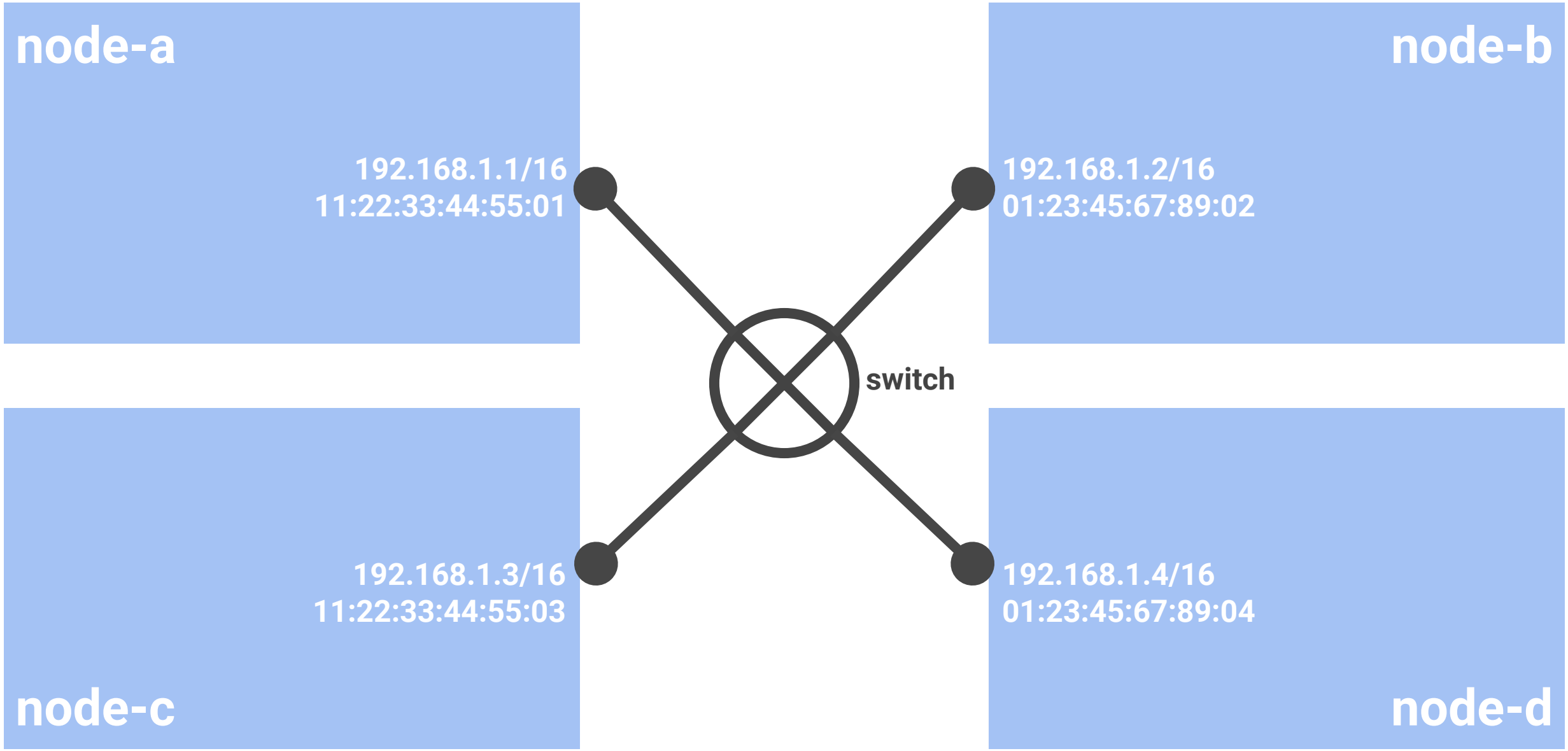
An Illustrated Guide to Kubernetes Networking

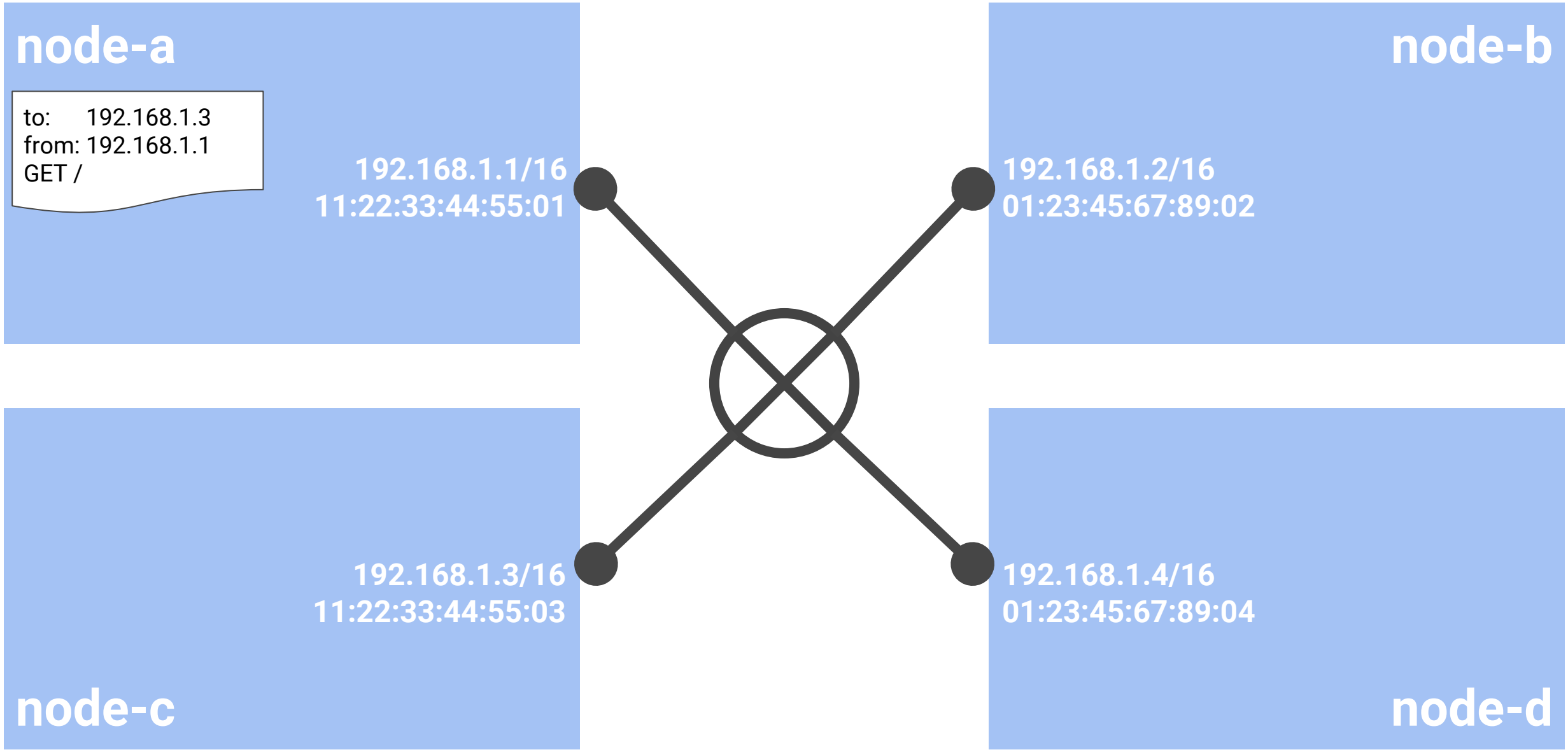
Tim Hockin <thockin@google.com>
Senior Staff Software Engineer
@thockin

Layer 2: ethernet

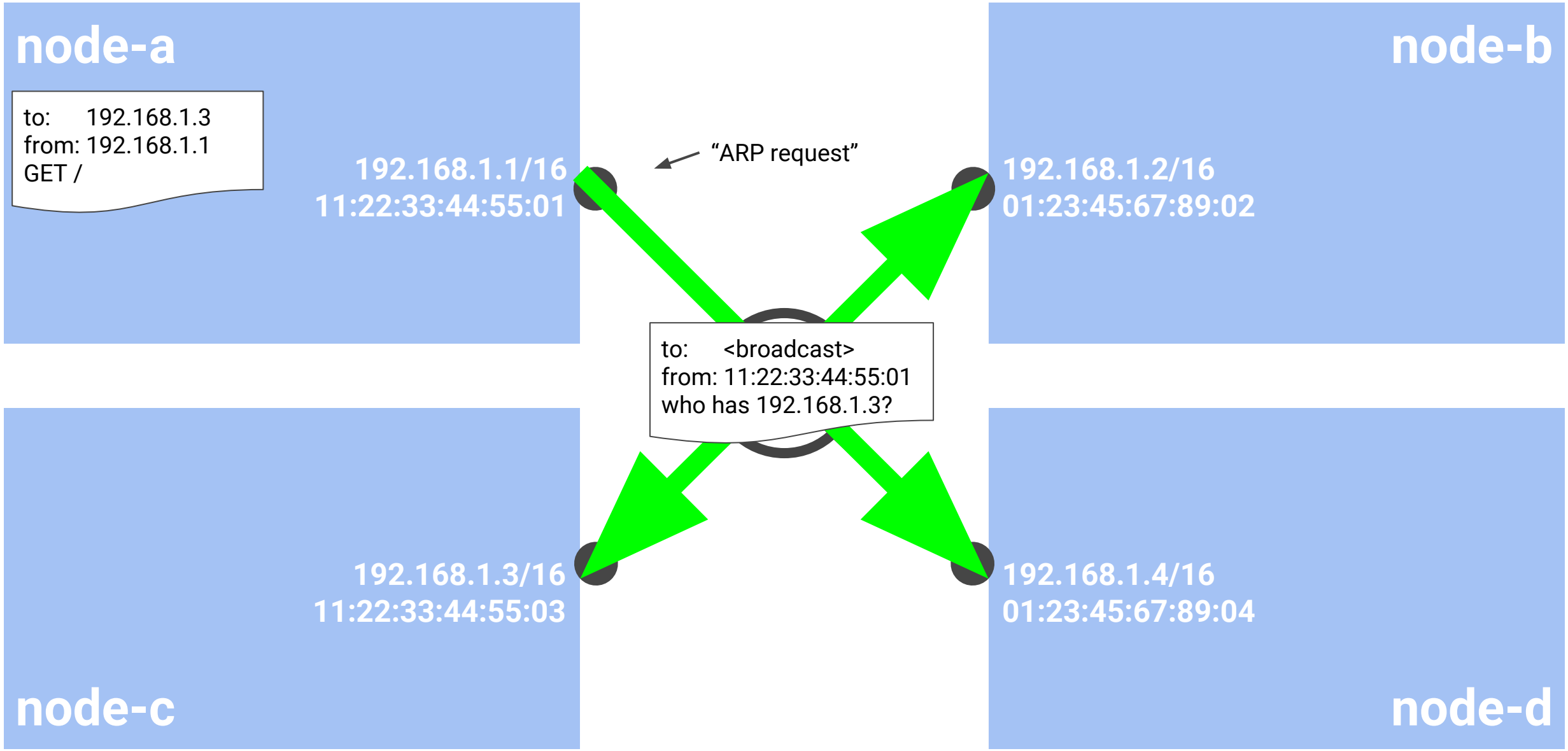
L2


node-a
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
switch
node-c
192.168.1.3/16
11:22:33:44:55:03
node-d
192.168.1.4/16
01:23:45:67:89:04

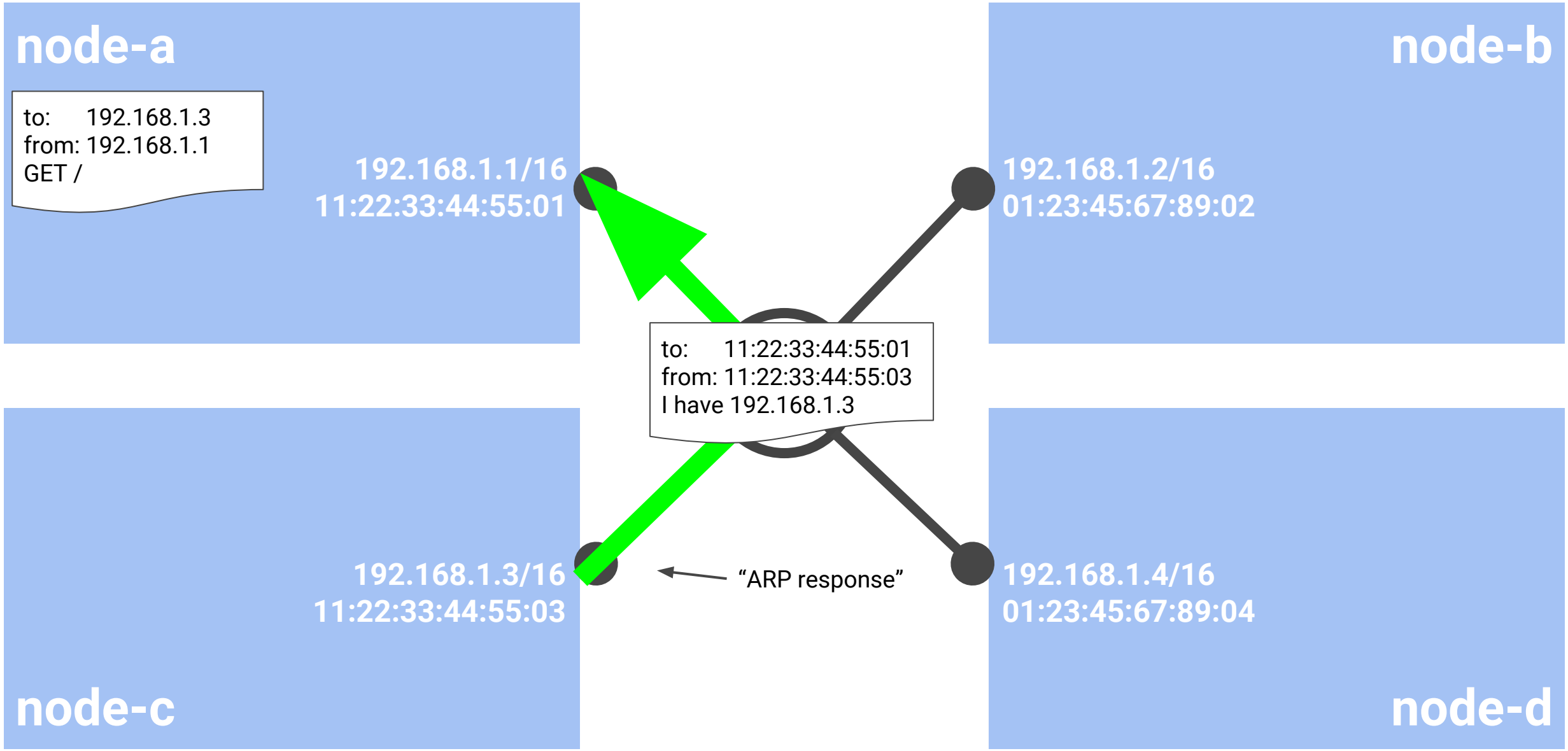
L2


node-a
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
192.168.1.3/16
11:22:33:44:55:03
node-c
192.168.1.4/16
01:23:45:67:89:04
node-d

