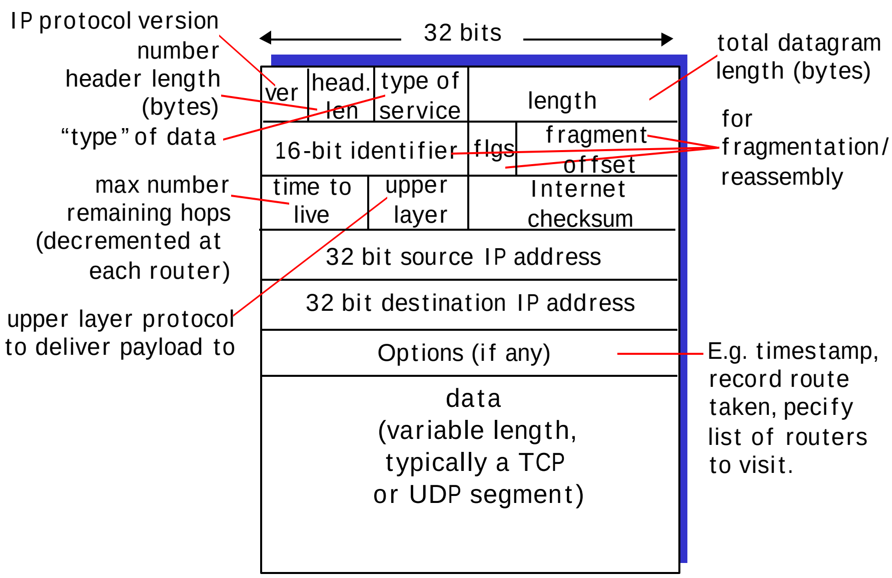
Routing in the Internet

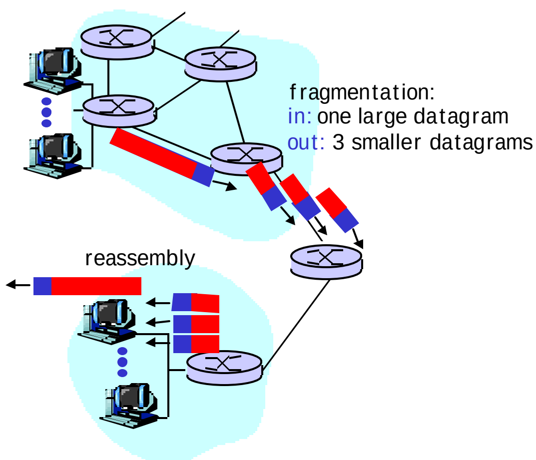
- The Global Internet consists of Autonomous Systems (AS) interconnected with each other:
  - **Stub AS**: small corporation
  - **Multihomed AS**: large corporation (no transit)
  - **Transit AS**: provider
- Two-level routing:
  - **Intra-AS:** administrator is responsible for choice
  - **Inter-AS:** unique standard

IP datagram format



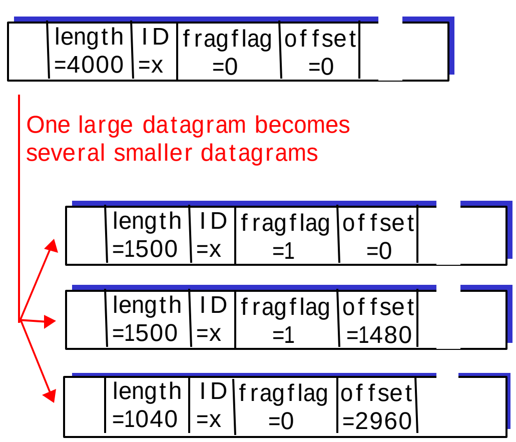
IP Fragmentation & Reassembly

- network links have MTU (max.transfer size) - largest possible link-level frame.
  - different link types, different MTUs
- large IP datagram divided (“fragmented”) within net
  - one datagram becomes several datagrams
  - “reassembled” only at final destination
  - IP header bits used to identify, order related fragments



4: Network Layer 4a-3

IP Fragmentation and Reassembly



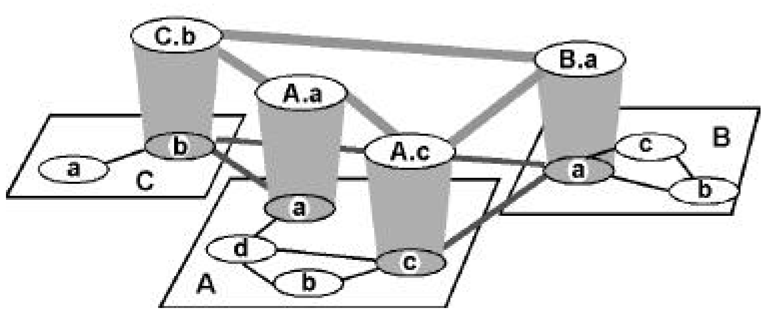
4: Network Layer 4a-4

ICMP: Internet Control Message Protocol

- used by hosts, routers, gateways to communication network-level information
  - error reporting: unreachable host, network, port, protocol
  - echo request/reply (used by ping)
- network-layer "above" IP:
  - ICMP msgs carried in IP datagrams
- ICMP message: type, code plus first 8 bytes of IP datagram causing error

