Management and Operations of Networks, Services, and Systems
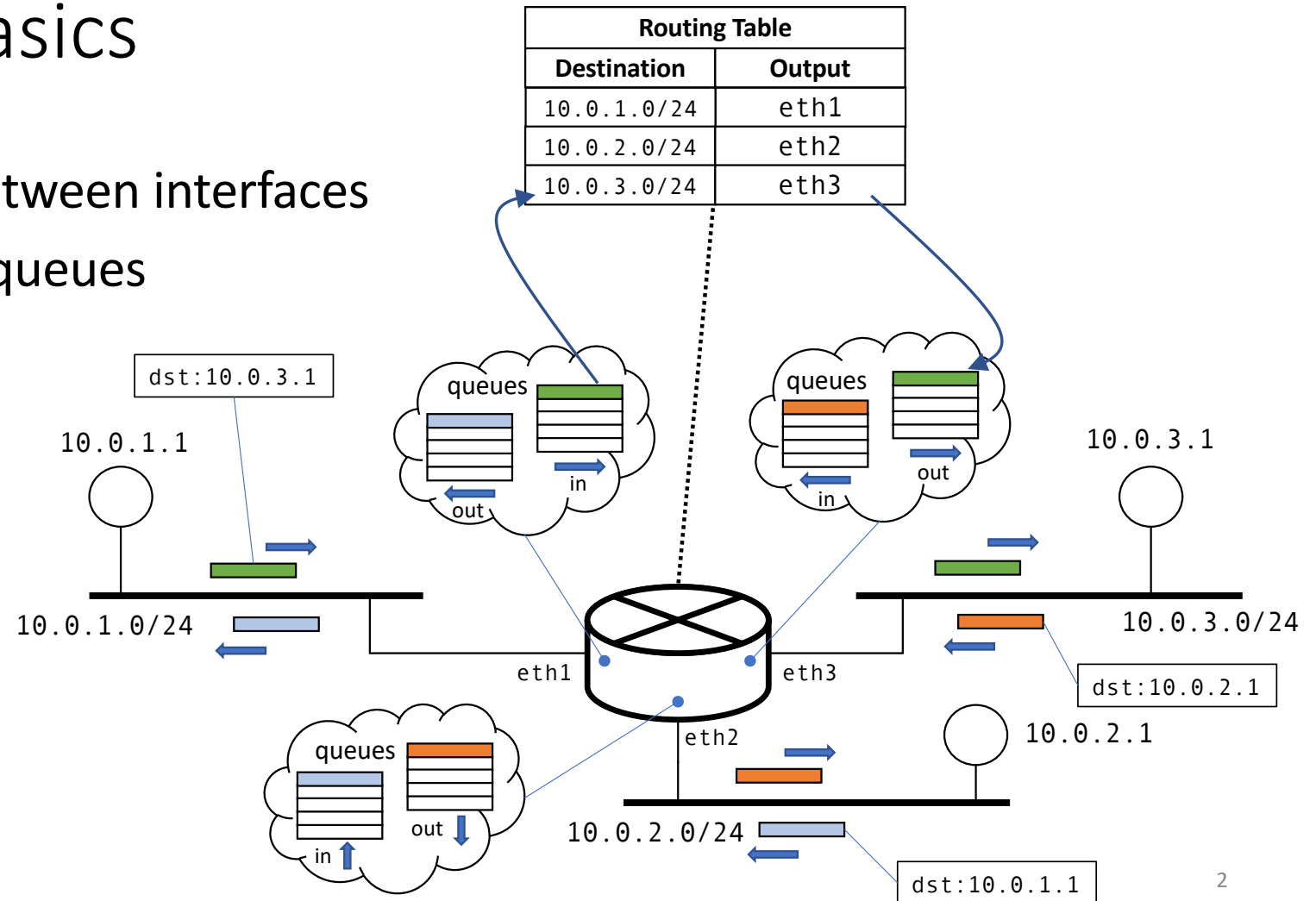
Routing

Ricardo Morla

FEUP – GORS/M.EEC, GRS/M.EIC
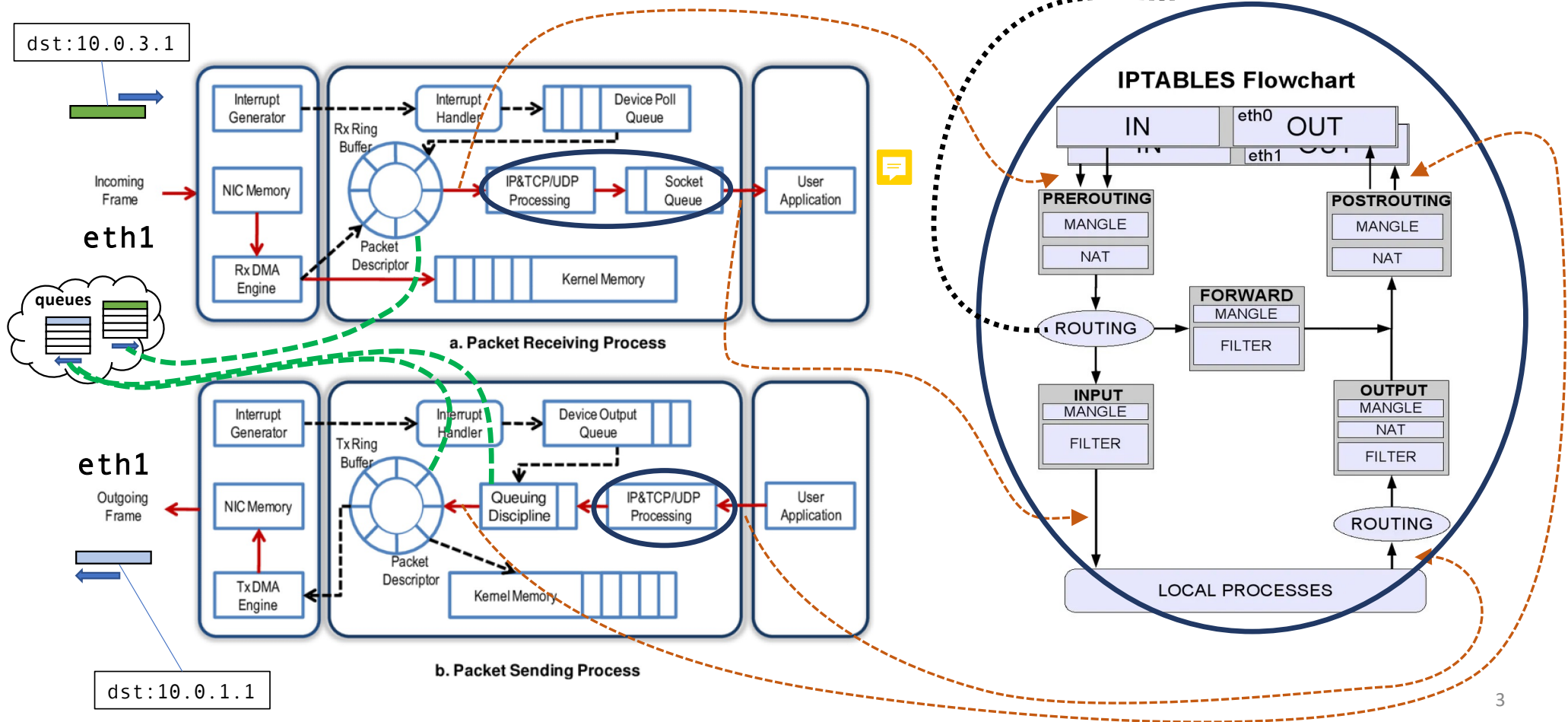
Routing basics

