

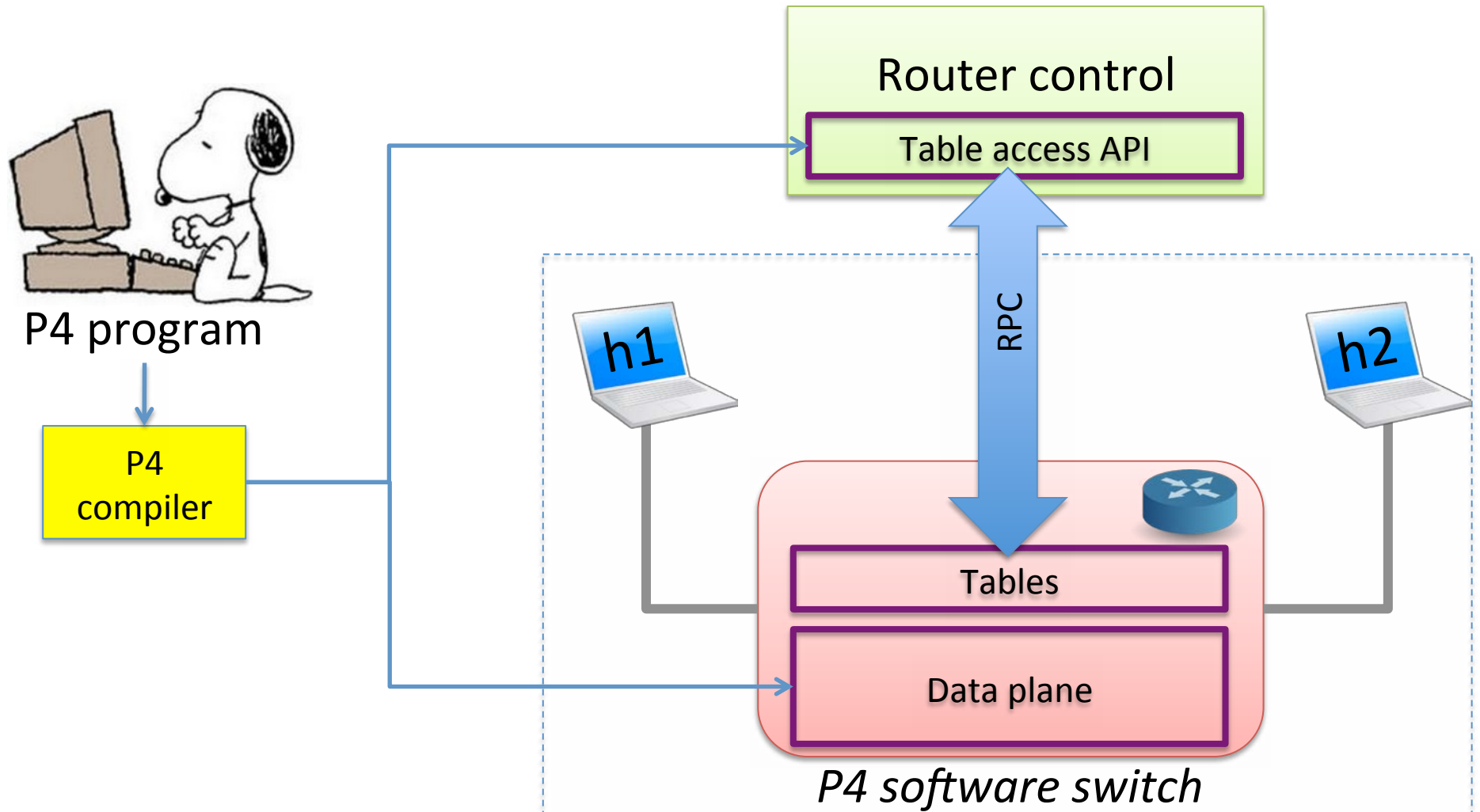
P4 demo:  
a basic L2/L3 switch in 170 LOC

