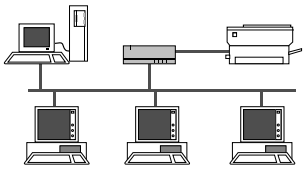


Distributed Systems
600.437
Multicast Protocols

Department of Computer Science
The Johns Hopkins University

Multicast Protocols

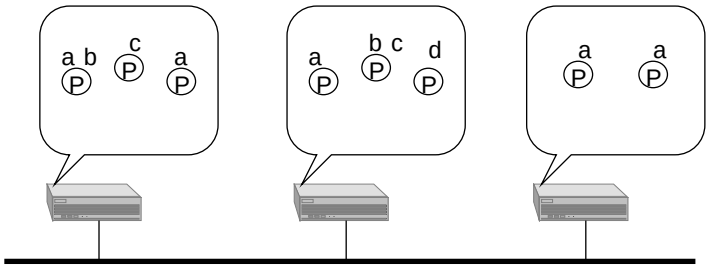
Lecture 4

Slides and Resources can be found at:

http://www.cnds.jhu.edu/courses/cs437/

Most of this work is documented in IETF RFC's and Internet-Drafts which can be found at: http://www.ietf.org/

The Multicast Paradigm



- Ordering (Unordered, FIFO, Causal, Agreed).
- Delivery guarantees (Unreliable, Reliable, Safe/Stable).
- Open groups versus close groups.
- Failure model (Omission, Fail-stop, Crash & Recovery, Network Partitions).
- Multiple groups.

Using Traditional Transport Protocols for Multicast

Point to point (TCP/IP)

- Automatic flow control ✓
- Reliable delivery ✓
- Connection service ✓
- Complexity (n^2)
- Linear degradation in performance

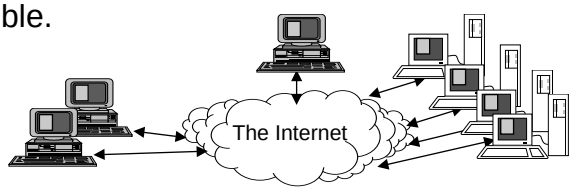
Using Traditional Transport Protocols for Multicast (cont.)

Unreliable broadcast/multicast (UDP, IP-Multicast)

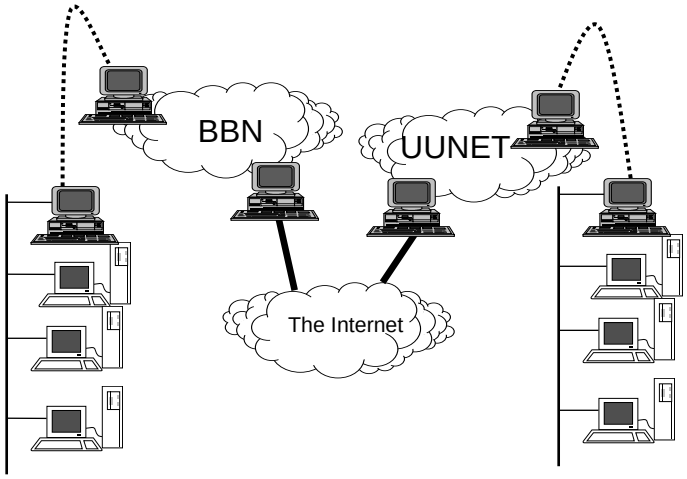
- Employs hardware support for broadcast and multicast. ✓
- Message losses : 0.01% at normal load, more than 30% at high load.
  - Buffers overflow (in the network and in the OS).
  - Interrupt misses.
- No connection service.

IP Multicast

- Multicast extension to IP.
- Best effort multicast service.
- No accurate membership.
- Class D addresses are reserved for multicast: 224.0.0.0 to 239.255.255.255 and are used as group addresses.
- The standard defines how hardware Ethernet multicast addresses can be used if these are possible.



IP-Multicast Logical Design



IP Multicast (cont.)

Extensions to IP inside a host:

- A host may send IP multicast by using a multicast address as the destination address.
- A host manages a table of groups and local application processes that belong to this group.
- When a multicast message arrives at the host, it delivers copies of it to all of the local processes that belong to that group.
- A host acts as a member of a group only if it has at least one active process that joined that group.