- Forwarding between interfaces
- Input/output queues
- Routing table

Routing Table	
Destination	**Output**
10.0.1.0/24	eth1
10.0.2.0/24	eth2
10.0.3.0/24	eth3

dst:10.0.3.1

10.0.1.1

queues

in

out

queues

out

in

10.0.3.1

10.0.1.0/24

eth1

eth3

10.0.3.0/24

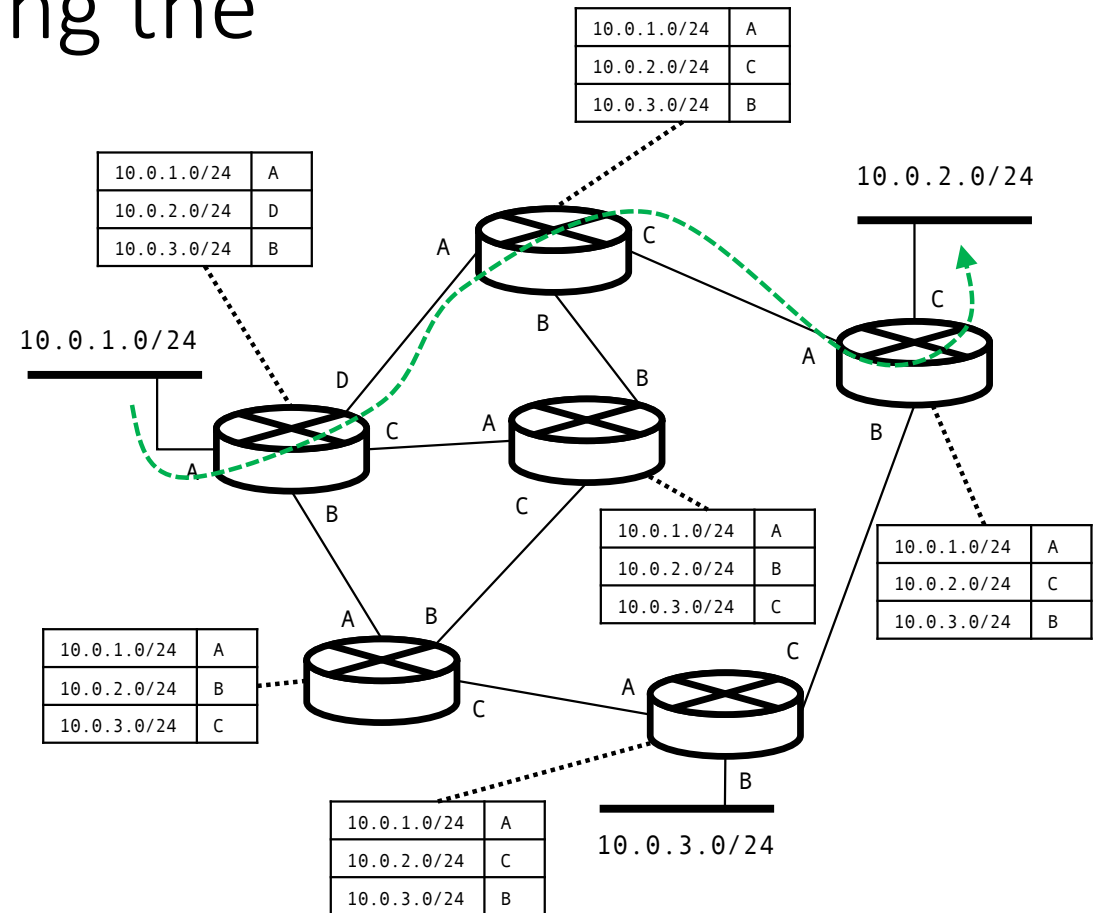dst:10.0.2.1

10.0.2.1

queues

out

in

10.0.2.0/24

eth2

dst:10.0.1.1

2

Inside the router

Routing Table	
Destination	Output
10.0.1.0/24	eth1
10.0.2.0/24	eth2
10.0.3.0/24	eth3

dst:10.0.3.1

eth1

Incoming Frame

queues

eth1

Outgoing Frame

dst:10.0.1.1

a. Packet Receiving Process

Interrupt Generator
NIC Memory
Rx DMA Engine
Rx Ring Buffer
Packet Descriptor
Interrupt Handler
Device Poll Queue
IP&TCP/UDP Processing
Socket Queue
Kernel Memory
User Application

b. Packet Sending Process

Interrupt Generator
NIC Memory
Tx DMA Engine
Tx Ring Buffer
Packet Descriptor
Interrupt Handler
Device Output Queue
Queuing Discipline
IP&TCP/UDP Processing
Kernel Memory
User Application

