_frame']
pipeline egress requires at least 1 stages

~/demo/p4factory/targets/simple_router$
```

```
listener 127.0.0.1:11111 --pd-server 127.0.0.1:22222 --
no-cli
switch has been started

**********
h1
default interface: eth0 10.0.0.10       00:04:00:00:00:
00
**********
**********
h2
default interface: eth0 10.0.1.10       00:04:00:00:00:
01
**********
Ready !
*** Starting CLI:
mininet>
```

2. Started switch running

1. Compiled a new switch from P4


3:31 PM
En
xterm
xterm
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
Expression...

```
~/demo/mininet-demo$ 
```

```
Ready !
*** Starting CLI:
mininet> h1 wireshark &
mininet> h1 ping h2

** (wireshark:7832): WARNING **: Couldn't connect to ac
cessibility bus: Failed to connect to socket /tmp/dbus-
DaZQDJ0Y2i: Connection refused
^CPING 10.0.1.10 (10.0.1.10) 56(84) bytes of data.

--- 10.0.1.10 ping statistics ---
4 packets transmitted, 0 received, 100% packet loss, ti
me 3000ms

mininet> 
Interrupt
mininet> 
```

Capturing from eth0   [Wireshark 1.10.6 (v1.10.6 from master-1.10)]

Filter: icmp

```
No.  Time   Source    Destinatior Protocol Lengtl Info
  26 13.966 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1el
  27 14.965 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1el
  28 15.966 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1el
  30 16.966 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1el
```