netdev0.1

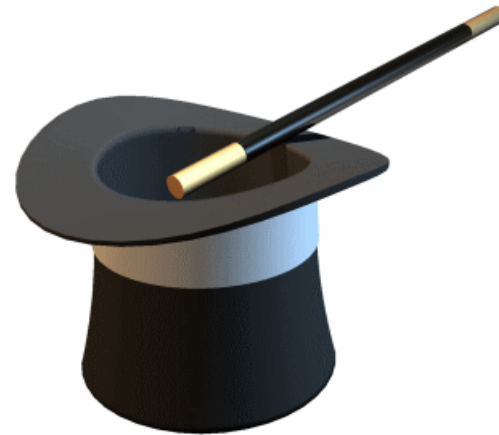
Ottawa, February 15, 2015

Mihai Budiu

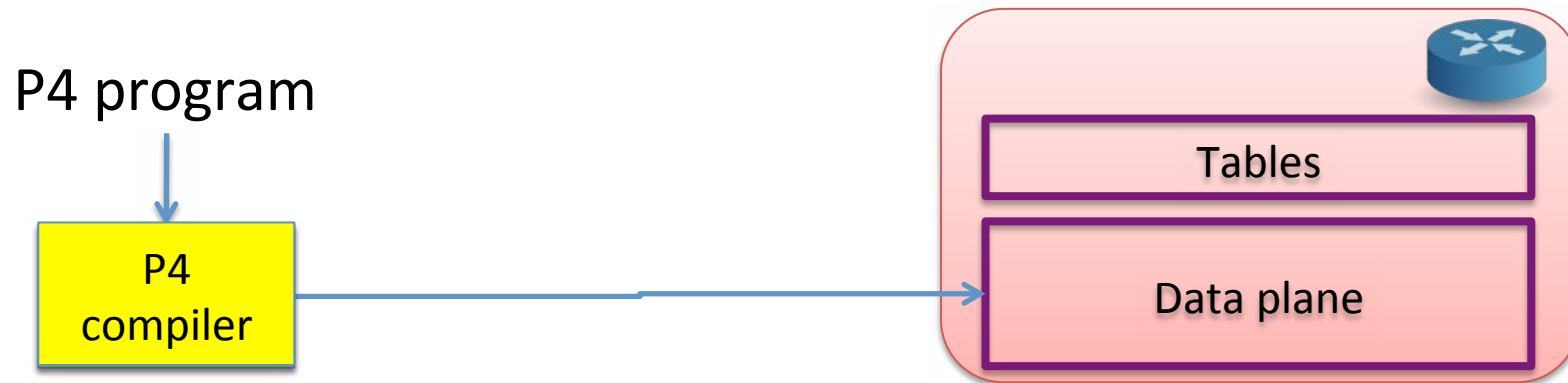
# Setup



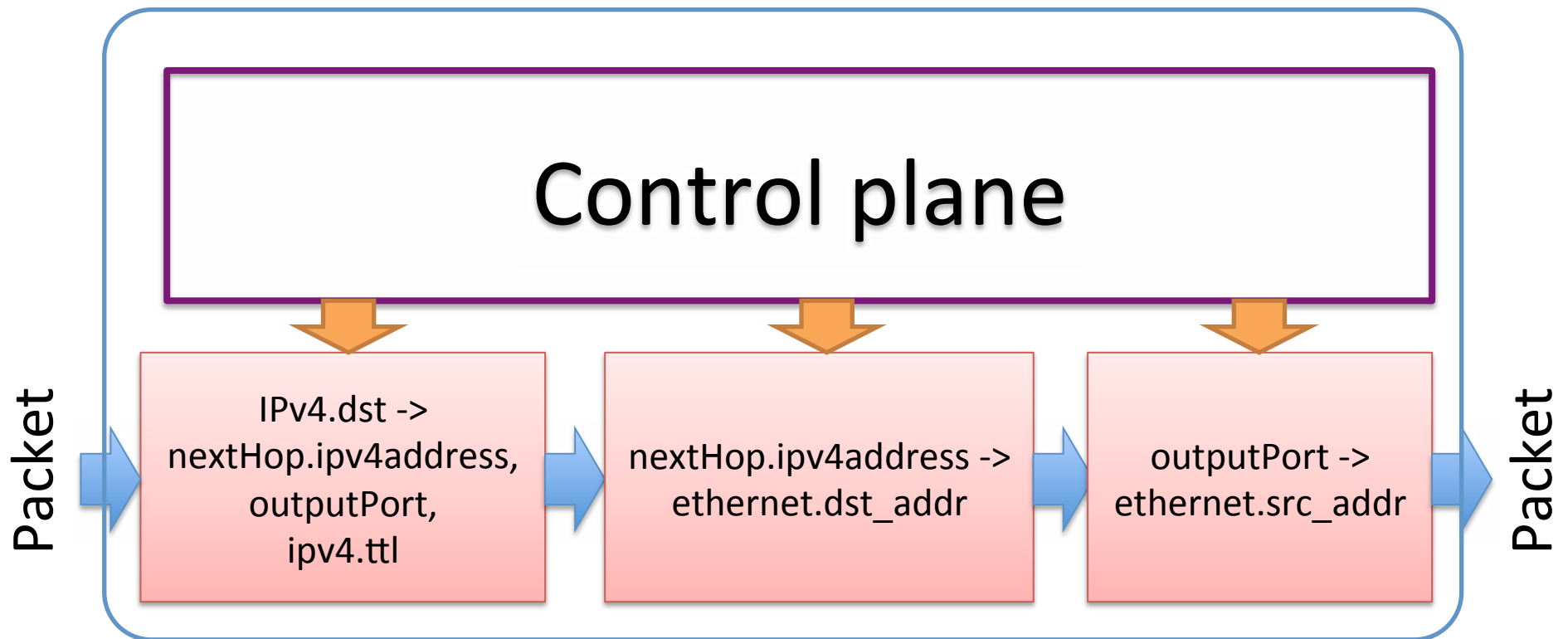
# DEMO PART 1



## Creating a basic ethernet+IPv4 switch from a P4 program

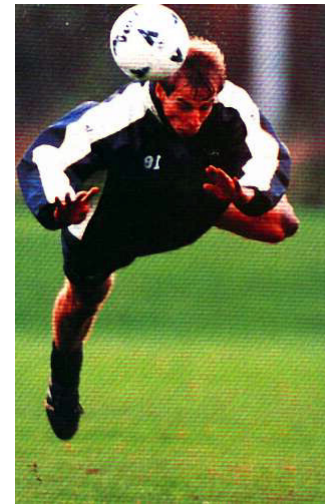


# Layer 3 packet forwarding



# Headers

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

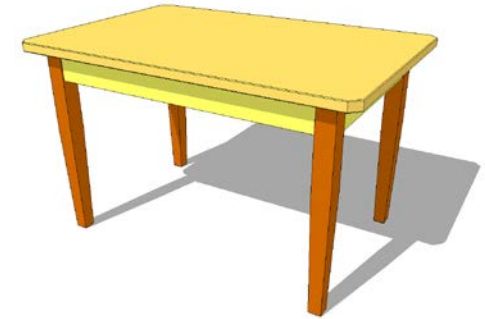


```
header_type ipv4_t { ... }  
// no options
```

# Parser

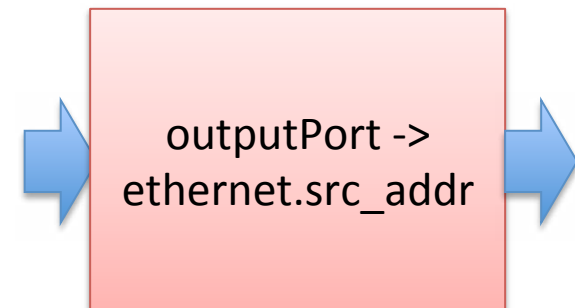
```
parser parse_ethernet {  
    extract(ethernet);  
    return select(latest.etherType) {  
        0x0800 : parse_ipv4;  
        default: ingress;  
    }  
}  
  
...  
  
parser parse_ipv4 { ... }  
  
calculated_field ipv4.hdrChecksum {  
    verify ipv4_checksum;  
    update ipv4_checksum;  
}
```

