

Reliable Multicast

Chalermek Intanagonwiwat

Slides courtesy of Haobo Yu and Christos Papadopoulos

1

Introduction: What is Multicast?

- Unicast: one source to one destination
- Multicast: one source to many destinations
- Two main functions:
 - Efficient data distribution
 - Logical naming of a group

2

Error detection

- Sender-reliable
- Receiver-reliable

3

Sender-reliable

- Wait for ACKs from all receivers. Re-send on timeout or selective ACK
 - +: easy resource management
 - -: wait for ACK
 - -: receiver state in sender not scalable
 - -: ACK implosion

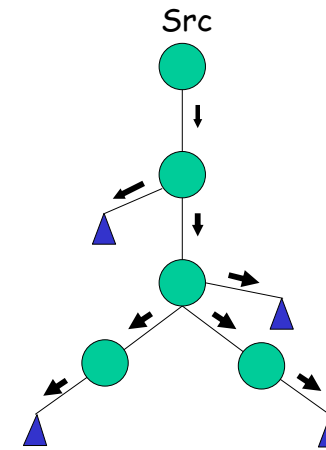
4

Receiver-reliable

- Receiver NACKs lost packet
 - +: no state at sender - good for mcast
 - -: does not provide 100% reliability
 - -: NACK implosion

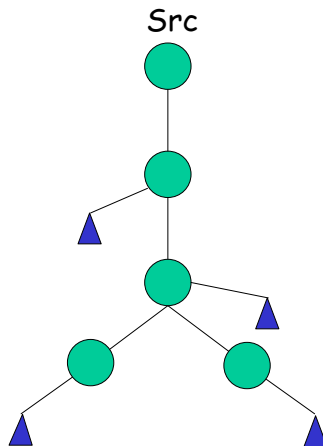
5

Implosion



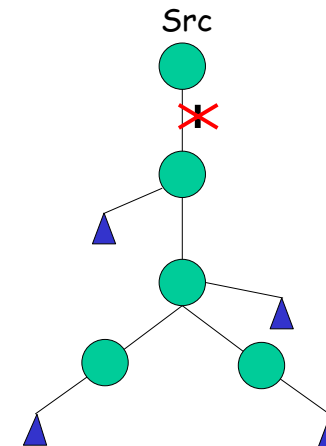
6

Implosion



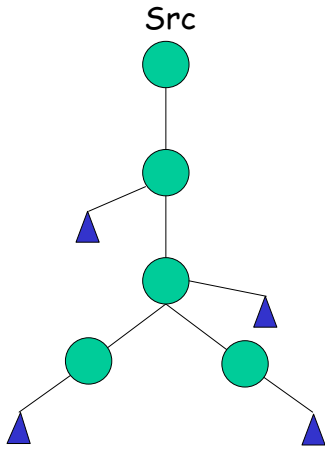
7

Implosion



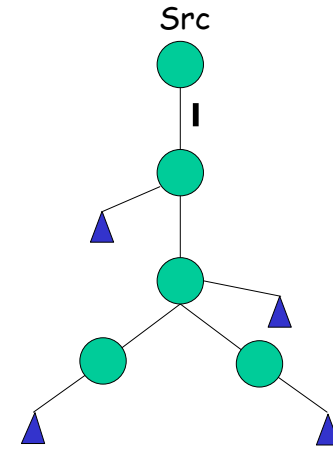
8

Implosion



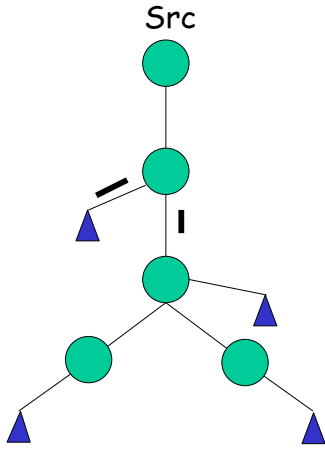
9

Implosion



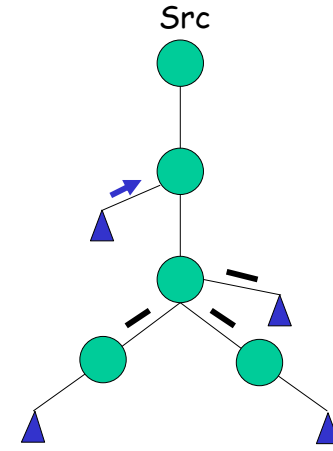
10

Implosion



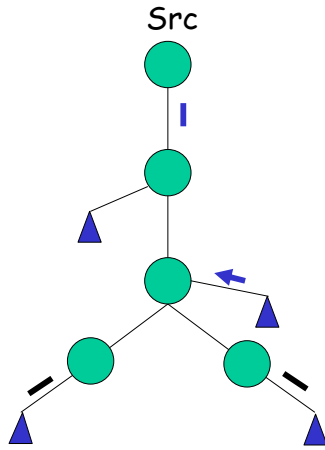
11

Implosion



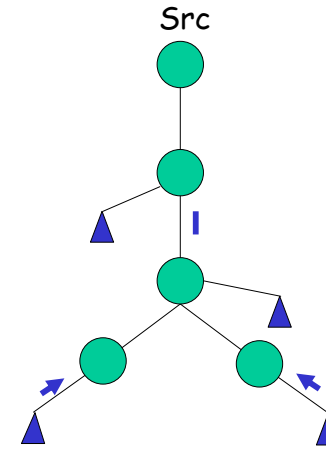
12

Implosion



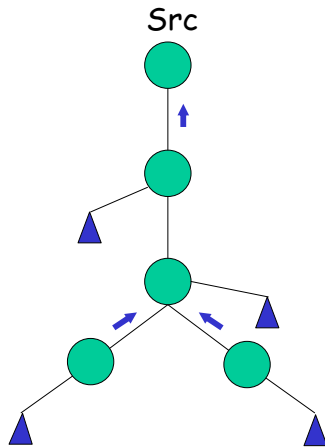
13

Implosion



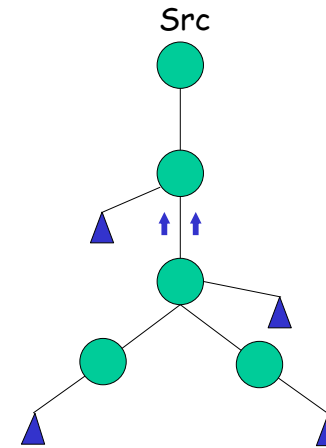
14

Implosion



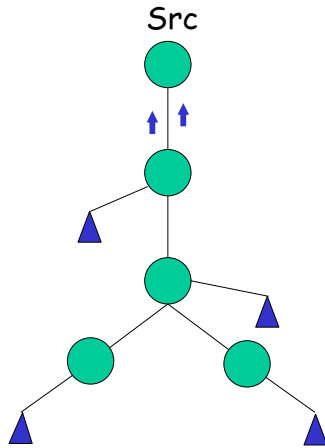
15

Implosion



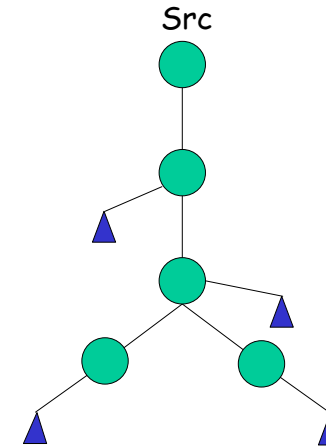
16

Implosion



17

Implosion



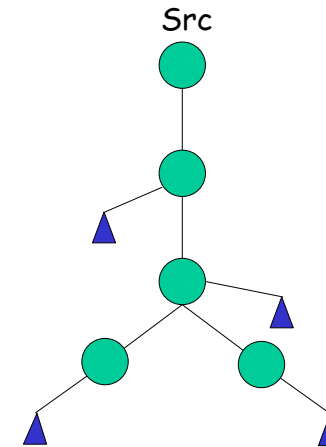
18

Retransmission

- Re-transmitter
 - options: sender, receiver
- How to retransmit
 - unicast, multicast, scoped multicast, retransmission group, ...
- Problem: Exposure