```
0000  00 aa bb 00 00 00 00 04  00 00 00 00 08 00 45 00   ........ ......E.
0010  00 54 91 6f 40 00 40 01  94 26 0a 00 00 0a 0a 00   .T.o@.@. .&......
0020  01 0a 08 00 d1 ba 1e b3  00 01 b5 09 ed 54 00 00   ........ .....T..
0030  00 00 a0 5f 06 00 00 00  00 00 10 11 12 13 14 15   ..._.... ........
0040  16 17 18 19 1a 1b 1c 1d  1e 1f 20 21 22 23 24 25   ........ .. !"#$%
0050  26 27 28 29 2a 2b 2c 2d  2e 2f 30 31 32 33 34 35   &'()*+,- ./012345
0060  36 37                                              67
```

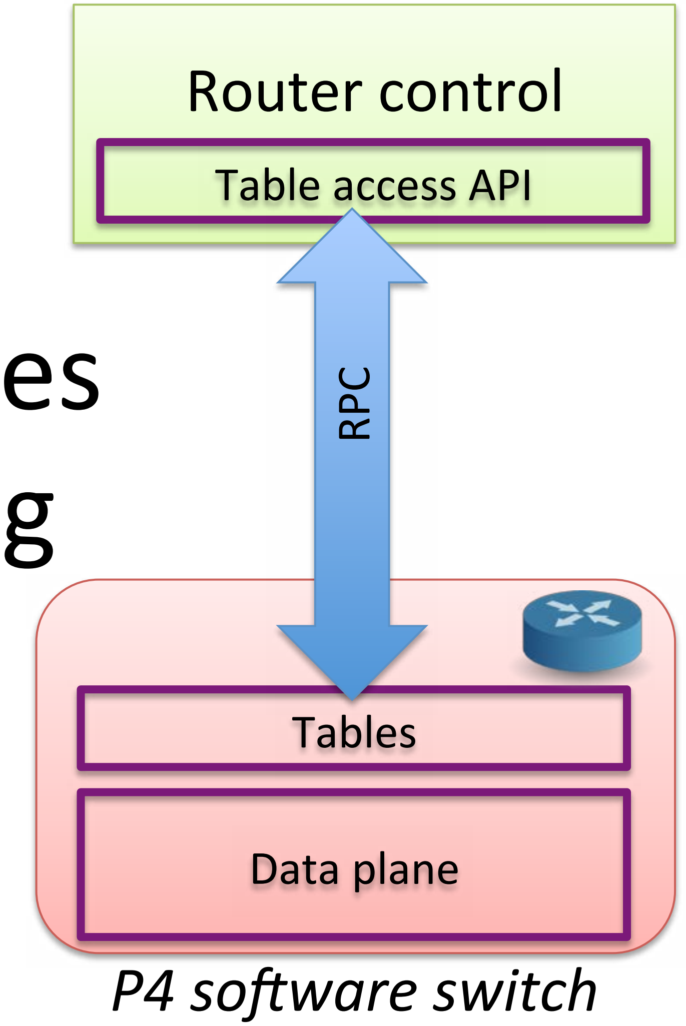
eth0: <live capture in progress> ...   ...   Profile: Default

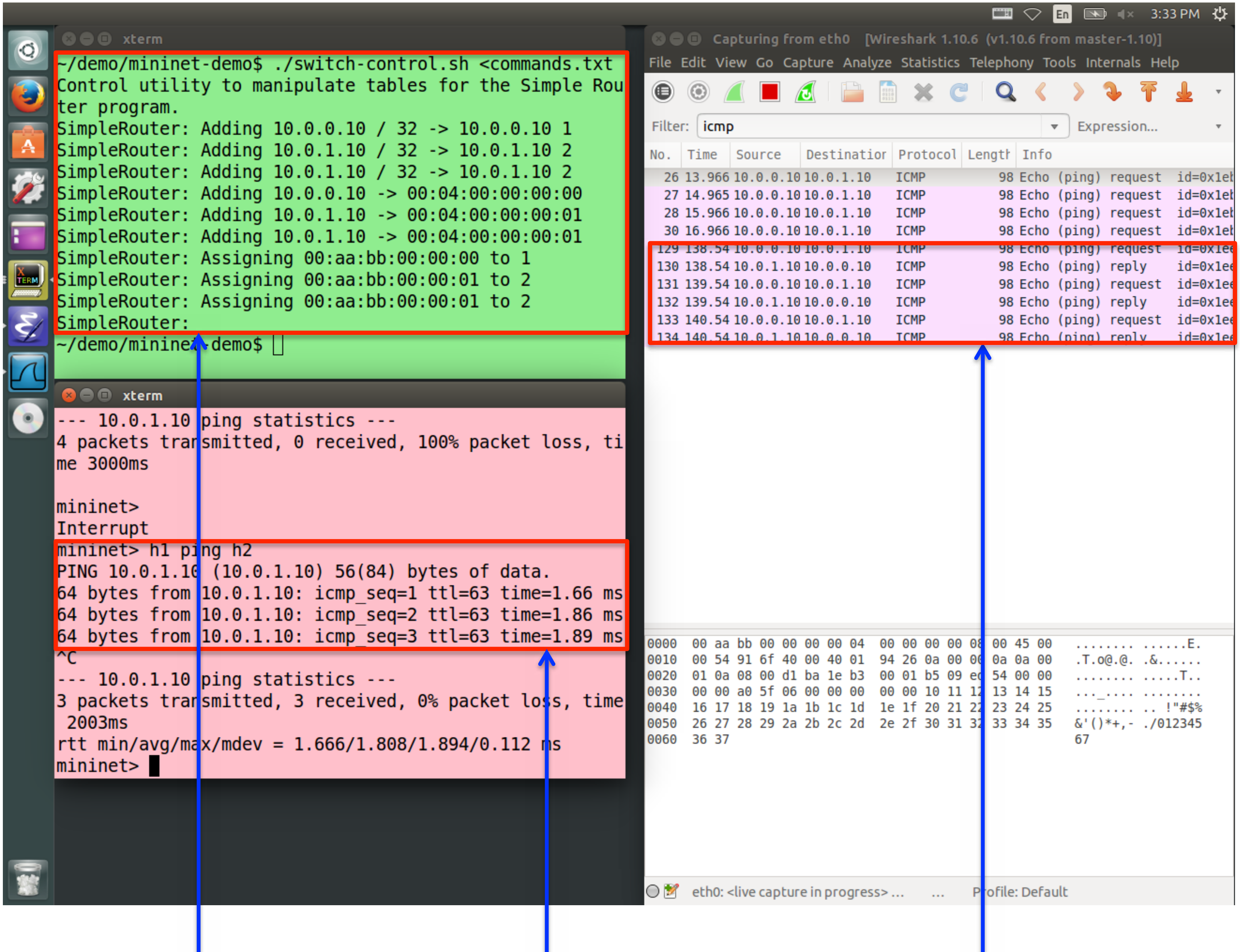
1. Ping from h1 to h2 does not work

2. Pings sent but no reply returned

DEMO PART 2

Populate switch tables to enable forwarding




3:33 PM
En
xterm
Capturing from eth0 [Wireshark 1.10.6 (v1.10.6 from master-1.10)]
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
Filter: icmp
Expression...
xterm
eth0: <live capture in progress> ... ... Profile: Default