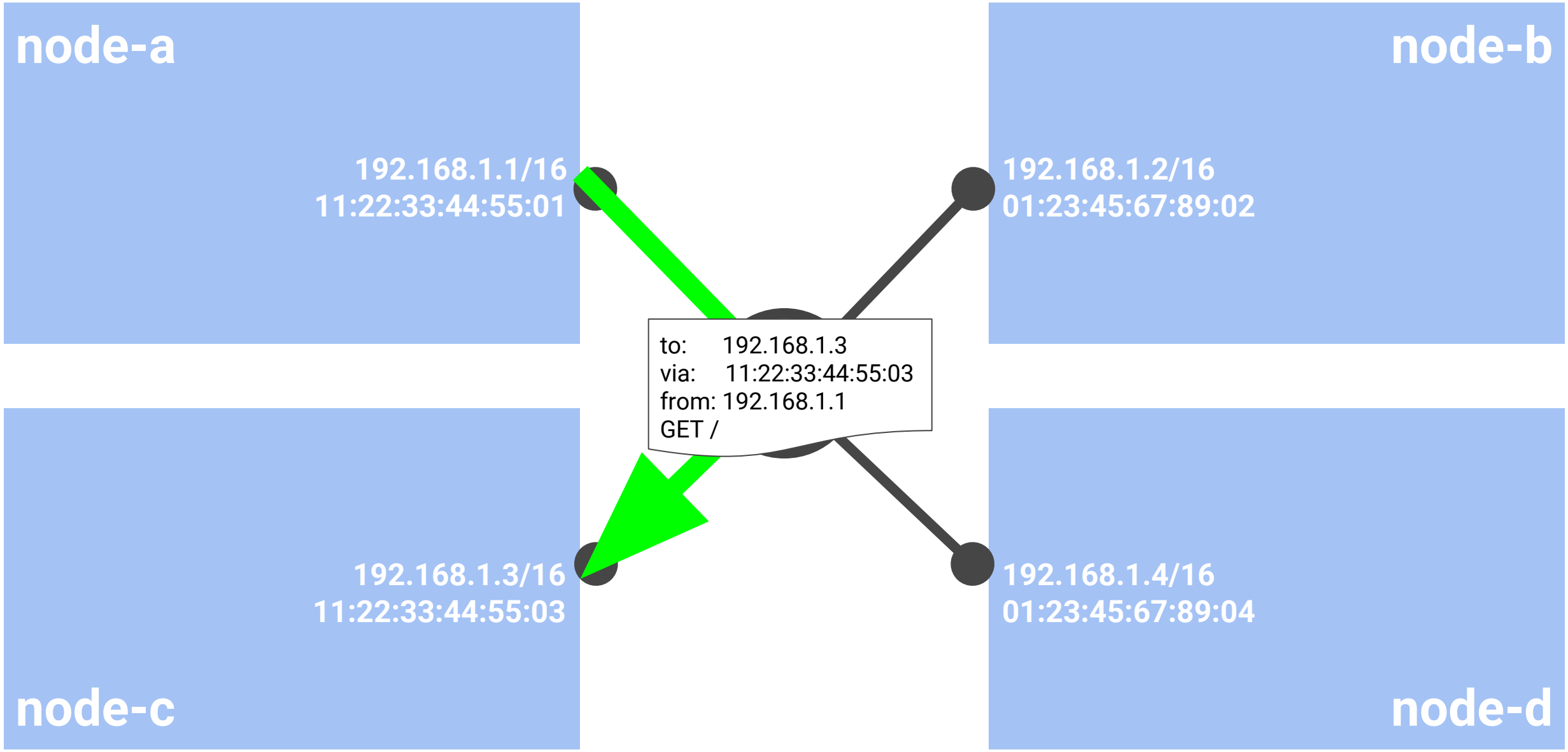
L2


node-a
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.1/16
11:22:33:44:55:01
"ARP request"
node-b
192.168.1.2/16
01:23:45:67:89:02
to: <broadcast>
from: 11:22:33:44:55:01
who has 192.168.1.3?
192.168.1.3/16
11:22:33:44:55:03
node-c
192.168.1.4/16
01:23:45:67:89:04
node-d

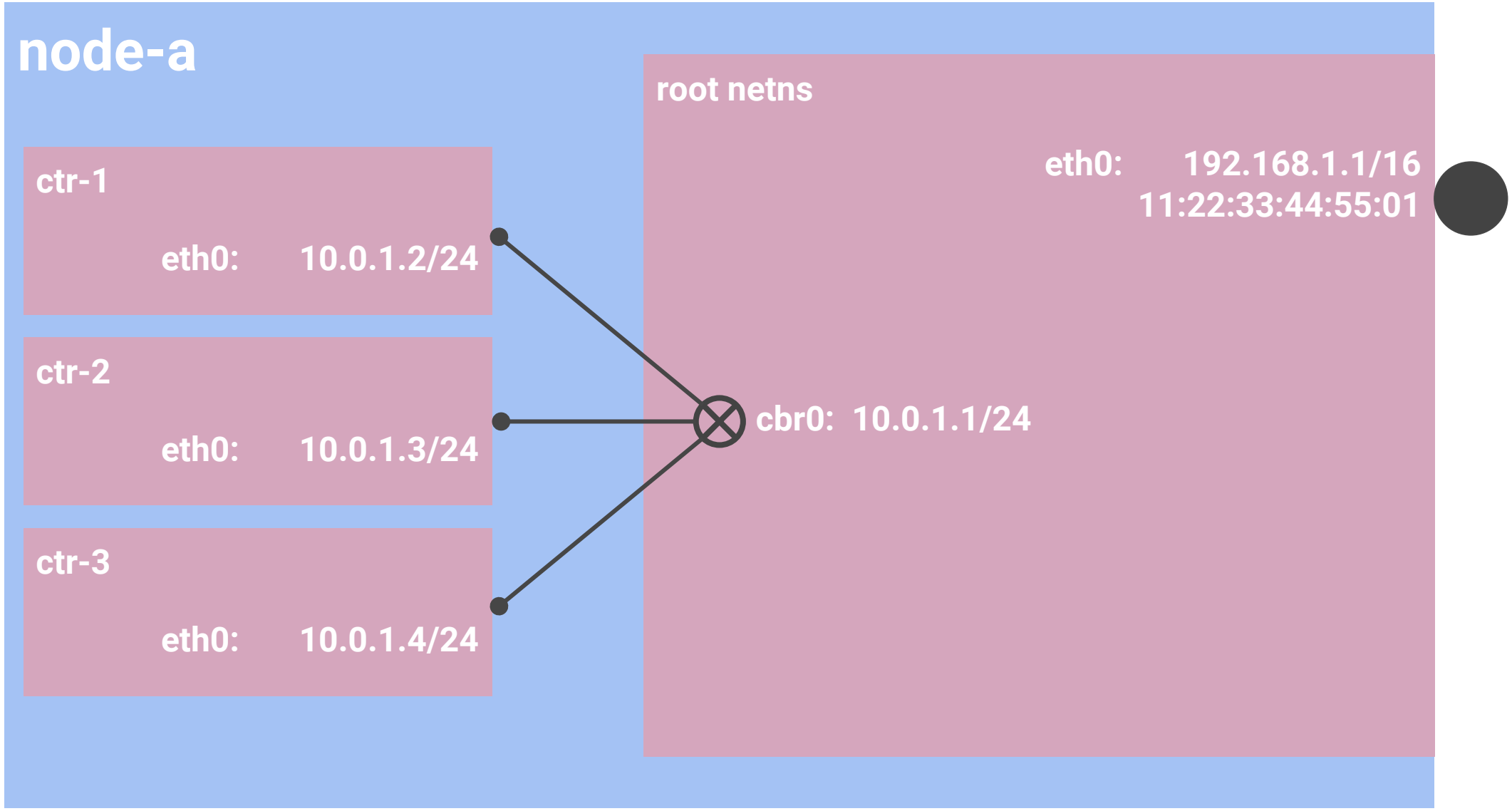
L2


node-a
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
to: 11:22:33:44:55:01
from: 11:22:33:44:55:03
I have 192.168.1.3
192.168.1.3/16
11:22:33:44:55:03
"ARP response"
192.168.1.4/16
01:23:45:67:89:04
node-c
node-d

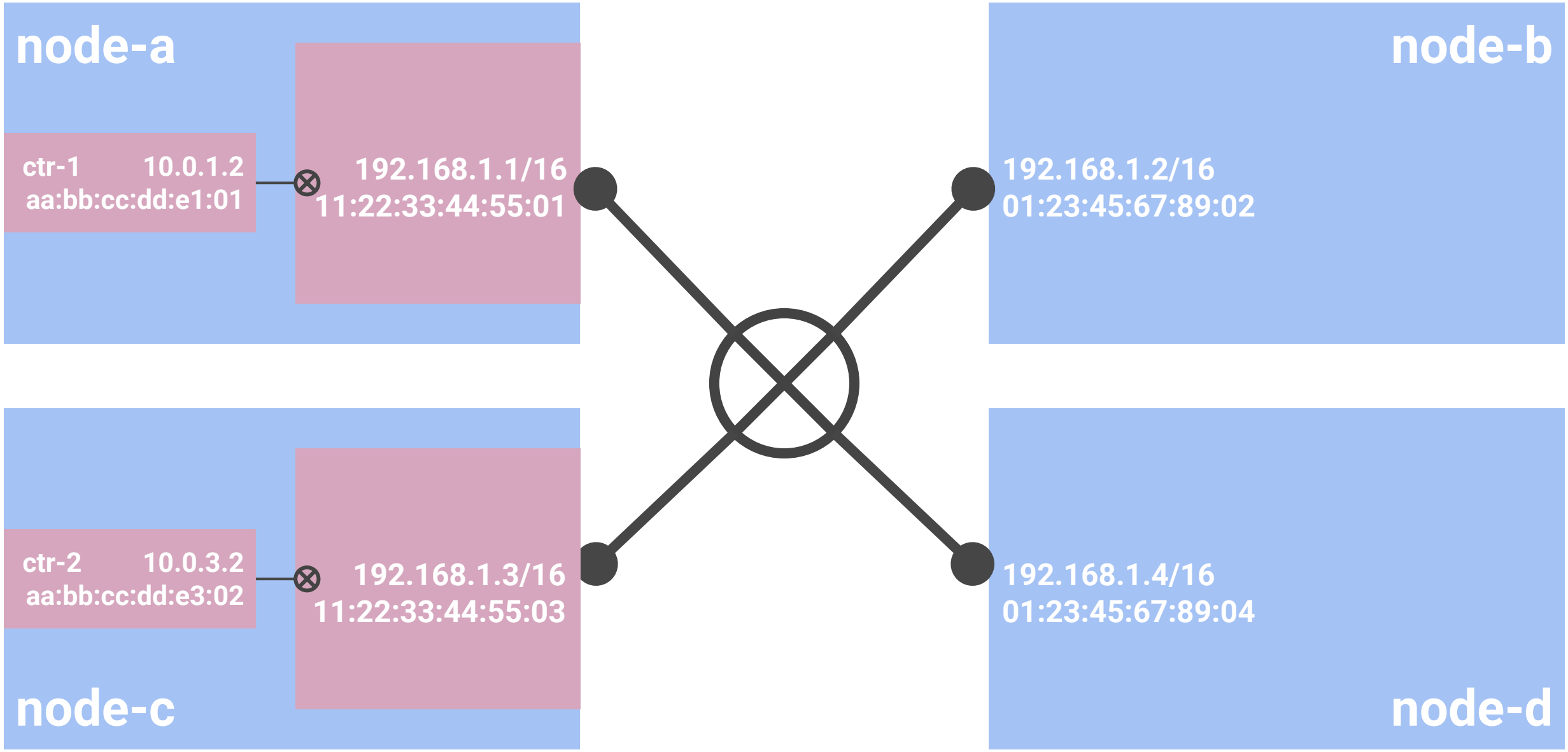
L2


node-a
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
to: 192.168.1.3
via: 11:22:33:44:55:03
from: 192.168.1.1
GET /
192.168.1.3/16
11:22:33:44:55:03
node-c
192.168.1.4/16
01:23:45:67:89:04
node-d

