E6998-02: Internet Routing

Lectures 11–14
OSPF

John Ioannidis
AT&T Labs – Research
ji+ir@cs.columbia.edu

Announcements

- Homework 2 out 9/29, due 10/15.
- Homework 3 out 10/15, due 10/29.
- Midterm on October 22.

OSPF

- More accurately: OSPFv2.
  - v1 was never really deployed.
- Link-state IGP, "open", based on Dijkstra's SPF algorithm.
- RFC2328 (and many others).
- Recommended IGP, esp. in a multivendor environment.
- Several features in common with other LS protocols
  - IS-IS, NLSP, PNNI.
  - We may look into IS-IS if time permits.
  - We'll point out some things that other protocols do better.

- Basis for other IETF LS protocols:
  - MOSPF.
  - OSPF for IPv6.

OSPF Properties

- Reduced LSA distribution overhead.
  - ***Areas*** limit the extent of flooding.
  - Multicast limits impact on broadcast networks.
  - OSPF goes (mostly) quiet when there are no route changes.
- 16-bit dimensionless metric.
- Equal-cost load balancing.
- Route aggregation.
  - CIDR, VLSM, etc.
- Route tagging.
- Authentication.

OSPF Overview

- Neighbor discovery:
  - ***Hello*** packets sent on all OSPF-enabled interfaces.
  - ***Neighbors***: routers on same link that agree on certain hello parameters.
  - ***Adjacencies***: virtual point-to-point links between certain neighbors over which routing information is exchanged.
- Link State Advertisements (***LSA***s):
  - Multiple LSA types.
  - Sent over all adjacencies.
  - List all of router's interfaces and the state of all links.
  - Flooded throughout an area.
  - Recorded in Link State Database and forwarded to neighbors.

OSPF Overview (cont'd)

- Designated Router / Backup Designated Router.
  - Two of the routers on a multiaccess link.
  - Used to reduce overall traffic on the link.
- When LSDB is complete:
  - Shortest Path Tree is computed on each router
    - (using Dijkstra's SPF algorithm).
  - Forwarding table built from SPT.
- Keep quiet:
  - Hellos are exchanged as link keepalives.
  - LSAs are retransmitted every 30 minutes.

OSPF Network Types

- Point to point links.
  - High and low speed PPP links.
- Broadcast networks.
  - Ethernet-like.
- Non-Broadcast Multiple Access (NBMA) networks.
  - ATM, Frame Relay, X.25, (tunnels).
- Point-to-multipoint.
  - Really, special configuration of NBMA networks.
  - Used on NMBA networks where not all stations on NMBA can talk directly to each other.
- Virtual links.
  - OSPF-specific meaning of the term.
  - Effectively, unnumbered point-to-point links.

Some Multicast Addresses

- 224.0.0.5 AllSPFRouters OSPF-ALL.MCAST.NET
- 224.0.0.6 AllDRouters OSPF-DSIG.MCAST.NET

- FF02::5 and FF02::6, respectively for OSPFv3.

- While we are at it:
- 224.0.0.1 ALL-SYSTEMS.MCAST.NET
- 224.0.0.2 ALL-ROUTERS.MCAST.NET
- 224.0.0.9 RIP2-ROUTERS.MCAST.NET
- 224.0.0.10 IGRP-ROUTERS.MCAST.NET

- Look up some more (with `dig -x` *address*).

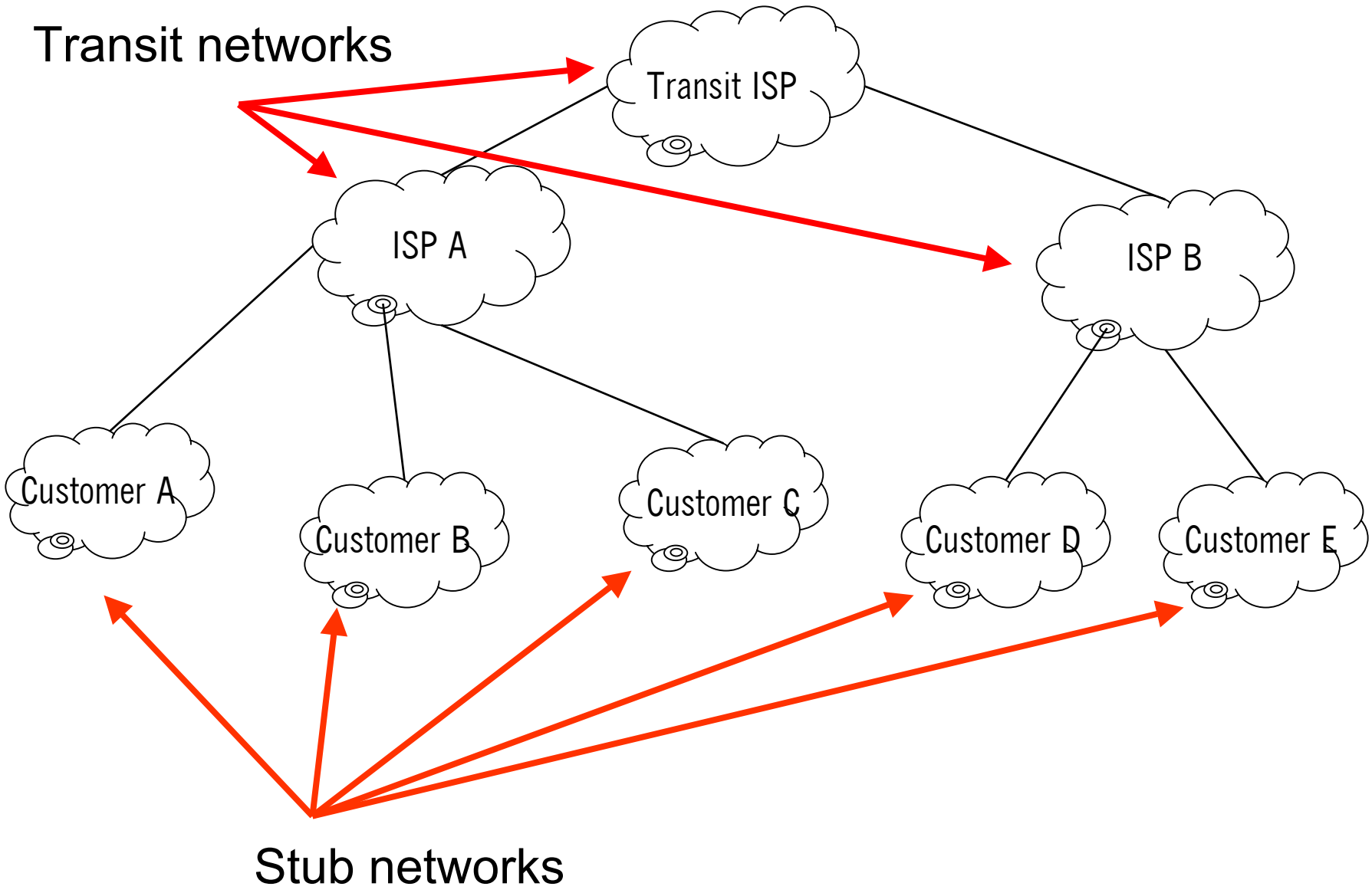
Destination Addresses Used

- On point-to-point networks:
  - No need to elect a DR.
  - Neighbors always become adjacent.
  - All OSPF packets except retransmitted LSAs sent to `AllSPFRouters` (224.0.0.5).
- On Broadcast networks:
  - DR and BDR are elected.
  - Packets sent to `AllSPFRouters` (224.0.0.5):
    - Hello packets.
    - All packets originating from the DR and BDR.
  - Packets sent to `AllDRouters` (224.0.0.6):
    - All packets sent by the rest of the routers.
    - Since these should only go to the DR/BDR.

Destination Addresses Used (cont'd)

- On NBMA networks:
  - DR and BDR are elected.
  - Extra configuration is needed to acquire neighbors.
  - All packets are unicast (no point in multicasting them).
- On Point-to-Multipoint networks:
  - These are treated as a collection of point-to-point links.
  - No DR/BDR are elected.
    - no need to.
  - Packets are multicast.
    - This way you don't have to find the address of the machine on the other side of the link.
- Virtual Links:
  - Packets are multicast.

Reminder: Transit *vs.* Stub Networks



Hello Protocol

- Sent every *HelloInterval* (default: 10s).
- Neighbor discovery.
- Parameter announcement/discovery.
  - No negotiation!
- Used as keepalive.
  - Dead after *RouterDeadInterval* (default: *4*HelloInterval*).
- Establishes bi-directional communication.
- On broadcast and NBMA networks:
  - Elects DRs and BDRs ([Backup] Designated Routers).

Hello Packet Contents

- ***Router ID*** of originating router (32 bits):
  - Highest IP address on loopback interfaces.
  - If no lb, highest IP address on regular interfaces.
  - Unchanged even if interfaces go down.

The rest of the fields pertain to the originating *interface*.

- ***Area ID*** (32 bits):
  - Area ID 0 is the ***backbone*** area.
- Checksum (16 bits).
- Authentication type (16 bits) and information (64 bits).
  - None, cleartext (bad!), or keyed hash.
  - The hash is appended to the packet and is not considered part of the packet for checksumming purposes.

Hello Packet Contents (cont'd)

- ***HelloInterval*** (16 bits).
- ***RouterDeadInterval*** (32bits).
- Options (6 of 8 bits).
- Router Priority (8 bits).
  - Used in DR election.
- DR and BDR (32 bits each).
  - 0.0.0.0 if no router has been elected.
- List of *neighbors*.
  - Router IDs of *neighbors*.