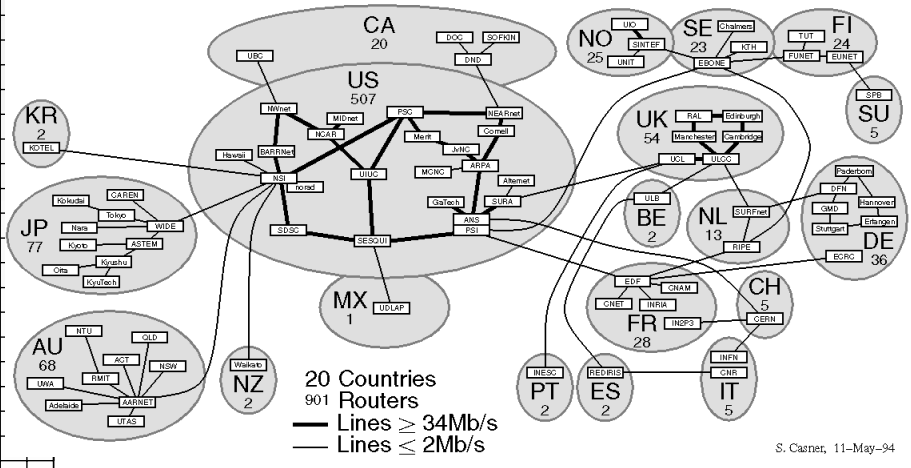
IP Multicast Group Management

Extensions to IP within one sub-net (IGMP)

- A multicast router periodically sends queries to all hosts participating in IP Multicast on the special 224.0.0.1 *all-hosts* group.
- Each relevant host sets a random timer for each group it is a member of. When the timer expires, it sends a report message on that group multicast address.
- Each host that gets a report message for a group cancel its local timer for that group.
- When a host joins a group it announces that on the group multicast address.

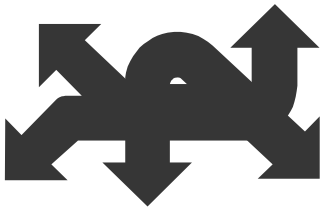
Multicast Backbone

Major MBONE Routers and Links



Multicast Backbone(cont)

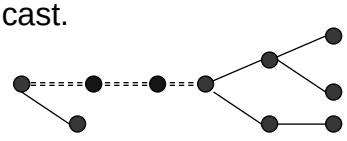
- BIG !! (International, inter-network,routers)
- Scalable?
  - Number of Sessions
  - Number of Recipients
- Variable Bandwidth (weakly connected islands)
- Challenges:
  - Routing
  - Reliability
  - Flow control



IP-Multicast Routing

Extensions to IP between the routers:
Distance Vector Multicast Routing Protocol (DVMRP, PIM-DM)

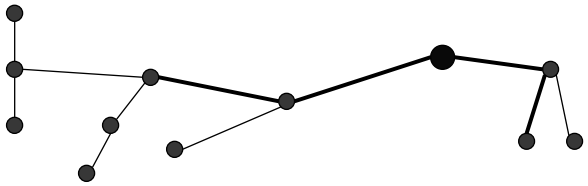
- Messages *ABOUT* groups are sent on the special *all hosts* group 224.0.0.1
- Time to live: limits the distance messages travel.
- Flood & Prune: All routers get packets initially, then prune out parts of the network that don't have group members.
- Tunneling: encapsulates multicast packets in regular packets in order to pass through routers that do not support IP Multicast.



IP-Multicast Routing(cont)

Multicast Routing Protocols
(MOSPF, PIM-SM, CBT)

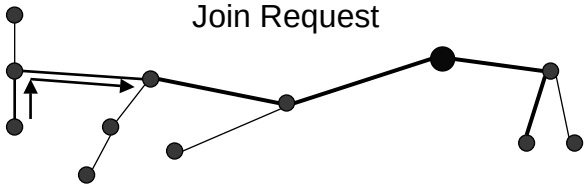
- No Tunneling
- Hierarchical Routing
- Dense vs. Sparse Routing
- Scalability(Router state only where needed, No floods, shared trees)
- Inter-AS Routing (MBGP+)



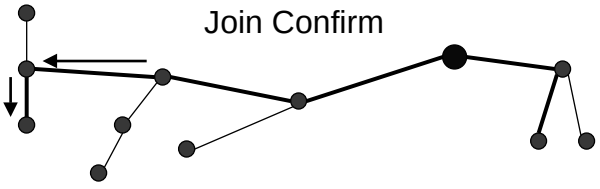
IP-Multicast Routing(cont)