IPTABLES Flowchart

IN | eth0 OUT
IN | eth1 OUT

PREROUTING
MANGLE
NAT

POSTROUTING
MANGLE
NAT

ROUTING

FORWARD
MANGLE
FILTER

INPUT
MANGLE
FILTER

OUTPUT
MANGLE
NAT
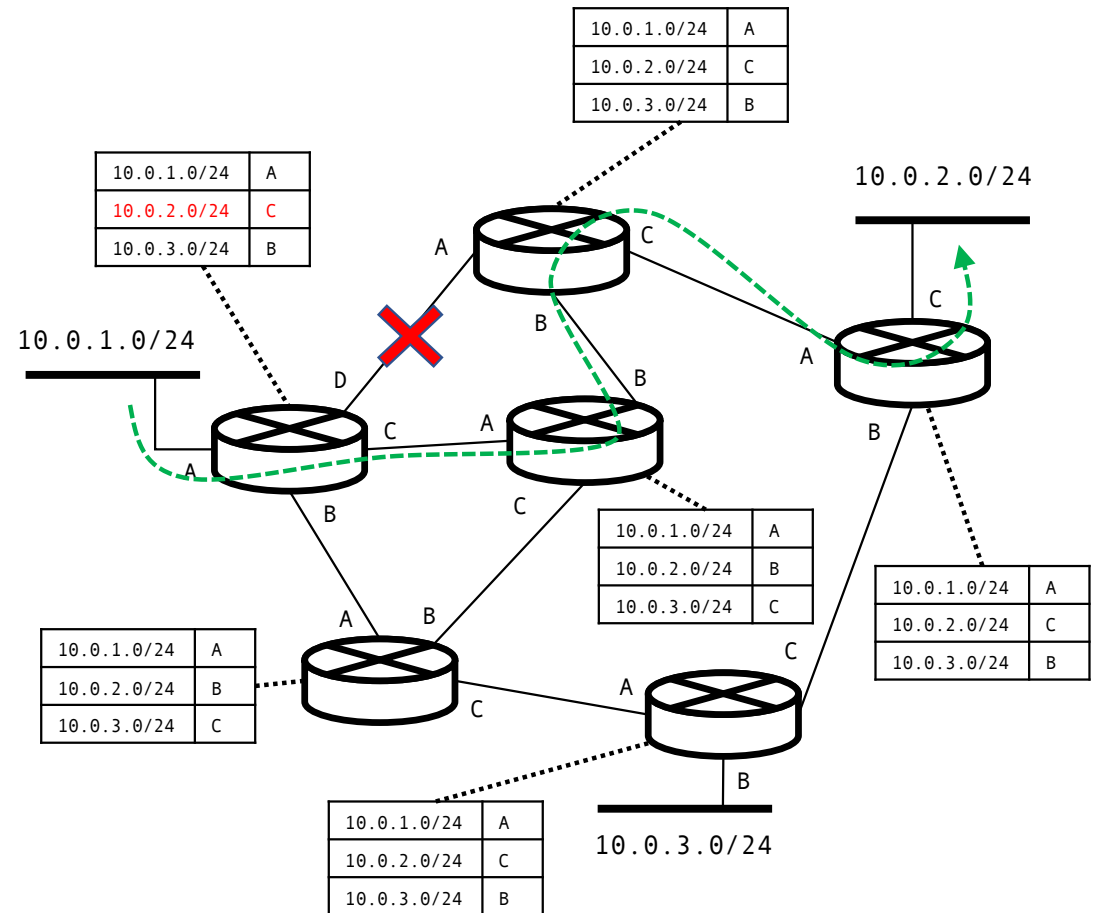FILTER

ROUTING

LOCAL PROCESSES

3

Routing – as in "choosing the paths for packets"

- Each router has its own routing table

- If routing tables are coherent, packets can go from any node in the network to any other node

- If they're not:
  - Loops
  - Blackholes

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.2.0/24

10.0.1.0/24	A
10.0.2.0/24	D
10.0.3.0/24	B

10.0.1.0/24

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

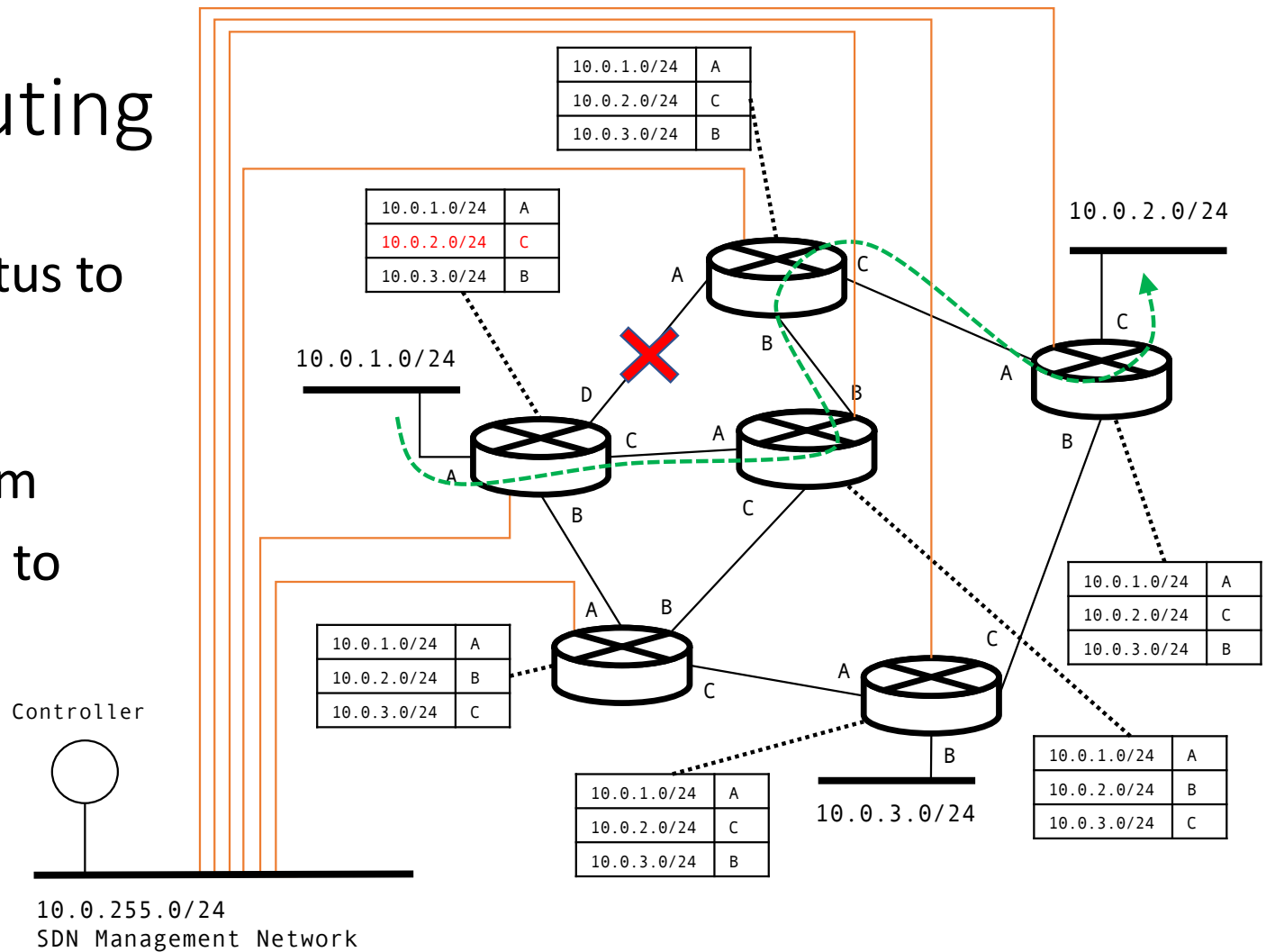| 10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.3.0/24