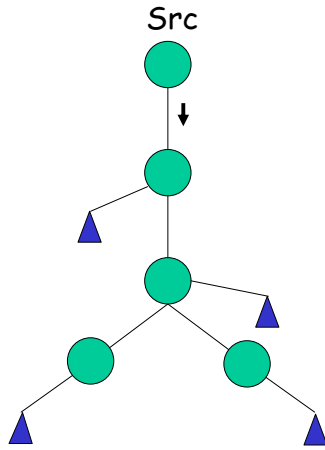
19

Exposure



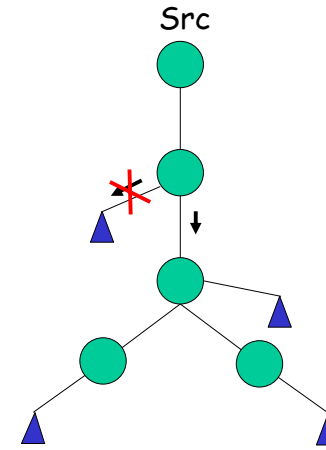
20

Exposure



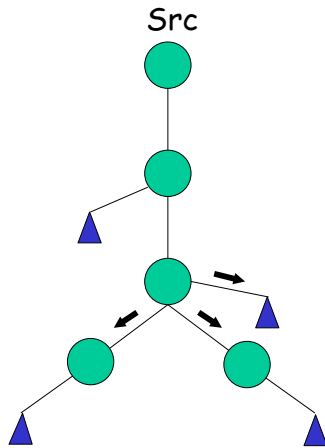
21

Exposure



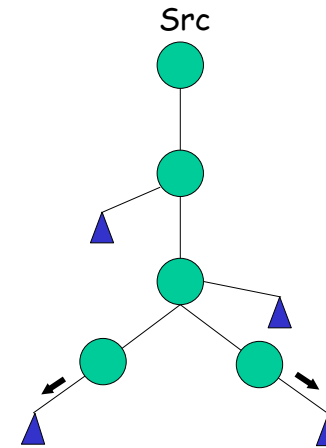
22

Exposure



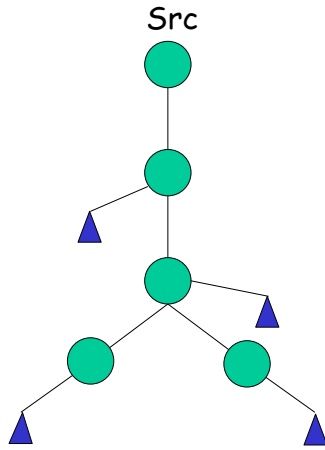
23

Exposure



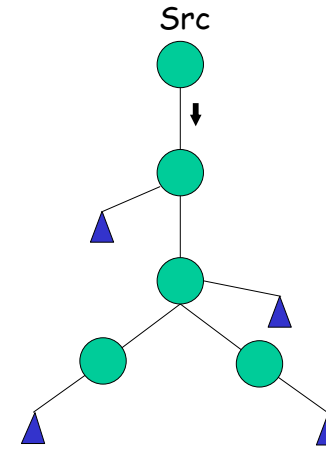
24

Exposure



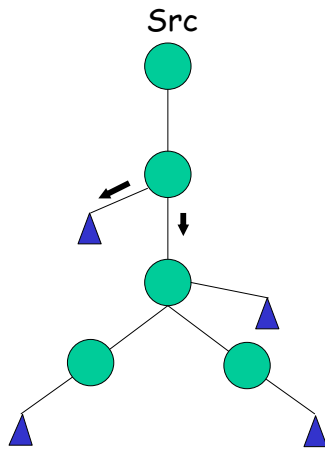
25

Exposure



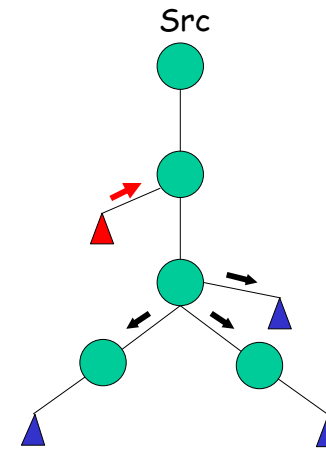
26

Exposure



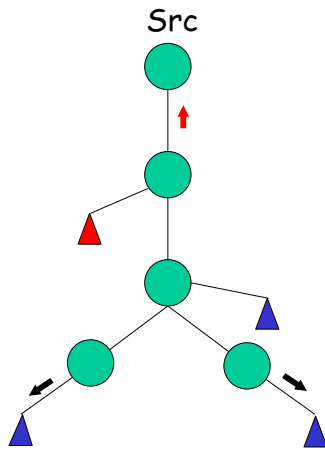
27

Exposure



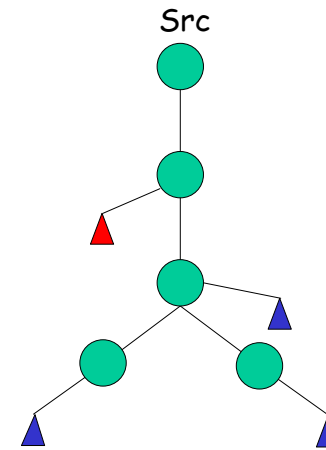
28

Exposure



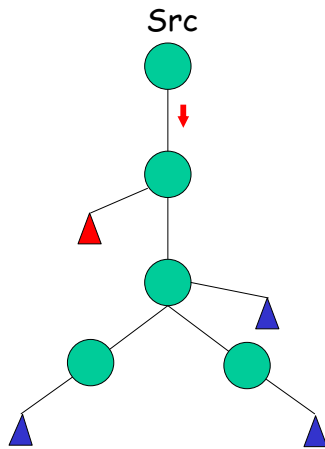
29

Exposure



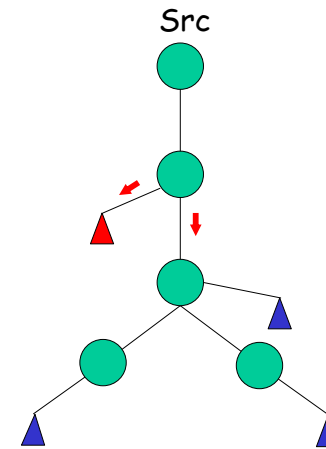
30

Exposure



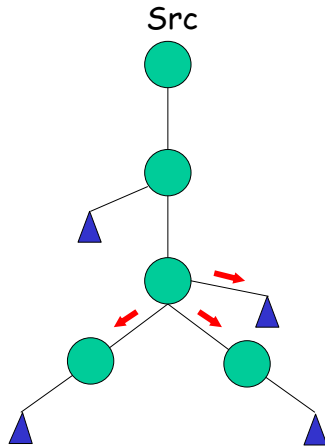
31

Exposure



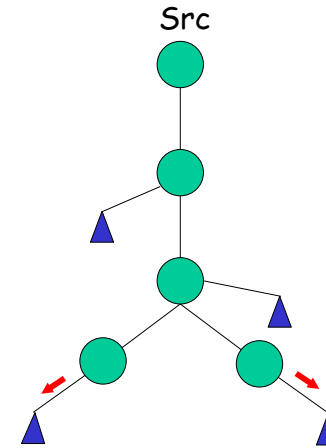
32

Exposure



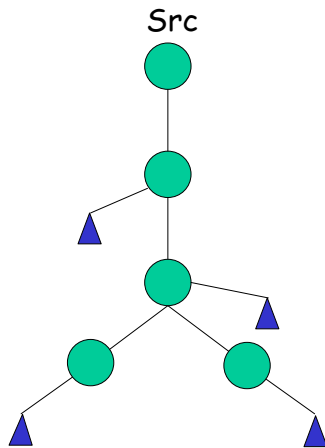
33

Exposure



34

Exposure

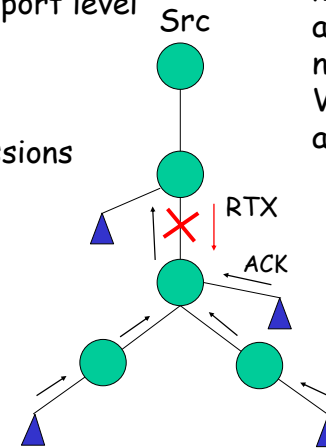


35

Aside: Using the routers

Routers do transport level processing:

- buffer packets
- fuse ACKs
- send retransmissions

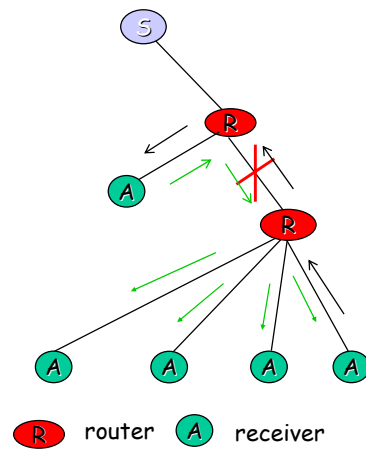


Model solves implosion and exposure, but not scalable. Violates end-to-end argument

36

Ideal recovery model

- "Ideal router"
- A single request is sent upstream
- A single repair is multicast from the nearest repairer to the subtree down the lossy link



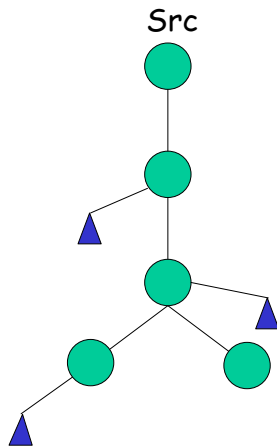
37

SRM

- Originally designed for *wb*
- Receiver-reliable
 - NACK-based
- Every member may multicast NACK or retransmission

