Subnet Bandwidth Manager: Admission Control over Ethernet

Raj Yavatkar, Intel

Don Hoffman, Sun

Ema Patki, Intel

Goals

- Step-by-step mapping of int-serv capabilities over Ethernet infrastructure
  - short-term solution should work with existing bridges, hubs/switches
  - work with IEEE 802.1? for Level-2 support
  - ex. traffic flow separation
- Step I -- admission control for RSVP traffic and policing at end systems
- Step II -- traffic flow separation at hubs/switches
- Step III -- traffic control support in hubs/switches

Objectives for Step I

- administrative control over max amount of multimedia traffic over any LAN segment
- Rely on end-system policing and rate-adaptive applications for best effort traffic
  - *slow start* type congestion avoidance
- Leverage existing RSVP-signaling as much as possible

Outline

- **Only architectural discussion**
  - **variations from RSVP message processing rules included**
  - **protocol details, packet formats, etc. NOT included**
- **Discuss possible alternatives**

Overview

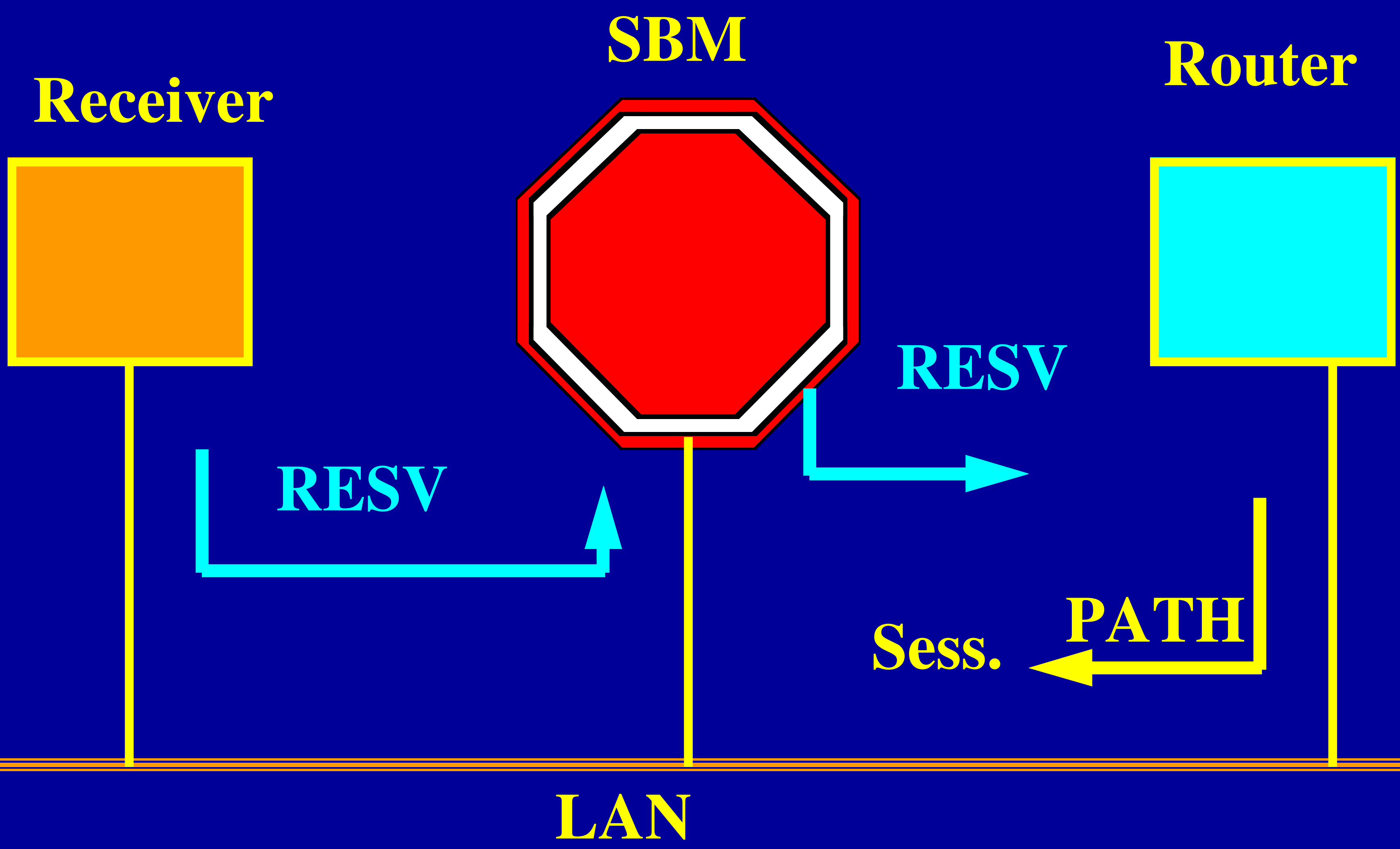
- **SBM (Subnet Bandwidth Manager) responsible for admission control**
- **a designated SBM (DSBM) for each LAN segment**
  - **an SBM may act as DSBM for many segments**
- **SBM is a UDP-based server**
- **soft state in SBM with refresh ala RSVP**
  - **recovery from SBM restart and failures**
- **dynamic binding between SBM and end-systems**
  - **IP multicast based**