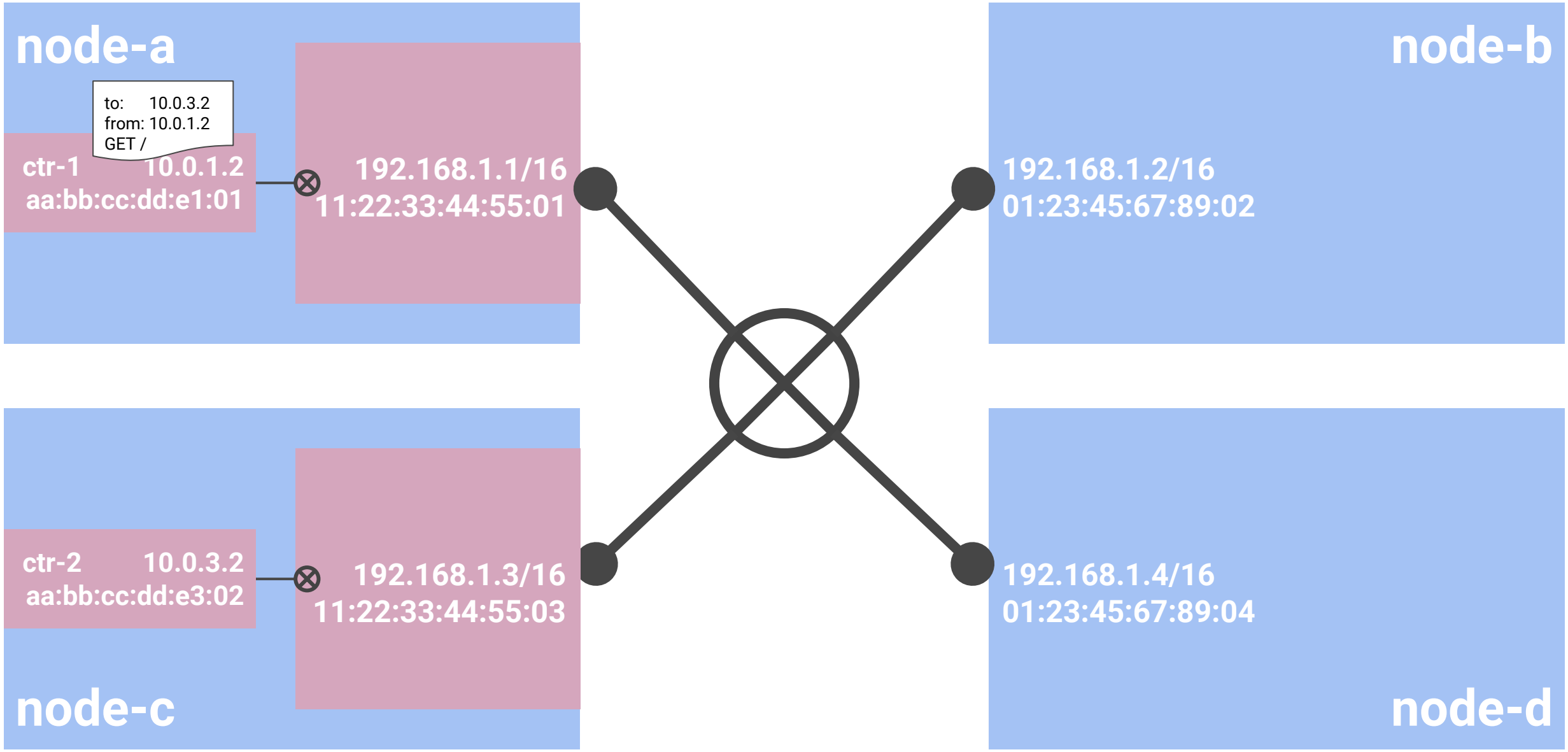
L2 with containers



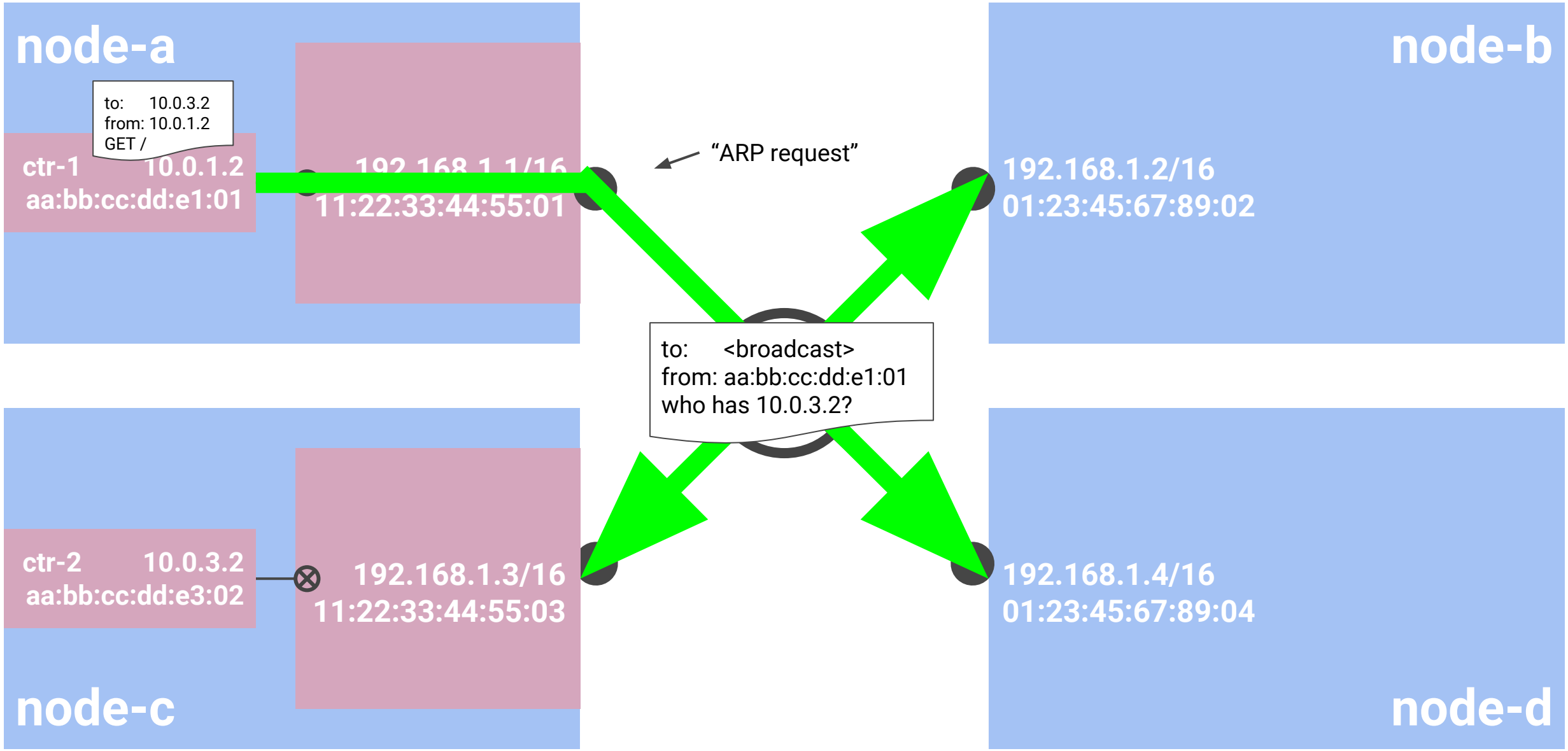
L2


node-a
ctr-1 10.0.1.2
aa:bb:cc:dd:e1:01
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
ctr-2 10.0.3.2
aa:bb:cc:dd:e3:02
192.168.1.3/16
11:22:33:44:55:03
node-c
192.168.1.4/16
01:23:45:67:89:04
node-d

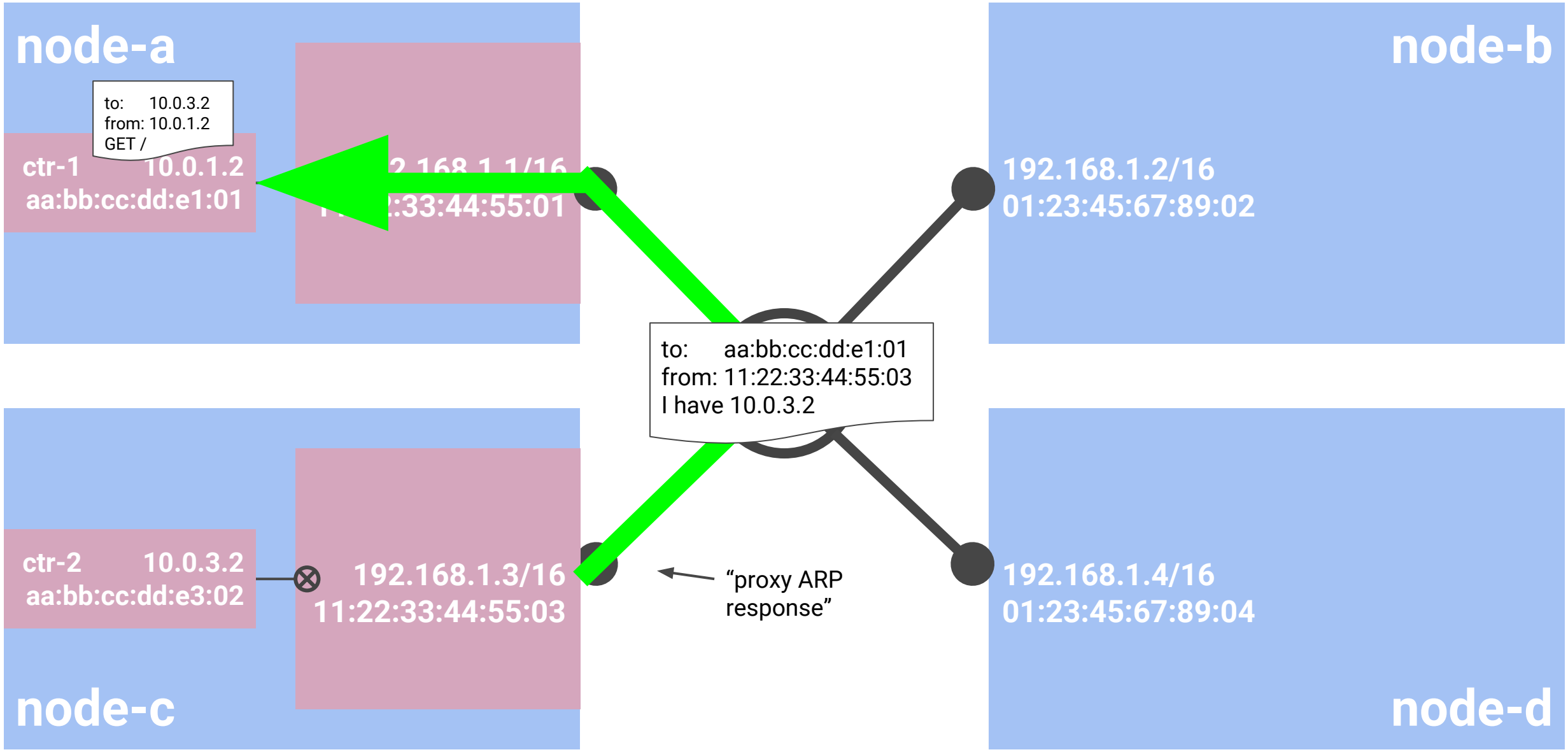
L2

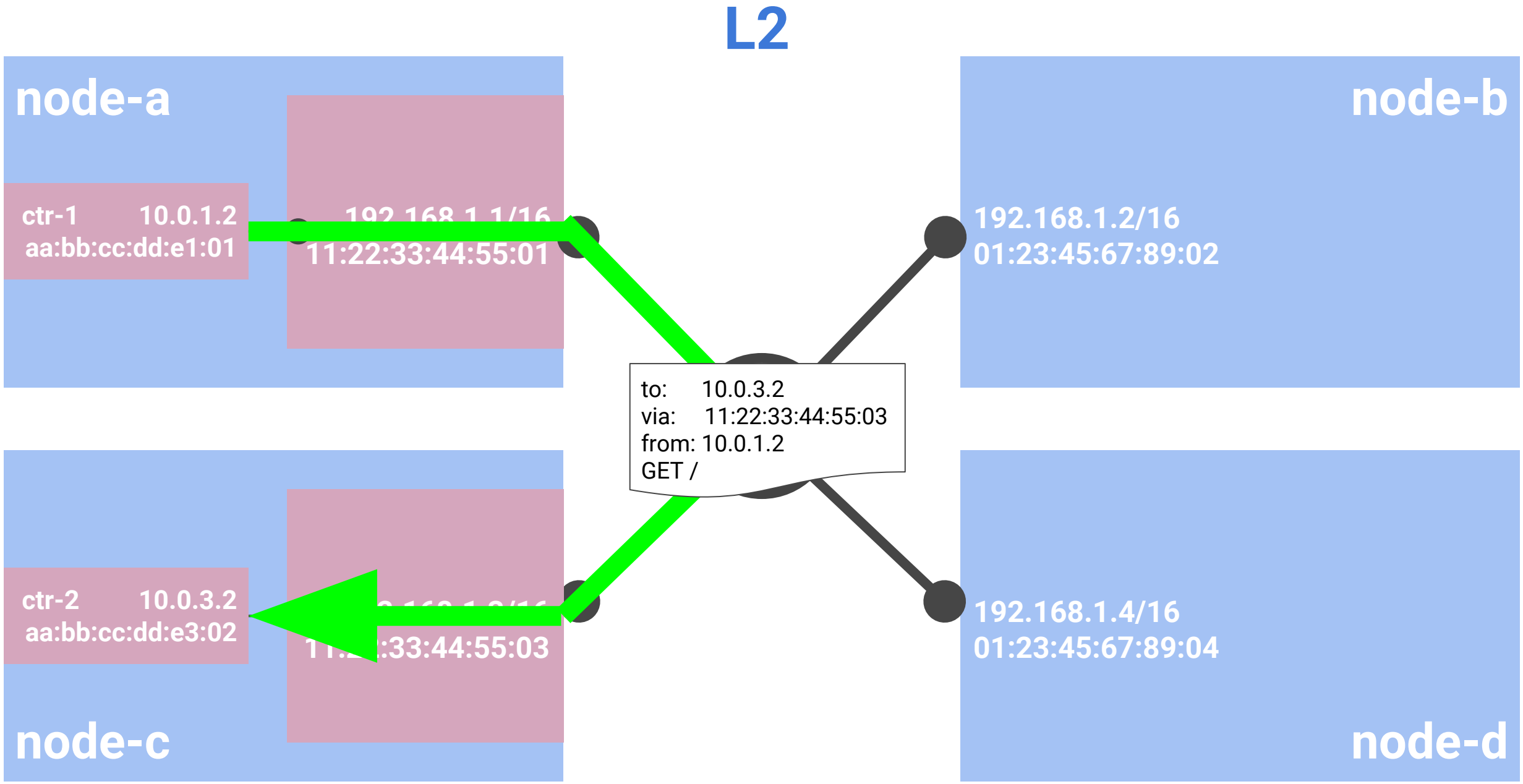

node-a
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-1 10.0.1.2
aa:bb:cc:dd:e1:01
192.168.1.1/16
11:22:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
ctr-2 10.0.3.2
aa:bb:cc:dd:e3:02
192.168.1.3/16
11:22:33:44:55:03
node-c
192.168.1.4/16
01:23:45:67:89:04
node-d

L2


node-a
node-b
node-c
node-d
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-1 10.0.1.2
aa:bb:cc:dd:e1:01
192.168.1.1/16
11:22:33:44:55:01
"ARP request"
192.168.1.2/16
01:23:45:67:89:02
to: <broadcast>
from: aa:bb:cc:dd:e1:01
who has 10.0.3.2?
ctr-2 10.0.3.2
aa:bb:cc:dd:e3:02
192.168.1.3/16
11:22:33:44:55:03
192.168.1.4/16
01:23:45:67:89:04

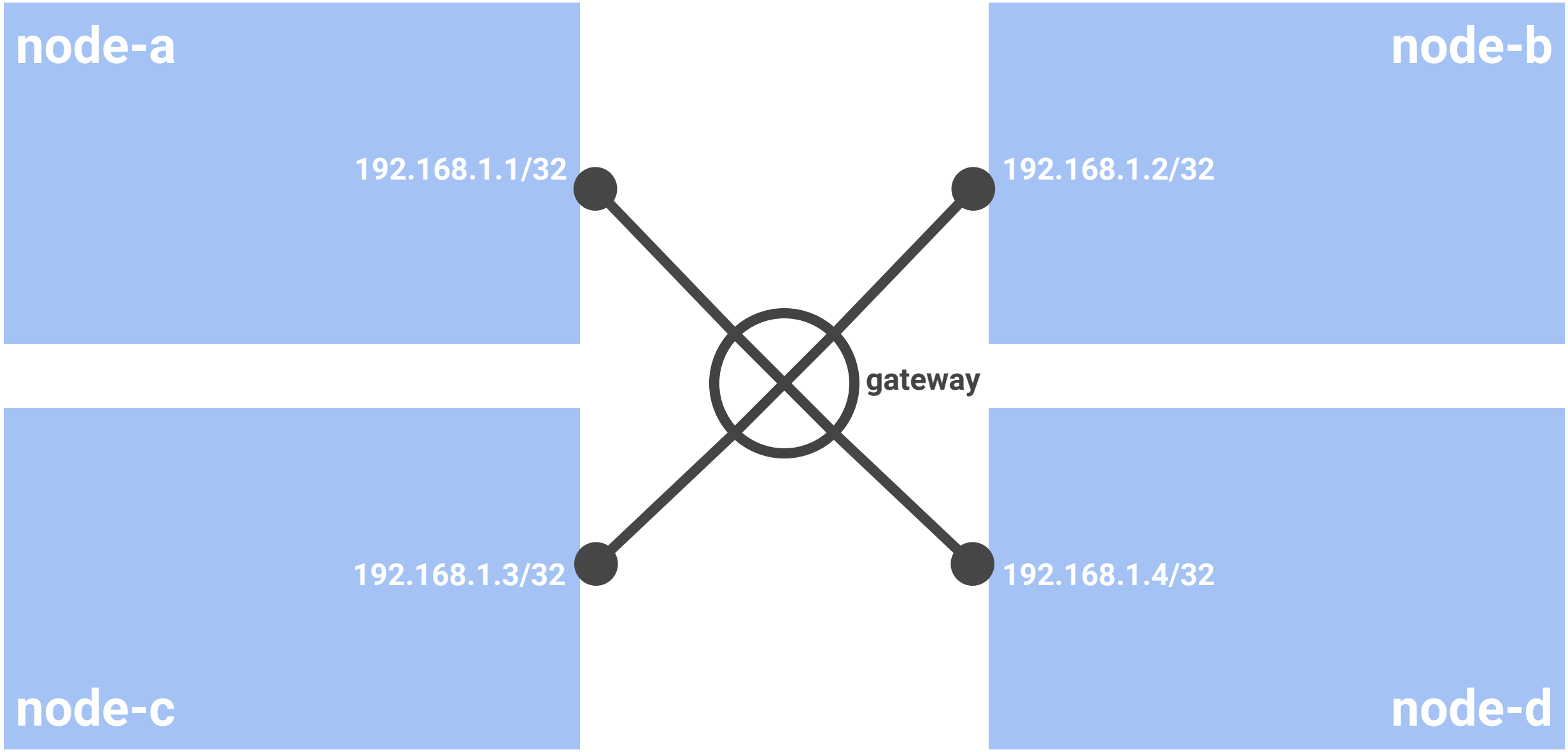
L2


node-a
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-1 10.0.1.2
aa:bb:cc:dd:e1:01
:33:44:55:01
node-b
192.168.1.2/16
01:23:45:67:89:02
to: aa:bb:cc:dd:e1:01
from: 11:22:33:44:55:03
I have 10.0.3.2
ctr-2 10.0.3.2
aa:bb:cc:dd:e3:02
192.168.1.3/16
11:22:33:44:55:03
"proxy ARP response"
192.168.1.4/16
01:23:45:67:89:04
node-c
node-d