# 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;  
}
```



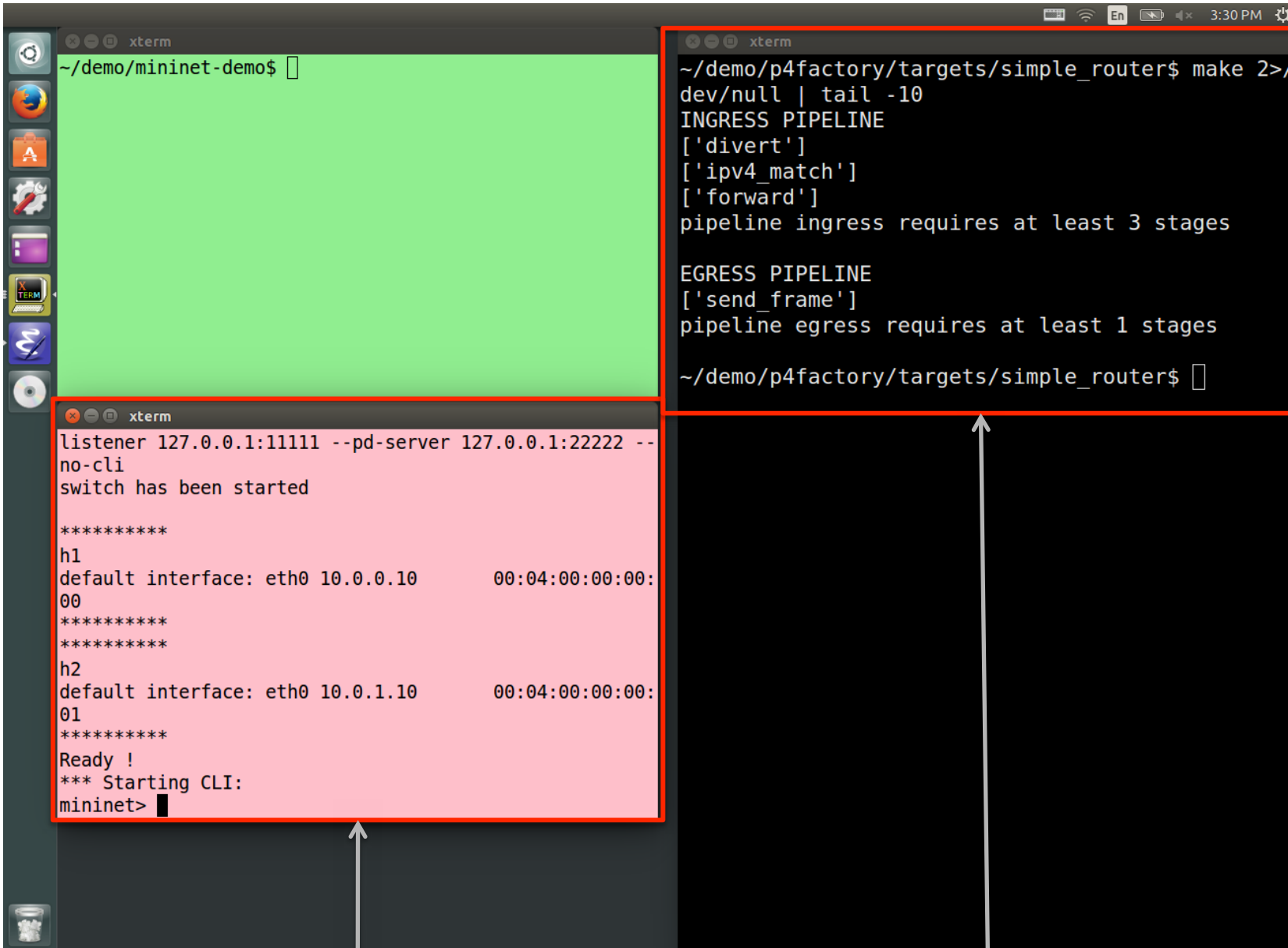
# Complete pipeline

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

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







2. Started switch running

1. Compiled a new switch from P4

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

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

** (wireshark:7832): WARNING **: Couldn't connect to accessibility 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, time 3000ms

mininet>
Interrupt
mininet>
```

```
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...

No. Time Source Destination Protocol Length Info
---
26 13.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
27 14.965 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
28 15.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
30 16.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1

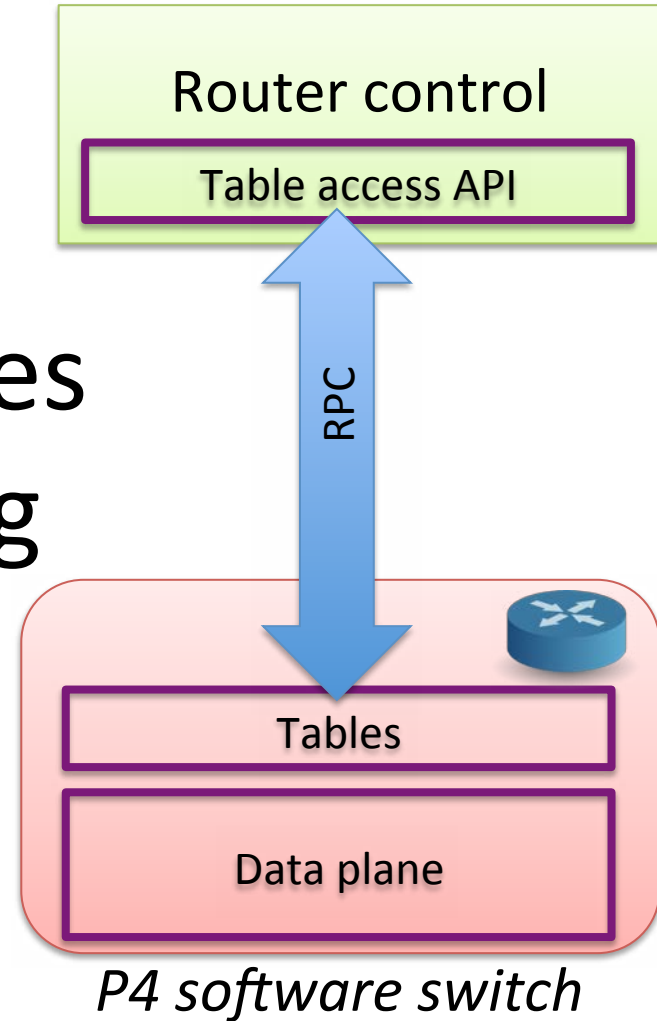
0000 00 aa bb 00 00 00 04 00 00 00 08 00 45 00 .....E.
0010 00 54 91 6f 40 00 01 94 26 0a 00 00 0a 0a 00 .T.o@.@. &.....
0020 01 0a 08 00 d1 ba 1e b3 00 01 b5 09 ed 04 00 00 .....T..
0030 00 00 a0 5f 06 00 00 00 00 10 11 12 13 14 15 .....
0040 16 17 18 19 1a 1b 1c 1d 1e 1f 20 21 22 23 24 25 .....!#$%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

1. Ping from h1 to h2 does not work

2. Pings sent but no reply returned

# DEMO PART 2

Populate switch tables to enable forwarding