Overview (contd.)

- **At the start, an RSVP node discovers and binds to its DSBM using IP multicast-based protocol**
- **PATH messages sent/forwarded to the session address (NO CHANGE)**
- **Outgoing RESERVE over Ethernet interface unicast to DSBM**
  - **a new LAN_PHOP object specifies the PHOP**
- **DSBM performs admission control and forwards RESERVE *toward* PHOP**

Overview (contd.)

- **DSBM processing similar to conventional RSVP processing**
  - **merges reservations appropriately**
  - **avoids *killer reservations***
  - **returns RESV_ERROR or RESV_CONFIRM (if necessary)**


Receiver
SBM
Router
RESV
RESV
PATH
Sess.
LAN

Important Notes

- NOT a centralized scheme
- Many SBMs can exist per LAN, each responsible for a separate portion of the LAN
- Does NOT require maintenance of consistent, distributed state across Hosts and SBMs
- Distribution of responsibility among SBMs allows scalability and fault tolerance

An Alternative Proposal

- Assume one SBM per IP subnet
- PATH message sent to special SBM address
  - mcast to SBM group address (a new encapsulation)
- SBM inserts itself as a PHOP between sender and receiver
- RESERVE automatically lands up at the SBM
- Presented to IEEE 801.P
  - does not allow for multiple SBMs within a subnet
  - an SBM per hub/switch should be allowed

Supporting Mechanisms

- **Discovering DSBM and binding to it**
  - **DSBM listens to a well-known UDP mcast address**
- **DSBM Election**
  - **more than one active SBM for same segment(s)**
  - **single SBM is a DSBM at any time, elected using an election protocol**
  - **Use of *I_AM_DSBM* declarations via IP multicast and some tiebreaking**
  - **peers step in when DSBM fails or terminates**

Application Behavior

- Sender on a LAN not to transmit any traffic until at least one successful RESERVE reaches it
- Outgoing flow to be policed to be within maximum RESERVE made
- For multicast flows, receivers must leave the session mcast group in case of RESV_ERR or PATH_TEAR
  - problematic in case of multiple senders
- Best-effort traffic must be rate-adaptive