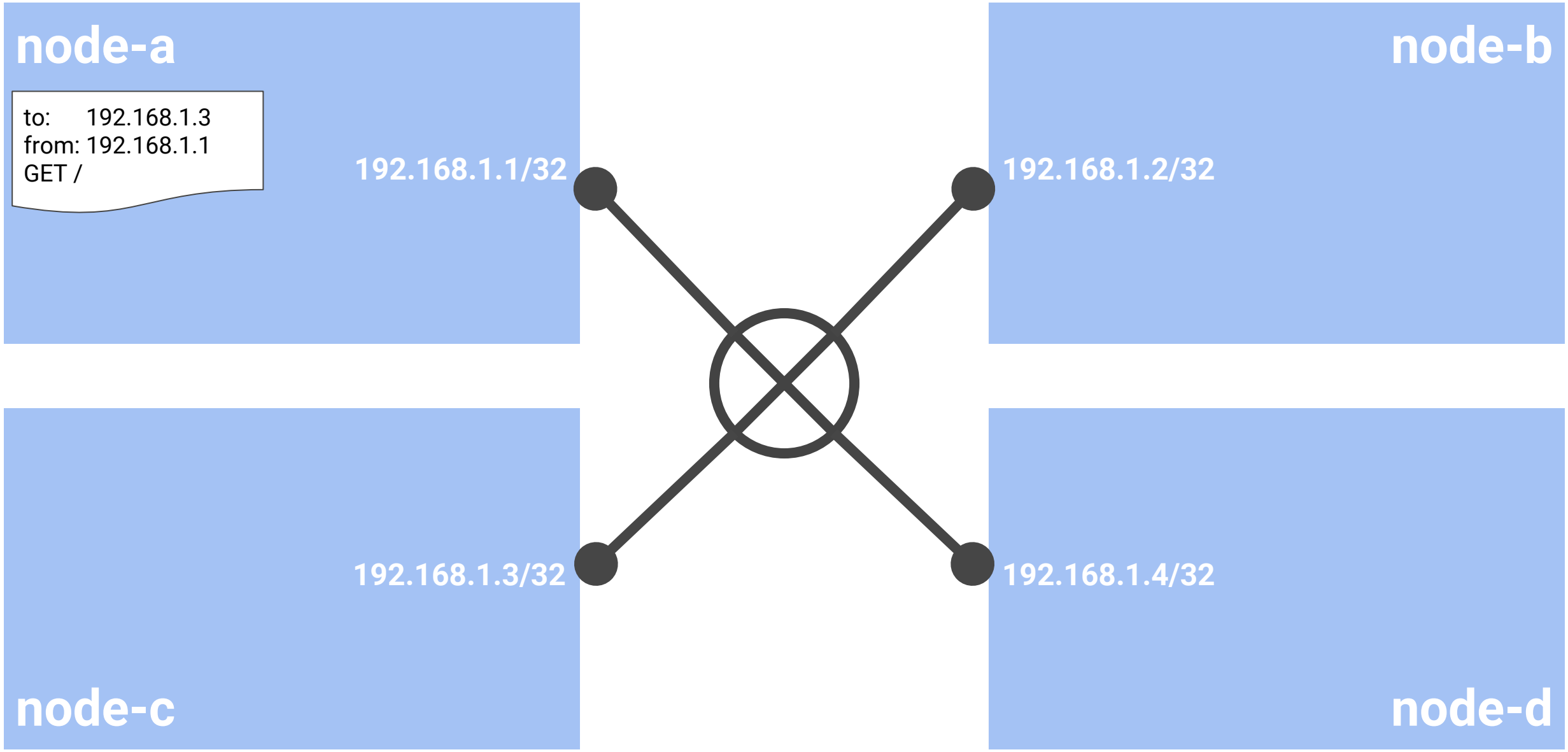

L2
node-a
node-b
ctr-1 10.0.1.2
aa:bb:cc:dd:e1:01
11:22:33:44:55:01
192.168.1.2/16
01:23:45:67:89:02
to: 10.0.3.2
via: 11:22:33:44:55:03
from: 10.0.1.2
GET /
ctr-2 10.0.3.2
aa:bb:cc:dd:e3:02
192.168.1.4/16
01:23:45:67:89:04
node-c
node-d

Layer 3 - IP

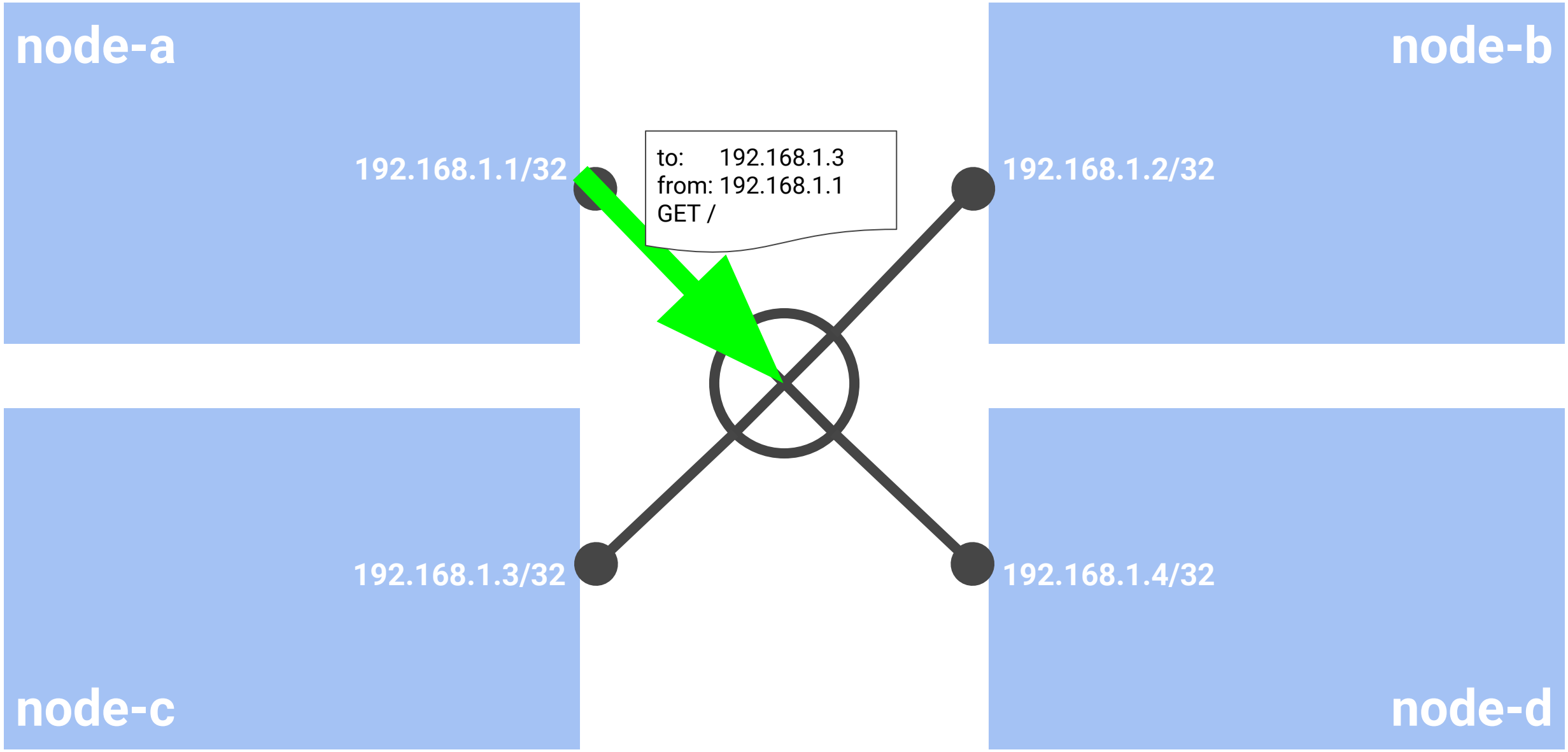
L3


node-a
node-b
192.168.1.1/32
192.168.1.2/32
gateway
192.168.1.3/32
192.168.1.4/32
node-c
node-d

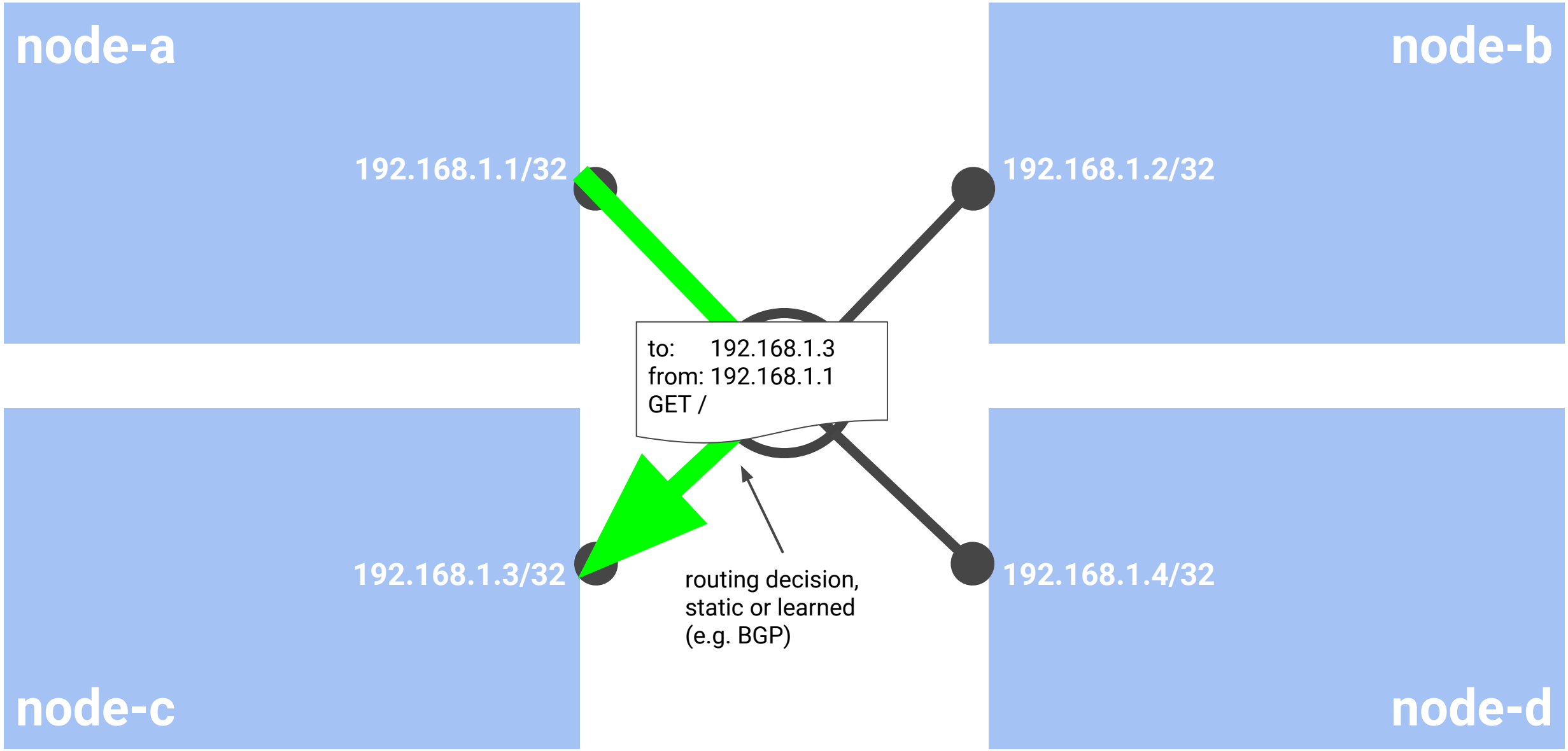
L3


node-a
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.1/32
node-b
192.168.1.2/32
192.168.1.3/32
192.168.1.4/32
node-c
node-d

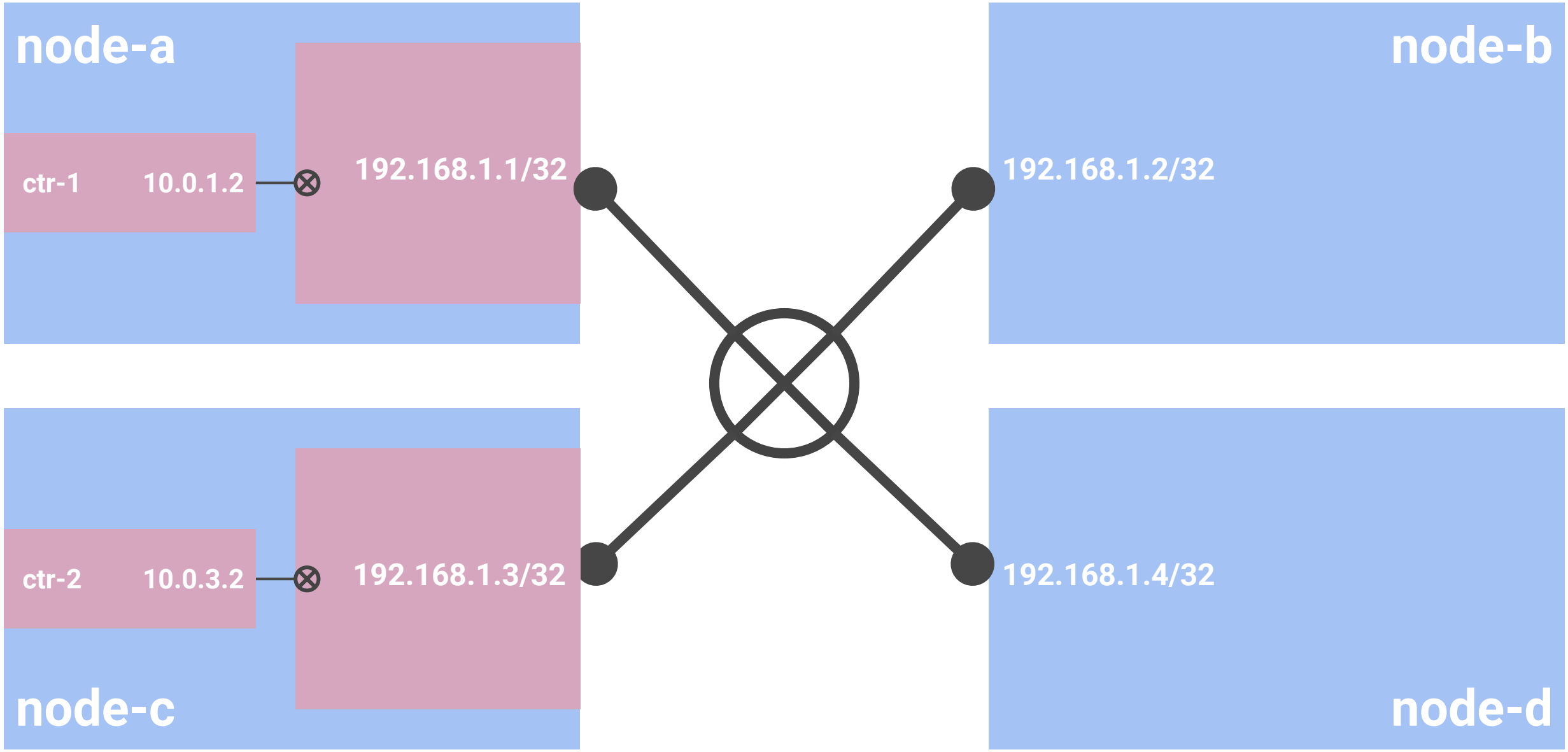
L3


node-a
node-b
192.168.1.1/32
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.2/32
192.168.1.3/32
192.168.1.4/32
node-c
node-d