Options

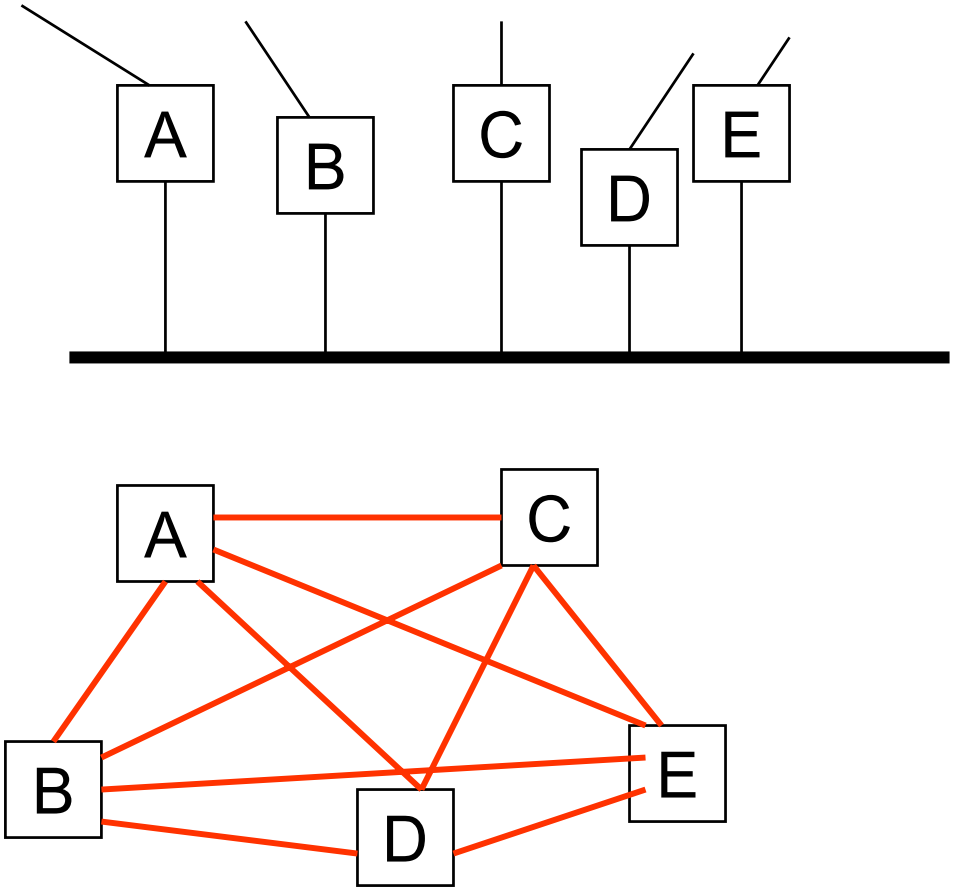
- DC bit: Router is capable of supporting OSPF over demand circuits.
- EA bit: Router is capable of sending and receiving External Attributes (type 8) LSAs.
- N bit: Router can support NSSA LSAs. N=1 implies E=0.
- P bit: (Same position as N bit). ABR should translate a type 7 into a type 5 LSA.
- MC: Used by MOSPF.
- E: Router is capable of accepting AS External LSAs.
  - In hello packets, indicates ability to send/receive Type 5.
- T: capable of supporting TOS.

Hello Packet Processing

- Receiving routers (on same link) check:
  - AreaID, Authentication, Netmask, HelloInterval, RouterDeadInterval, and Options.
  - If they don't match its own, packet is dropped.
- If RouterID is known to the receiving interface:
  - RouterDeadInterval timer is reset.

  else
  - RouterID is added to the table of known neighbors.
- If receiving router sees its own ID in the list of of neighbors in the hello packet, it knows that it has bi-directional communication with the sender.
- Adjacencies may now be formed, if appropriate.
  - Depends on network type.

Adjacencies on Broadcast Networks



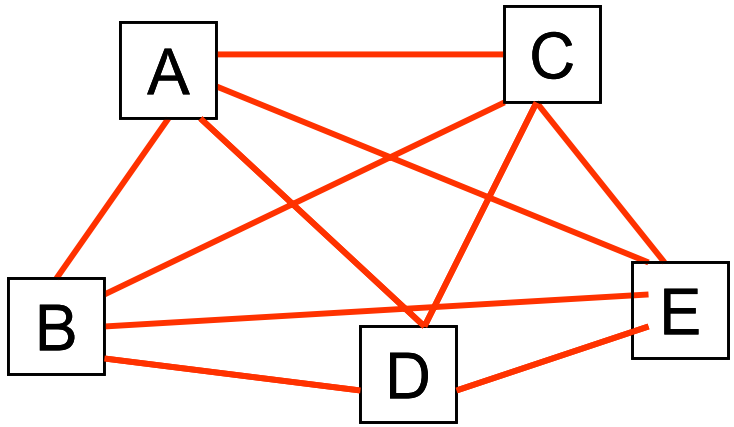
- If n routers are on a bc link, n(n-1)/2 adjacencies could be formed.
- n^2 LSAs would be originating from this network (why?).

Adjacencies, cont'd

- If routers formed pairwise adjacencies:
  - Each would originate (n-1)+1=n LSAs for the link.
  - Out of the network, n*n LSAs would be emanating.
- Routers would also send received LSAs to their adjacencies.
  - Multiple (n-1) copies of each LSA present on the network.
  - Even with multicast, (n-1) responses would still result.

- To prevent this, a Designated Router is elected.
  - Routers form adjacencies only with DR.
  - Link acts as a (multi-interface) virtual router as far as the rest of the area is concerned.

Adjacencies, cont'd



- One router is selected as the DR.
- Actually, another is selected as the BDR.
  - If the DR fails, we want the BDR to take over within RouterDeadInterval rather than go over a new election.
    - During which no traffic would be forwarded.
- Routers form adjacencies with both DR and BDR.
- DR and BDR also form adjacencies with each other.

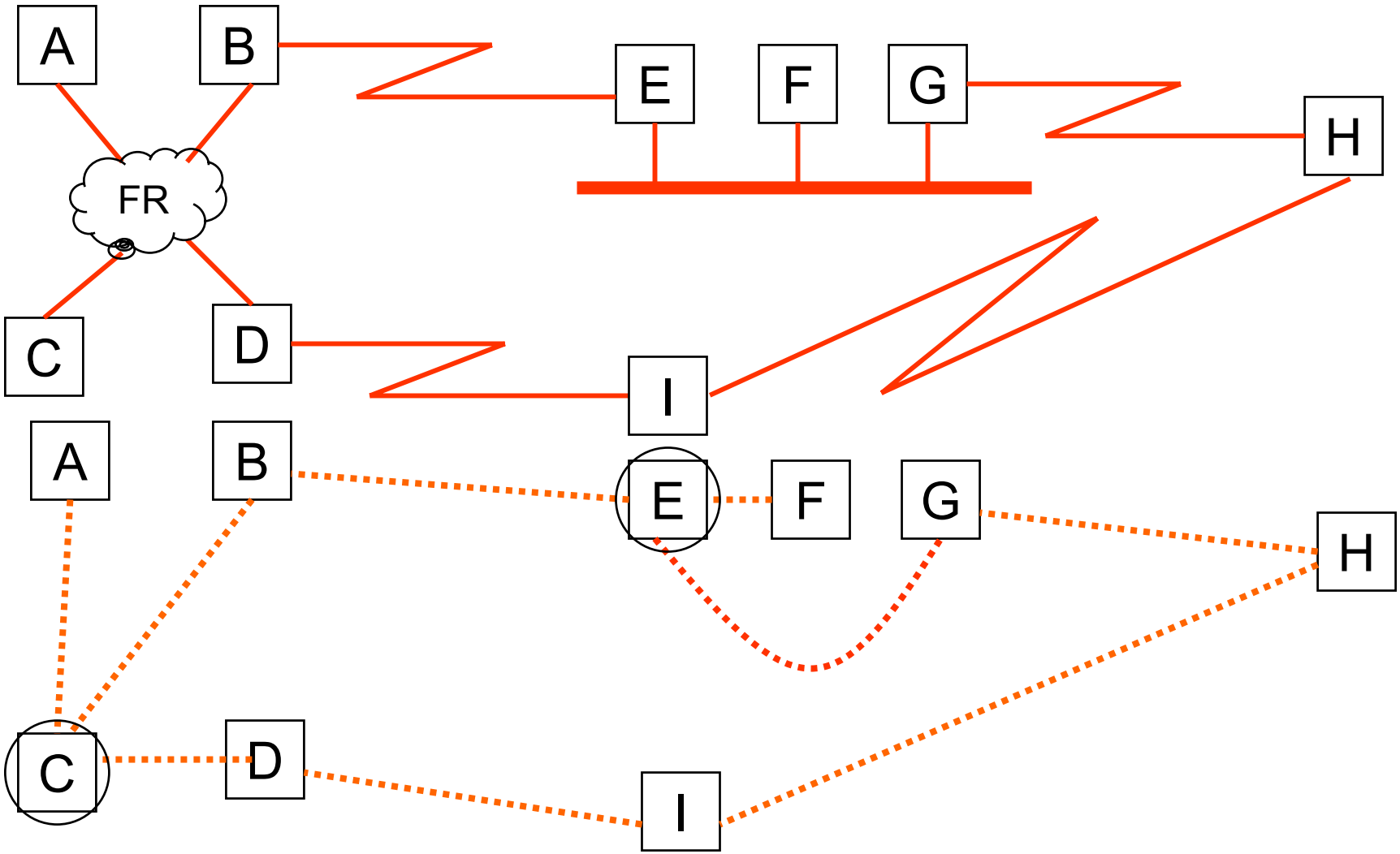
DR Election

- When router joins in:
  - Listen to hellos; if DR and BDR advertised, accept it.
    - This is the case if all Hello packets agree on who the DR and BDR are.
    - Unlike IS-IS, status quo is not disturbed!
- If there is no elected BDR, router with highest priority becomes BDR.
- Ties are broken by highest RouterID.
  - RouterIDs are unique (IP address of lb if).
- If there is no DR, BDR is promoted to DR.
- New BDR is elected.