Multicast Routing Protocols
(PIM-SM, CBT) Join Operation

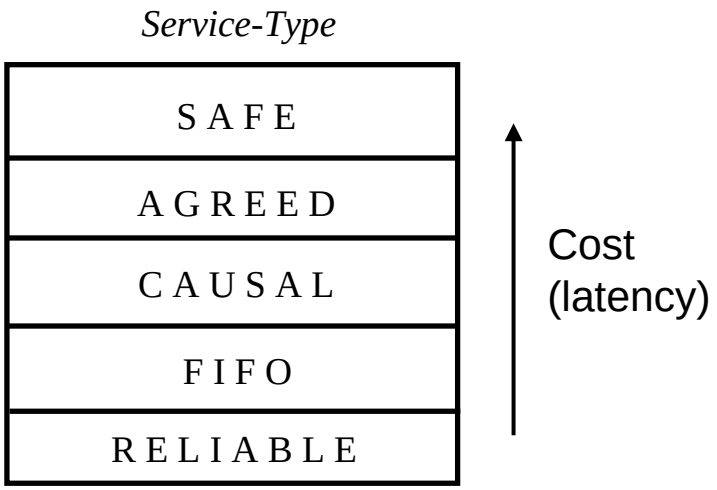
Join Request



Join Confirm



Reliable Multicast Services



Reliable Multicast Services (cont.)

Fifo Order

$m \overset{cause}{\dashrightarrow} m'$ if $send_q(m) \dashrightarrow send_q(m')$

Causal Order

$m \overset{cause}{\dashrightarrow} m'$ if $receive_q(m) \dashrightarrow send_q(m')$

Agreed Order

- Total order
- Consistent with Causal order and overlapping groups

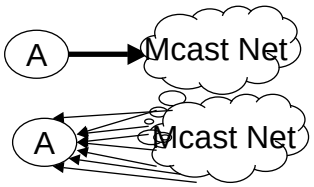
Safe Delivery

* Not ordering

- Consistent with Agreed order.
- Message is delivered after received by all processors.

Reliable IP-Multicast

- Hard Problem!
  - Implosion
  - Latency
  - Redundant traffic

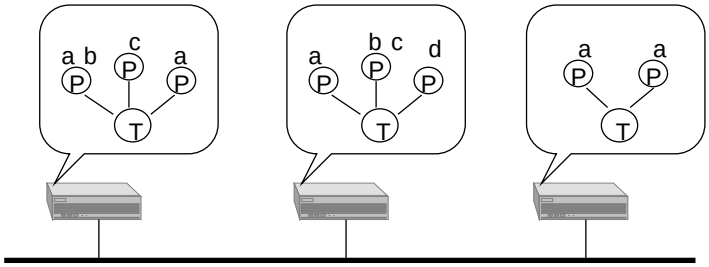


- Ack vs. Nack?
- Unicast or Multicast repairs and requests?
- Localized recovery or source based?
- Stability of data
- Router or end-host based?

Transis: A Group Communication Layer

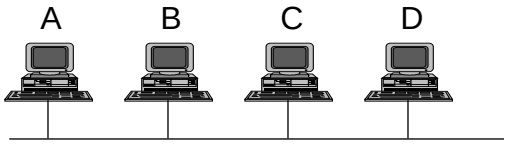
- Utilizes hardware broadcast ✓
- Reliable delivery ✓
- Automatic flow control ✓
- Connection service ✓
- Scales up. ✓
- Consistent semantics. ✓

Process groups in Transis



- One Transis daemon in each machine
- Multiple destination groups per message
- Message ordering across groups
- Open group semantics

The Trans Protocol



A Emits: A_1 A_2 A_3 ...

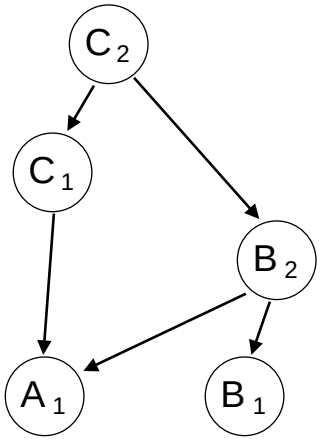
Scenario: A_1 a_1B_1 b_1B_2 b_2C_1 ...