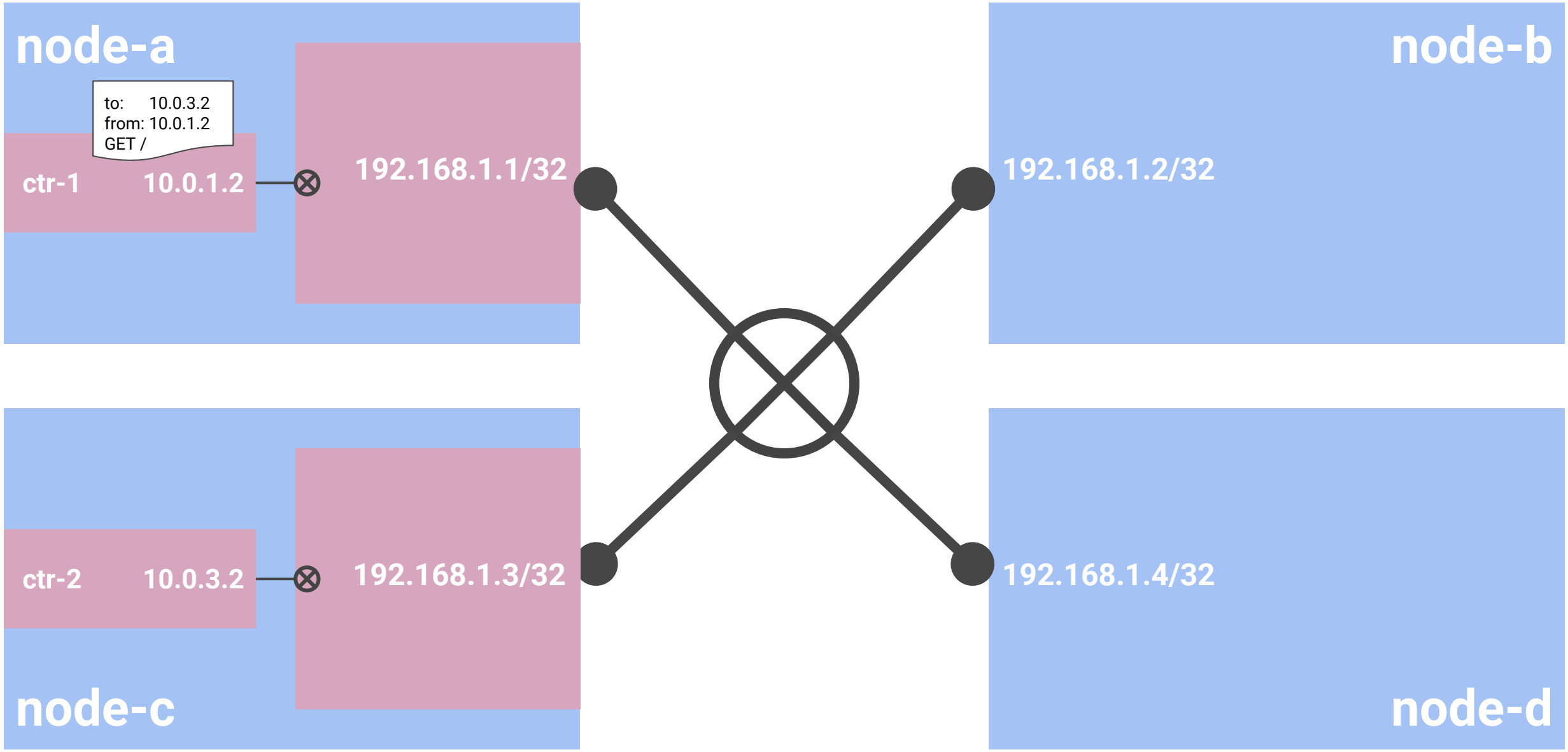
L3


node-a
node-b
192.168.1.1/32
192.168.1.2/32
to: 192.168.1.3
from: 192.168.1.1
GET /
192.168.1.3/32
192.168.1.4/32
routing decision,
static or learned
(e.g. BGP)
node-c
node-d

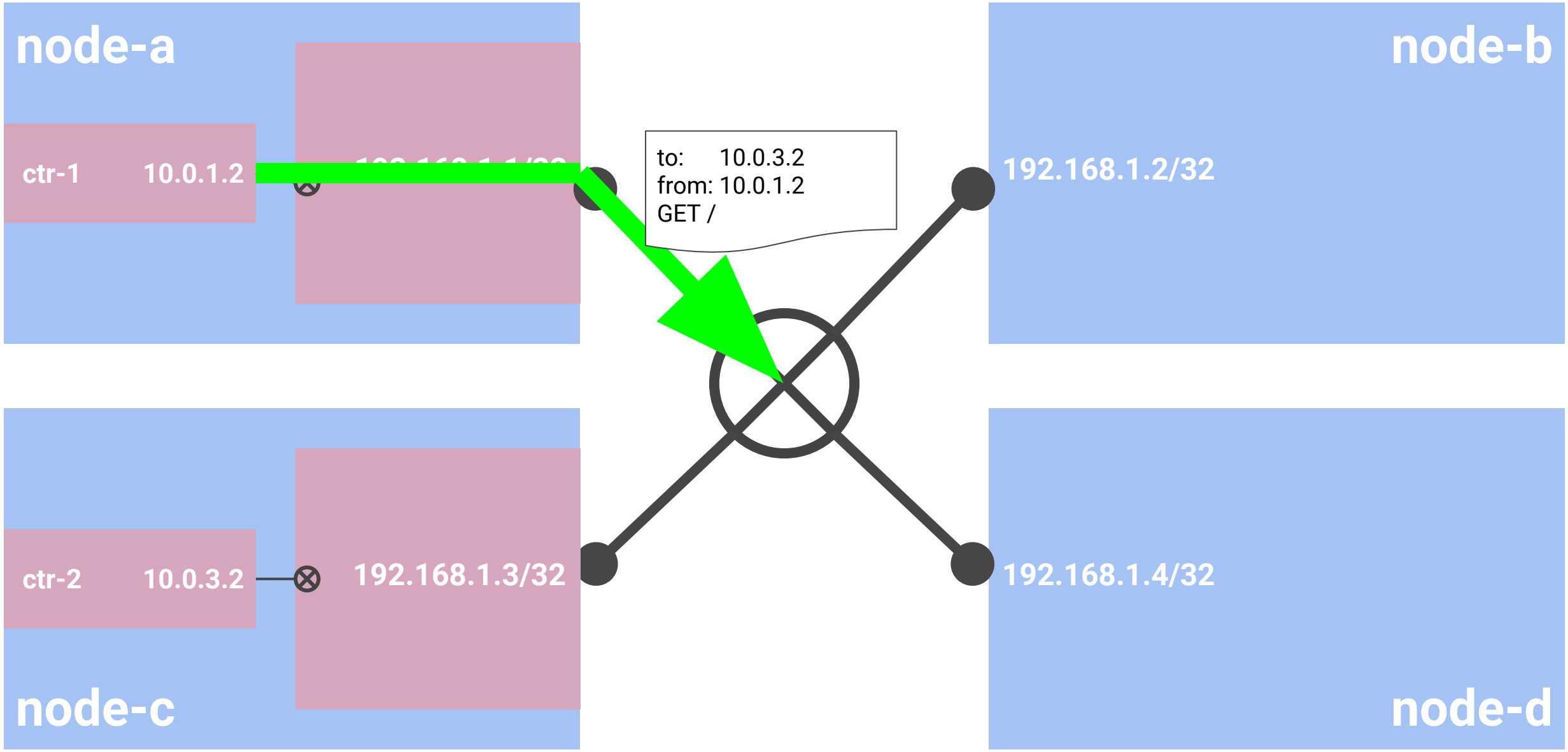
L3


node-a
ctr-1
10.0.1.2
192.168.1.1/32
node-b
192.168.1.2/32
ctr-2
10.0.3.2
192.168.1.3/32
node-c
192.168.1.4/32
node-d

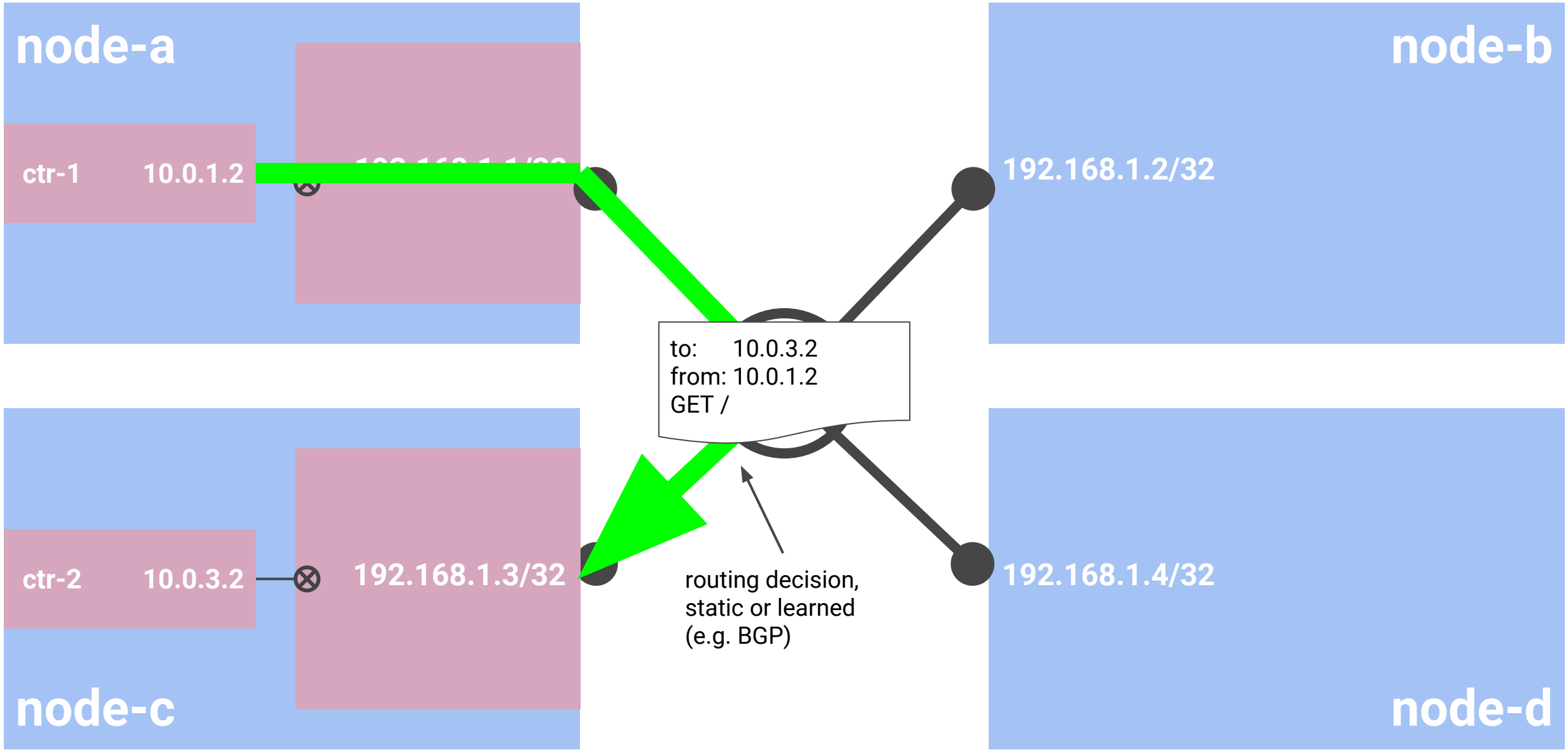
L3


node-a
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-1
10.0.1.2
192.168.1.1/32
node-b
192.168.1.2/32
ctr-2
10.0.3.2
192.168.1.3/32
node-c
192.168.1.4/32
node-d

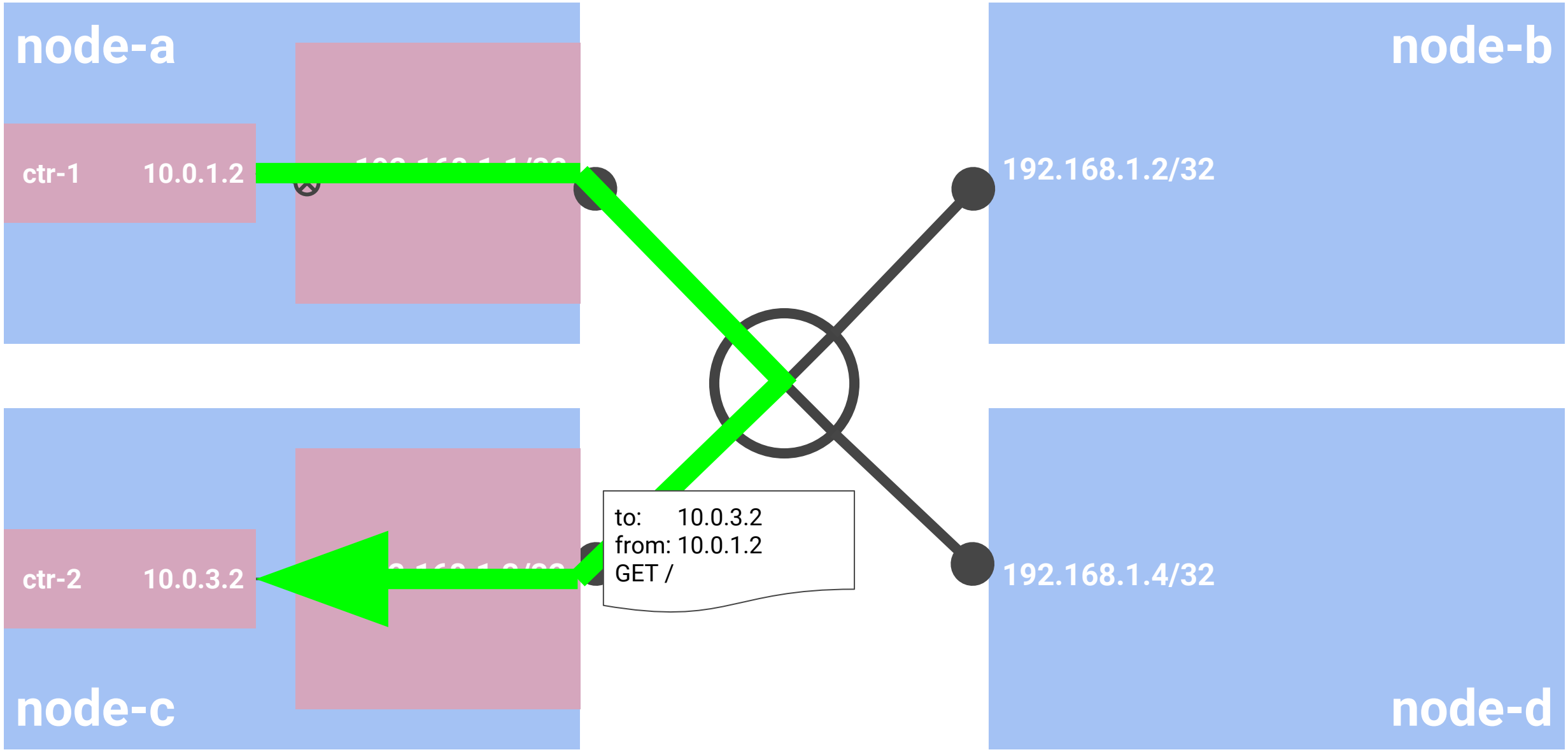
L3


node-a
ctr-1
10.0.1.2
to: 10.0.3.2
from: 10.0.1.2
GET /
node-b
192.168.1.2/32
ctr-2
10.0.3.2
192.168.1.3/32
192.168.1.4/32
node-c
node-d

L3


node-a
ctr-1
10.0.1.2
node-b
192.168.1.2/32
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-2
10.0.3.2
192.168.1.3/32
routing decision,
static or learned
(e.g. BGP)
192.168.1.4/32
node-c
node-d