4

Reacting to changes in topology

- Each router can sense if their links are up or down
  - Or if new networks are added
- Automatically update routing tables
  - New path bypasses down link
- Based on some objective to optimize:
  - Smallest number of hops
  - Smallest delay
  - Largest throughput
  - Balance load on links
  - Administrative distance (…)

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.2.0/24

10.0.1.0/24

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

| 10.0.1.0/24 | A |
| 10.0.2.0/24 | C |
| 10.0.3.0/24 | B |

10.0.3.0/24

Centralized routing

- Nodes report link status to central location
- Central location runs optimization algorithm
- Returns routing table to each router
- SDN-like solution
  - Data centers

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.2.0/24

10.0.1.0/24

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

Controller

10.0.1.0/24	A
10.0.2.0/24	C
10.0.3.0/24	B

10.0.3.0/24

10.0.1.0/24	A
10.0.2.0/24	B
10.0.3.0/24	C

10.0.255.0/24
SDN Management Network

Distributed routing

- Update routing tables dynamically
- According to information from other nodes
  - Each router knows the networks it is directly connected to
  - Routers send **control packets** to other routers
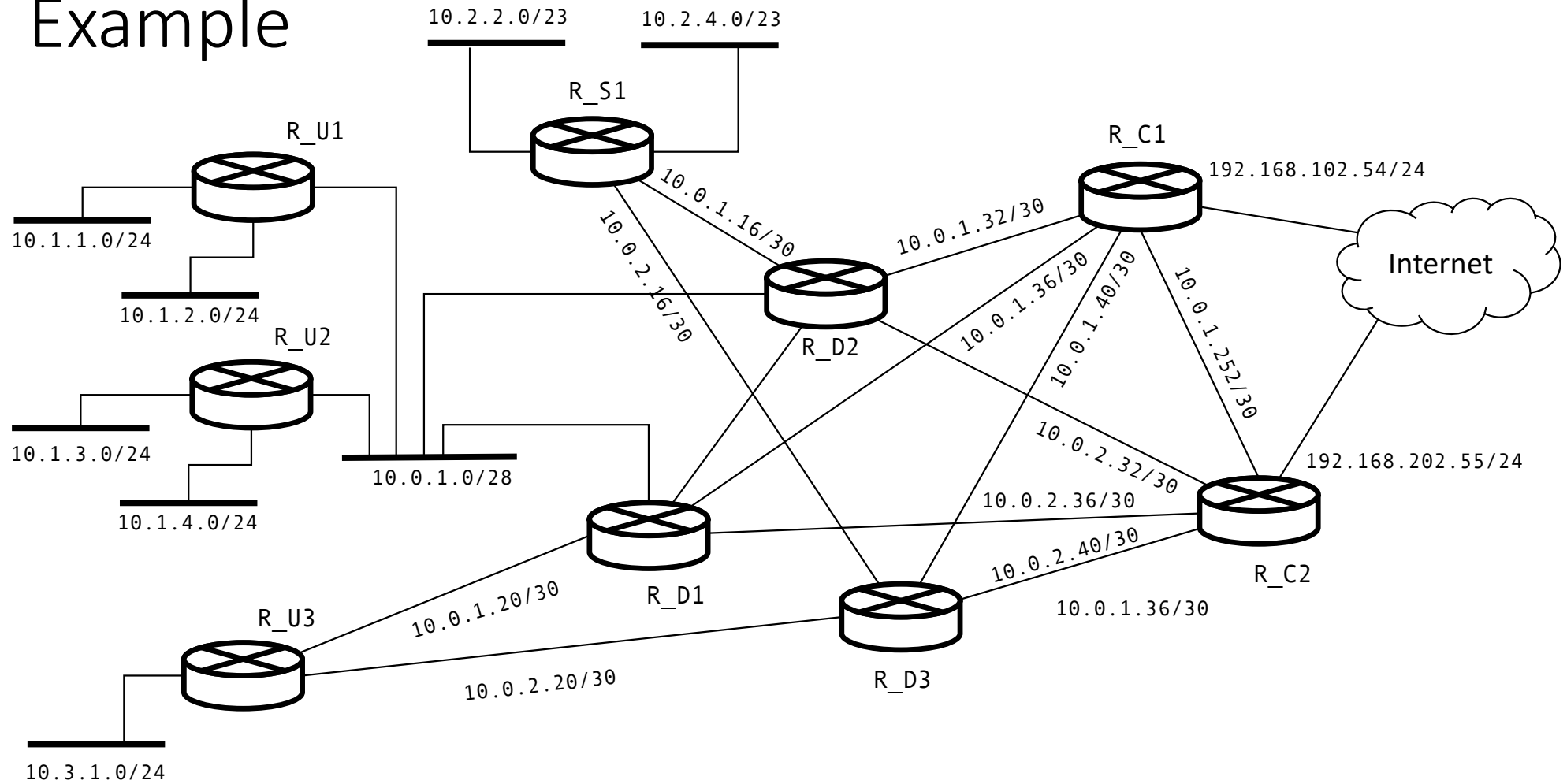- Geographically dispersed network
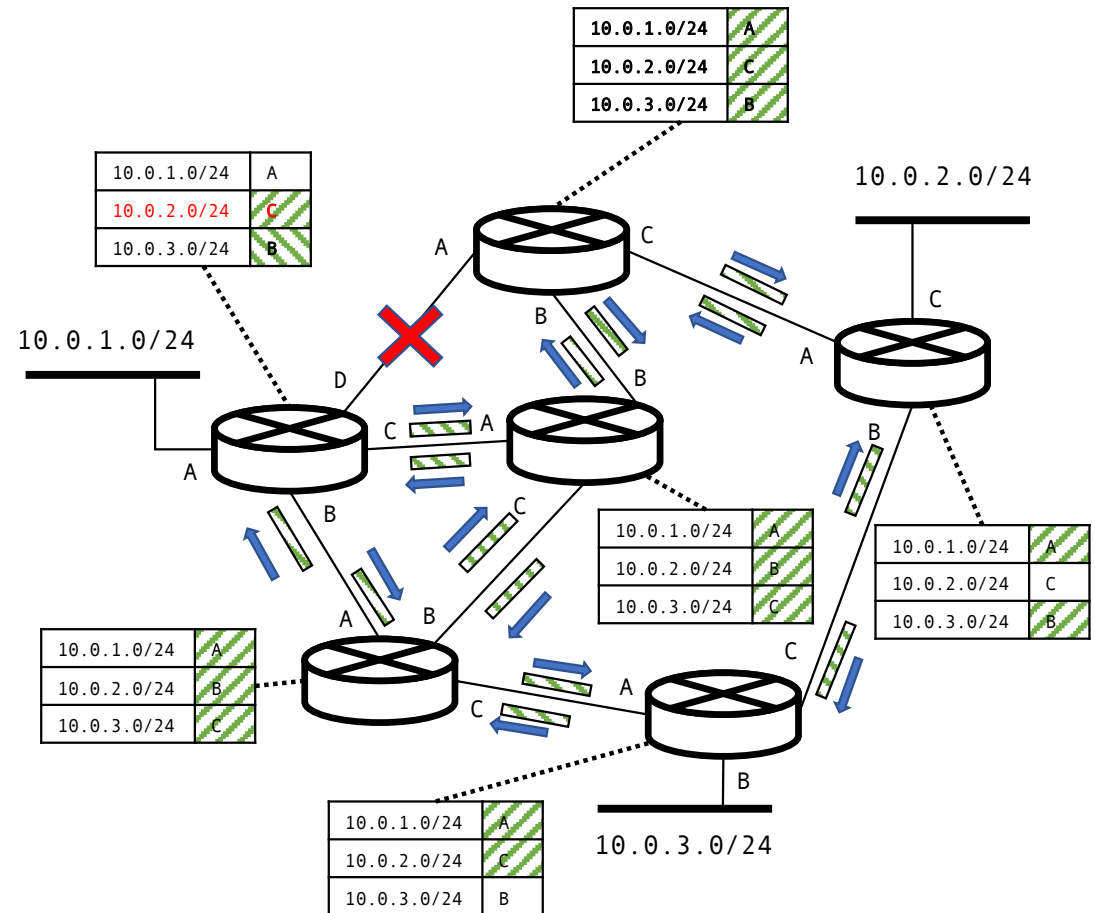
Routing algorithms