DR Election Details

- Routers who believe can be BDRs or DRs put their own IDs in their Hello packets.
- Once 2-way communication has been established, all routers know who the candidates are.
- They can now all pick a BDR.
  - Highest priority, then Router ID.
- And then a DR.
- If only one router claims he's the DR, he becomes the DR.
- First two routers to come up become the DR and BDR.

Adjacencies, cont'd



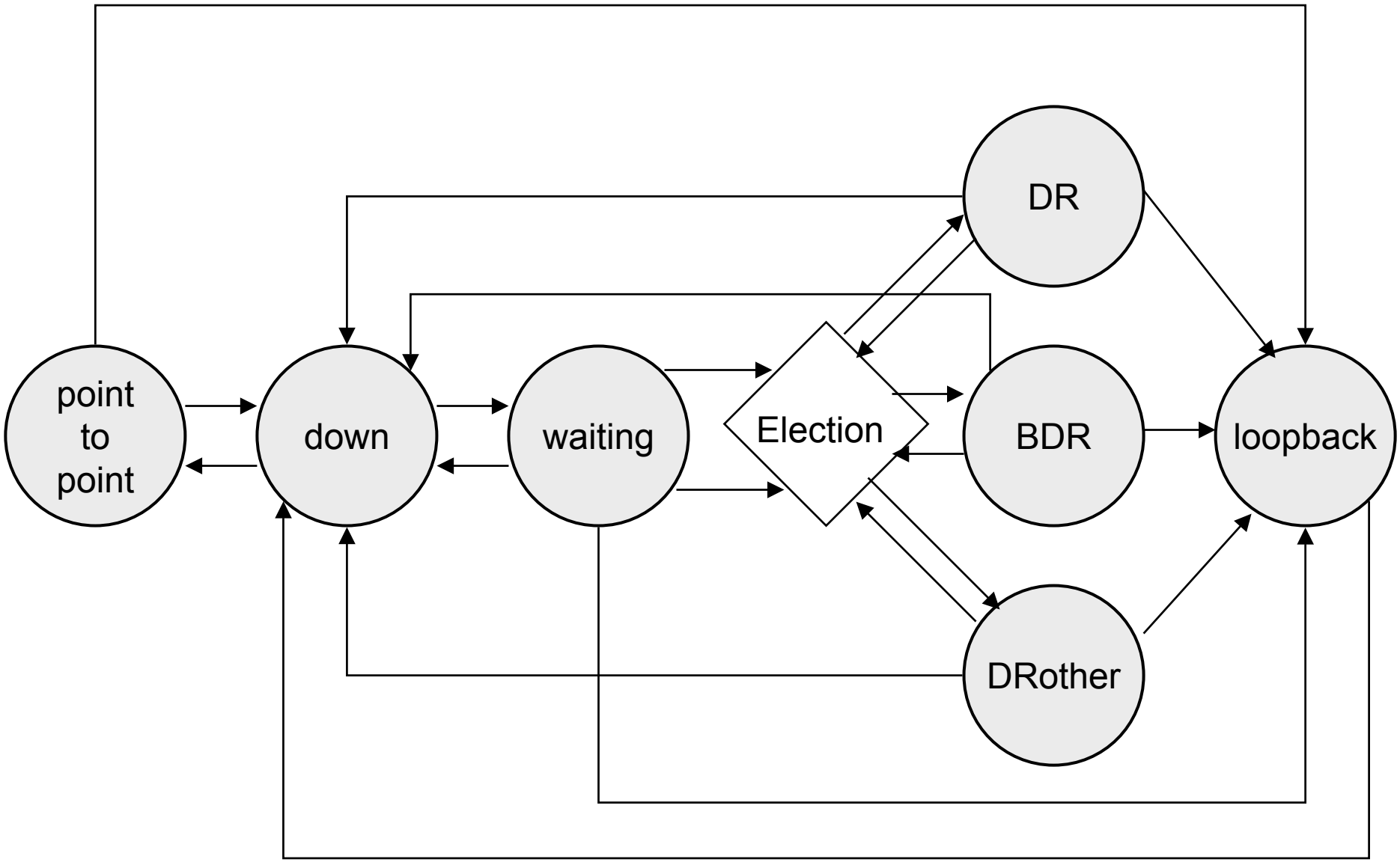
Routers connected by data links ⇔ nodes connected by adjacencies.

OSPF Interface Data Structure

- IP Address and Mask
- Area ID
- Router ID
- Network Type
- Cost
- Interface Transit Delay
- State
- Priority
- DR
- BDR

- Hello Interval
- Hello Timer
- Router Dead Interval
- Wait Timer
  - Before DR selection
- Rxmit Interval
  - Ack packets
- Neighbors
- Auth type
- Auth key

OSPF Interface State Machine



OSPF Neighbors

- Form adjacencies.
- Pass routing information over them.

- Adjacency establishment:
  - Neighbor discovery.
  - Bidirectional communication.
    - Neighbors listed in each other's Hello packets.
  - [DR election].
  - Database synchronization.
    - Ensure neighbors have identical LS information.
  - Full adjacency.
- Neighbor State Machine: read about it in RFC2328.

OSPF Neighbor Data Structure

Relationship of router with its neighbors.

- Interface
- Area ID
- Neighbor ID
- Neighbor IP Address
- Neighbor Priority
- Neighbor Options
- DR/BDR
- Master/Slave
- State
- Poll Interval (NBMA only)
- Inactivity Timer
- DD sequence number
- Last received DDP
- DB Summary list
- LS Retransmission list
- LS Request list

Database Synchronization

- Last step before full adjacency.
- Neighbors exchange *summaries* of each LSA they have.
- Master/Slave relationship to determine who starts:
  - Router with highest RouterID.
- Database Description packet:
  - OSPF Header: RouterID, AreaID, Checksum, Auth.
  - Interface MTU. Options.
  - I(nitial), M(ore), M(aster)/S(lave) bits.
  - DD Sequence Number.
  - LSA Header:
    - Age, Options, Type (of LSA).
    - Link State ID (meaning varies by LSA Type).
    - Advertising Router, Sequence Number.

Full Adjacency

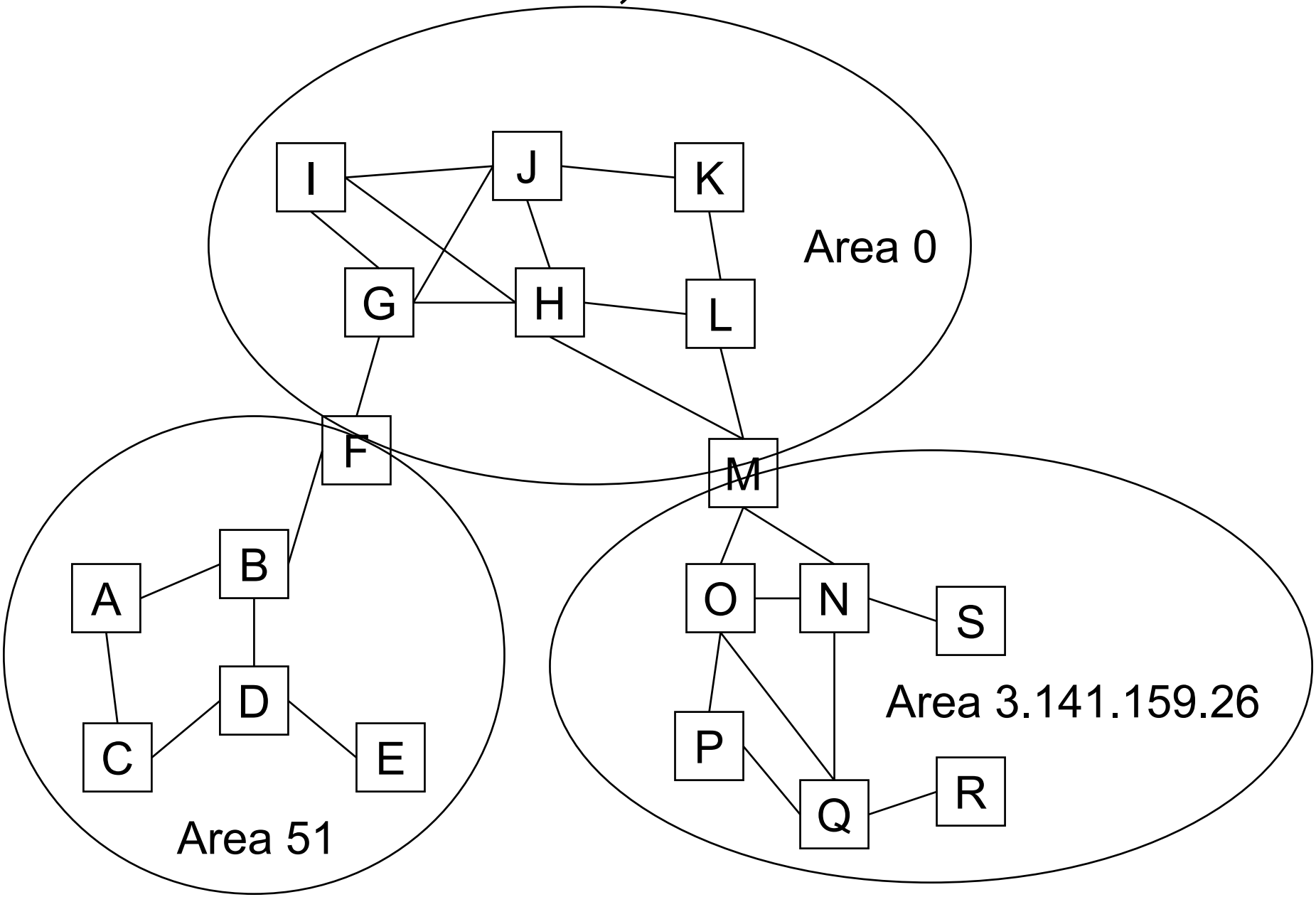
- After DDs have been exchanged, routers know what LSAs they are missing.