Type	Code	description
0	0	echo reply (ping)
3	0	dest. network unreachable
3	1	dest host unreachable
3	2	dest protocol unreachable
3	3	dest port unreachable
3	6	dest network unknown
3	7	dest host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

Internet AS Hierarchy



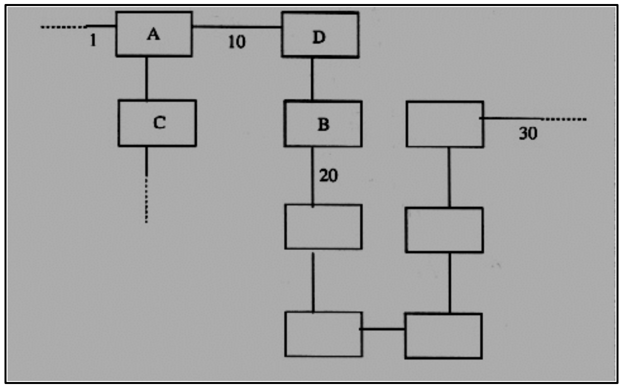
Intra-AS Routing

- Also known as Interior Gateway Protocols (IGP)
- Most common IGPs:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco propr.)

RIP (Routing Information Protocol)

- Distance vector type scheme
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
  - *Can you guess why?*
- Distance vector: exchanged every 30 sec via a Response Message (also called **Advertisement**)
- Each Advertisement contains up to 25 destination nets

RIP (Routing Information Protocol)



Destination Network	Next Router	Num. of hops to dest.
1	A	2
20	B	2
30	B	7
10	--	1
....	....	....

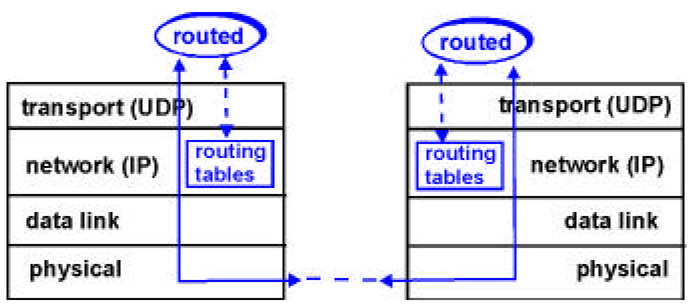
RIP: Link Failure and Recovery

- If no advertisement heard after 180 sec, neighbor/link dead
- Routes via the neighbor are invalidated; new advertisements sent to neighbors
- Neighbors in turn send out new advertisements if their tables changed
- Link failure info quickly propagates to entire net
- Poison reverse used to prevent ping-pong loops (infinite distance = 16 hops)

RIP Table processing

- RIP routing tables managed by an **application** process called route-d (daemon)
- Advertisements encapsulated in UDP packets (no reliable delivery required; advertisements are periodically repeated)

RIP Table processing



RIP Table example (continued)

RIP Table example
(at router *giroflee.eurocom.fr*):

- Three attached class C networks (LANs)
- Router only knows routes to attached LANs
- Default router used to “go up”
- Route multicast address: 224.0.0.0
- Loopback interface (for debugging)

RIP Table example

Destination	Gateway	Flags	Ref	Use	Interface
127.0.0.1	127.0.0.1	UH	0	26492	lo0
192.168.2.	192.168.2.5	U	2	13	fa0
193.55.114.	193.55.114.6	U	3	58503	le0
192.168.3.	192.168.3.5	U	2	25	qaa0
224.0.0.0	193.55.114.6	U	3	0	le0
default	193.55.114.129	UG	0	143454	