- Link state
  - Full topology of the network, then Dijkstra
  - Distributed: flood link state announcements – OSPF
- Distance vector
  - Distance to destination node, Bellman-Ford
  - Distributed: update distances, resend to neighbors
- Path vector
  - List of nodes to destination node, local preferences
  - Distributed: update path vectors, resend to neighbors

Interior routing

- Same administrative domain
  - All networks and devices administred by same entity
  - Typically the internal network of an organization
- Protocols
  - RIP (outdated) – distance vector
  - EIGRP (Cisco) – distance vector, optimized, hybrid
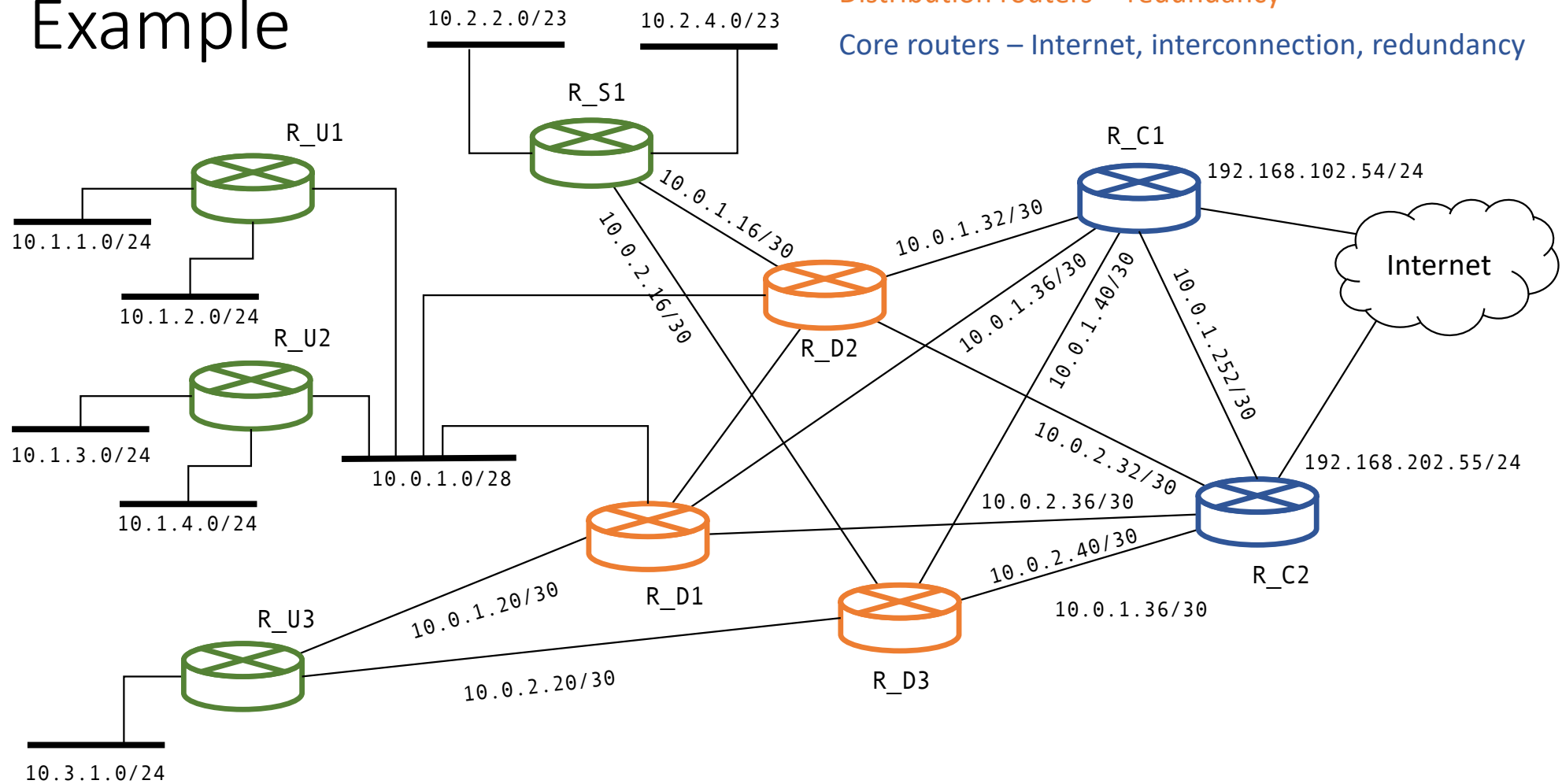 - OSPF – link state