- LSA Requests.
- LSA Updates.
- LSA Acknowledgements (implicit or explicit).

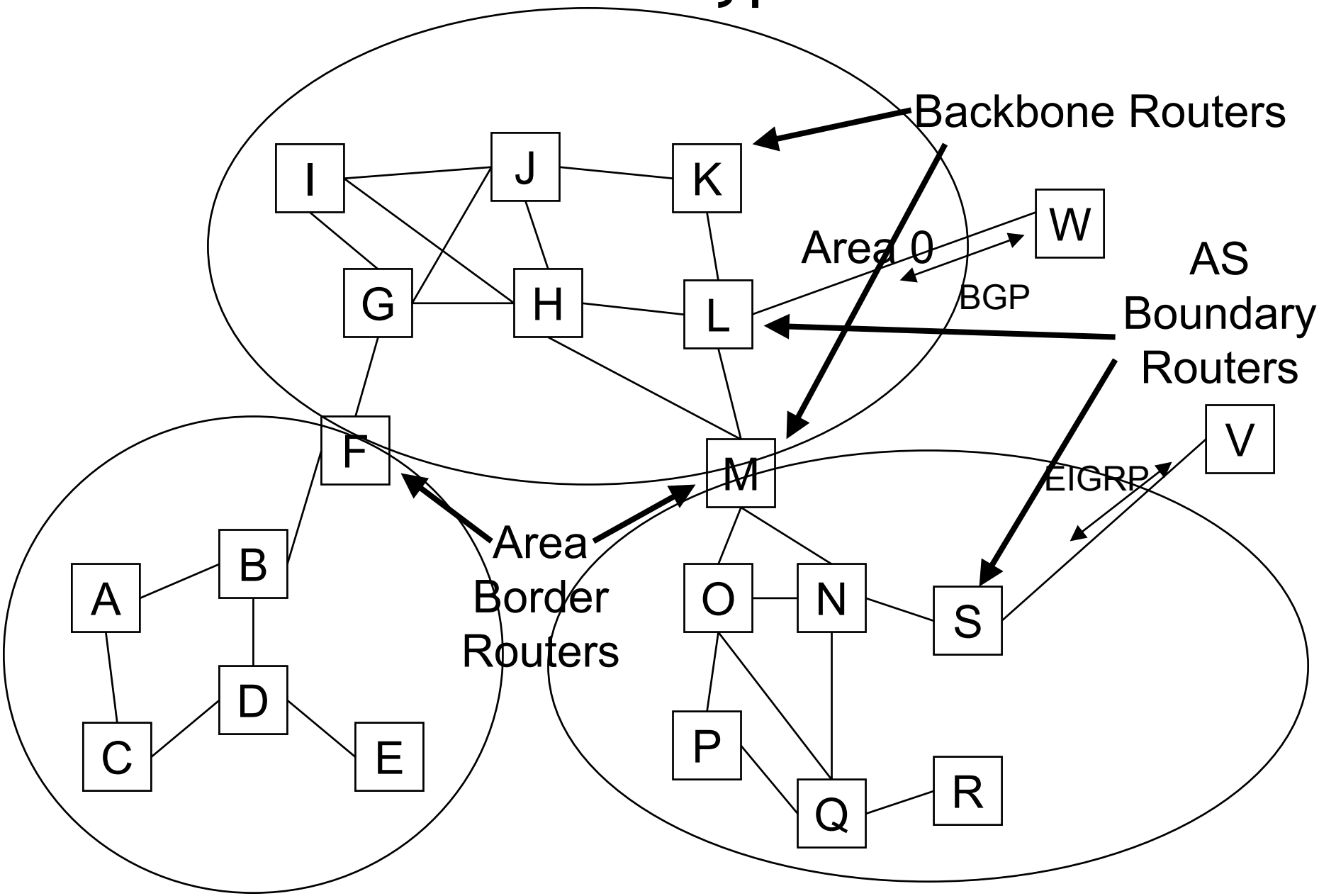
Areas

- An AS (or Routing Domain) is divided into Areas.
- Group of routers.
- “Close” to each other.
- Reduce the extent of LSA flooding.
- Intra-area traffic.
- Inter-area traffic.
- External traffic.
  - Injected from a different AS.