L3

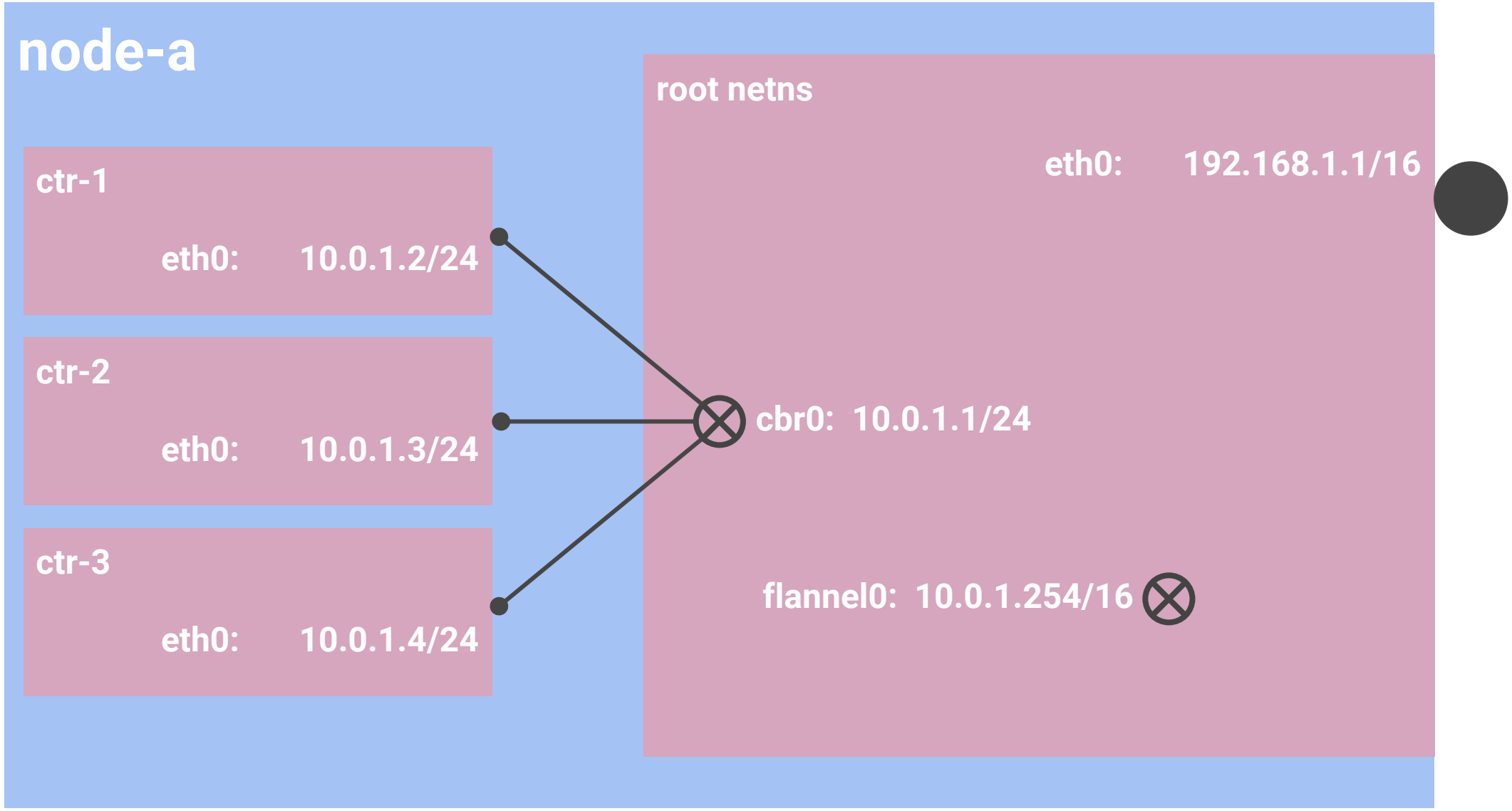

node-a
ctr-1
10.0.1.2
node-b
192.168.1.2/32
to: 10.0.3.2
from: 10.0.1.2
GET /
ctr-2
10.0.3.2
192.168.1.4/32
node-c
node-d

Overlays

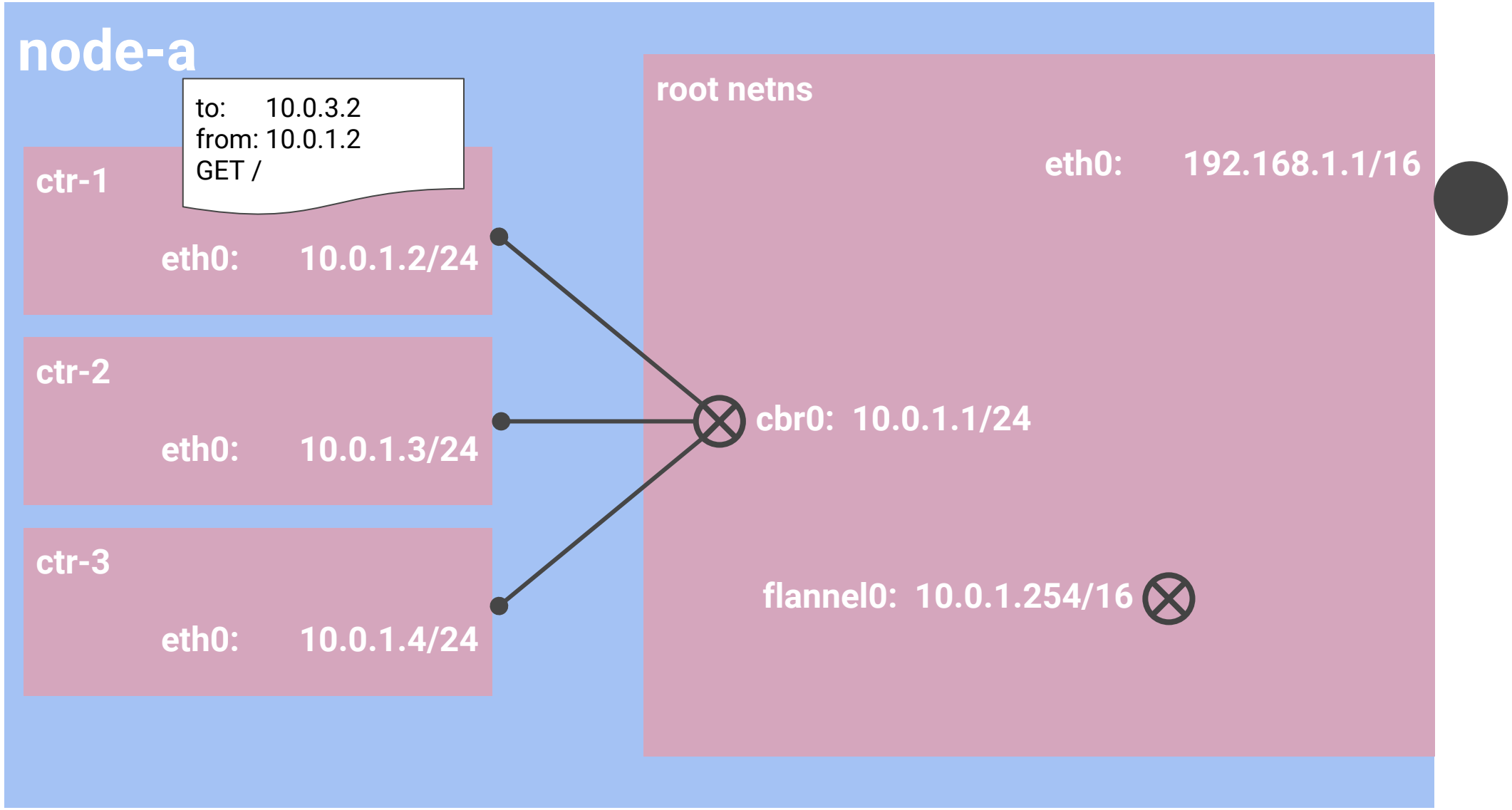
Q: When should I use an overlay?

A: When nothing else works, *or* when you have specific reasons to want it (e.g. the added value of management)

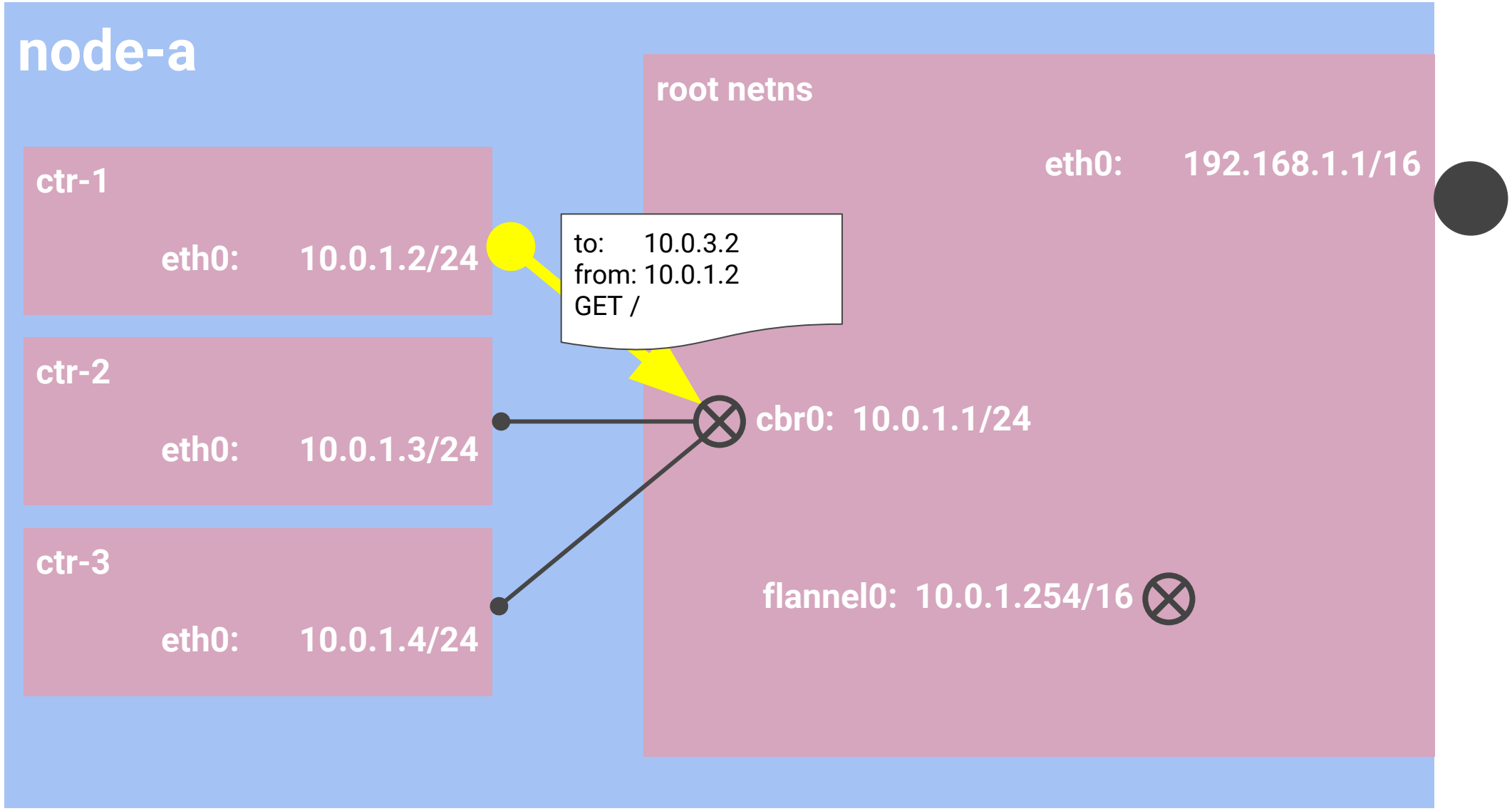
Overlay (e.g. flannel, weave)



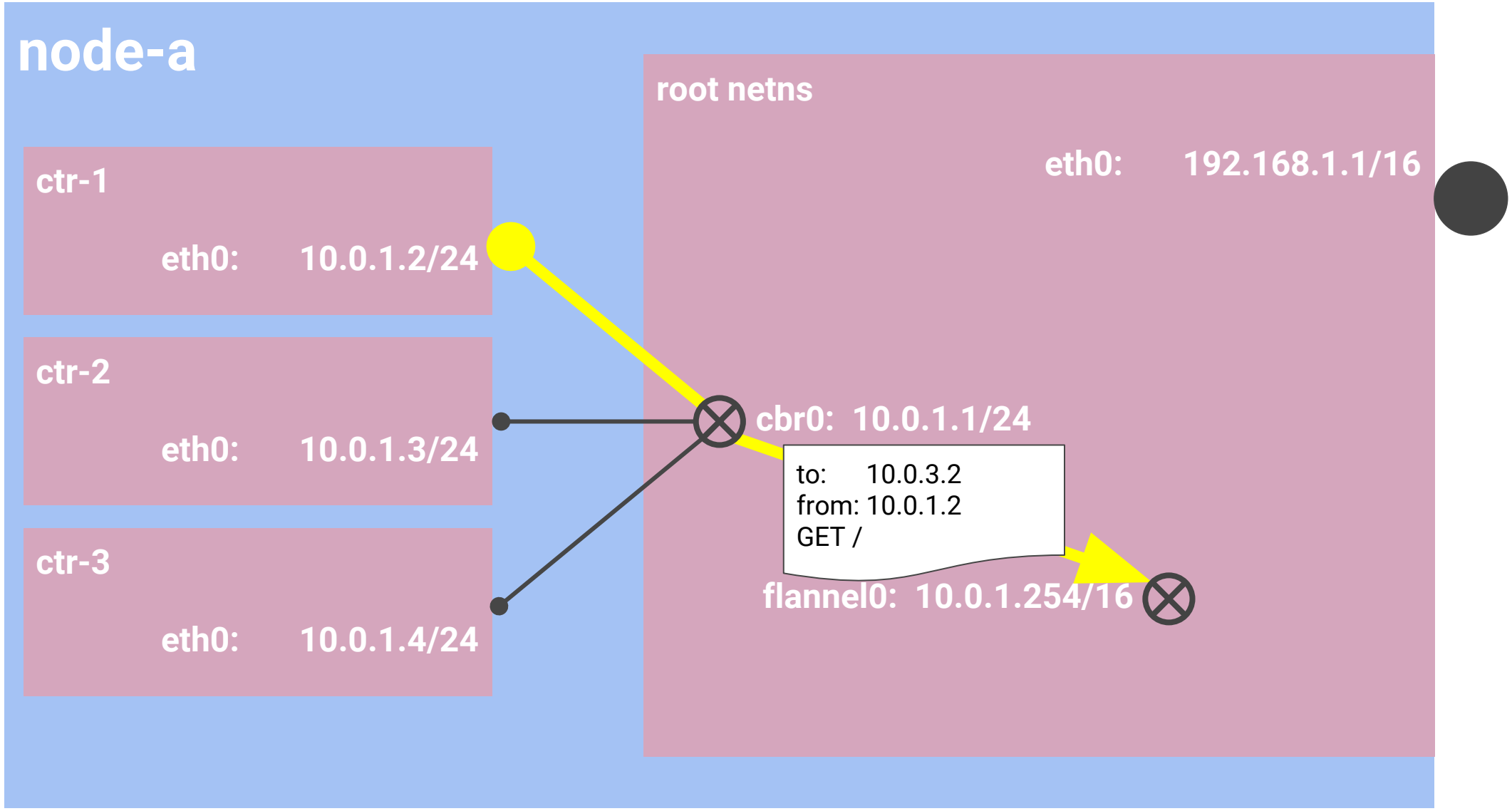
Overlay (e.g. flannel, weave)



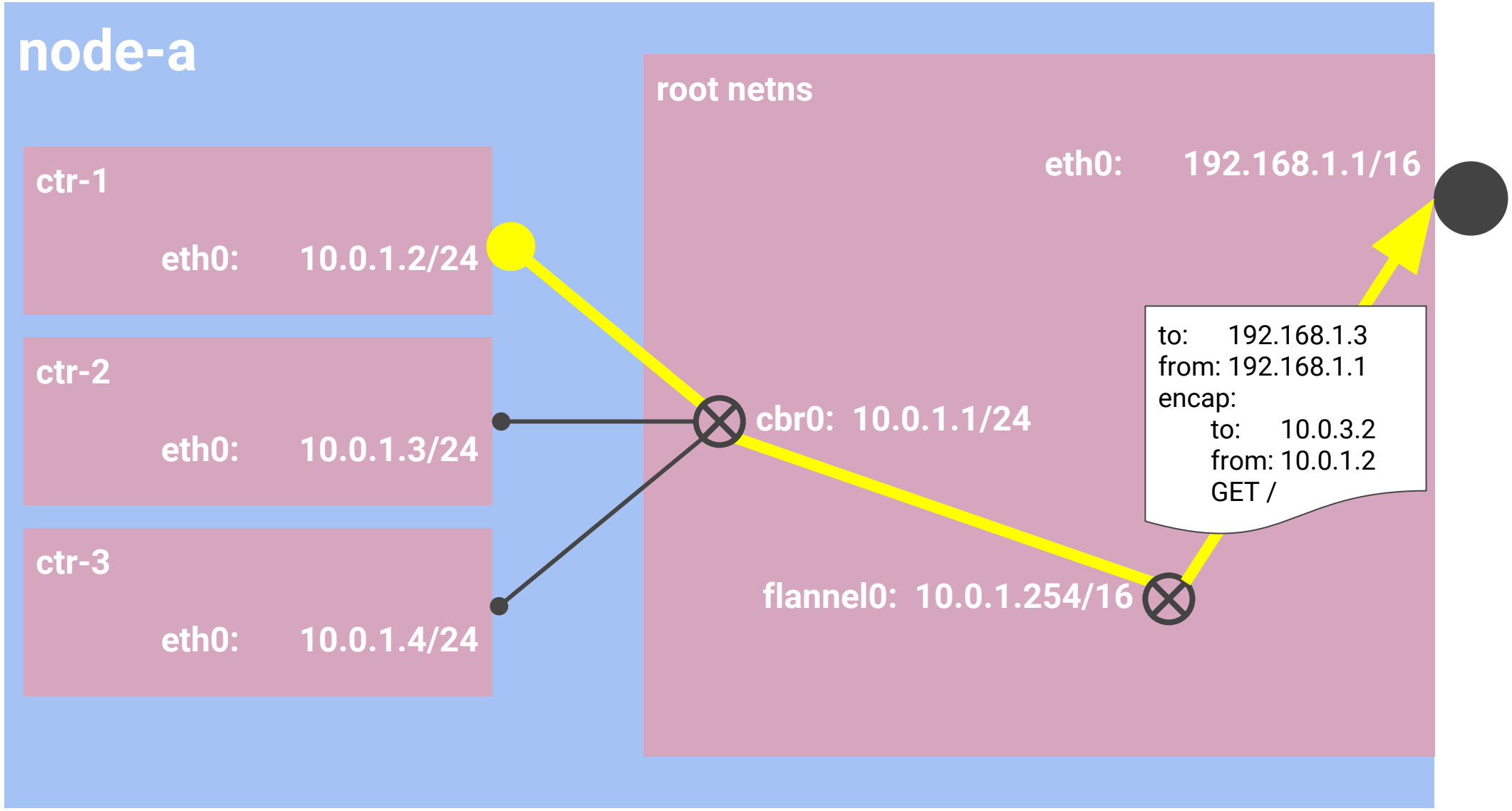
Overlay (e.g. flannel, weave)



