Survey of ALM, OM, Hybrid Technologies

John Buford
Panasonic Princeton Laboratory

July 13, 2006

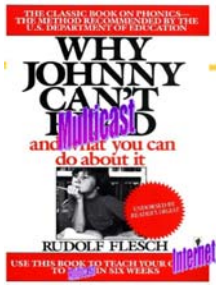
Topics

- Problem statement
- Terminology
- ALM
- OM
- Hybrid
- Summary of ALM and OM
- Next steps

Problem Statement

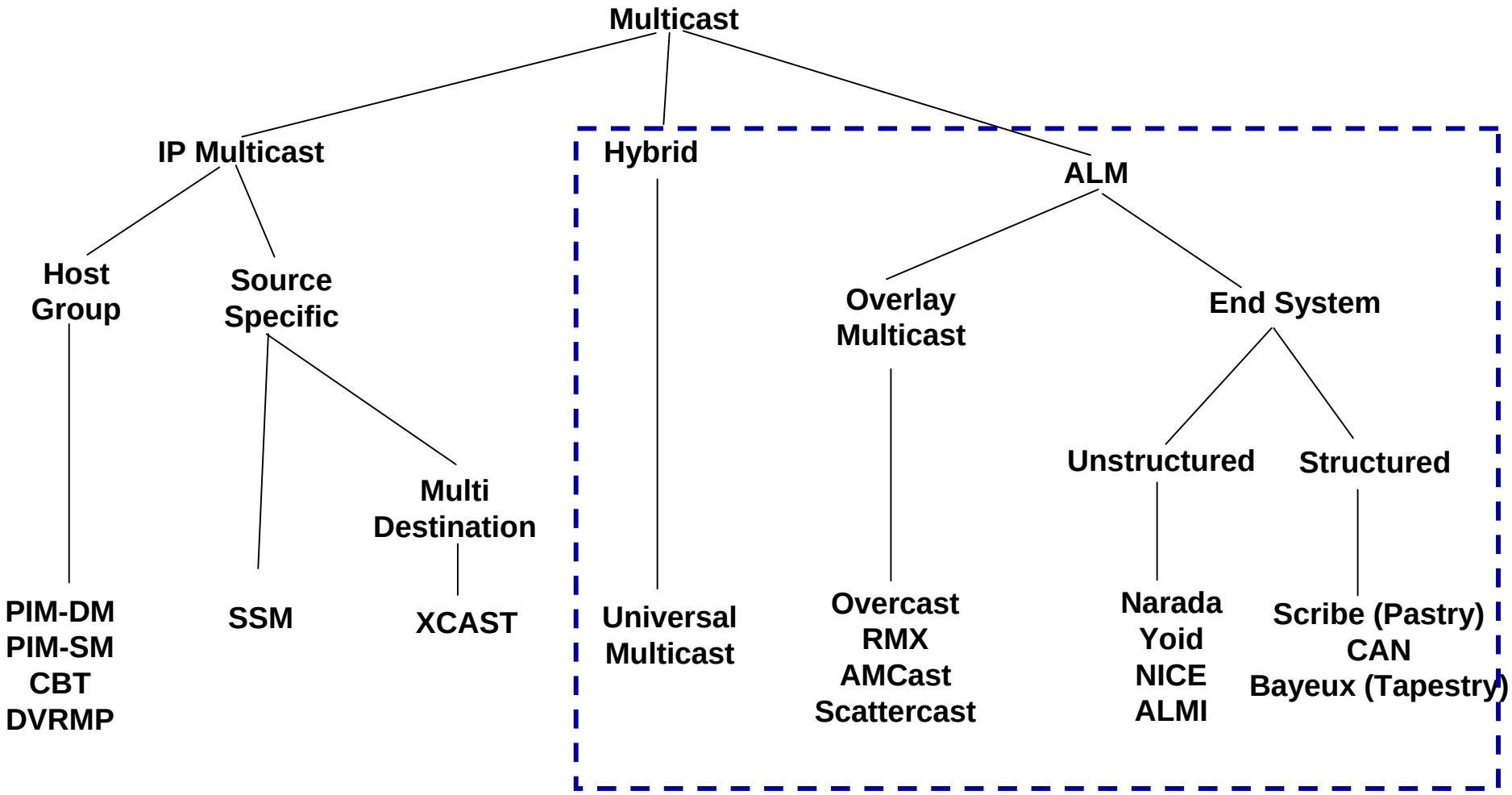
- IP Multicast:
  - Many possible applications but slow deployment
- Factors frequently cited:
  - Number of network devices that need modification
  - Inter-domain deployment issues
  - Hardware lifetimes
  - Global deployment requirement
  - Pricing model
  - Need for a scalable inter-domain multicast routing protocol

- Problem statement
  - Offer more flexible deployment options
  - Accelerate deployment of native multicast
    - Allow use with incremental deployment
    - Enable growth of multicast applications to create market demand to drive business case for network upgrade
  - Address other dimensions of multicast scalability
    - Highly dynamic group membership
    - Millions of small groups
  - Address other network environments
    - Concatenated VPNs (I.e., GIG), Mobile Networks



Mostafa Ammar. Why Johnny Can't Multicast Lessons about the Evolution of the Internet. Keynote - NOSDAV 03.