Areas, cont'd



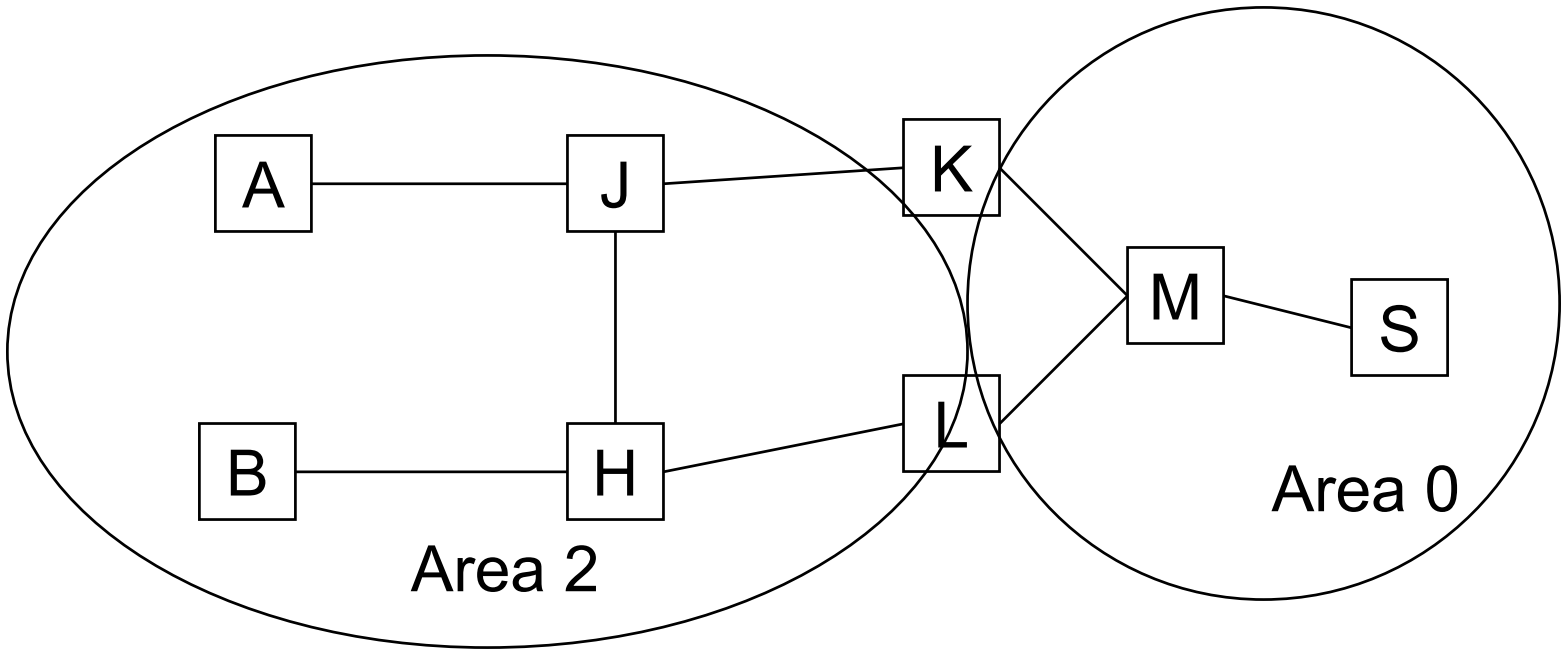
Router Types



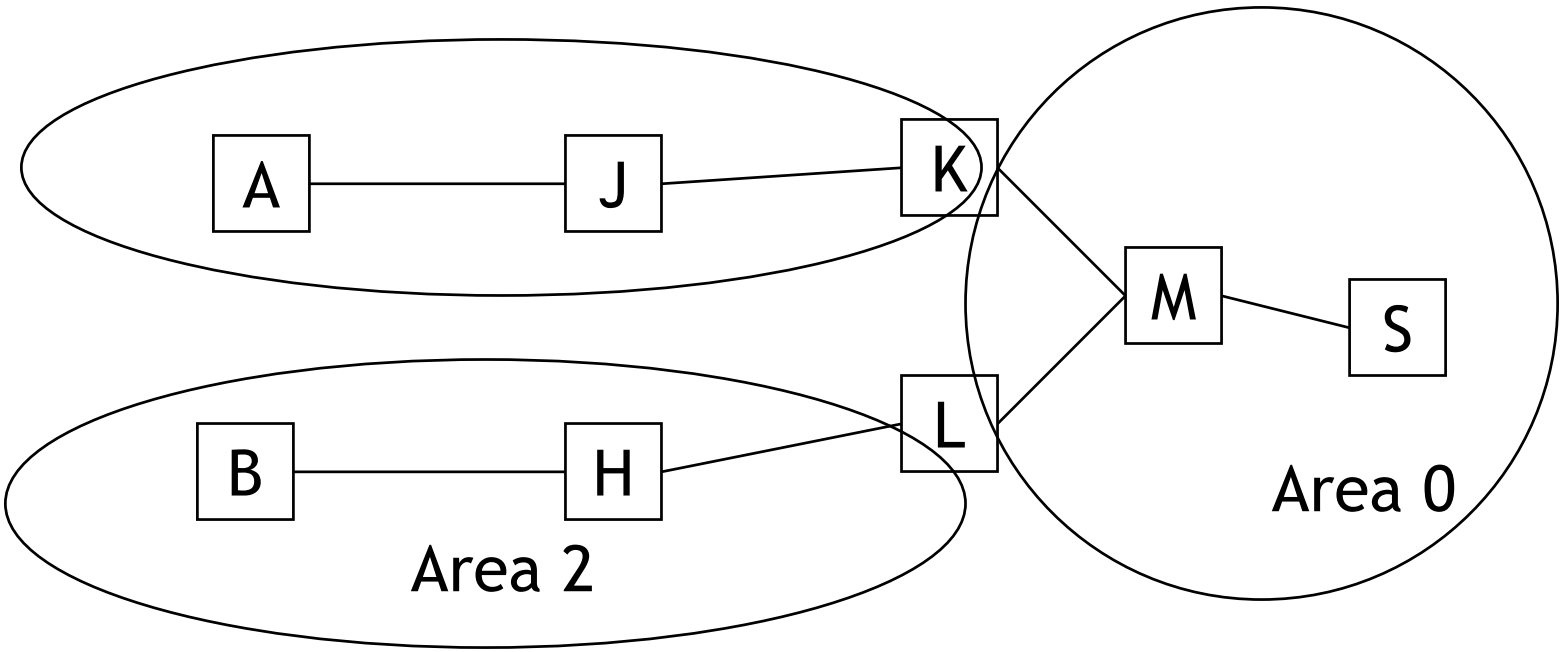
Area Partitions

- Link and router failures can cause areas to *partition*.
- Some partitions are *healed* automatically.
- Some need manual intervention.
  - Virtual Links.
- Isolated area: link failure results in no path to the rest of the network.
  - Obviously, cannot be healed at all.
  - Redundancy is important!

Partitions Include an ABR

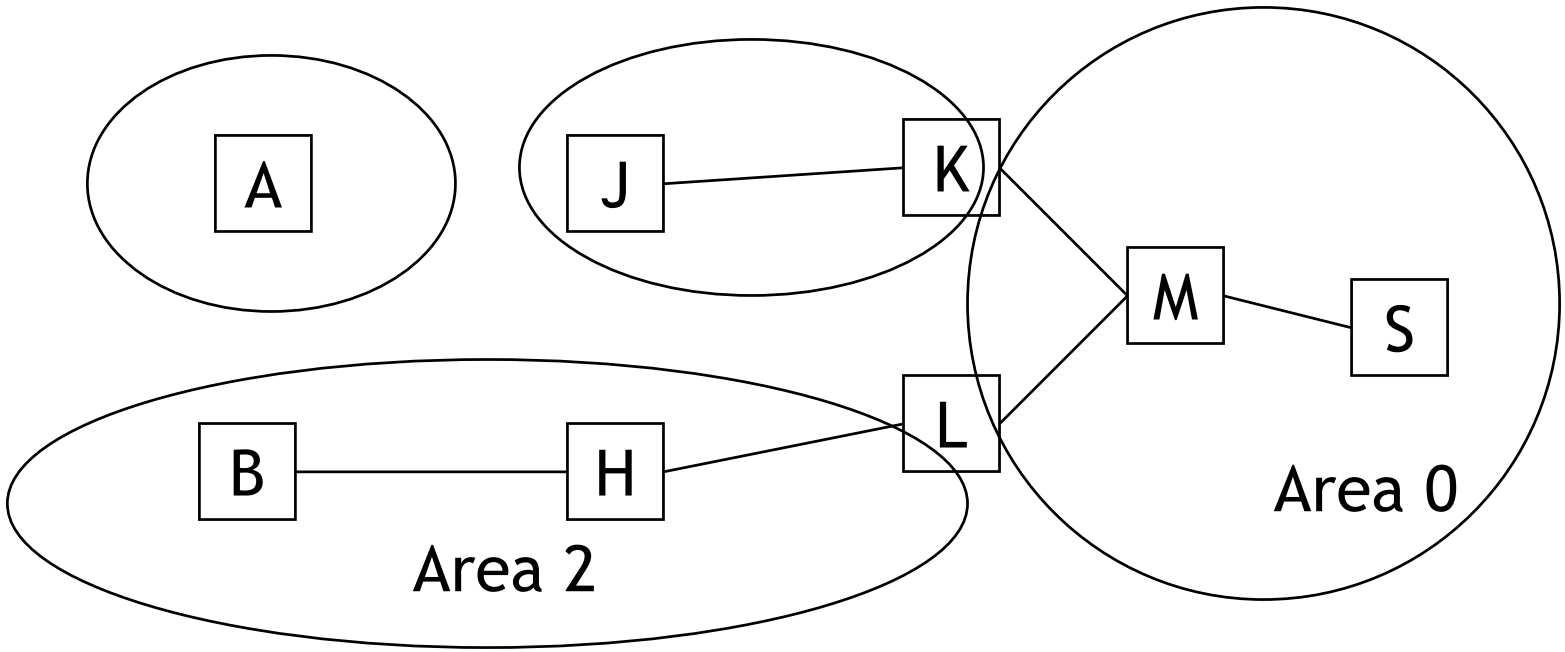


Partitions Include an ABR



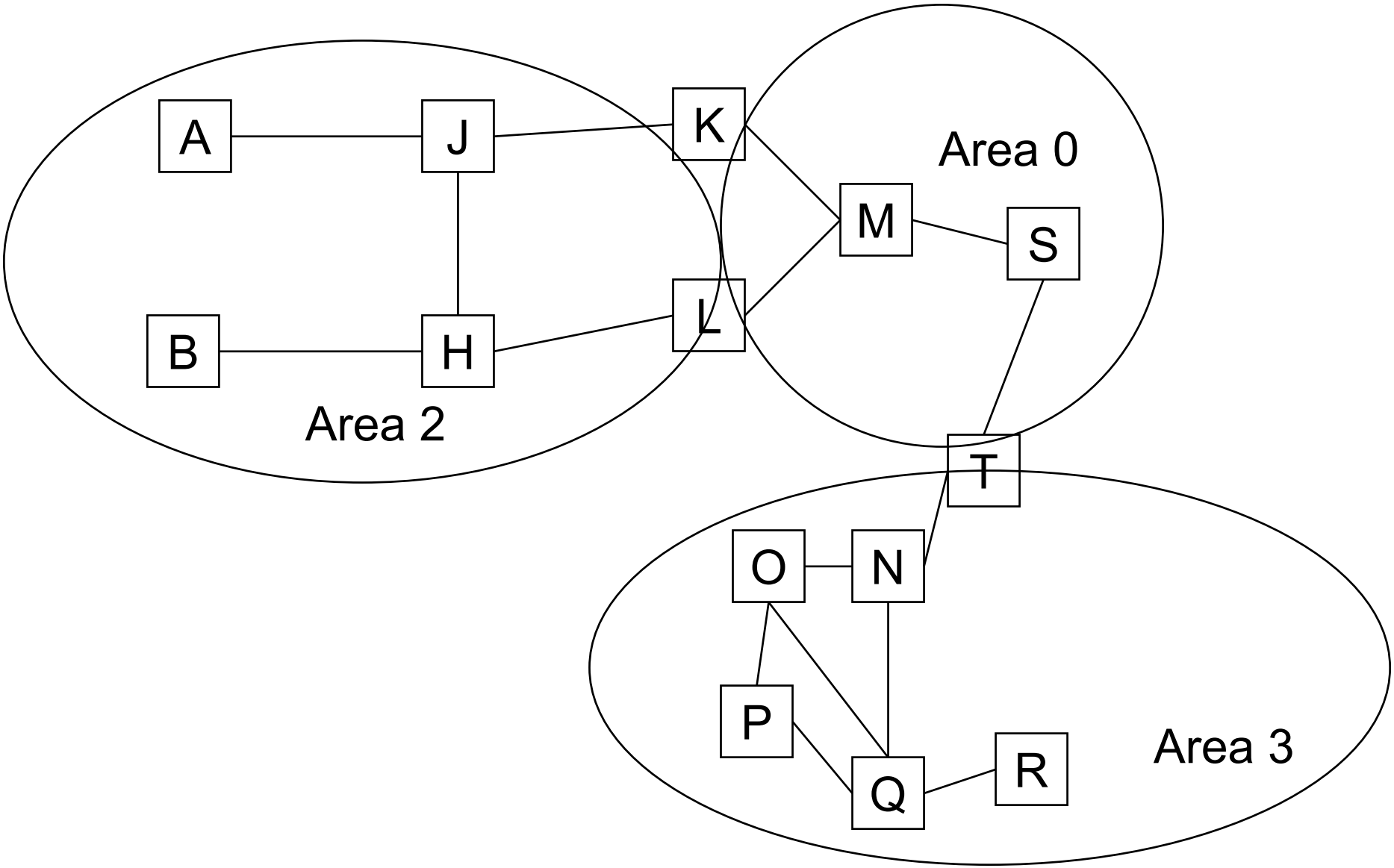
Area 2 gets partitioned, but all its routers can reach an ABR, so traffic is not disrupted.