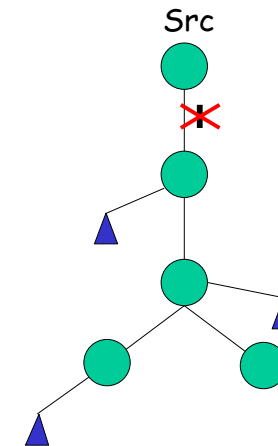
38

SRM Request Suppression



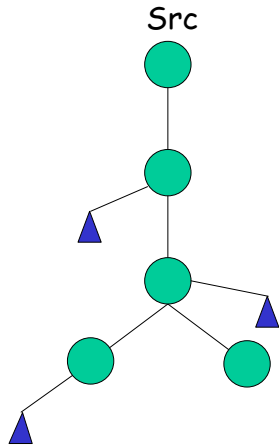
39

SRM Request Suppression



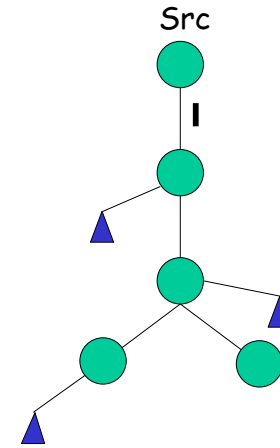
40

SRM Request Suppression



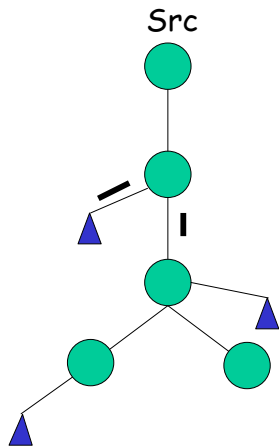
41

SRM Request Suppression



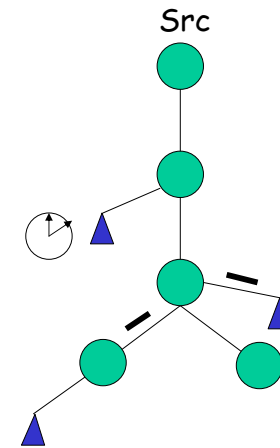
42

SRM Request Suppression



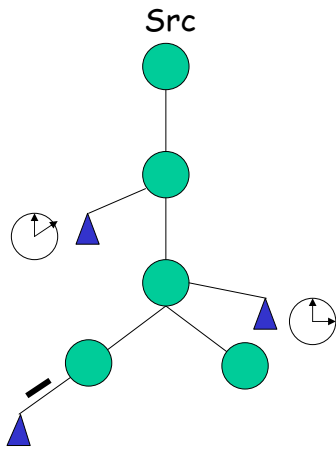
43

SRM Request Suppression



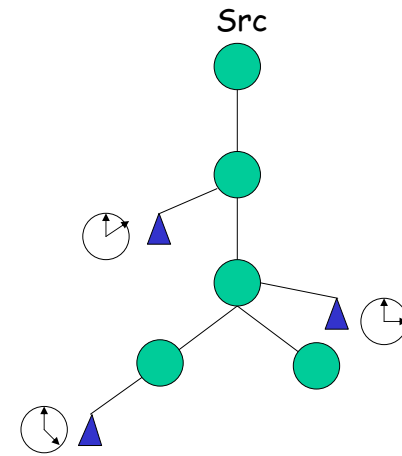
44

SRM Request Suppression



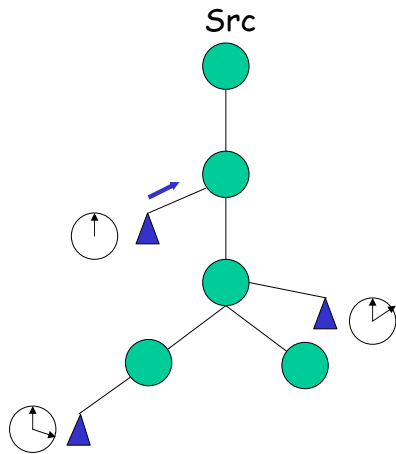
45

SRM Request Suppression



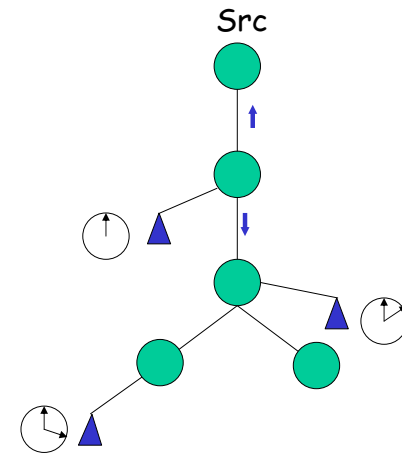
46

SRM Request Suppression



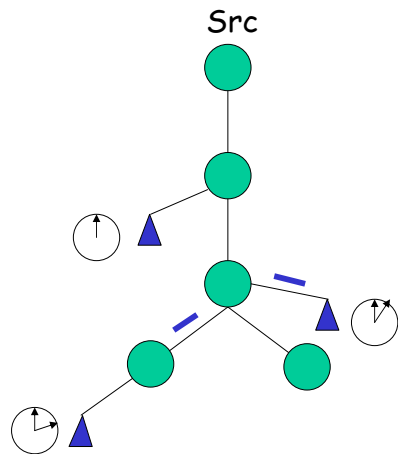
47

SRM Request Suppression

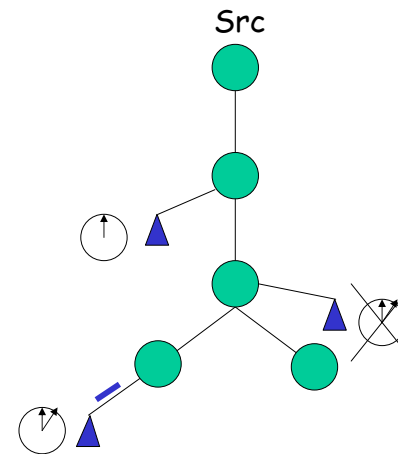


48

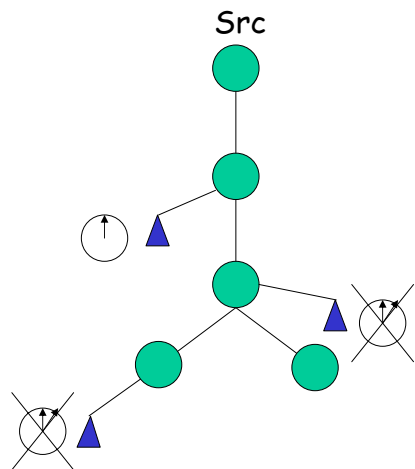
SRM Request Suppression



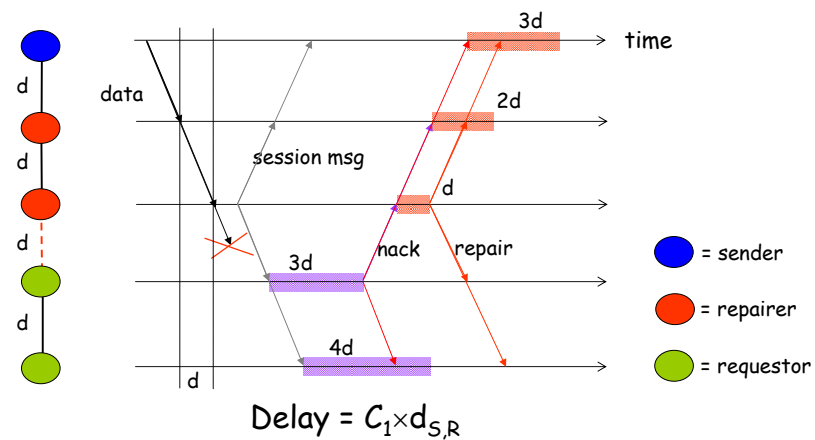
SRM Request Suppression



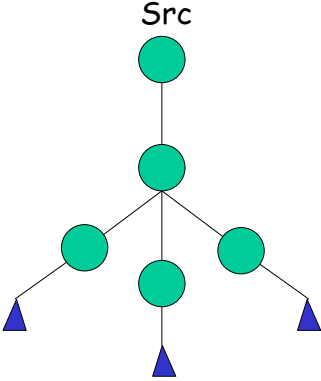
SRM Request Suppression



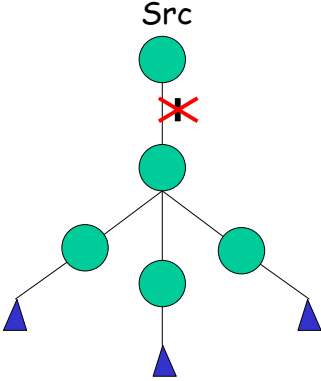
Deterministic Suppression



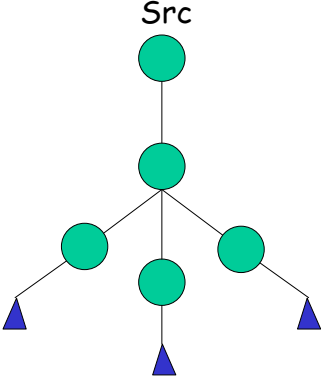
SRM Star Topology



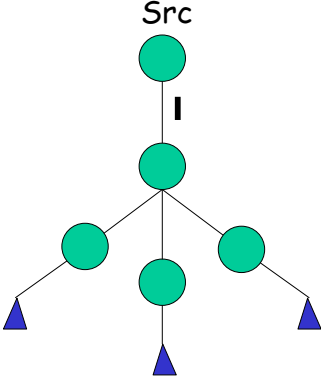
SRM Star Topology



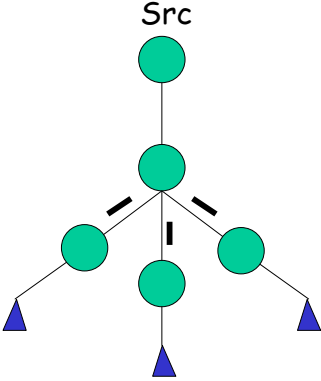
SRM Star Topology



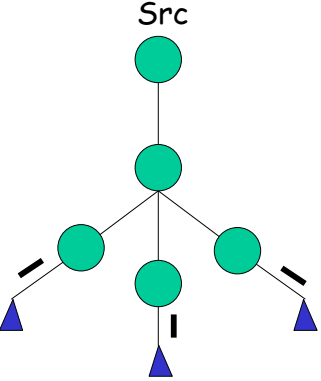
SRM Star Topology



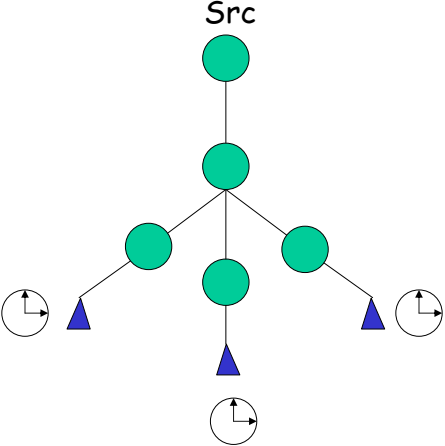
SRM Star Topology



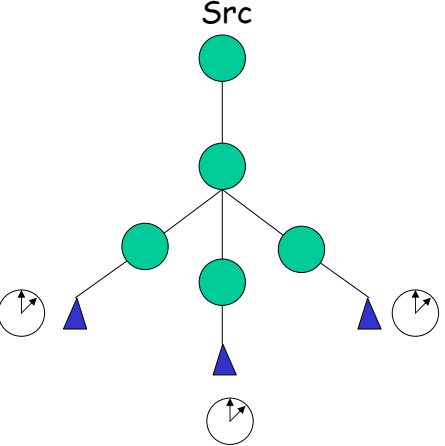
SRM Star Topology



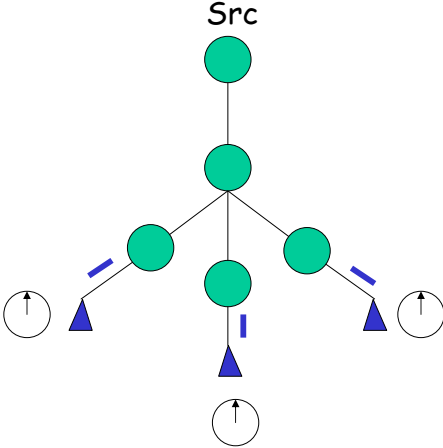