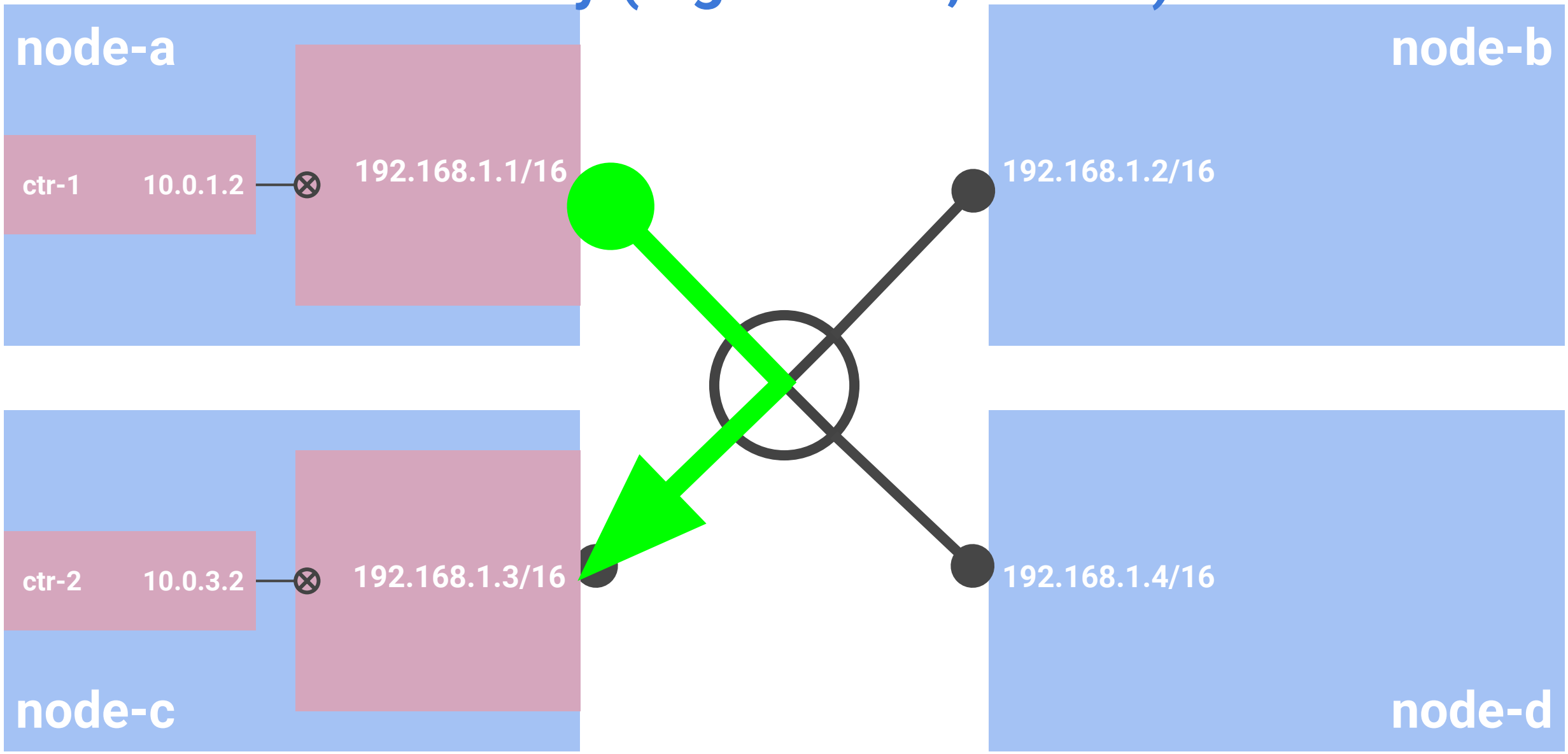
Overlay (e.g. flannel, weave)



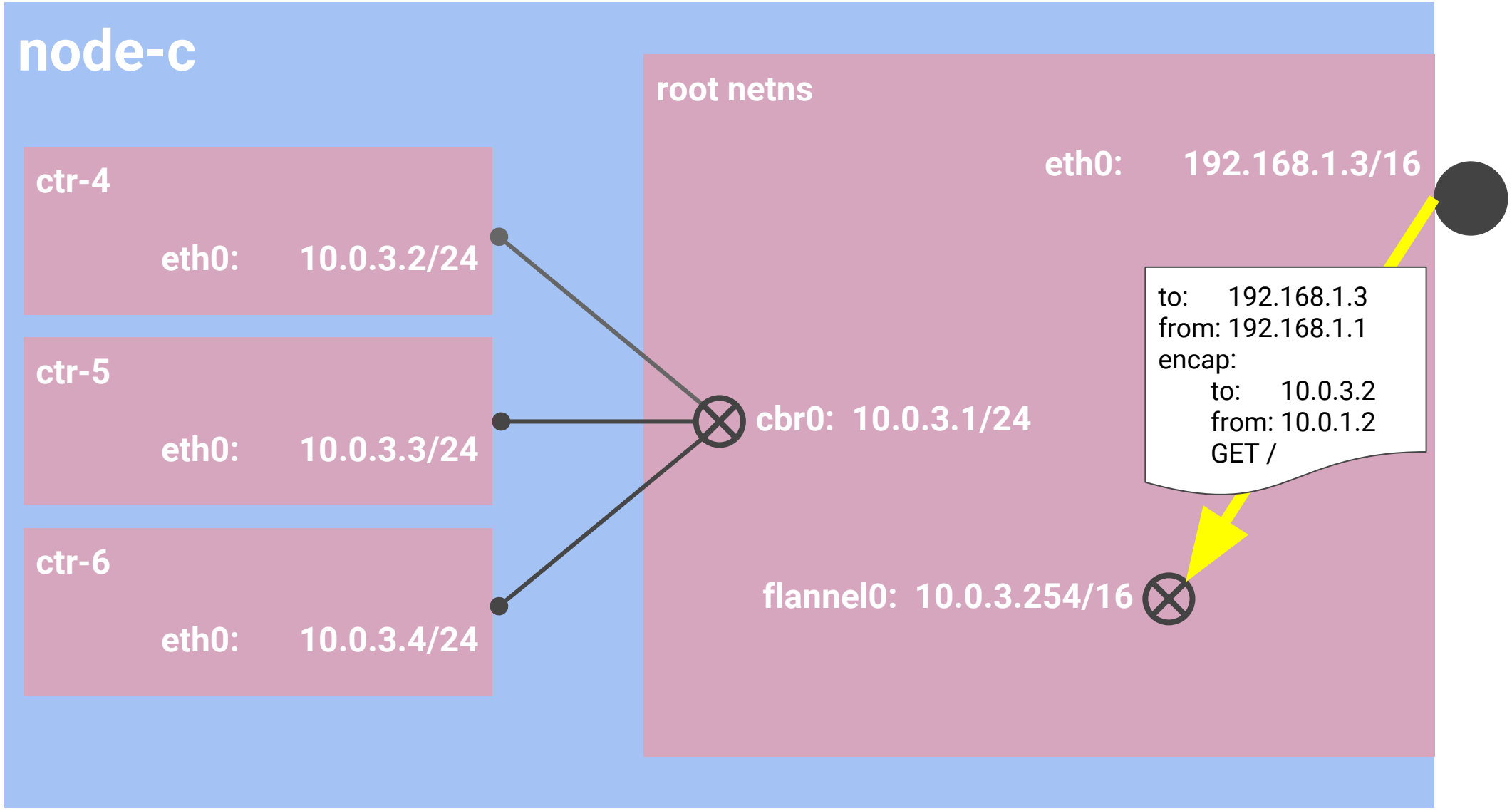
Overlay (e.g. flannel, weave)



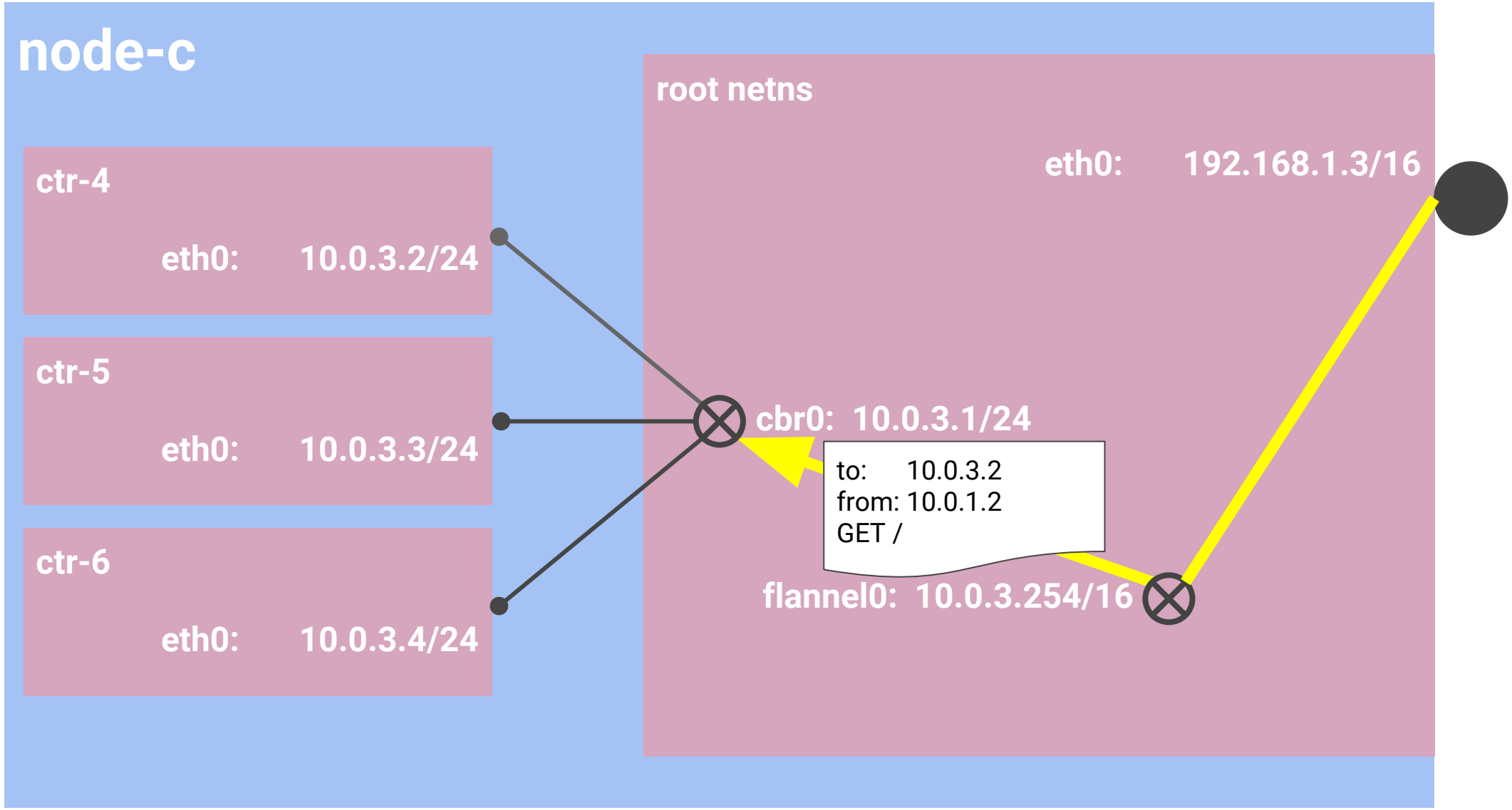
Overlay (e.g. flannel, weave)



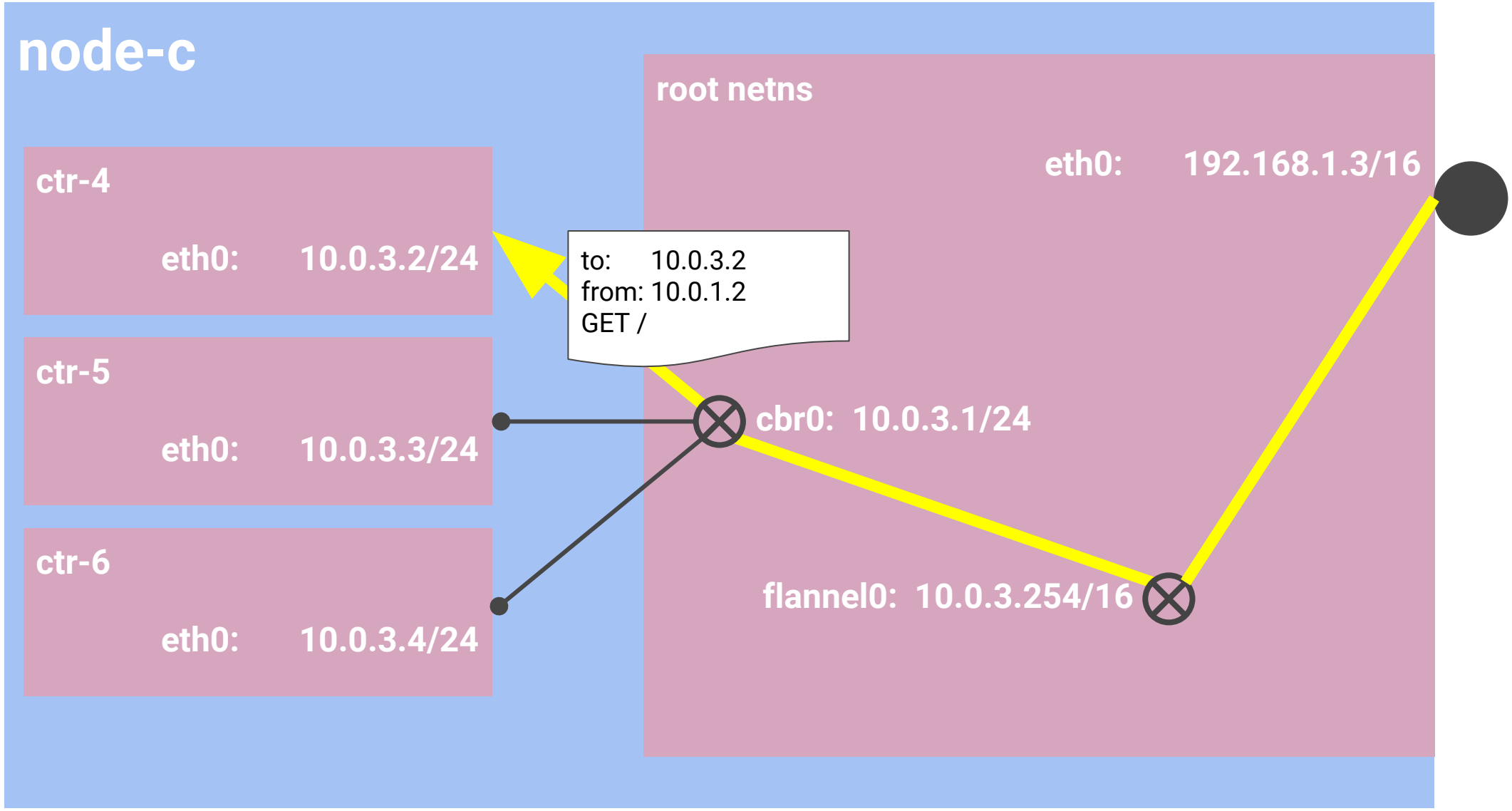
Overlay (e.g. flannel, weave)