Isolated area

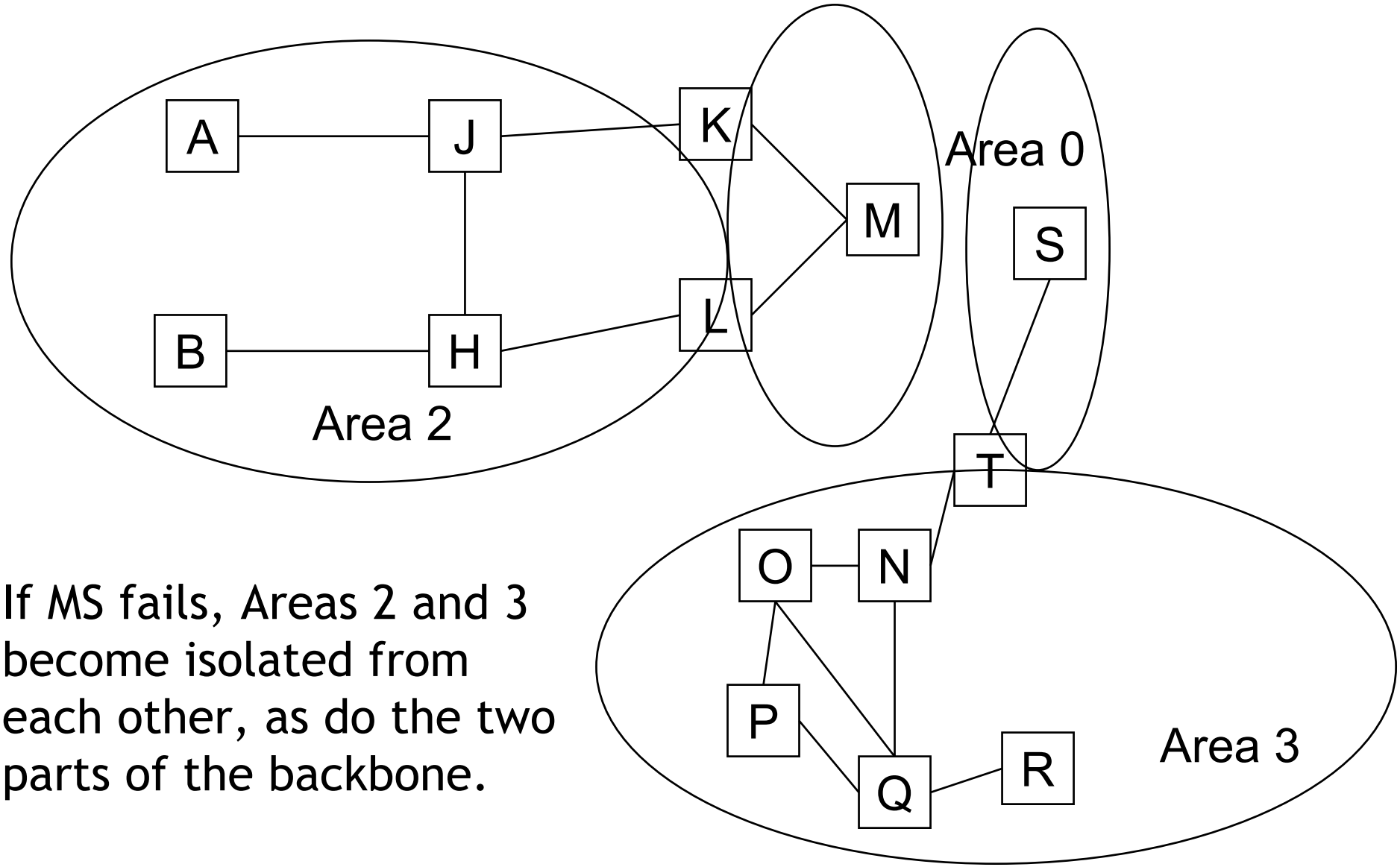


If AJ fails, A becomes isolated.

Backbone Partition?

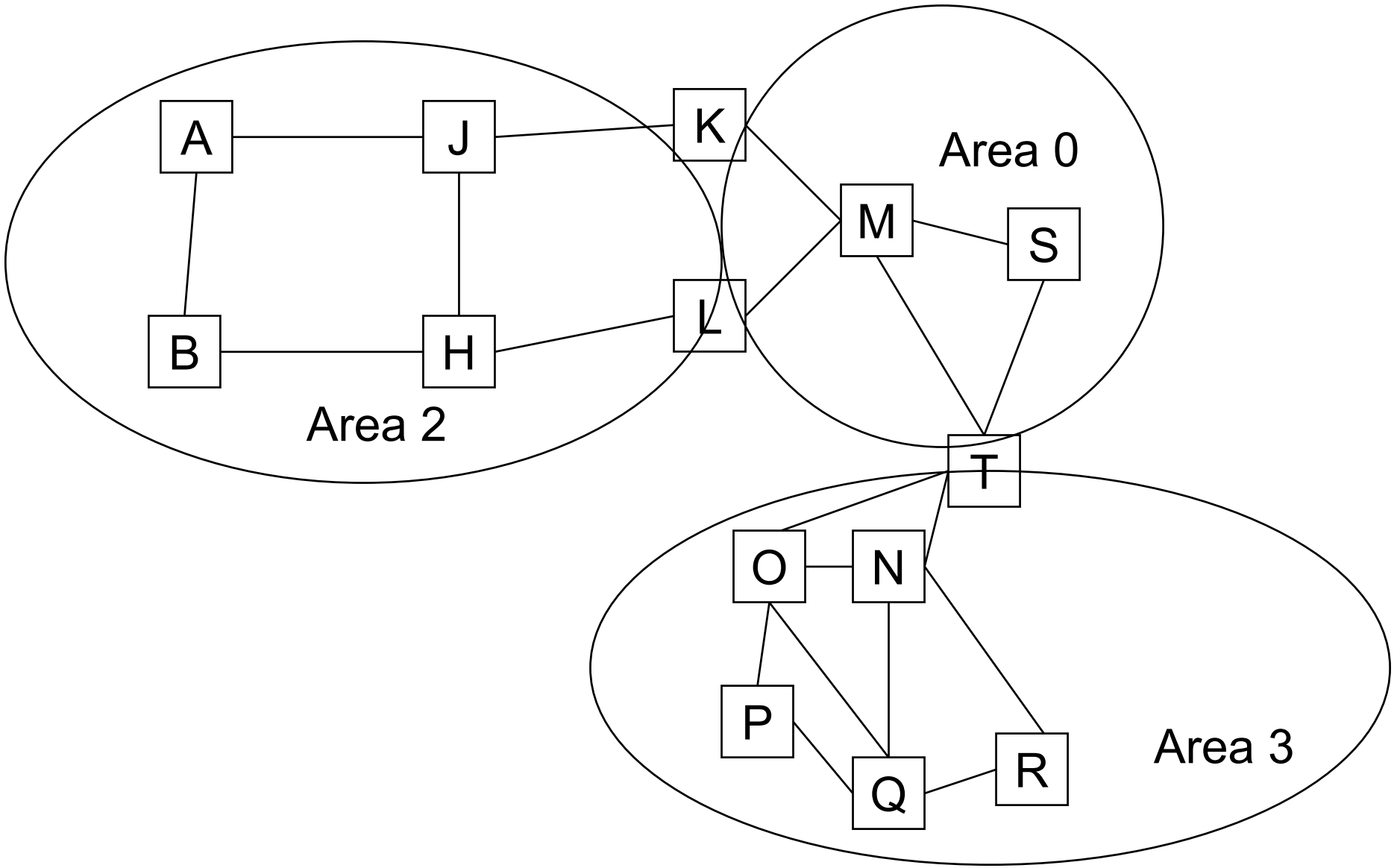


Backbone Partition



If MS fails, Areas 2 and 3 become isolated from each other, as do the two parts of the backbone.