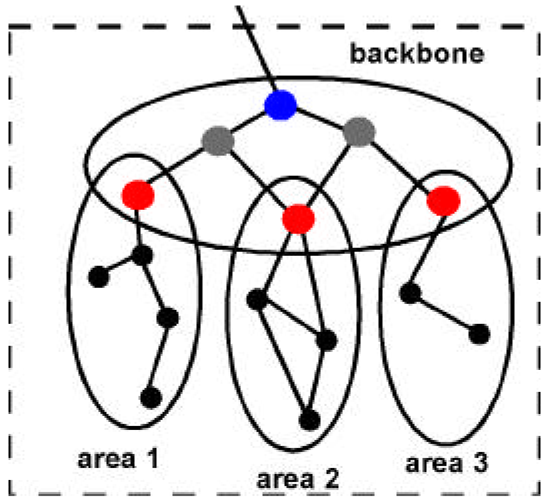
OSPF (Open Shortest Path First)

- "open": publicly available
- Uses the Link State algorithm
  - LS packet dissemination
  - Topology map at each node
  - Route computation using Dijkstra's alg
- OSPF advertisement carries one entry per neighbor router
- Advertisements disseminated to entire Autonomous System (via flooding)

OSPF "advanced" features (not in RIP)

- Security: all OSPF messages are authenticated (to prevent malicious intrusion); TCP connections used
- Multiple same-cost paths allowed (only one path in RIP)
- For each link, multiple cost metrics for different TOS (eg, satellite link cost set "low" for best effort; high for real time)
- Integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- Hierarchical OSPF in large domains.

Hierarchical OSPF



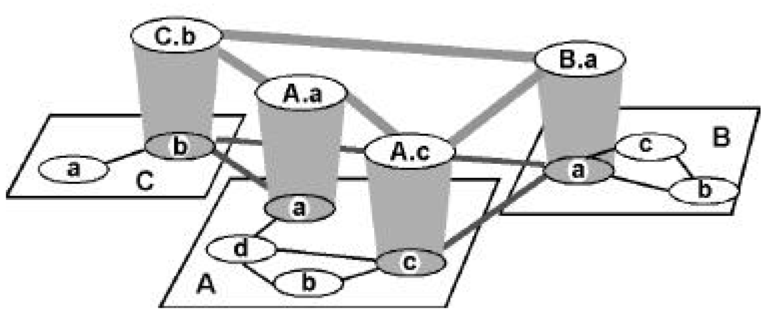
Hierarchical OSPF

- Two-level hierarchy: local area and backbone.
- Link-state advertisements do not leave respective areas.
- Nodes in each area have detailed area topology; they only know direction (shortest path) to networks in other areas.
- **Area Border routers** "summarize" distances to networks in the area and advertise them to other Area Border routers.
- **Backbone routers** run an OSPF routing alg limited to the backbone.
- **Boundary routers** connect to other ASs.

IGRP (Interior Gateway Routing Protocol)

- CISCO proprietary; successor of RIP (mid 80s)
- Distance Vector, like RIP
- several cost metrics (delay, bandwidth, reliability, load etc)
- uses TCP to exchange routing updates
- routing tables exchanged only when costs change
- Loop-free routing achieved by using a Distributed Updating Alg. (DUAL) based on *diffused computation*
- In DUAL, after a distance increase, the routing table is *frozen* until all affected nodes have learned of the change.

Inter-AS routing



Inter-AS routing (cont)

- BGP (Border Gateway Protocol): the de facto standard
- **Path Vector** protocol: and extension of Distance Vector
- Each Border Gateway broadcast to neighbors (peers) the entire path (ie, sequence of ASs) to destination
- For example, Gateway X may store the following path to destination Z:

Path (X,Z) = X,Y1,Y2,Y3,… ,Z

Inter-AS routing (cont)

- Now, suppose Gwy X send its path to peer Gwy W
- Gwy W may or may not select the path offered by Gwy X, because of cost, policy ($$$$) or loop prevention reasons.
- If Gwy W selects the path advertised by Gwy X, then:

Path (W,Z) = w, Path (X,Z)

Note: path selection based not so much on cost (eg,# of AS hops), but mostly on administrative and policy issues (e.g., do not route packets through competitor's AS)

Inter-AS routing (cont)

- Peers exchange BGP messages using TCP.
- OPEN msg opens TCP connection to peer and authenticates sender
- UPDATE msg advertises new path (or withdraws old)
- KEEPALIVE msg keeps connection alive in absence of UPDATES; it also serves as ACK to an OPEN request
- NOTIFICATION msg reports errors in previous msg; also used to close a connection

Why different Intra- and Inter-AS routing ?

- **Policy**: Inter is concerned with policies (which provider we must select/avoid, etc). Intra is contained in a single organization, so, no policy decisions necessary
- **Scale**: Inter provides an extra level of routing table size and routing update traffic reduction above the Intra layer
- **Performance**: Intra is focused on performance metrics; needs to keep costs low. In Inter it is difficult to propagate performance metrics efficiently (latency, privacy etc). Besides, policy related information is more meaningful.

We need **BOTH**!