Overlay (e.g. flannel, weave)

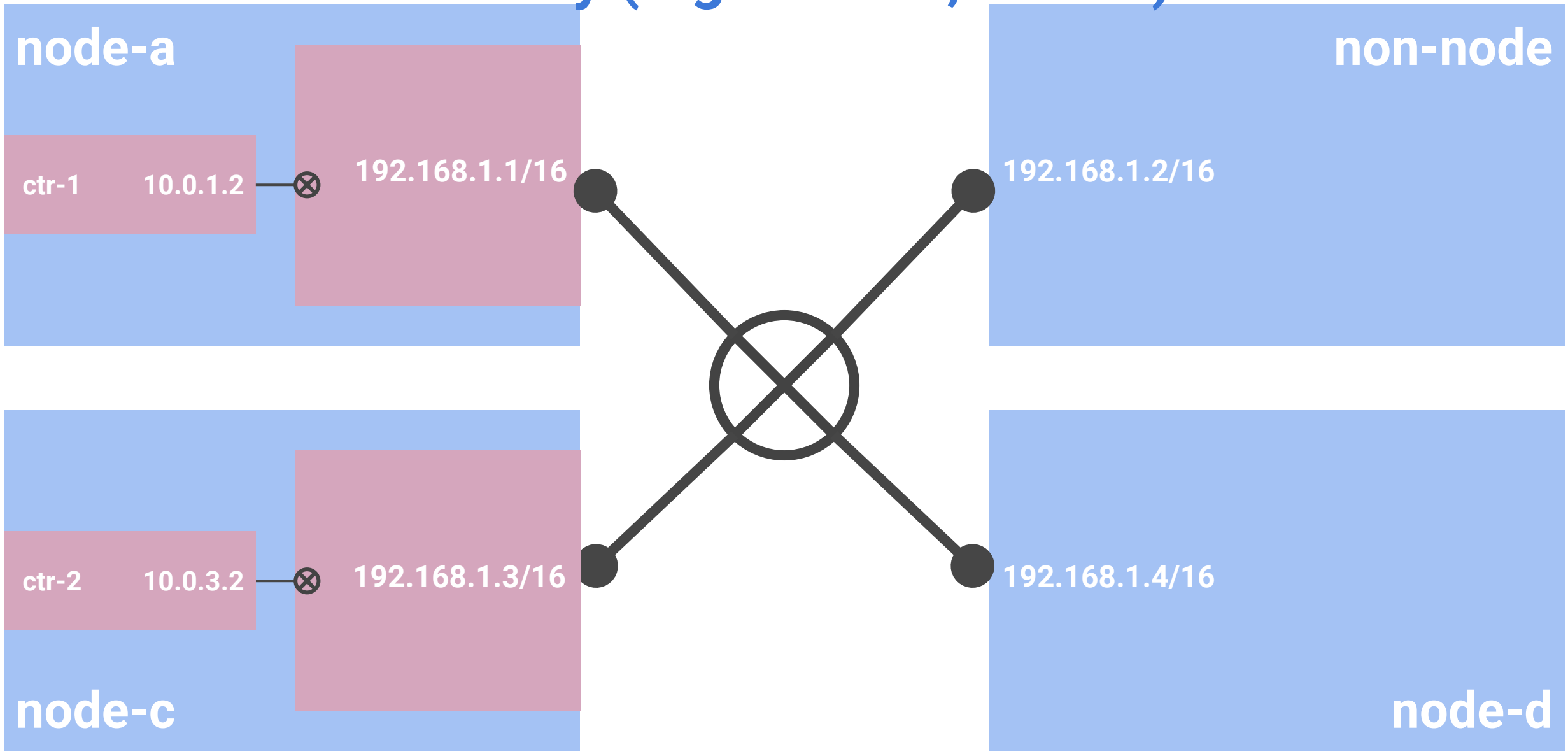


Overlay (e.g. flannel, weave)

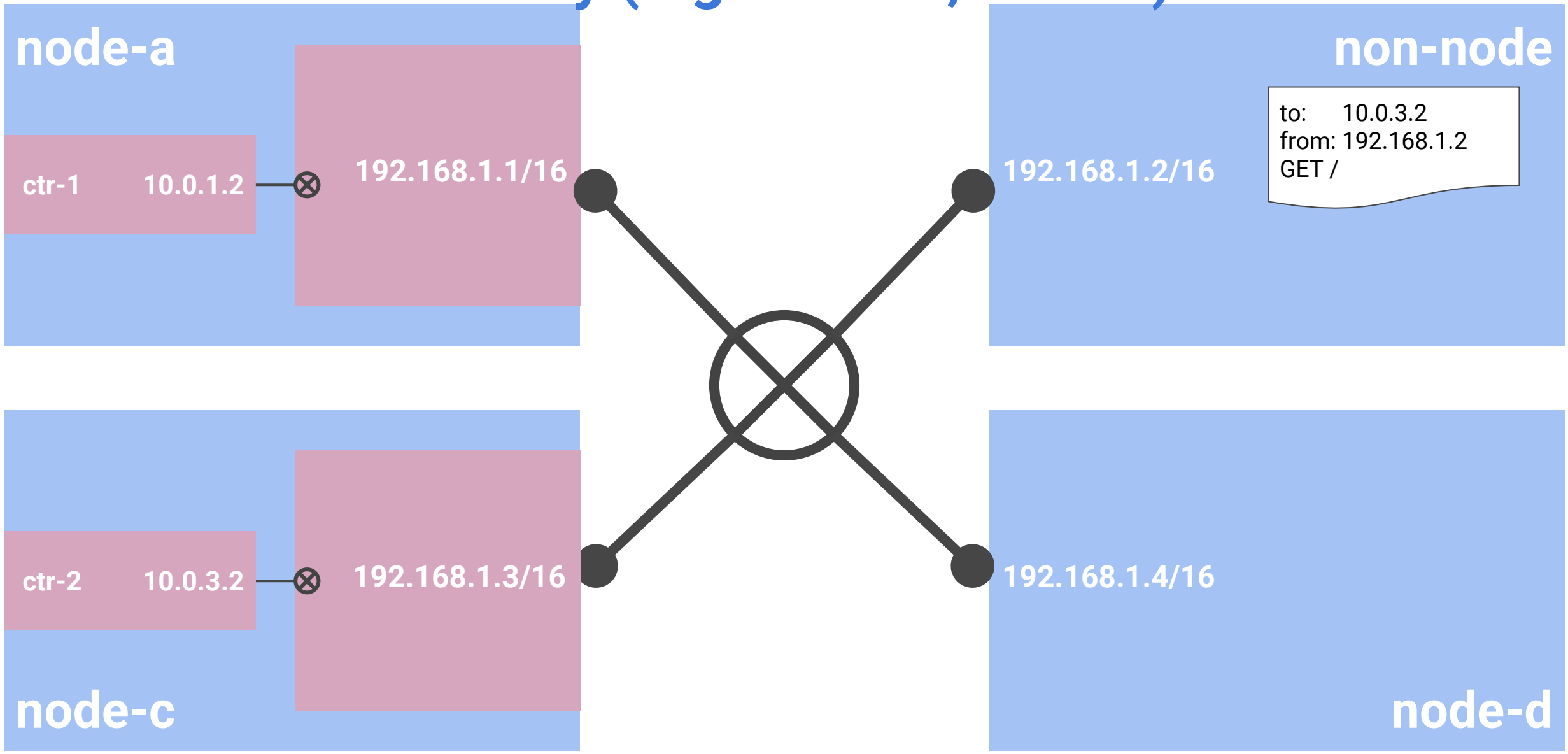


Overlays - the hard part

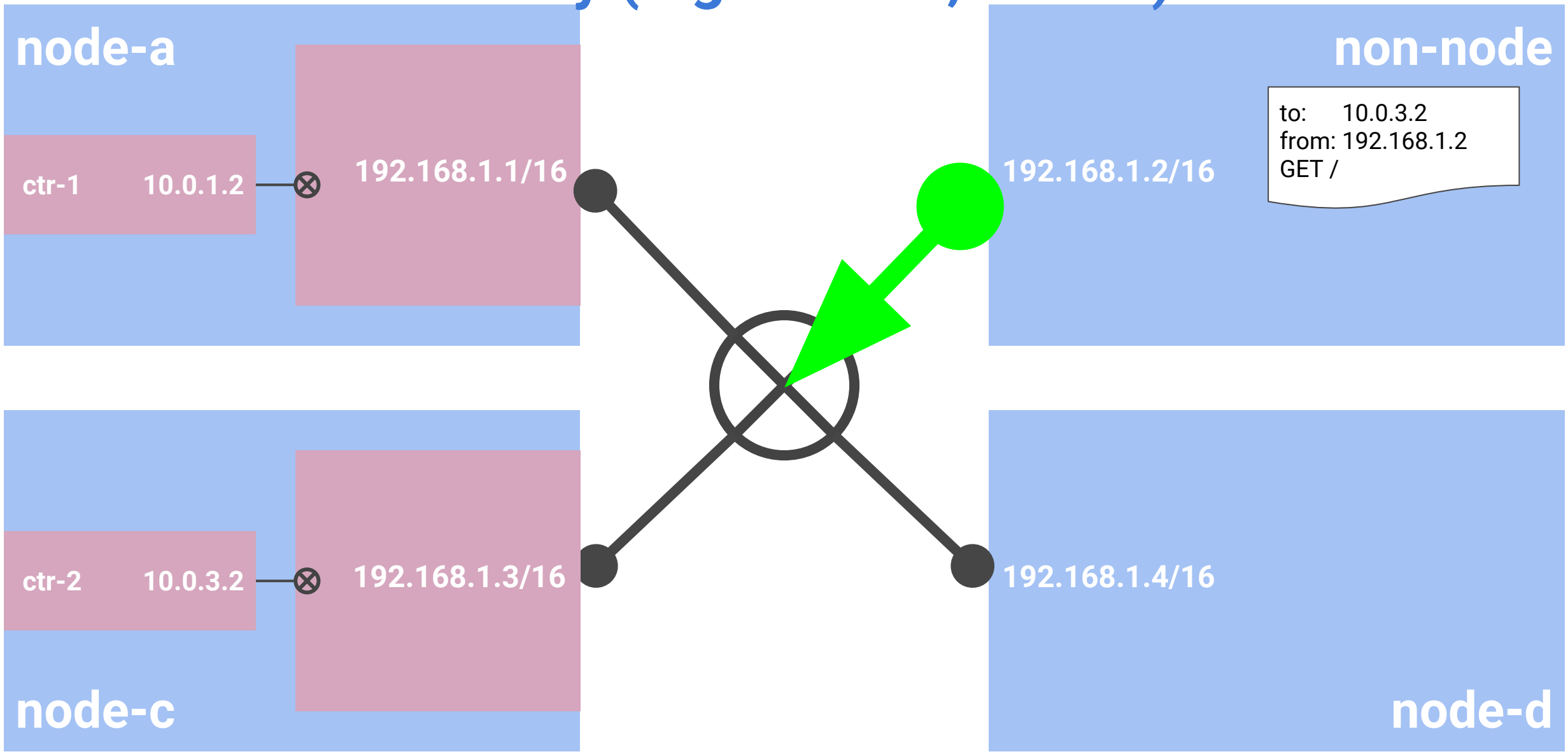
Overlay (e.g. flannel, weave)



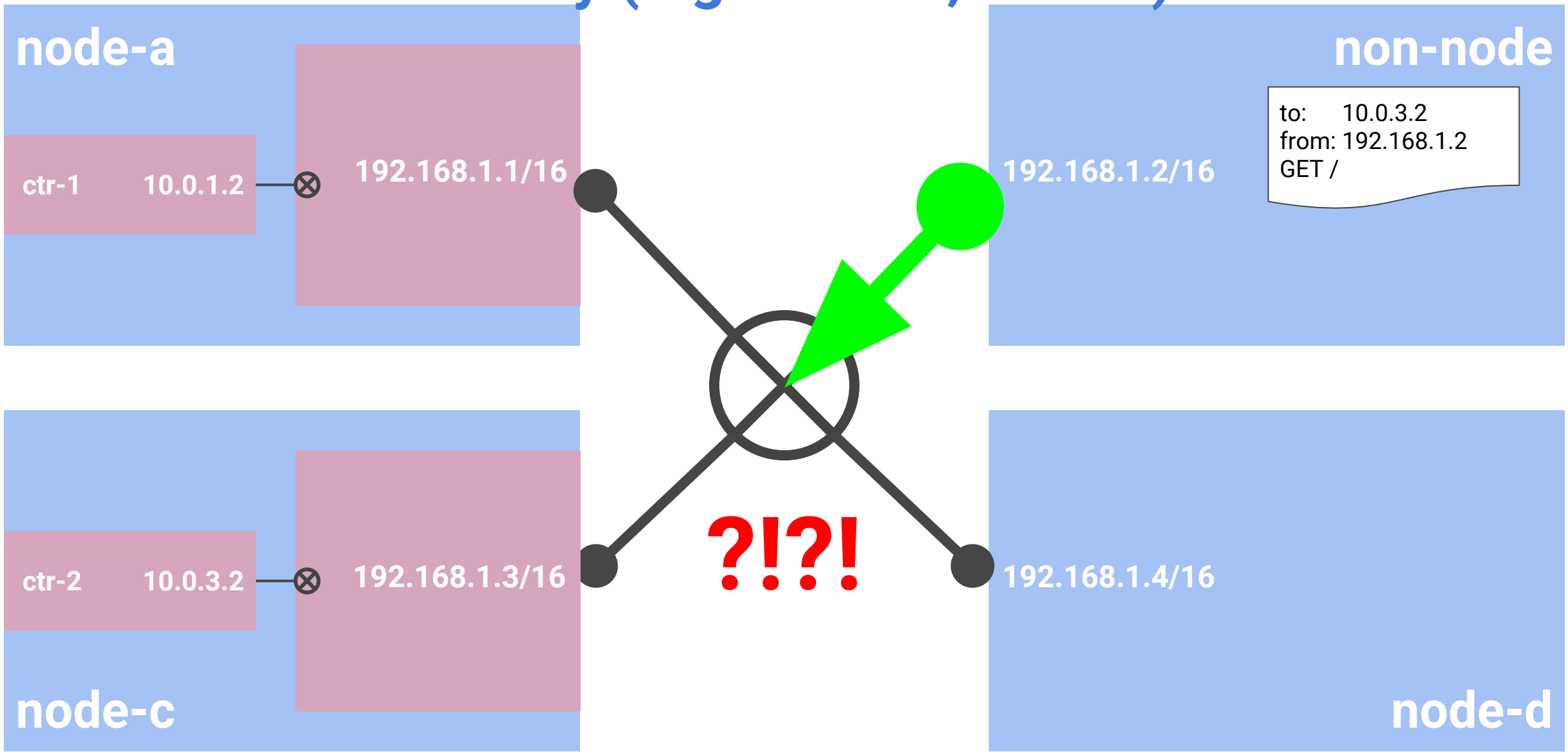
Overlay (e.g. flannel, weave)



Overlay (e.g. flannel, weave)



Overlay (e.g. flannel, weave)



We need a bridge between the physical and overlay networks...

We need a bridge between the physical and overlay networks...

- could: route to nodes
- could: route to 1 or more bridge machines
- could: run flannel on client machines

We need a bridge between the physical and overlay networks...

- could: route to nodes
- could: route to 1 or more bridge machines
- could: run flannel on client machines
- see “When should I use an overlay?”