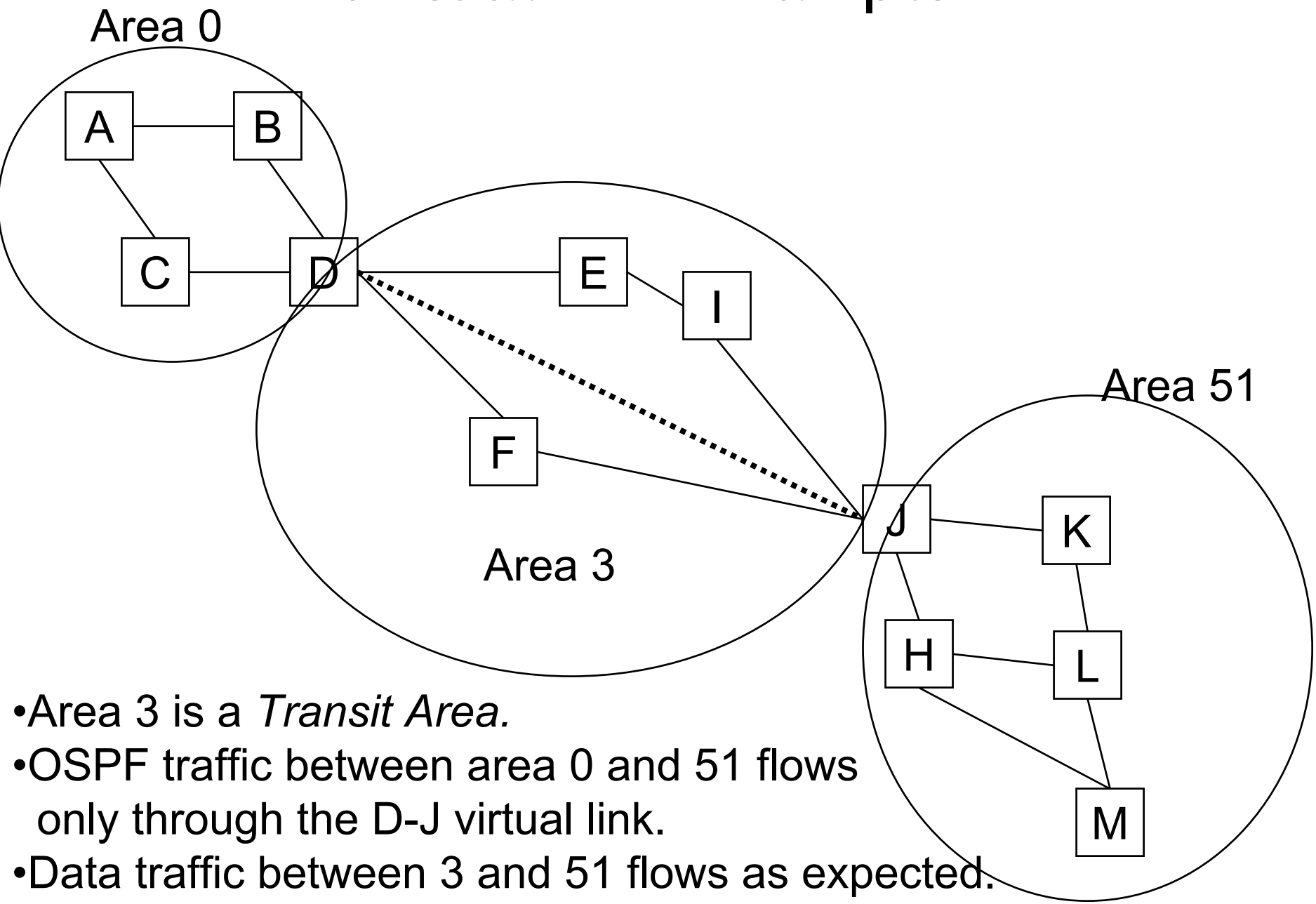
Redundancy is good



Virtual Links

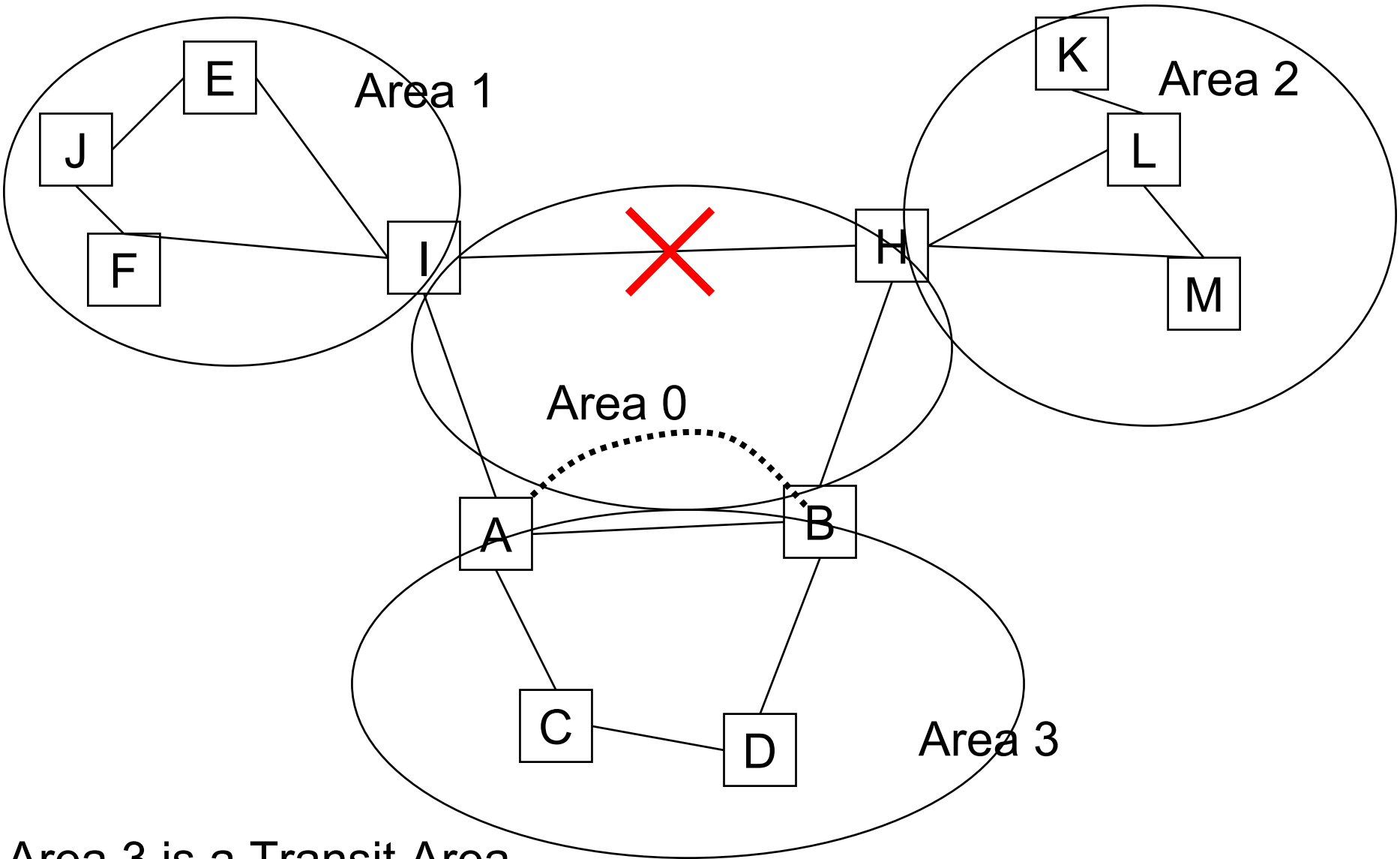
- Link to the backbone through a non-backbone area.
- Unnumbered (unaddressed).
- Connect an area to the BB through a non-BB area.
- Heal a partitioned BB through a non-BB area.
- No physical wires.
  - Exists solely as a result of configuration.
  - An example of a tunnel implemented without encapsulation.
- Configured between two ABRs.
- Transit Area: area through which VL is configured.
- Routers "connected" with VLs become adjacent.

Virtual Link Example 1



- Area 3 is a *Transit Area.*
- OSPF traffic between area 0 and 51 flows only through the D-J virtual link.
- Data traffic between 3 and 51 flows as expected.

Virtual Link Example 2



Area 3 is a Transit Area

Flooding

- Link State Database: list of all LSAs the router has heard (and sent).
- Change in topology results in new or changed LSAs.
- Changed LSAs are flooded throughout the network:
  - Link State Updates.
  - Link State Acknowledgements.
- Each LSA reaches every router.
- Updates/Acks only flow between adjacent routers
  - i.e., *it's not the update packets that get flooded, it's their contents (the LSAs).*

Updates

- On point-to-point networks, multicast to `AllSPFRouters`.
- On broadcast networks:
  - DRothers multicast updates to `AllDRouters`.
  - The DR then multicasts an update to `AllSPFRouters`.
  - If the DR fails to do that, BDR takes over, otherwise BDR stays silent.
- On NBMA networks:
  - DRothers unicast updates to DR and BDR.
  - DR unicasts updates to all adjacent routers.
    - (multicast/broadcast, if present, is simulated in NBMA networks).

Reliable flooding

- Transmitted LSAs must be acked.
- Implicit acks: send the same LSA back.
  - Used when you would have sent it anyway.
- Explicit acks: OSPF packet type 5.
  - Carry only LSA header.
- When sending an LSA, put it in a retransmission queue in the neighbor data structure.
  - Retransmitted every RxmtInterval (or until adj. is broken).
- Delayed acks: more LSAs acked in a single update packet.
- Direct acks: sent immediately and are unicast.
  - When duplicate LSA received from neighbor.
  - Rxed LSA has MaxAge and router has no copy of it.

Sequence numbers

- Linear sequence number space.
  - Signed 32bit integers.
  - Start at `InitialSequenceNumber` (0x80000001).
  - End at `MaxSequenceNumber` (0x7fffffff).
- First LSA goes out with `InitialSequenceNumber`.
- Each new LSA adds 1 to the previous sequence number.
- If is `MaxSequenceNumber` reached:
  - LSA must be flushed out of other routers' list.
  - LSA is sent out with MaxAge.
  - When all neighbors (adj.) have acked, flush LSA and create new one.

Age

- Age of LSA in seconds.
- Unsigned 16-bit integer.
  - From 0 to `MaxAge` (3600).
- Set to 0 by originating router.
- At each router transit, incremented by `InfTransDelay`.
- Also incremented as it resides in database.
- When LSA reaches `MaxAge`, it is reflooded so it can be eliminated from the network.
- When the originating router wants to flush an LSA, it sets the age to `MaxAge` and floods it.
- LSAs are refreshed every `LSRefreshTime` (1800s).
  - With Sequence Number incremented by 1.
  - LSA group pacing.

LSA Comparison

- Highest sequence number is newest.
- Else highest checksum is newest.
  - Differing checksums with same sequence number imply corruption.
  - If the "newest" LSA is corrupt, it will make it back to the originating router, which will then flood a new LSA with the next sequence number .
  - Else if one of the ages is `MaxAge`, it is newest.
- Else if ages differ by more than 15 minutes (`MaxAgeDiff`), lowest age is newest.
  - Router clocks may berunning at slightly different speeds.
  - Different paths cause same LSA to arrive with slightly different ages.
- Else LSAs are the same.
- An LSA in a router is replaced when a "newer" one is received.

LSA Types

1. Router
2. Network
3. Network Summary
4. ASBR Summary
5. AS External
6. Group Membership
7. NSSA External
8. External Attributes
9. Opaque (link-local scope)
10. Opaque (area-local scope)
11. Opaque (AS scope)

Router LSA

- Produced by every router.
- Flooded within an area.

- List of all of router's links (interfaces)
  - Point-to-point links (real or virtual).
  - Stub networks (networks the router serves).
- Type (=1)
- RouterID
- Number of links
- Link Descriptions (i/f address, link type, metric).

Network LSA

- Produced by the DR on MA networks.
- Flooded within an area.

- Represent the multiaccess network.
  - (MA network acts as a pseudonode).

- Type (=2)
- Network address and netmask.
- Addresses of attached routers.