Example

10.2.2.0/23   10.2.4.0/23

R_S1

R_U1

R_C1

192.168.102.54/24

10.1.1.0/24

10.0.1.16/30   10.0.1.32/30

Internet

10.1.2.0/24

10.0.2.16/30

10.0.1.36/30

10.0.1.40/30

10.0.1.252/30

R_U2

R_D2

10.1.3.0/24

10.0.1.0/28

10.0.2.32/30   192.168.202.55/24

10.1.4.0/24

10.0.2.36/30

R_C2

10.0.1.20/30   R_D1

10.0.2.40/30

10.0.1.36/30

R_U3

10.0.2.20/30   R_D3

10.3.1.0/24

10

Example

Access routers – end user traffic to Internet and servers

Distribution routers -- redundancy

Core routers – Internet, interconnection, redundancy

10.2.2.0/23   10.2.4.0/23

R_S1

R_U1

10.1.1.0/24

10.1.2.0/24

R_U2

10.1.3.0/24

10.1.4.0/24

10.0.1.16/30

10.0.2.16/30

10.0.1.32/30

R_C1

192.168.102.54/24

Internet

R_D2

10.0.1.36/30

10.0.1.40/30

10.0.1.252/30

10.0.1.0/28

10.0.2.32/30

10.0.2.36/30

192.168.202.55/24

R_D1

10.0.1.20/30

R_U3

10.0.2.20/30

10.0.2.40/30

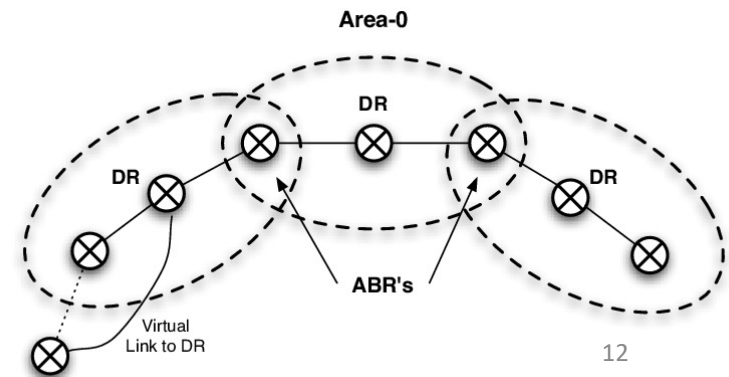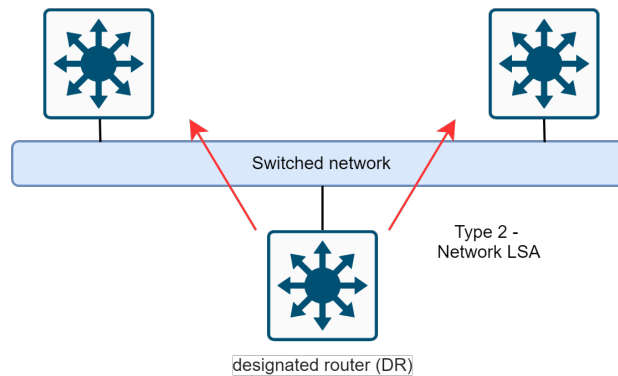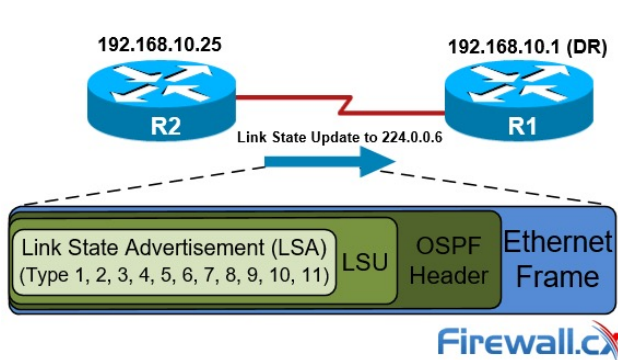R_D3

10.0.1.36/30

R_C2

10.3.1.0/24

11

OSPF

- Messages
  - Hello packets between routers to establish adjacency
  - Link state advertisement, link-state database
- Designated routers, network adjacency vs. point-to-point link
- Cost metric manually defined, or multiple of reference bandwidth
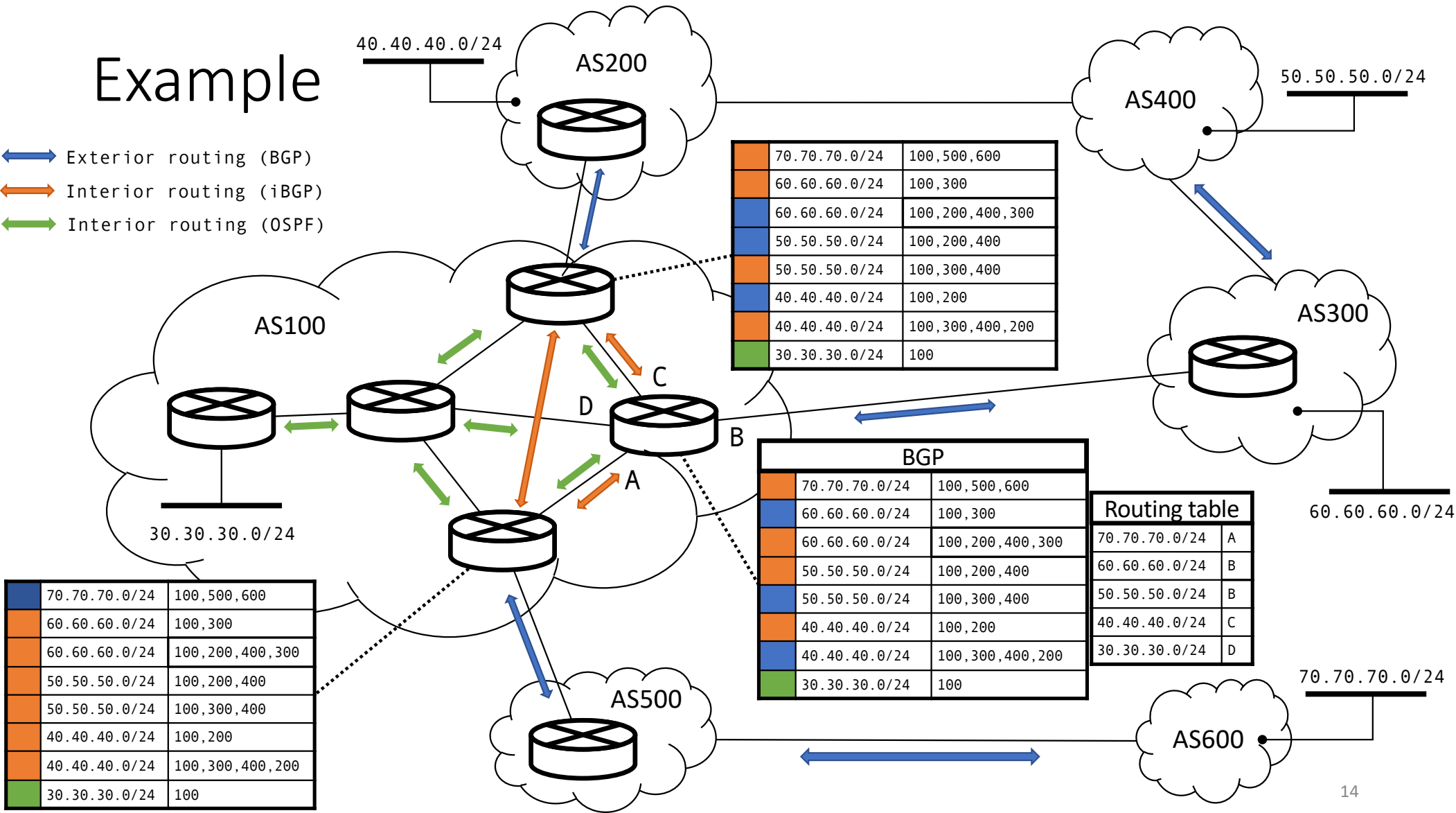- Backbone area, separate link-state database, Area Border Router