```
~/demo/mininet-demo$ ./switch-control.sh <commands.txt
Control utility to manipulate tables for the Simple Rou
ter program.
SimpleRouter: Adding 10.0.0.10 / 32 -> 10.0.0.10 1
SimpleRouter: Adding 10.0.1.10 / 32 -> 10.0.1.10 2
SimpleRouter: Adding 10.0.1.10 / 32 -> 10.0.1.10 2
SimpleRouter: Adding 10.0.0.10 -> 00:04:00:00:00:00
SimpleRouter: Adding 10.0.1.10 -> 00:04:00:00:00:01
SimpleRouter: Adding 10.0.1.10 -> 00:04:00:00:00:01
SimpleRouter: Assigning 00:aa:bb:00:00:00 to 1
SimpleRouter: Assigning 00:aa:bb:00:00:01 to 2
SimpleRouter: Assigning 00:aa:bb:00:00:01 to 2
SimpleRouter:
~/demo/mininet-demo$
```

```
--- 10.0.1.10 ping statistics ---
4 packets transmitted, 0 received, 100% packet loss, ti
me 3000ms

mininet>
Interrupt
mininet> h1 ping h2
PING 10.0.1.10 (10.0.1.10) 56(84) bytes of data.
64 bytes from 10.0.1.10: icmp_seq=1 ttl=63 time=1.66 ms
64 bytes from 10.0.1.10: icmp_seq=2 ttl=63 time=1.86 ms
64 bytes from 10.0.1.10: icmp_seq=3 ttl=63 time=1.89 ms
^C
--- 10.0.1.10 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time
 2003ms
rtt min/avg/max/mdev = 1.666/1.808/1.894/0.112 ms
mininet>
```

```
No.  Time   Source    Destinatior Protocol Lengtt Info
 26 13.966 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1eb
 27 14.965 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1eb
 28 15.966 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1eb
 30 16.966 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1eb
129 138.54 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1ee
130 138.54 10.0.1.10 10.0.0.10   ICMP      98 Echo (ping) reply    id=0x1ee
131 139.54 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1ee
132 139.54 10.0.1.10 10.0.0.10   ICMP      98 Echo (ping) reply    id=0x1ee
133 140.54 10.0.0.10 10.0.1.10   ICMP      98 Echo (ping) request  id=0x1ee
134 140.54 10.0.1.10 10.0.0.10   ICMP      98 Echo (ping) reply    id=0x1ee
```