Handling Complex Physical Topologies

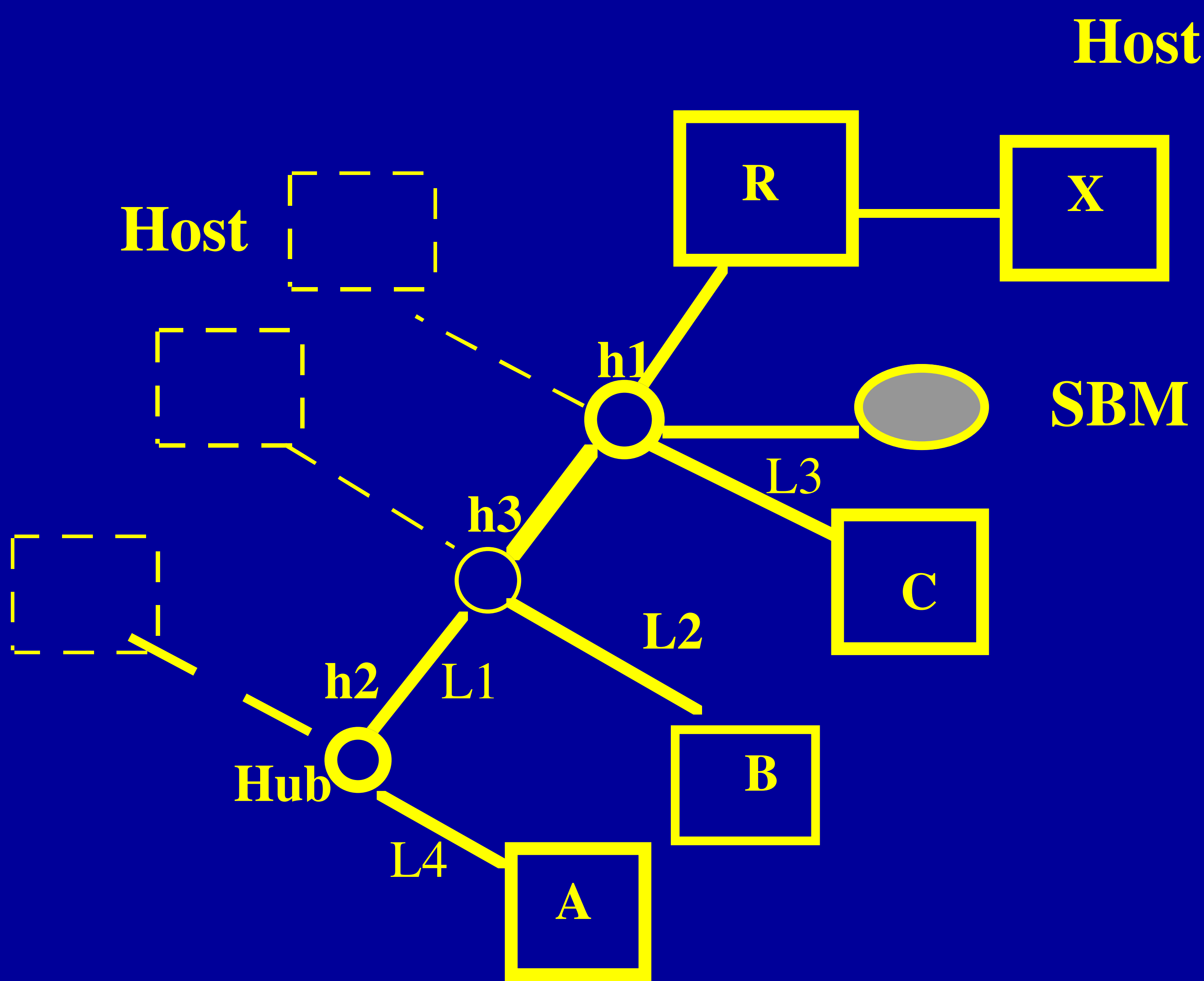
- Multi-hop topology consisting of bridges, hubs, and switches
  - data flows traverse only a subset of segments
  - multiple DSBMs, each for separate portion of LAN desirable
- How to discover topology information?
- How to discover peer DSBMs and communicate with them?

Discovering LAN Topology

- **techniques used by network management utilities**
  - **static configuration info works only in case of non-redundant paths**
- **an interface to spanning tree topology info needed**
  - **IEEE 802.1 area**
- **IETF hub MIB working group**
  - ***topology mapping* section**

Topology Discovery Protocol

- **Placing two endpoints on a map and identifying LAN segments between them**
  - **endpoints identified using MAC addresses**
  - **tell managed hubs in the collision domain to watch for packets with the endpoint MAC address**
  - **send a *PING* to the endpoint to get it to transmit**
  - **SBM then uses hub MIB interface to read the group/ port of target MAC address from managed hubs**
  - **SBM then identifies affected segments**


Host
Host
R
X
h1
SBM
L3
h3
C
L2
h2
L1
Hub
B
L4
A

Peer-to-Peer DSBM Communication

- **Peers discovered using SBM_QUERY to SBM_GRP address**
  - **information is cached with time-out**
- **After successful admission control, DSBM forwards a RESERVE to the peer on next hop towards LAN_PHOP**
- **An error is sent back hop-by-hop using reservation state in intermediate DSBMs**

Areas of Co-operation With IEEE 801.P

- **definition of a standard interface for accessing spanning tree (routing) information in MAUs**
- **mechanisms for traffic flow separation**
  - **ex. priority mechanism**
- **RSVP-based admission control combined with traffic flow separation**
  - **a good approximation to *Controlled Load*?**