SRM Star Topology



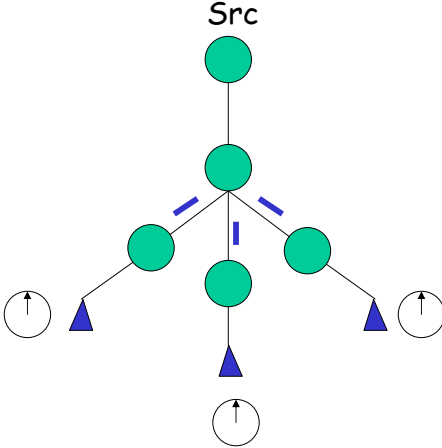
SRM Star Topology



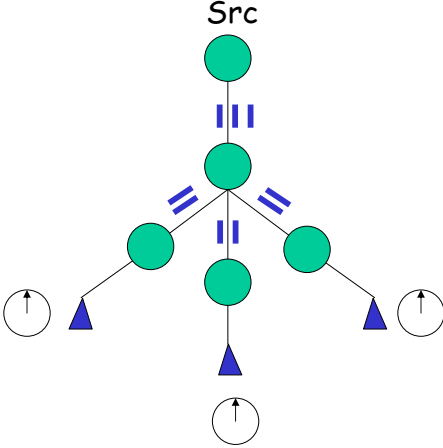
SRM Star Topology



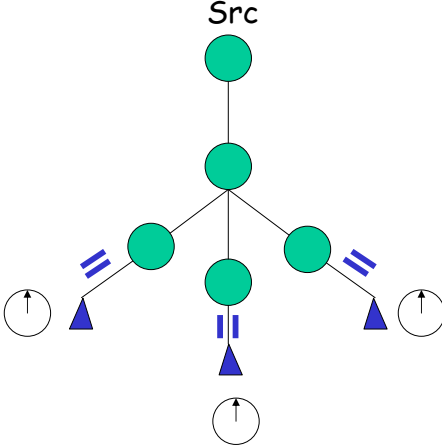
SRM Star Topology



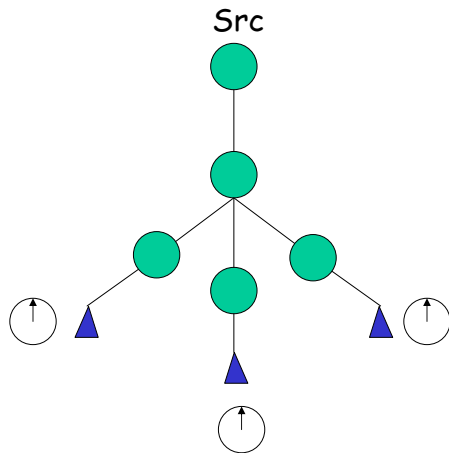
SRM Star Topology



SRM Star Topology

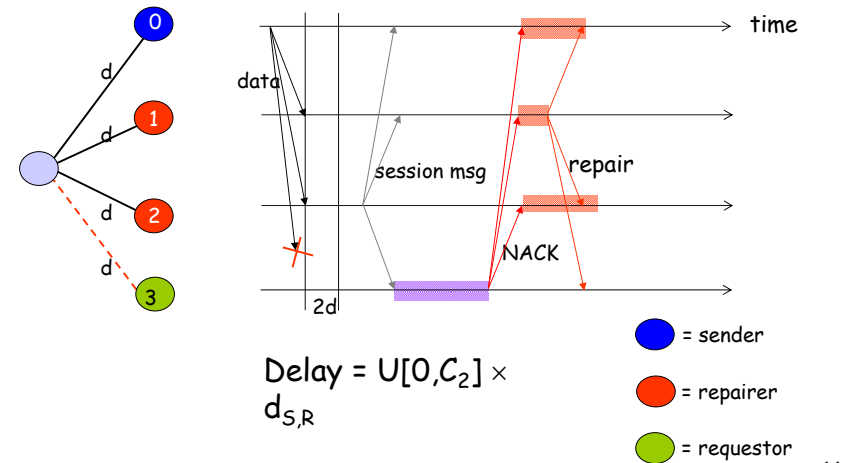


SRM Star Topology



65

SRM: Stochastic Suppression



66

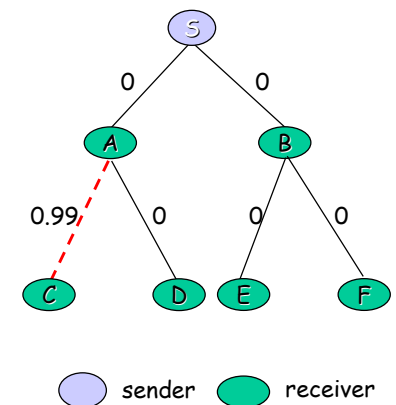
SRM (summary)

- NACK/Retransmission suppression
 - Delay before sending
 - Delay based on RTT estimation
 - Deterministic + Stochastic components
- Periodic session messages
 - Full reliability
 - Estimation of distance matrix among members

67

What's missing?

- Losses at link (A,C) causes retransmission to the whole group
- Only retransmit to those members who lost the packet
- [Only request from the nearest responder]



68

Local Recovery

- ❑ Application-level hierarchy
 - Fixed v.s. dynamic
- ❑ TTL scoped multicast
- ❑ Router supported

69

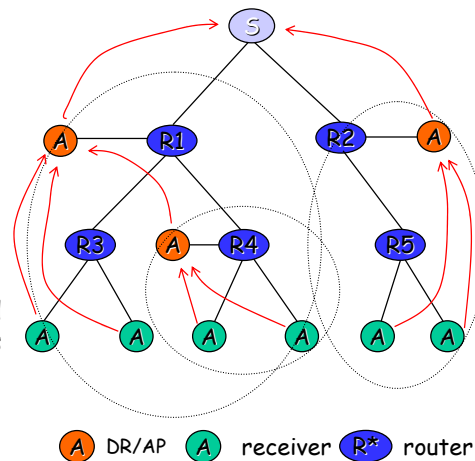
RMTP

- ❑ Reliable Multicast Transport Protocol by Purdue and AT&T Research Labs
- ❑ Designed for file dissemination (single-sender)
- ❑ Deployed in AT&T's billing network

70

RMTP: Fixed hierarchy

- ❑ Rcvrs grouped into local regions
- ❑ Rcvr unicasts periodic ACK to its ACK Processor (AP), AP unicasts its own ACK to its parent
- ❑ Rcvr dynamically chooses closest statically configured Designated Receiver (DR) as its AP