```
0000  00 aa bb 00 00 00 00 04  00 00 00 00 08 00 45 00   ........ ......E.
0010  00 54 91 6f 40 00 40 01  94 26 0a 00 00 0a 0a 00   .T.o@.@. .&......
0020  01 0a 08 00 d1 ba 1e b3  00 01 b5 09 ec 54 00 00   ........ .....T..
0030  00 00 a0 5f 06 00 00 00  00 00 10 11 12 13 14 15   ..._.... ........
0040  16 17 18 19 1a 1b 1c 1d  1e 1f 20 21 22 23 24 25   ........ .. !"#$%
0050  26 27 28 29 2a 2b 2c 2d  2e 2f 30 31 32 33 34 35   &'()*+,- ./012345
0060  36 37                                              67
```

1. Populated all tables

2. Ping starts running

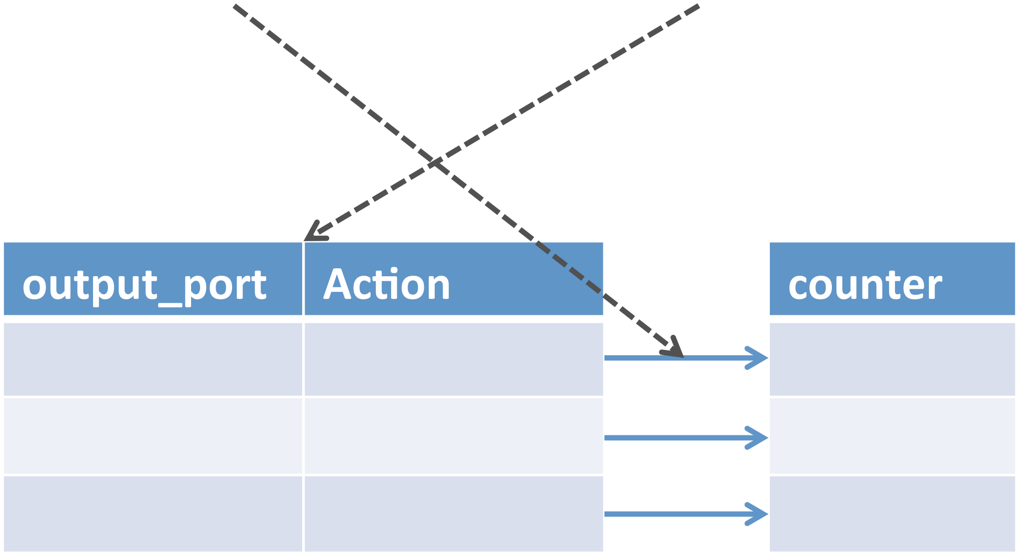
3. Pings sent and reply received

DEMO PART 3

Counters in the datapath

Adding counters to table entries

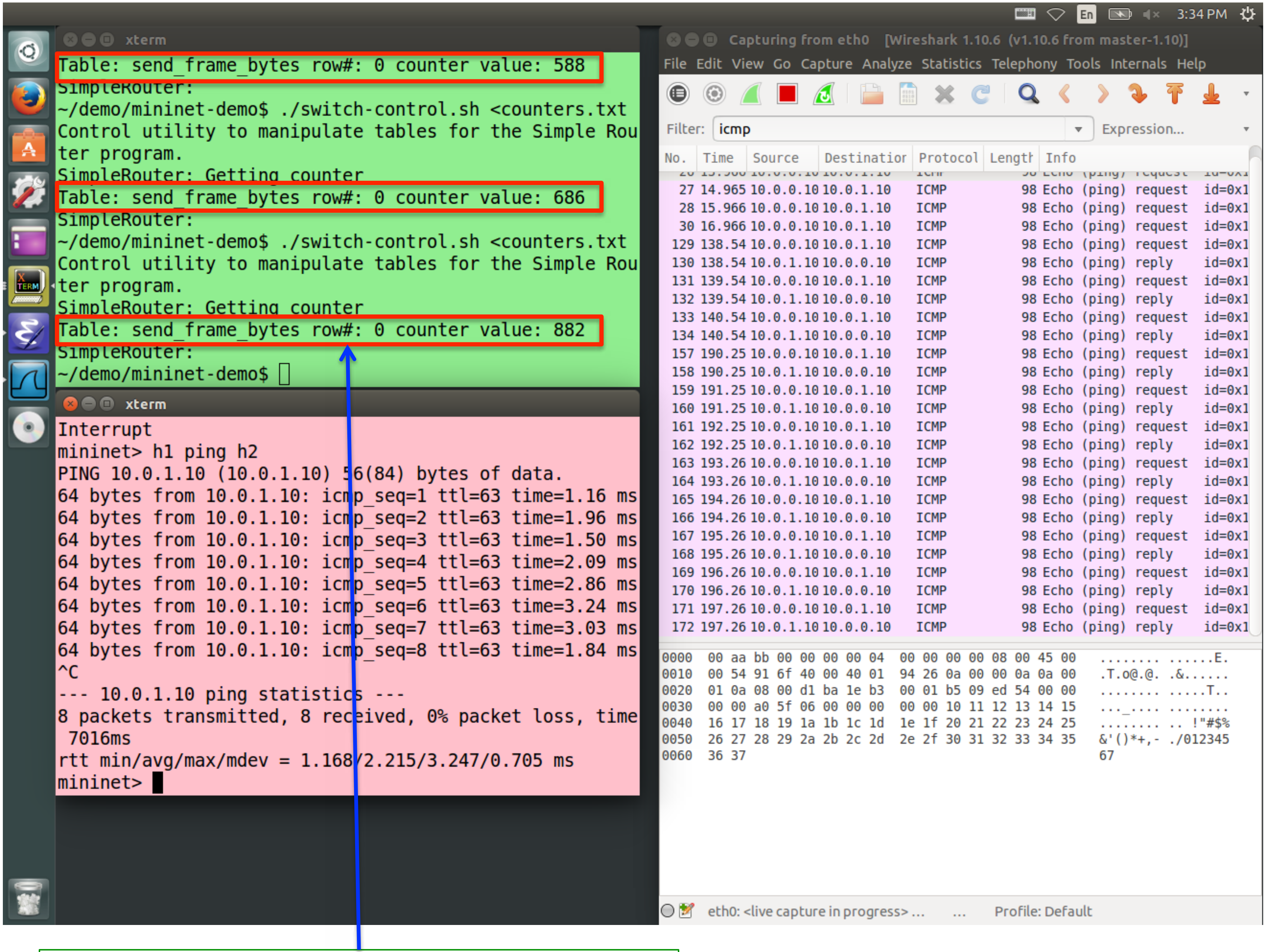
```
counter send_frame_bytes
{
   type : bytes;
   direct : send_frame;
}
```



output_port	Action

counter

Table send_frame


xterm
Capturing from eth0 [Wireshark 1.10.6 (v1.10.6 from master-1.10)]
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
3:34 PM
En