```
xterm
~/demo/mininet-demo$ ./switch-control.sh <commands.txt
Control utility to manipulate tables for the Simple Router 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$

xterm
--- 10.0.1.10 ping statistics ---
4 packets transmitted, 0 received, 100% packet loss, time 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>
```

```
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...

No. Time Source Destination Protocol Length Info
26 13.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
27 14.965 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
28 15.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
30 16.966 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
129 138.54 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
130 138.54 10.0.1.10 10.0.0.10 ICMP 98 Echo (ping) reply id=0x1e1
131 139.54 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
132 139.54 10.0.1.10 10.0.0.10 ICMP 98 Echo (ping) reply id=0x1e1
133 140.54 10.0.0.10 10.0.1.10 ICMP 98 Echo (ping) request id=0x1e1
134 140.54 10.0.1.10 10.0.0.10 ICMP 98 Echo (ping) reply id=0x1e1

0000 00 aa bb 00 00 00 04 00 00 00 00 00 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 e0 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;
}
```

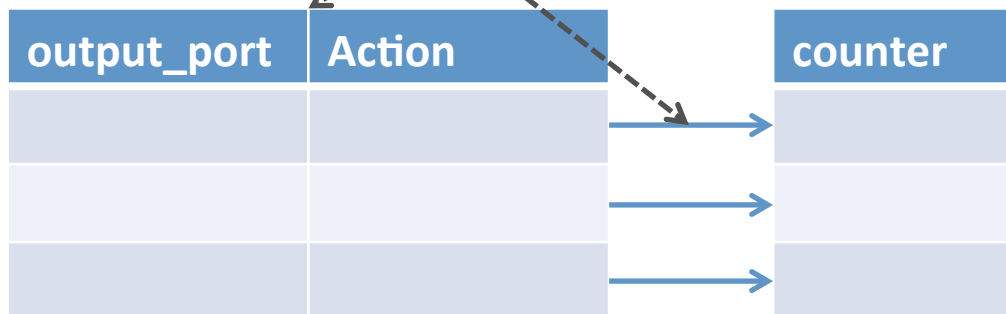


Table send\_frame