71

RMTP: Error control

- ❑ DR checks retx "request" periodically
- ❑ Mcast or unicast retransmission
 - Based on percentage of requests
 - Scoped mcast for local recovery
- ❑ Immediate transmission request
 - Used for late join

72

RMTP: Comments

- +: Heterogeneity
 - Lossy link or slow receiver will only affect a local region
- -: Position of DR critical
 - Static hierarchy cannot adapt local recovery zone to loss points

73

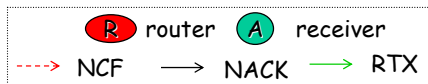
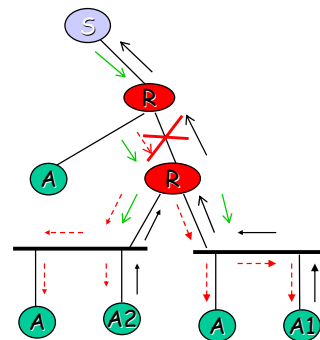
PGM

- Cisco's reliable multicast protocol
- NACK-based, with suppression
- Repair only forwarded to the NACKers

74

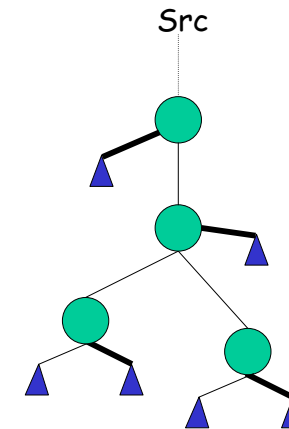
PGM: Request forwarding

- NACK + random delay
- Forwarded upstream towards the source
- Only one NACK is forwarded for every packet loss
- NCF: NACK suppression and hop-by-hop NACK reliability