```
Table: send_frame_bytes row#: 0 counter value: 588
SimpleRouter:
~/demo/mininet-demo$ ./switch-control.sh <counters.txt
Control utility to manipulate tables for the Simple Rou
ter program.
SimpleRouter: Getting counter
Table: send_frame_bytes row#: 0 counter value: 686
SimpleRouter:
~/demo/mininet-demo$ ./switch-control.sh <counters.txt
Control utility to manipulate tables for the Simple Rou
ter program.
SimpleRouter: Getting counter
Table: send_frame_bytes row#: 0 counter value: 882
SimpleRouter:
~/demo/mininet-demo$
```

```
Interrupt
mininet> h1 ping h2
PING 10.0.1.10 (10.0.1.10) 56(84) bytes of data.
64 bytes from 10.0.1.10: icmp_seq=1 ttl=63 time=1.16 ms
64 bytes from 10.0.1.10: icmp_seq=2 ttl=63 time=1.96 ms
64 bytes from 10.0.1.10: icmp_seq=3 ttl=63 time=1.50 ms
64 bytes from 10.0.1.10: icmp_seq=4 ttl=63 time=2.09 ms
64 bytes from 10.0.1.10: icmp_seq=5 ttl=63 time=2.86 ms
64 bytes from 10.0.1.10: icmp_seq=6 ttl=63 time=3.24 ms
64 bytes from 10.0.1.10: icmp_seq=7 ttl=63 time=3.03 ms
64 bytes from 10.0.1.10: icmp_seq=8 ttl=63 time=1.84 ms
^C
--- 10.0.1.10 ping statistics ---
8 packets transmitted, 8 received, 0% packet loss, time
 7016ms
rtt min/avg/max/mdev = 1.168/2.215/3.247/0.705 ms
mininet>
```

Filter: icmp

```
No.  Time   Source    Destinatior Protocol Lengtl Info
 27 14.965 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
 28 15.966 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
 30 16.966 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
129 138.54 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
130 138.54 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
131 139.54 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
132 139.54 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
133 140.54 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
134 140.54 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
157 190.25 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
158 190.25 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
159 191.25 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
160 191.25 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
161 192.25 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
162 192.25 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
163 193.26 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
164 193.26 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
165 194.26 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
166 194.26 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
167 195.26 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
168 195.26 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
169 196.26 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
170 196.26 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
171 197.26 10.0.0.10 10.0.1.10   ICMP        98 Echo (ping) request  id=0x1
172 197.26 10.0.1.10 10.0.0.10   ICMP        98 Echo (ping) reply    id=0x1
```