Exterior routing

- Different administrative domains
  - Each node in this network has a different administrator
  - The Internet
  - Each node is an Autonomous System – and a network by itself
- Protocols
  - BGP – path vector

Example

40.40.40.0/24

AS200

AS400

50.50.50.0/24

🟧	70.70.70.0/24	100,500,600
🟧	60.60.60.0/24	100,300
🟦	60.60.60.0/24	100,200,400,300
🟦	50.50.50.0/24	100,200,400
🟧	50.50.50.0/24	100,300,400
🟦	40.40.40.0/24	100,200
🟧	40.40.40.0/24	100,300,400,200
🟩	30.30.30.0/24	100

AS100

AS300

C

D

B

A

30.30.30.0/24

60.60.60.0/24

BGP

🟧	70.70.70.0/24	100,500,600
🟦	60.60.60.0/24	100,300
🟧	60.60.60.0/24	100,200,400,300
🟧	50.50.50.0/24	100,200,400
🟦	50.50.50.0/24	100,300,400
🟧	40.40.40.0/24	100,200
🟦	40.40.40.0/24	100,300,400,200
🟩	30.30.30.0/24	100

Routing table

70.70.70.0/24	A
60.60.60.0/24	B
50.50.50.0/24	B
40.40.40.0/24	C
30.30.30.0/24	D

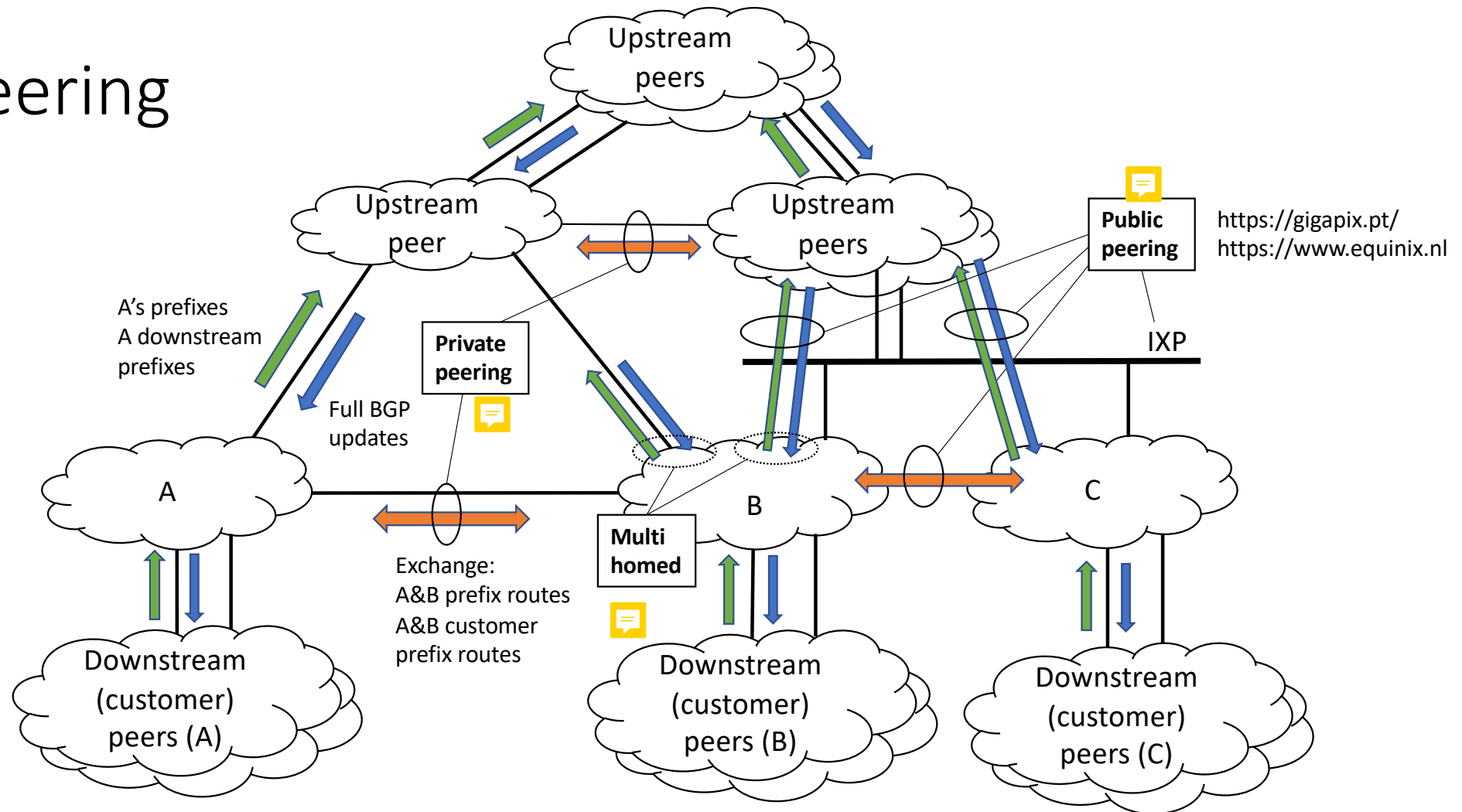|		
🟦	70.70.70.0/24	100,500,600
🟧	60.60.60.0/24	100,300
🟧	60.60.60.0/24	100,200,400,300
🟧	50.50.50.0/24	100,200,400
🟧	50.50.50.0/24	100,300,400
🟧	40.40.40.0/24	100,200
🟧	40.40.40.0/24	100,300,400,200
🟩	30.30.30.0/24	100

AS500

70.70.70.0/24

AS600

BGP protocol

- RFC 4271, BGP4
- Keep-alive messages, BGP update messages
- Exchange path-vector routes to other BGP routers
  - For each network aka network prefix
  - Select one route, announce it to neighbors, insert it in routing table
- Route map: set of rules to check what to do with route
  - Drop route, modify route and add it to routing table
- Route selection criteria
  - Weight (local to each router); AS-local preference for route (local to each AS), local route; length of AS path; MED multiple exit hint to other AS's, ...