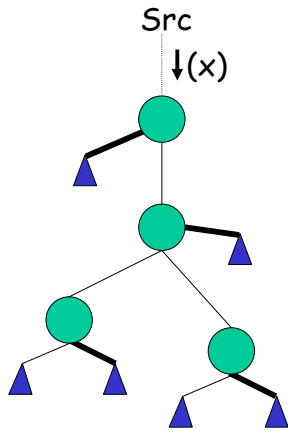
75

PGM Summary



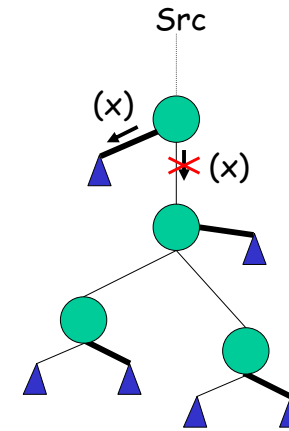
76

PGM Summary



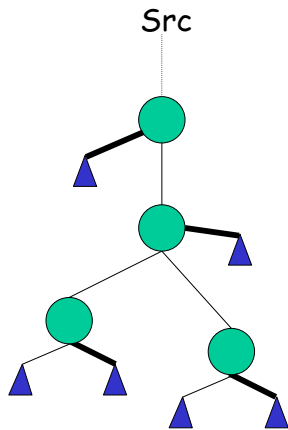
77

PGM Summary



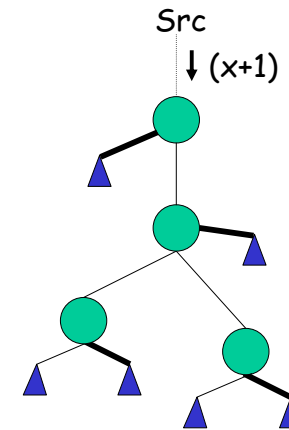
78

PGM Summary



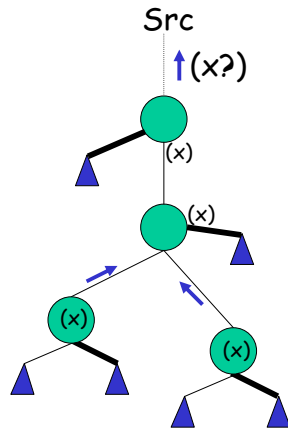
79

PGM Summary



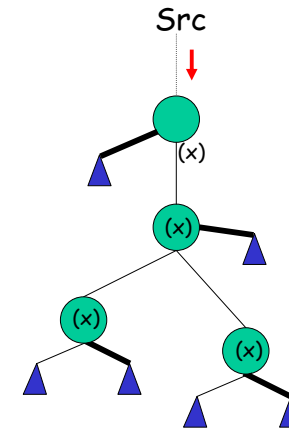
80

PGM Summary



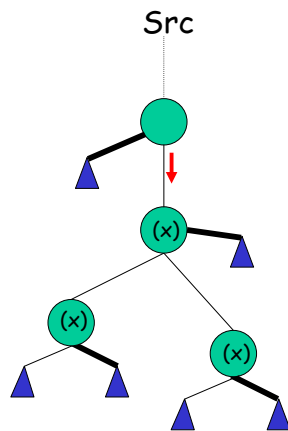
85

PGM Summary



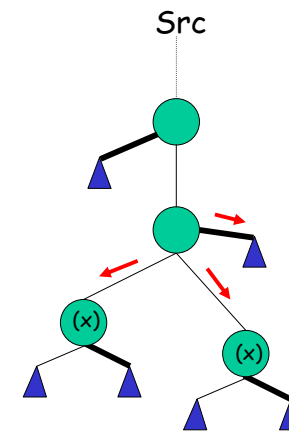
86

PGM Summary



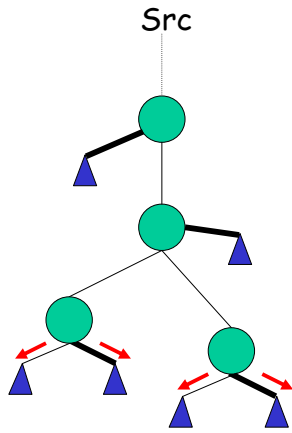
87

PGM Summary



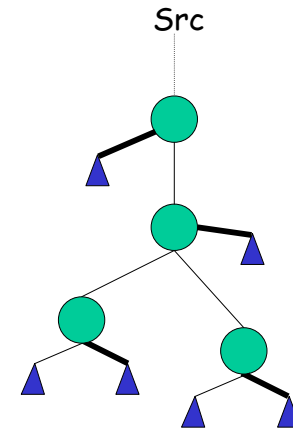
88

PGM Summary



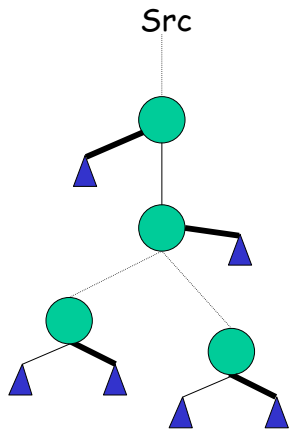
89

PGM Summary



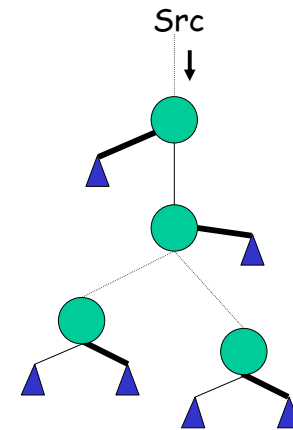
90

Problem: Repeated Retransmissions



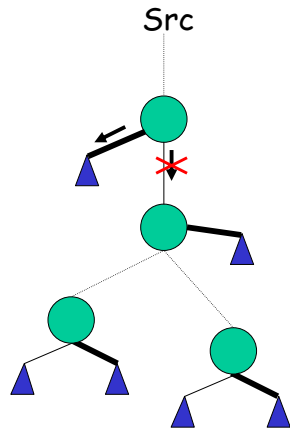
91

Repeated Retransmissions

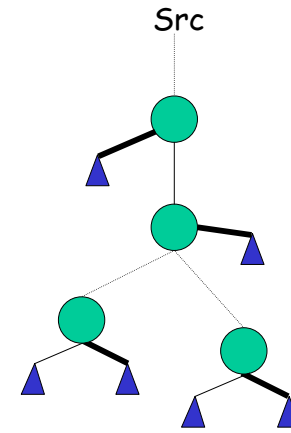


92

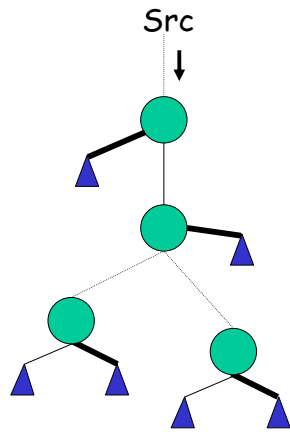
Repeated Retransmissions



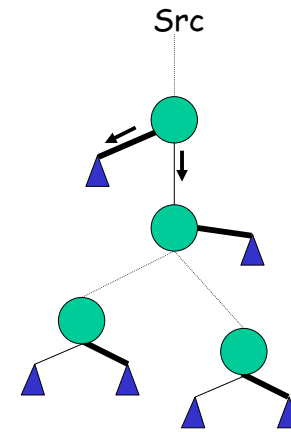
Repeated Retransmissions



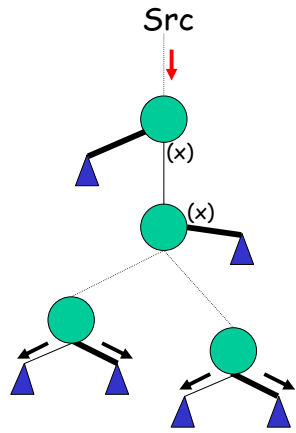
Repeated Retransmissions



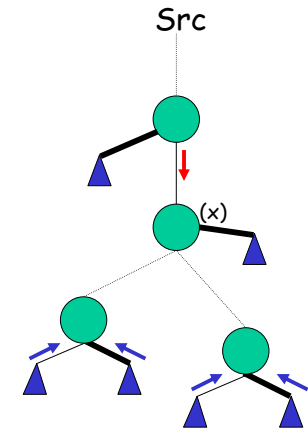
Repeated Retransmissions



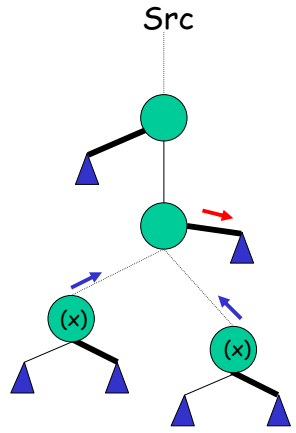
Repeated Retransmissions



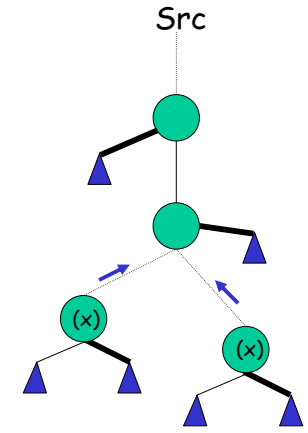
Repeated Retransmissions



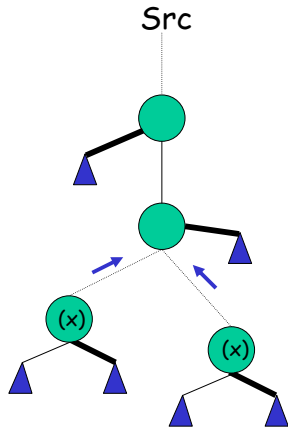
Repeated Retransmissions



Repeated Retransmissions

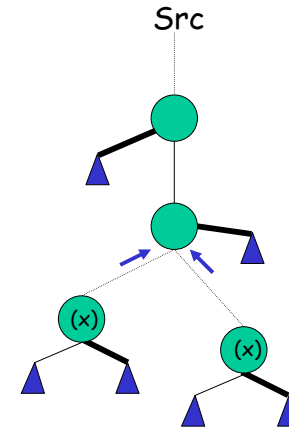


Repeated Retransmissions



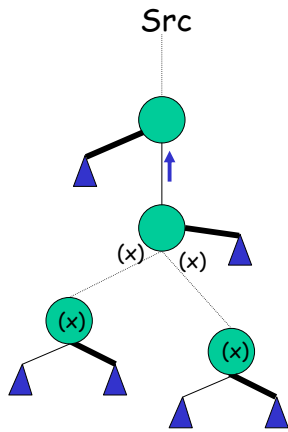
105

Repeated Retransmissions



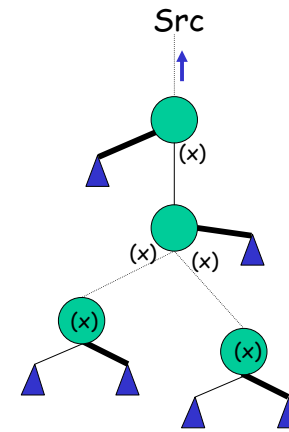
106

Repeated Retransmissions



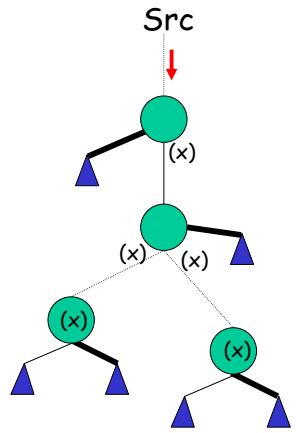
107

Repeated Retransmissions



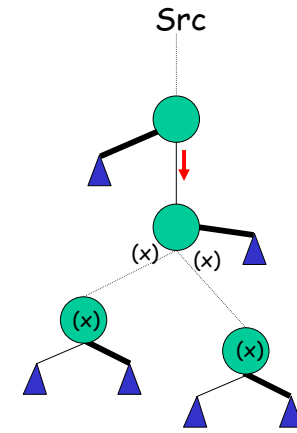
108

Repeated Retransmissions



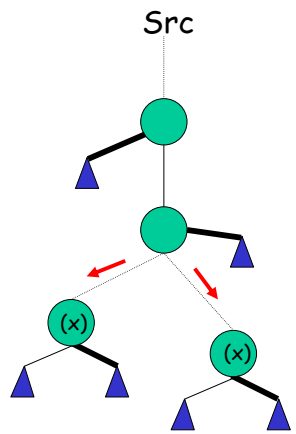
109

Repeated Retransmissions



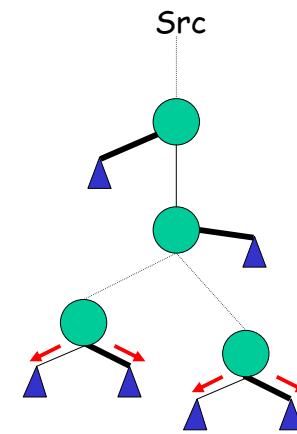
110

Repeated Retransmissions



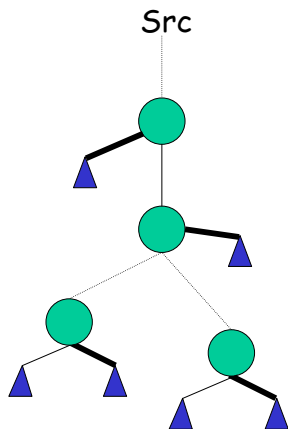
111

Repeated Retransmissions



112

Repeated Retransmissions



LMS

- Light-weight Multicast Service
- Enhance multicast routing with **selective forwarding**, nothing beyond that

LMS

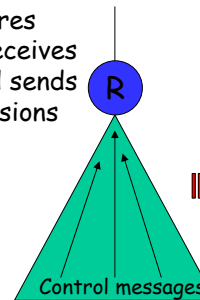
LMS extends router forwarding - what routers are meant to do in the first place

- No packet storing or processing at routers
- Strictly IP: no peeking into higher layers

The LMS concept

Heavy-weight model

Router stores packets, receives NACKs and sends retransmissions

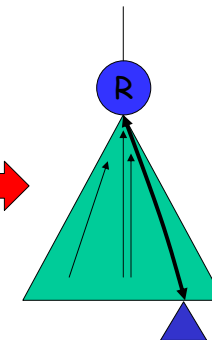


LMS

Router chooses a receiver as a surrogate

Router steers all control messages to surrogate

Router relays messages from surrogate to the subtree



Receiver acting as a surrogate

Core Ideas

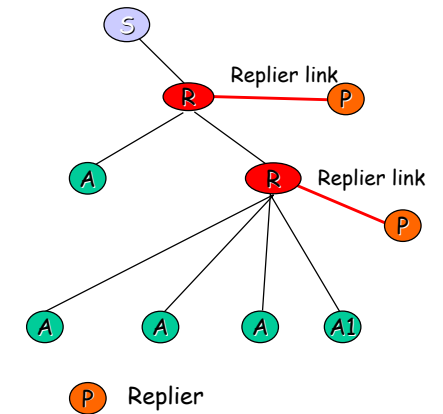
- Each router selects a replier (surrogate)
- Routers steer requests to repliers
- Routers help repliers multicast replies to loss subtree

LMS achieves the efficiency of the heavy-weight model, but without the weight

117

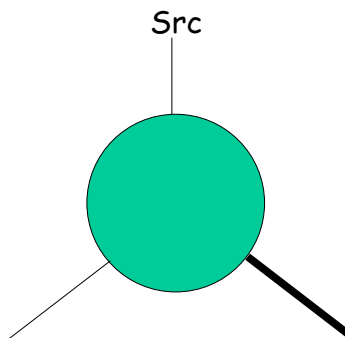
LMS: Concepts

- Replier
 - Receiver volunteered to answer requests
- Turning point
 - Where requests start to move downstream
- Directed mcast
 - Mcast to a subtree