Direct Ack: $a_1B_1 \rightarrow A_1$

Indirect Ack: $b_2C_1 \rightarrow A_1, B_1$

Example

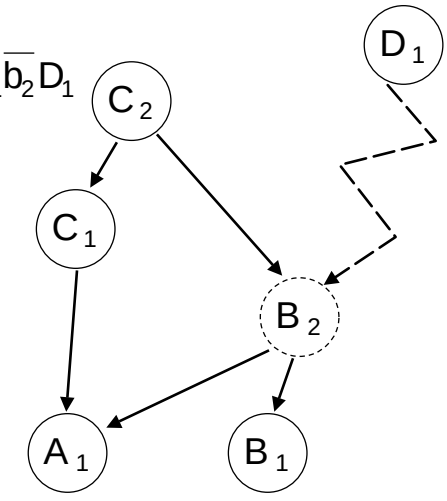
Scenario: A_1 B_1 $a_1b_1B_2$ a_1C_1 $c_1b_2C_2$...



Example (Cont.)

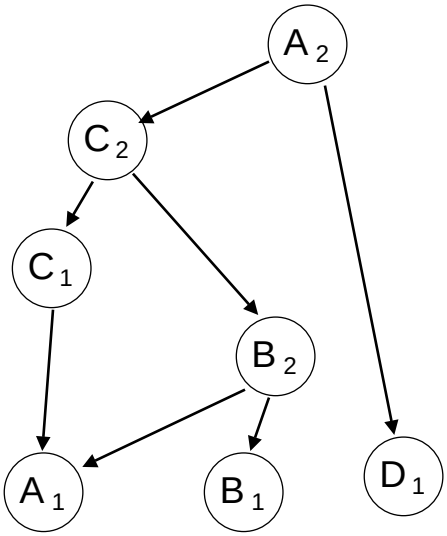
D received: A_1 B_1 a_1C_1 $c_1b_2C_2$...

Nack:: $c_1b_1\overline{b_2}D_1$

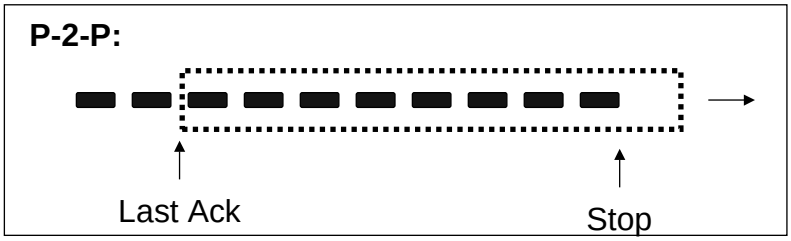


The DAG

The DAG may be revealed in a different way, but its structure will be identical at all the processors!!

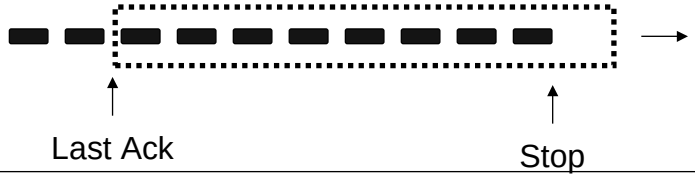


Flow Control

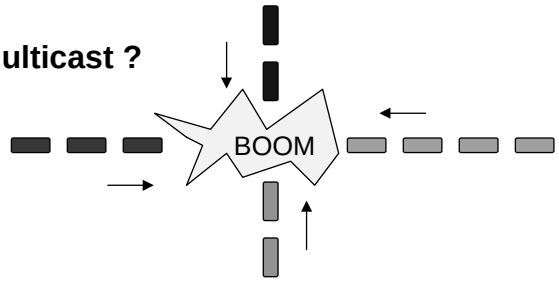


Flow Control

P-2-P:



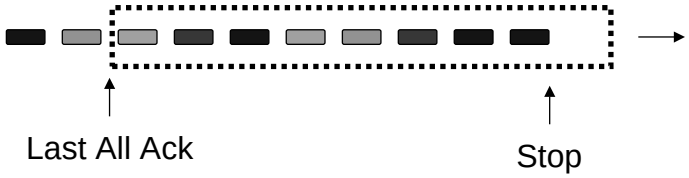
Multicast ?



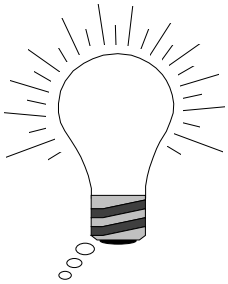
Transis Flow Control

Multicast:

Network Sliding Window



Other Services



- Membership
- Ordering
- Inter-operability
- Relaxed semantics
  - Reliability (Probabilistic,Soon Enough,FEC)
  - Policy Routing
- Real Time Guarantees and QoS