```
0000  00 aa bb 00 00 00 00 04  00 00 00 00 08 00 45 00   ........ ......E.
0010  00 54 91 6f 40 00 40 01  94 26 0a 00 00 0a 0a 00   .T.o@.@. .&......
0020  01 0a 08 00 d1 ba 1e b3  00 01 b5 09 ed 54 00 00   ........ .....T..
0030  00 00 a0 5f 06 00 00 00  00 00 10 11 12 13 14 15   ..._.... ........
0040  16 17 18 19 1a 1b 1c 1d  1e 1f 20 21 22 23 24 25   ........ .. !"#$%
0050  26 27 28 29 2a 2b 2c 2d  2e 2f 30 31 32 33 34 35   &'()*+,- ./012345
0060  36 37                                              67
```

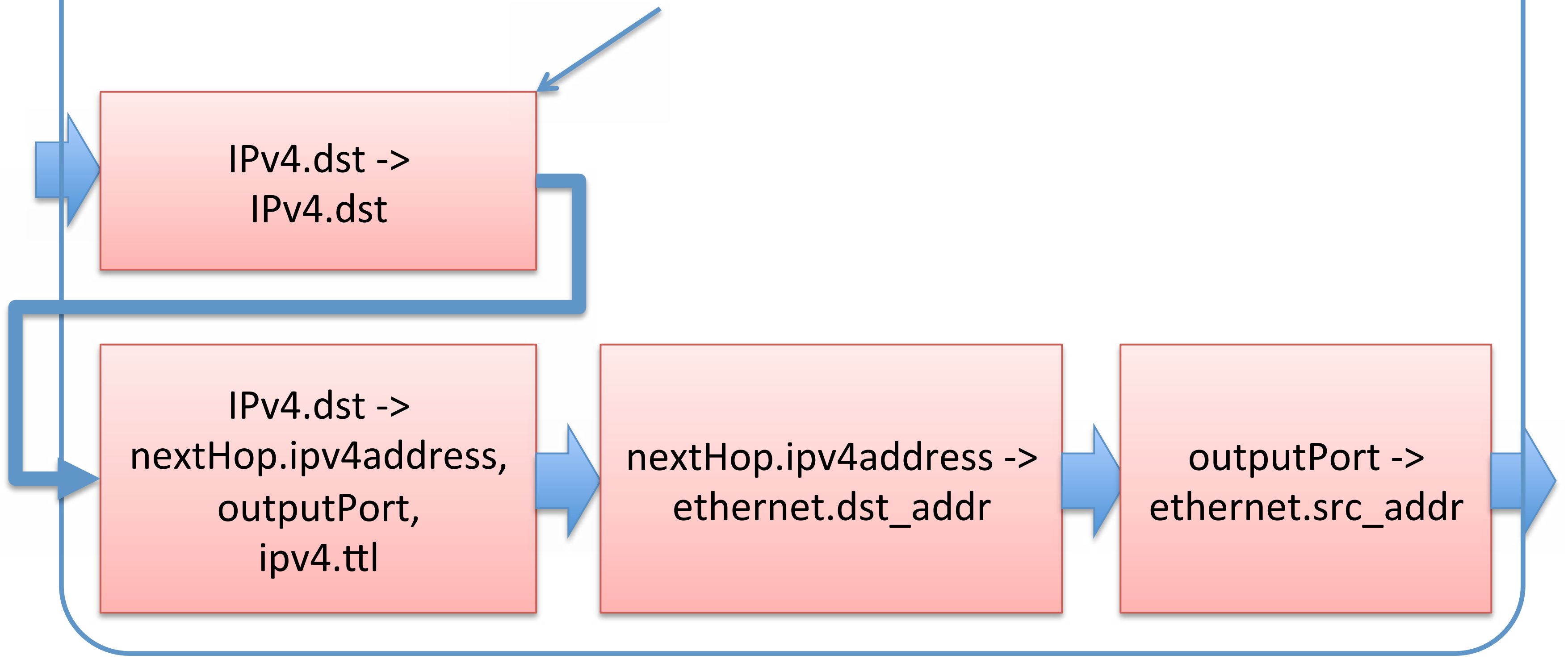
Reading byte counters while ping is running

DEMO PART 4

Divert traffic

Diverting traffic

Divert traffic for specific destinations to another destination



IPv4.dst -> IPv4.dst

IPv4.dst -> nextHop.ipv4address, outputPort, ipv4.ttl

nextHop.ipv4address -> ethernet.dst_addr

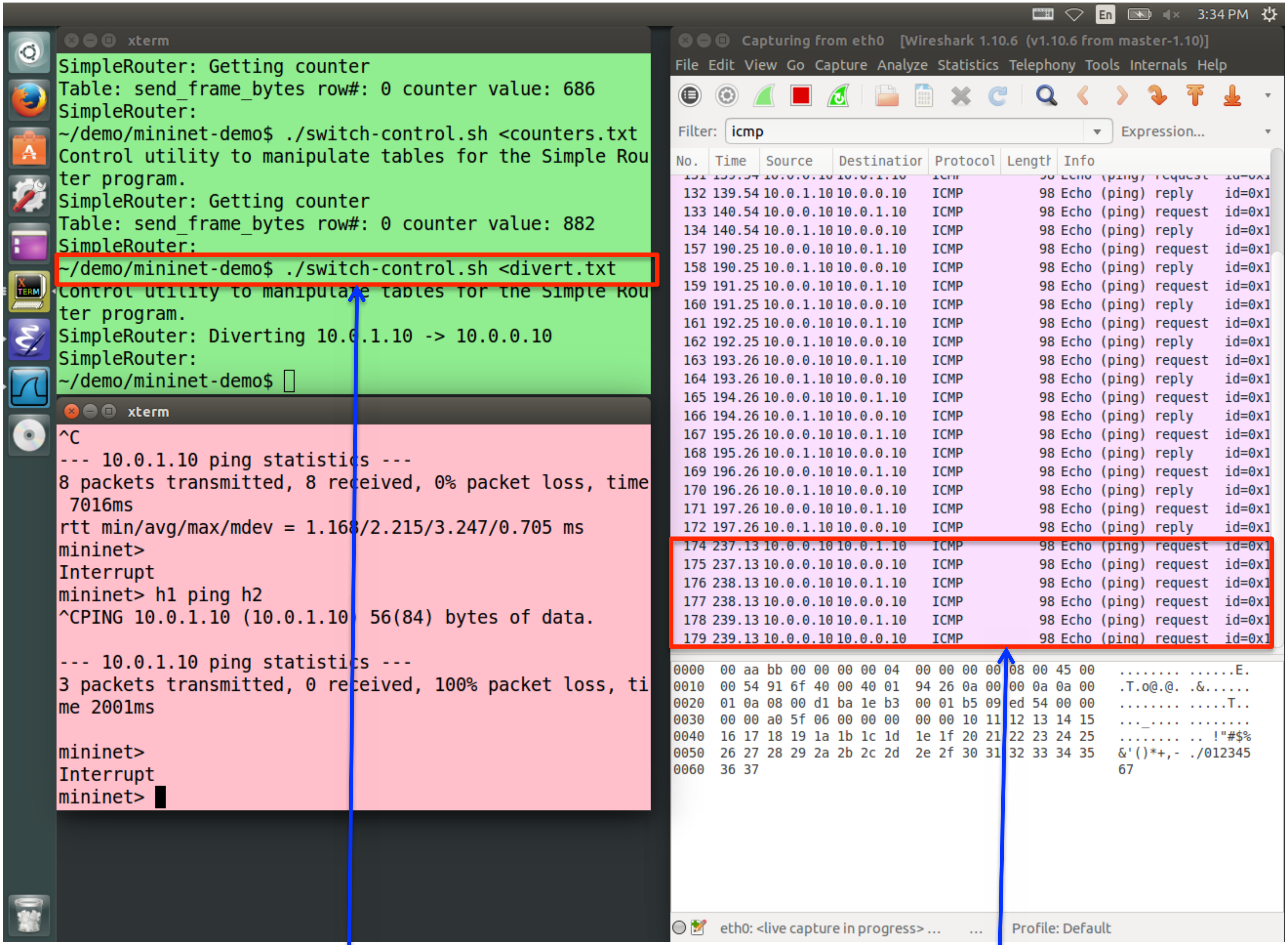
outputPort -> ethernet.src_addr

Modified pipeline