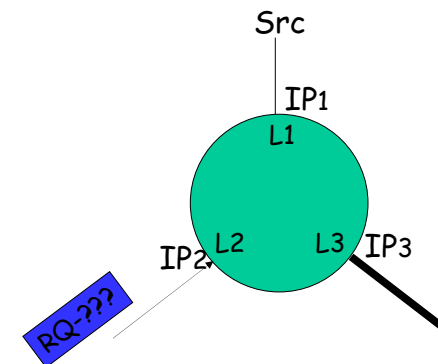
118

Request Handling



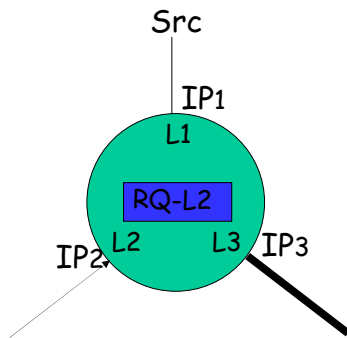
119

Request from non-replier link



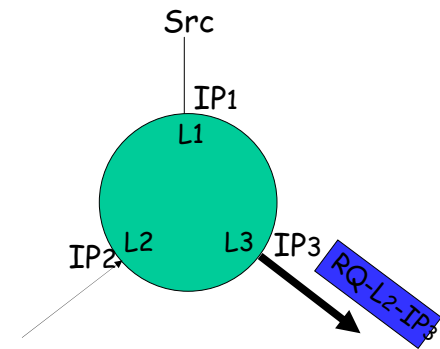
120

The Turning Point: stamp incoming iface



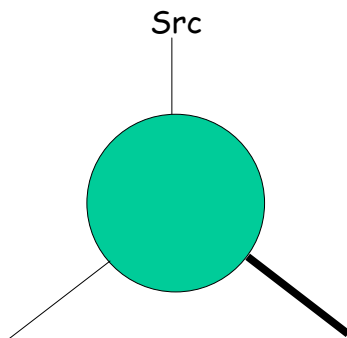
121

Out to replier link



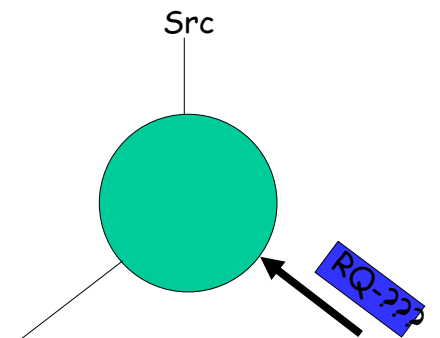
122

Request handling



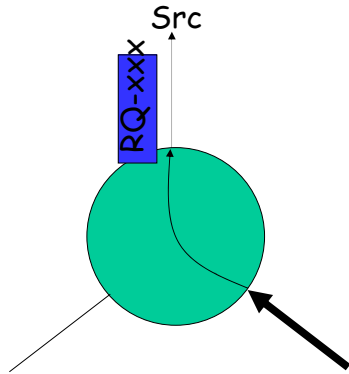
123

Request from replier link



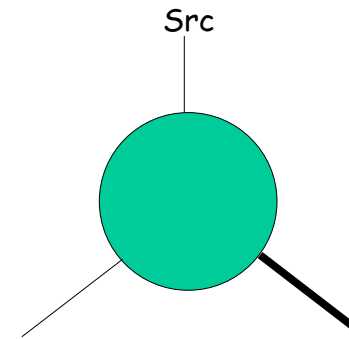
124

Send upstream unchanged



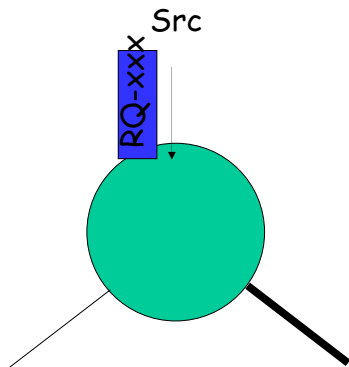
125

Request handling



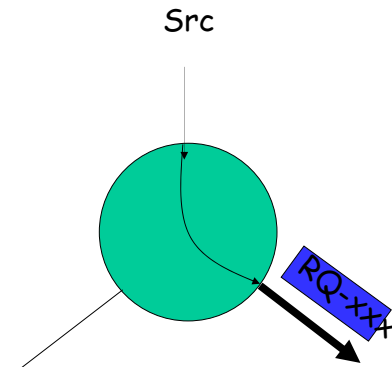
126

Request from upstream



127

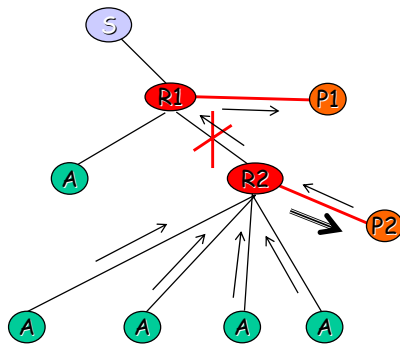
To replier link unchanged



128

LMS: Request forwarding

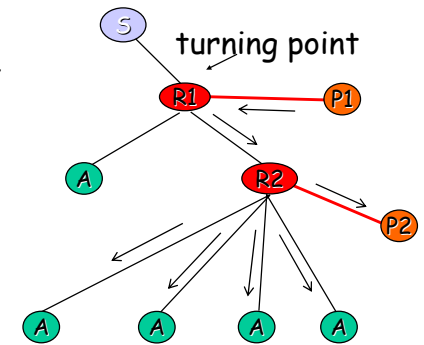
- Mcast to the group
- If a request reaches a turning point, it's forwarded towards the replier
- No request suppression or merging, but scope of requests is limited



129

LMS: Reply forwarding

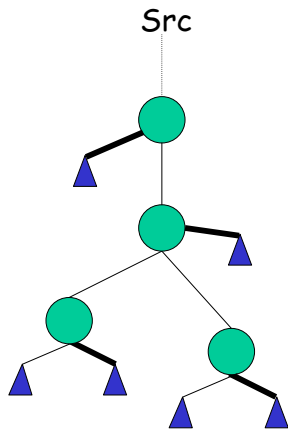
- At turning point, `<router_addr:link_id>` is added into request packet
- Replier includes it into its retx packet
- The specified router forwards packet downstream at `link_id`
- No retx suppression necessary