```
xterm
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 Router 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 Router 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>
```

```
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
No. Time Source Destination Protocol Length 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

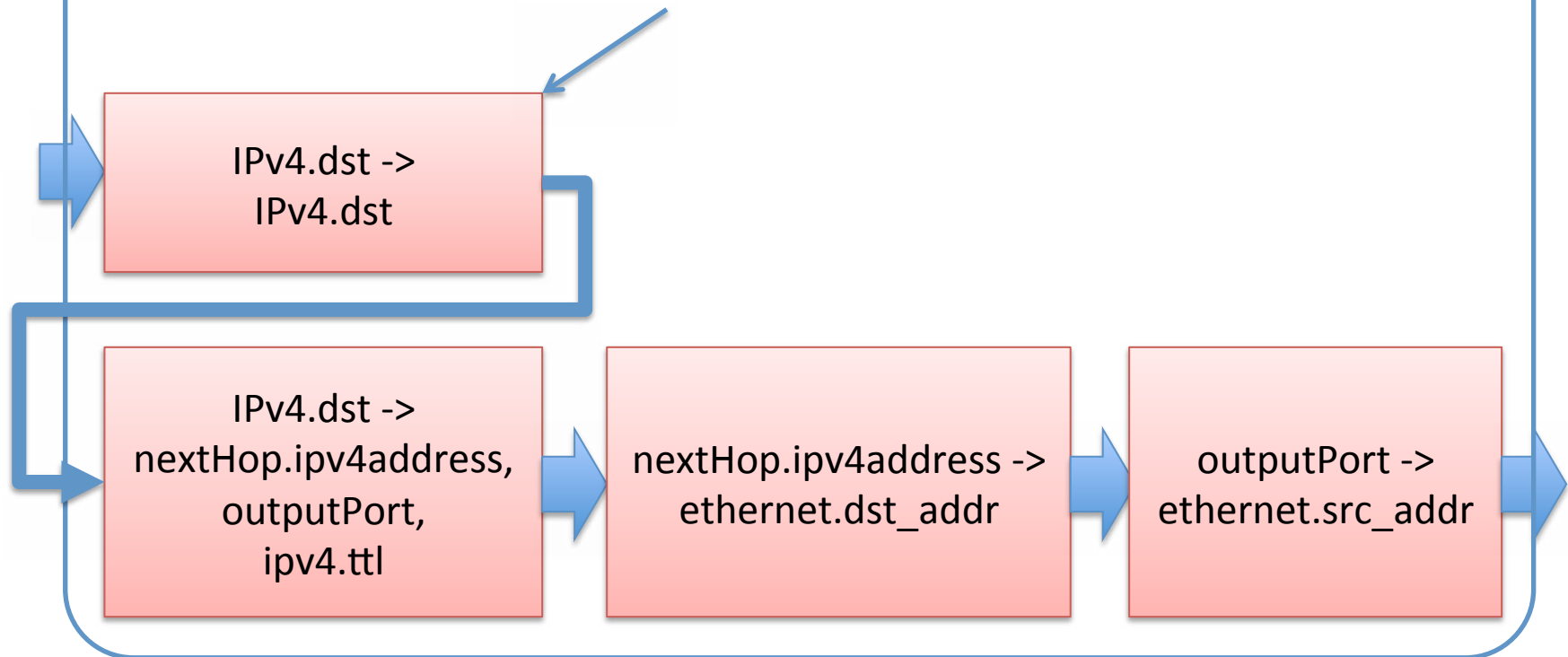
Divert traffic





# Diverting traffic

*Divert traffic for specific destinations to another destination*



# 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;
}
```

```

xterm
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 Router 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 Router program.
SimpleRouter: Diverting 10.0.1.10 -> 10.0.0.10
SimpleRouter:
~/demo/mininet-demo$

xterm
^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, time
2001ms

mininet>
Interrupt
mininet>

```

```

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...
No. Time Source Destination Protocol Length Info
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 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. 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>