```
control ingress {
  apply(divert);
  apply(ipv4_match);
  apply(forward);
}

control egress {
  apply(send_frame);
}
```

Diverting traffic

```
action replaceIp(ipdest) {
  modify_field(ipv4.dstAddr, ipdest);
}
table divert {
  reads { ipv4.dstAddr: exact;}
  actions {
    replaceIp;
    nop;
  }
  size: 256;
}
```


3:34 PM
En
xterm
xterm
Capturing from eth0 [Wireshark 1.10.6 (v1.10.6 from master-1.10)]
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
Expression...
eth0: <live capture in progress> ...
Profile: Default

```
SimpleRouter: Getting counter
Table: send_frame_bytes row#: 0 counter value: 686
SimpleRouter:
~/demo/mininet-demo$ ./switch-control.sh <counters.txt
Control utility to manipulate tables for the Simple Rou
ter program.
SimpleRouter: Getting counter
Table: send_frame_bytes row#: 0 counter value: 882
SimpleRouter:
~/demo/mininet-demo$ ./switch-control.sh <divert.txt
Control utility to manipulate tables for the Simple Rou
ter program.
SimpleRouter: Diverting 10.0.1.10 -> 10.0.0.10
SimpleRouter:
~/demo/mininet-demo$ 
```

```
^C
--- 10.0.1.10 ping statistics ---
8 packets transmitted, 8 received, 0% packet loss, time
 7016ms
rtt min/avg/max/mdev = 1.168/2.215/3.247/0.705 ms
mininet>
Interrupt
mininet> h1 ping h2
^CPING 10.0.1.10 (10.0.1.10) 56(84) bytes of data.

--- 10.0.1.10 ping statistics ---
3 packets transmitted, 0 received, 100% packet loss, ti
me 2001ms

mininet>
Interrupt
mininet> 
```

Filter: icmp

```
No.  Time   Source    Destinatior Protocol Lengtl Info
131 139.54 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
132 139.54 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
133 140.54 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
134 140.54 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
157 190.25 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
158 190.25 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
159 191.25 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
160 191.25 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
161 192.25 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
162 192.25 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
163 193.26 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
164 193.26 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
165 194.26 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
166 194.26 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
167 195.26 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
168 195.26 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
169 196.26 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
170 196.26 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
171 197.26 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
172 197.26 10.0.1.10 10.0.0.10   ICMP       98 Echo (ping) reply    id=0x1
174 237.13 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
175 237.13 10.0.0.10 10.0.0.10   ICMP       98 Echo (ping) request  id=0x1
176 238.13 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
177 238.13 10.0.0.10 10.0.0.10   ICMP       98 Echo (ping) request  id=0x1
178 239.13 10.0.0.10 10.0.1.10   ICMP       98 Echo (ping) request  id=0x1
179 239.13 10.0.0.10 10.0.0.10   ICMP       98 Echo (ping) request  id=0x1
```

```
0000  00 aa bb 00 00 00 00 04  00 00 00 00 08 00 45 00   ........ ......E.
0010  00 54 91 6f 40 00 40 01  94 26 0a 00 00 0a 0a 00   .T.o@.@. .&......
0020  01 0a 08 00 d1 ba 1e b3  00 01 b5 09 ed 54 00 00   ........ .....T..
0030  00 00 a0 5f 06 00 00 00  00 00 10 11 12 13 14 15   ..._.... ........
0040  16 17 18 19 1a 1b 1c 1d  1e 1f 20 21 22 23 24 25   ........ .. !"#$%
0050  26 27 28 29 2a 2b 2c 2d  2e 2f 30 31 32 33 34 35   &'()*+,- ./012345
0060  36 37                                              67
```

1. Inserted table entries to divert packets

2. Pings sent and received by same host

Availability

All this code will be available
as FOSS
before March 31, 2015
on http://p4.org