BGP protocol

- TCP port 179
- Authentication, RFC 2385
  - MD5 hash of password and BGP data
- IP header TTL set to 1
  - Or more, if BGP connection over multiple IP hops
- Route aggregation
  - 32.32.32.0/24, 32.32.33.0/24 announced as 32.32.32.0/23
- eBGP vs. iBGP
  - e: between routers of different AS
  - i: between routers of same AS in multiple peer AS's, full mesh, route reflectors

Peering

Upstream peers

Upstream peer

Upstream peers

Public peering

https://gigapix.pt/
https://www.equinix.nl

A's prefixes
A downstream prefixes

Private peering

IXP

Full BGP updates

A

B

C

Exchange:
A&B prefix routes
A&B customer prefix routes

Multi homed

Downstream (customer) peers (A)

Downstream (customer) peers (B)

Downstream (customer) peers (C)

http://www.bgphelp.com/2017/01/05/bgp-types-of-peering/

Internet Tiers

- Tier 1
  - Settlement-free peering only
- Tier 2
  - Some free, other paying
- Tier 3
  - Paying only peering

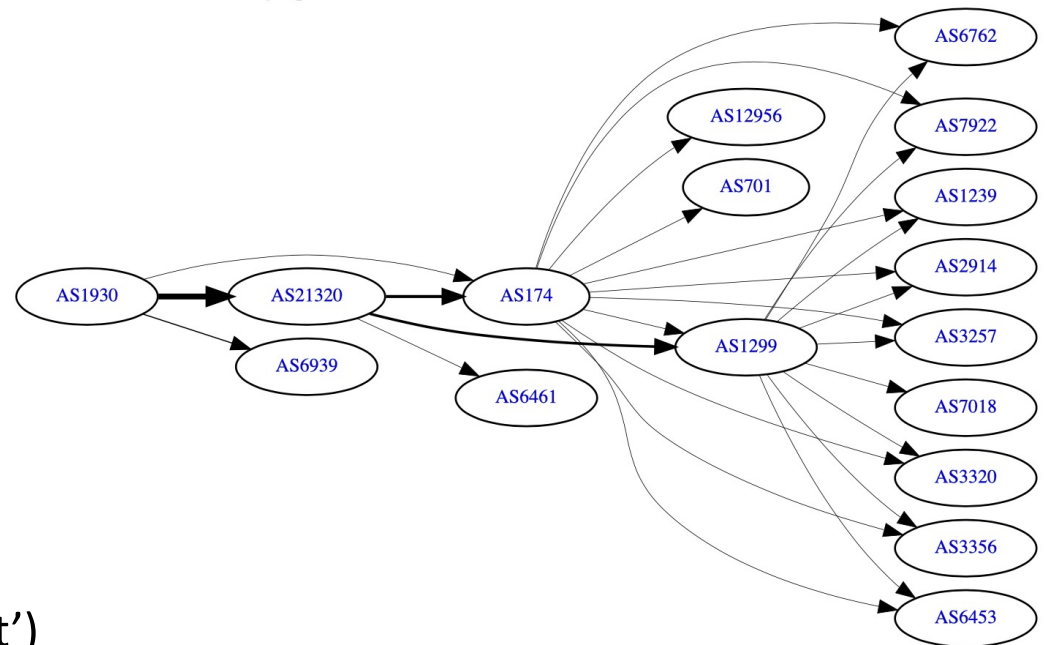https://en.wikipedia.org/wiki/Tier_1_network

Looking glass
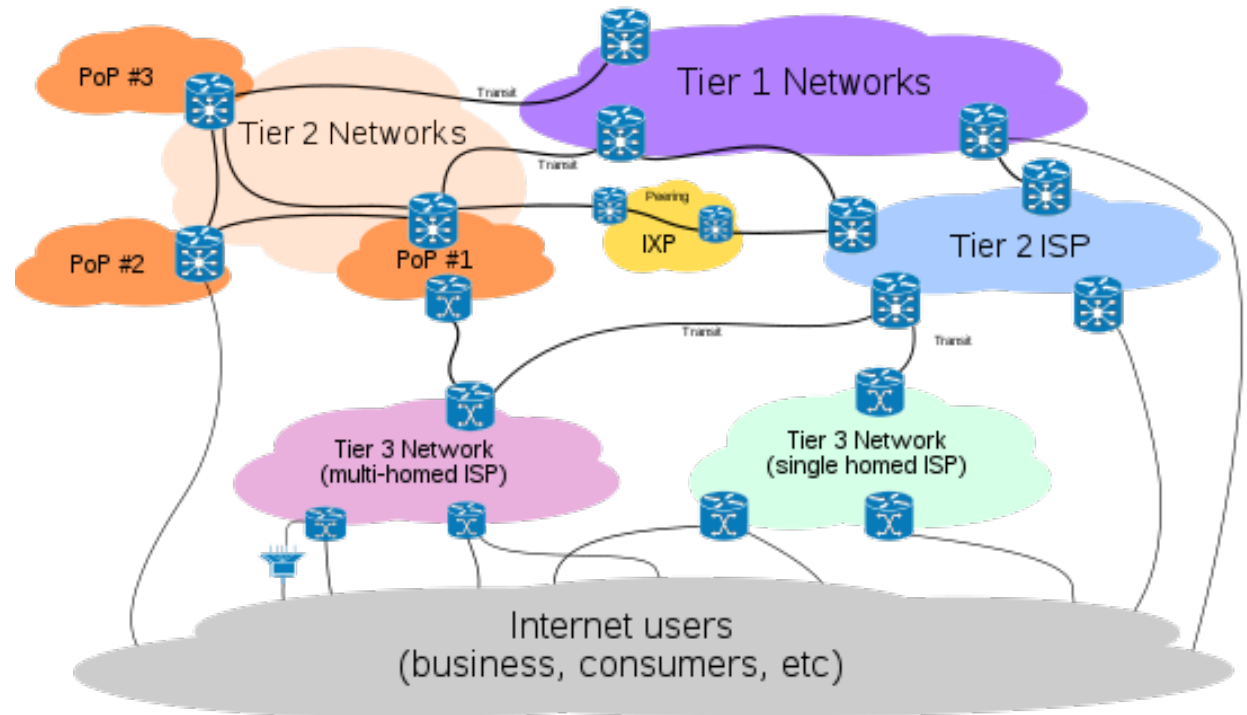
- http://www.routeviews.org
- https://bgp.he.net
- RCCN
  - AS 1930
  - https://bgp.he.net/AS1930
- Archive
  - RIBS and UPDATES
- pyasn
  import pyasn
  asndb = pyasn.pyasn('ipasn_20140513.dat')
  asndb.lookup('8.8.8.8')

https://bgp.he.net/AS1930#_graph4

AS1930 IPv4 Route Propagation

Management and Operations of Networks, Services, and Systems
Routing

Ricardo Morla

FEUP – GORS/M.EEC, GRS/M.EIC