130

LMS Summary

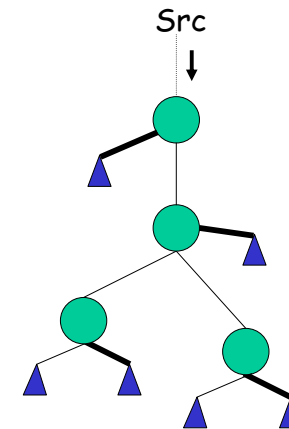
- DATA
- NAK
- RTX



131

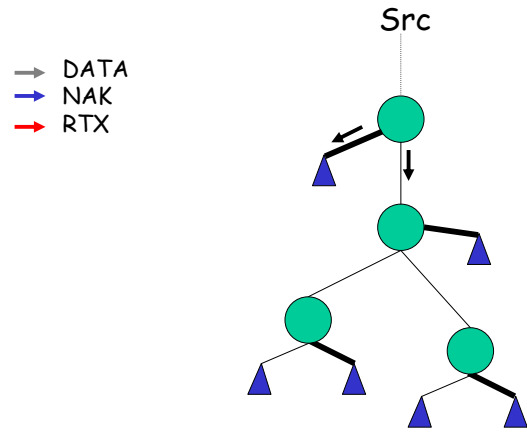
LMS Summary

- DATA
- NAK
- RTX



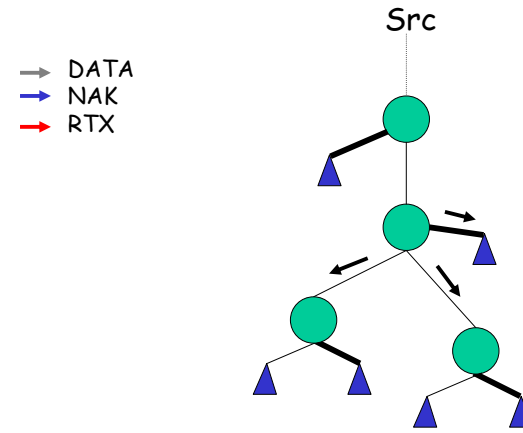
132

LMS Summary



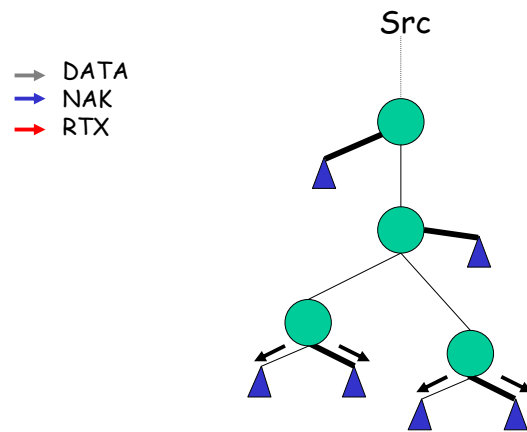
133

LMS Summary



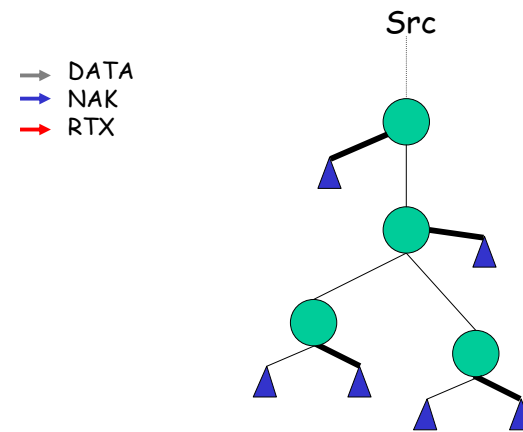
134

LMS Summary



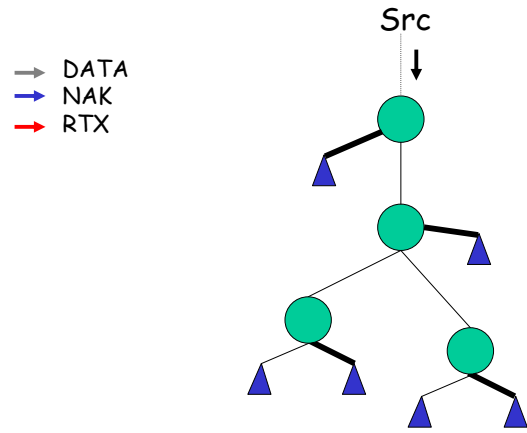
135

LMS Summary



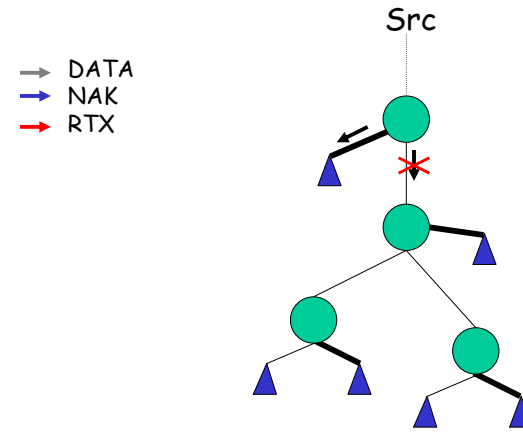
136

LMS Summary



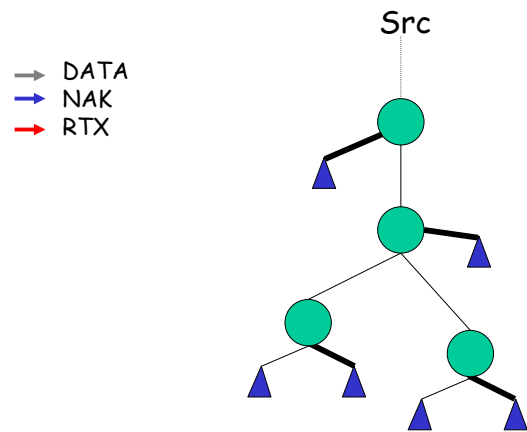
137

LMS Summary



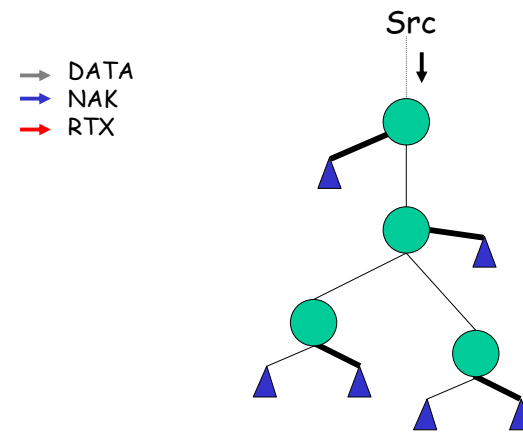
138

LMS Summary



139

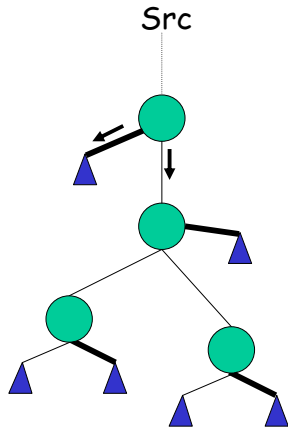
LMS Summary



140

LMS Summary

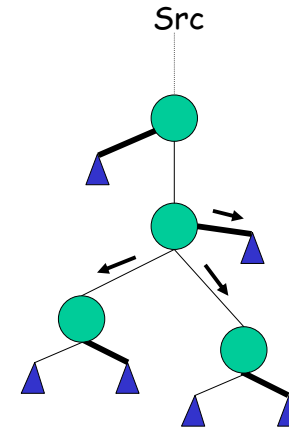
→ DATA
→ NAK
→ RTX



141

LMS Summary

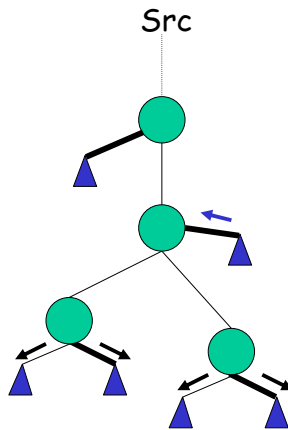
→ DATA
→ NAK
→ RTX



142

LMS Summary

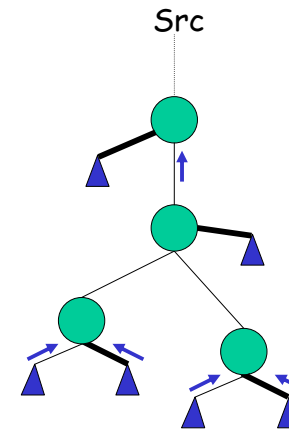
→ DATA
→ NAK
→ RTX



143

LMS Summary

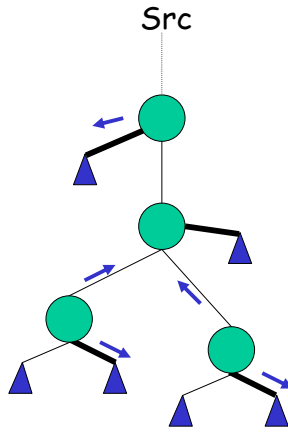
→ DATA
→ NAK
→ RTX



144

LMS Summary

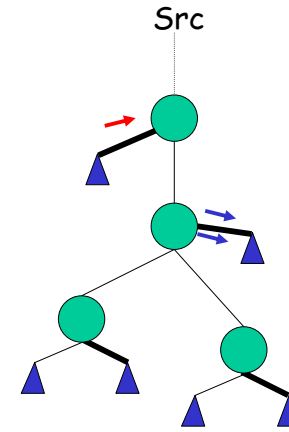
→ DATA
→ NAK
→ RTX



145

LMS Summary

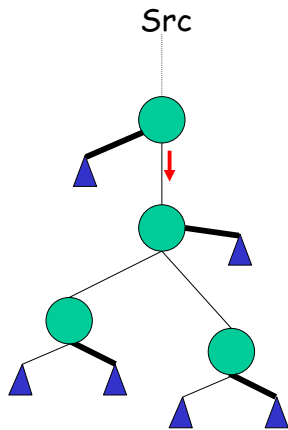
→ DATA
→ NAK
→ RTX



146

LMS Summary

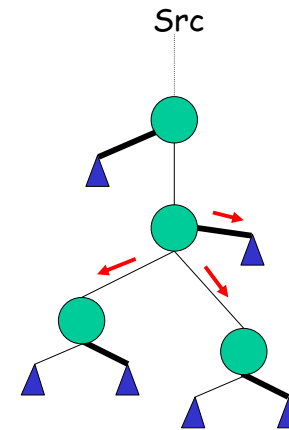
→ DATA
→ NAK
→ RTX



147

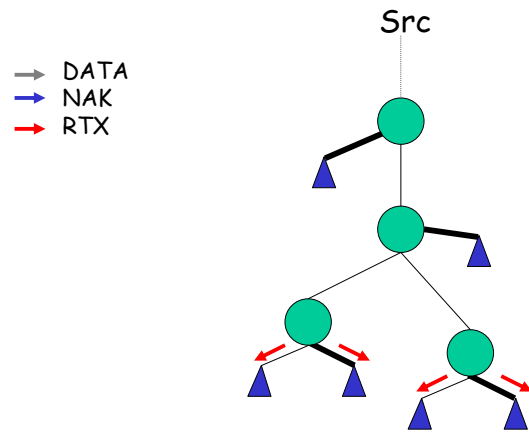
LMS Summary

→ DATA
→ NAK
→ RTX



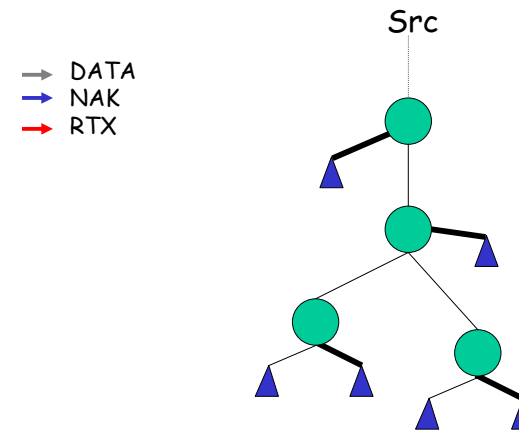
148

LMS Summary



149

LMS Summary



150

LMS: Comments

- Replier problems
 - Selection? Fault tolerance?
 - How well will repliers scale w.r.t. |group|?
- Works with unidirectional shared trees (PIM)
 - Needs to relay requests from core/RP to sender
- Difficulties with bi-directional shared trees
 - Needs per-source state

151