Network Summary LSA

- Produced by Area Border Routers.
- Sent into an area to advertise prefixes outside that area.
  - One per destination (prefix).
  - If multiple paths known, lowest-cost LSA is advertised.
- When a NS LSA is received, the cost of the route to the ABR is added to the cost advertised in the NS LSA.
  - Distance-vector behavior!

- Type (=3)
- Prefix
- Metric

AS Boundary Router Summary LSA

- Produced by ABRs.
- Identical to NS (type 3) LSAs.
  - Advertise (host) routes to ASBRs.
  - Destination is a host address, prefix length is 32.

- Type (=4)
- ASBR IP address and mask (all-ones).
- Metric.

AS External LSA

- Produced by ASBRs.
- Advertise a destination (or a default route) external to the AS.
- Flooded throughout the AS (but not stub areas).
  - Since they are not associated with a particular area!

- Type (=5)
- Advertised prefix.
- Forwarding address (of external router).
  - A type 4 LSA has already informed us of how to reach the ASBR!
- Metric.

Other LSAs

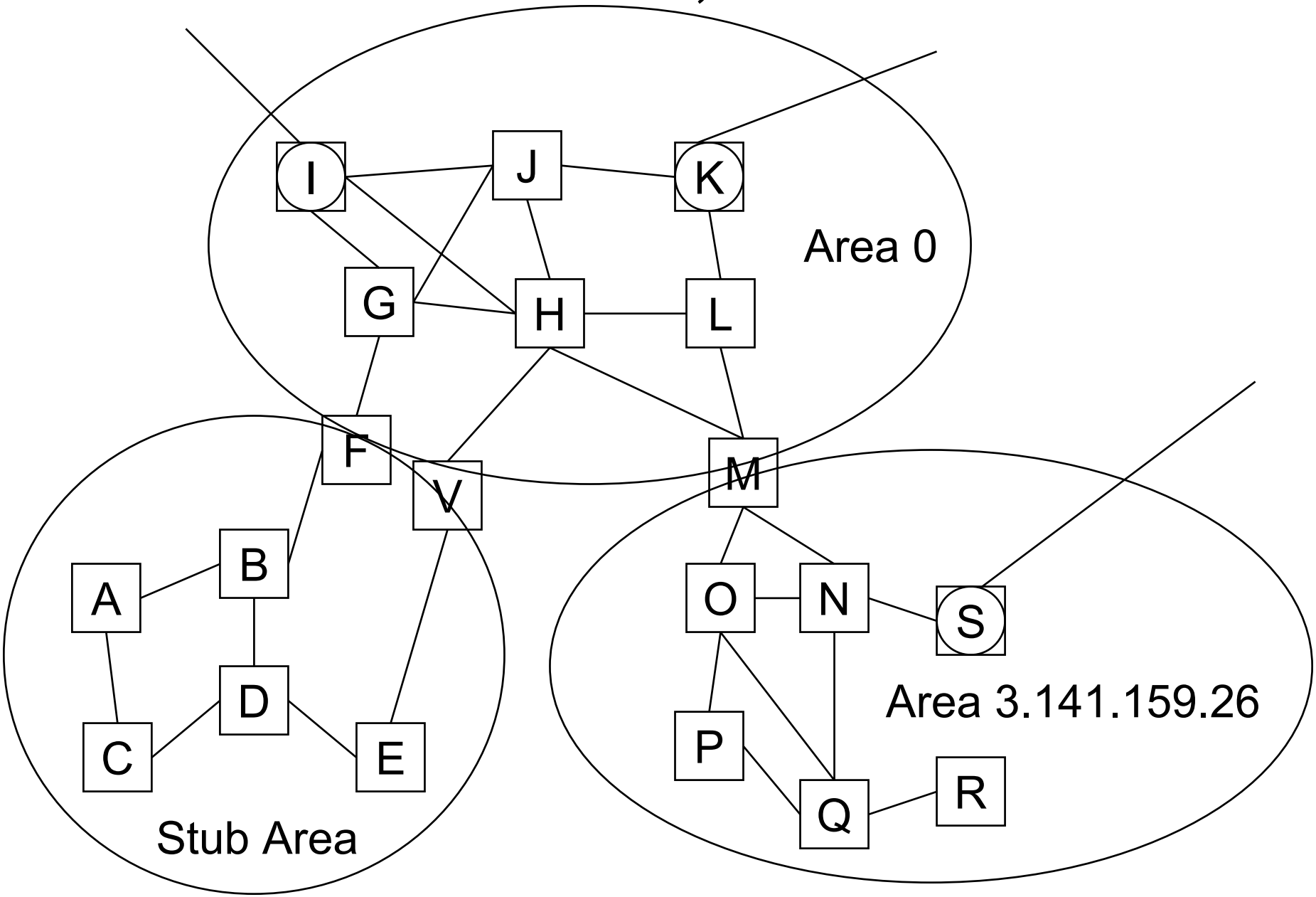
- Group membership.
  - Used for MOSPF.
- NSSA External.
  - Like AS External, but only flooded within the NSSA.
- External attributes.
  - Proposed as an alternative to IBGP.
- Opaque.
  - Proposed so that OSPF can be used to carry app-specific data to all routers in an AS.

Stub Areas

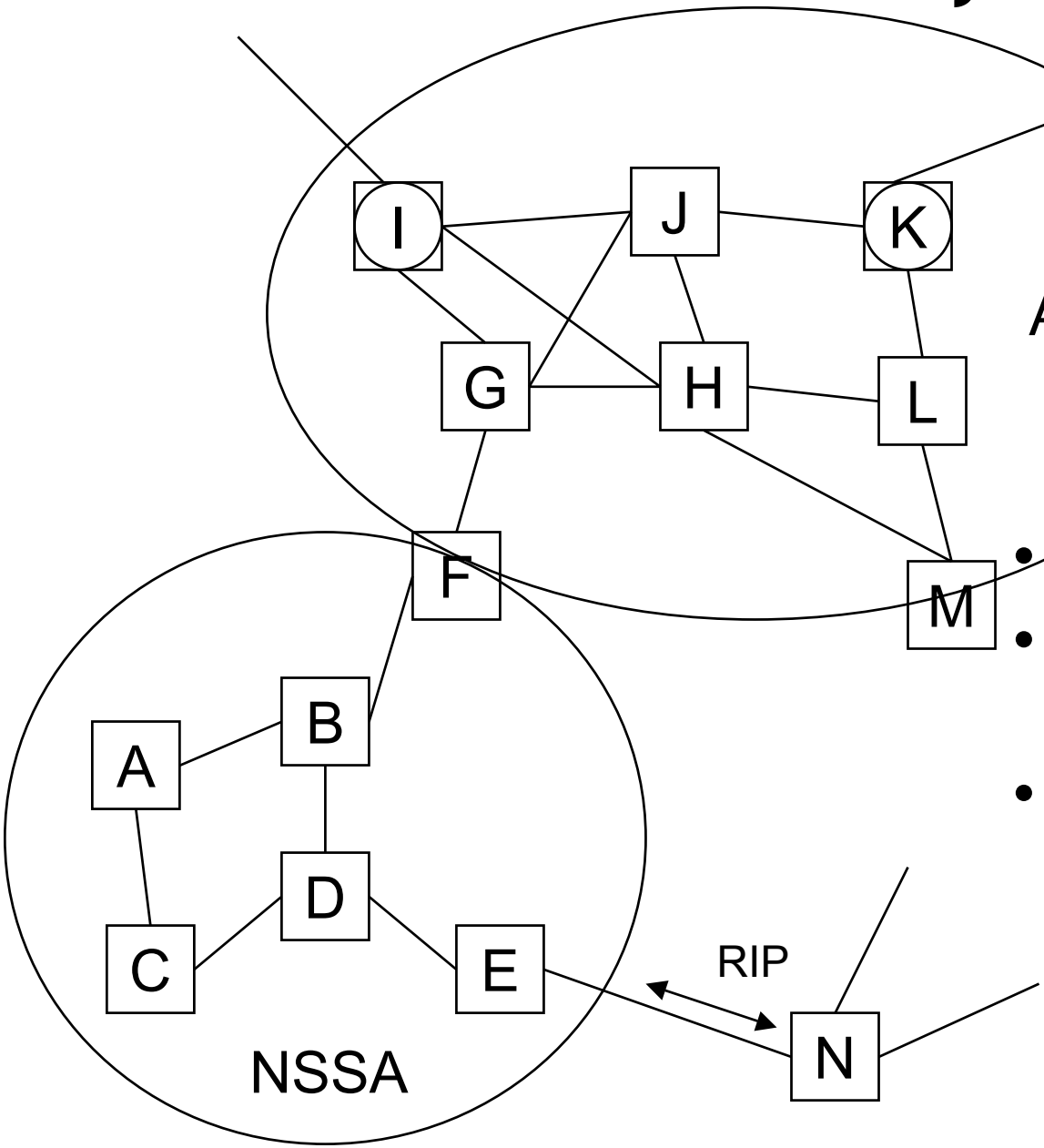
- Areas with no ASBRs.
- To reach ASBRs, you have to go through the ABR anyway.
- No point in advertising type 5 (AS External) LSAs.
  - No point in advertising type 4 (ASBR Summary) LSAs either.
- Just advertise Network Summary routes into the Stub area.
  - Appropriate ABR gets picked to reach prefix (nothing special here).
- No virtual links can be configured through a Stub Area.

- Totally-stubby areas: type 3 (Network summary) LSAs are not advertised, except for a default route.
- May not pick optimal routes.
  - (*e.g.*, F and V both advertise a single default route; but optimal path may not be taken).

Stub Areas, cont'd



Not-So-Stubby Areas



- N has a default route.
- E advertises type 7 LSAs to the area.
- F may block it (P=0) or translate into type 5 (P=1).