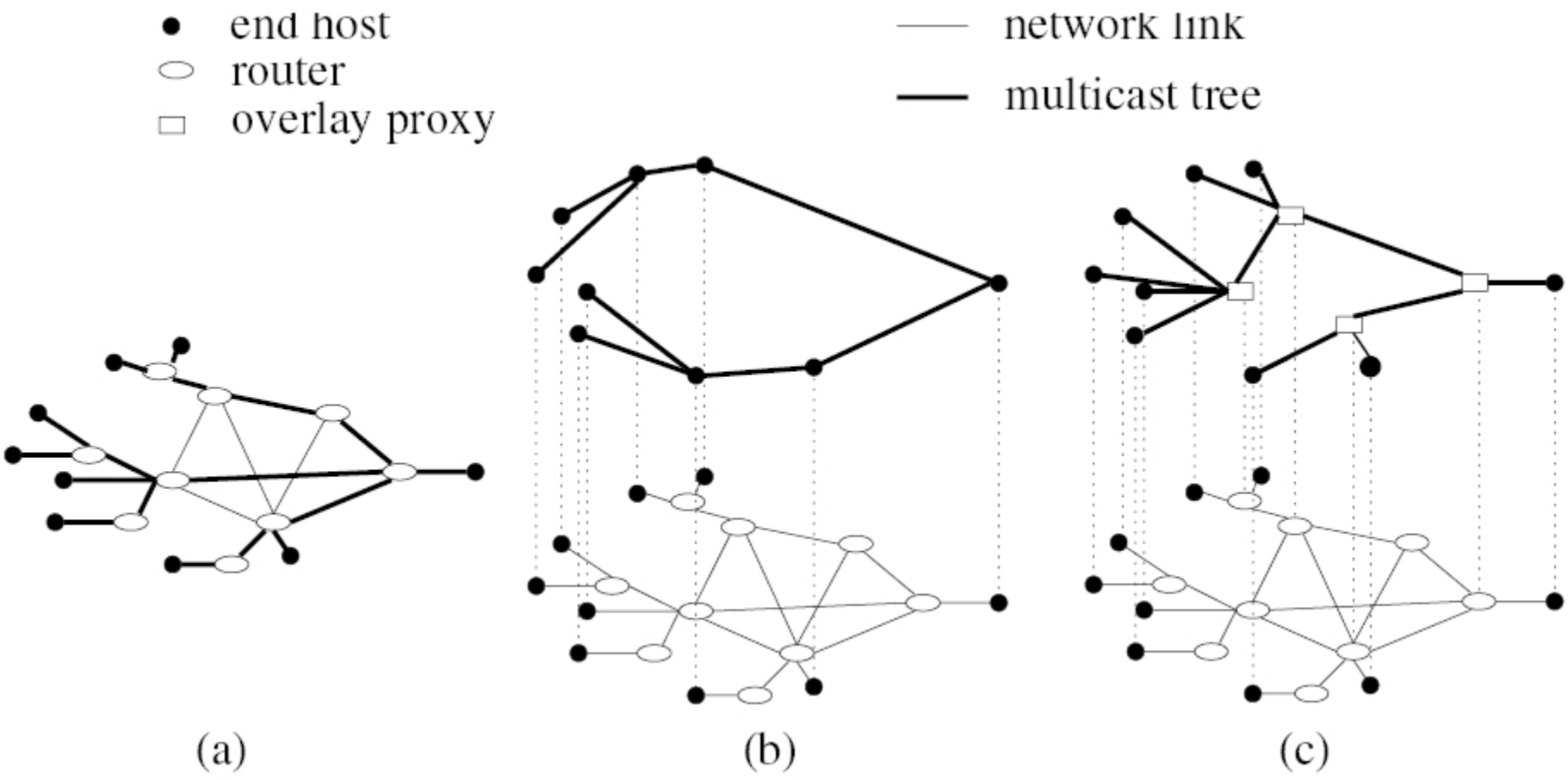
Taxomony


Multicast
IP Multicast
Hybrid
ALM
Host Group
Source Specific
Multi Destination
Overlay Multicast
End System
Unstructured
Structured
PIM-DM
PIM-SM
CBT
DVRMP
SSM
XCAST
Universal Multicast
Overcast
RMX
AMCast
Scattercast
Narada
Yoid
NICE
ALMI
Scribe (Pastry)
CAN
Bayeux (Tapestry)

Terminology

- Application Layer Multicast
  - Multicasting functionality is implemented at the application layer, i.e. at the end-hosts instead of the network routers
- Overlay Multicast
  - Construct a backbone overlay by deploying special intermediate proxies, proxies create multicast trees among themselves
  - End hosts communicate with proxies via unicast or native multicast
  - (Might this better be called Proxied Overlay Multicast?)
- Hybrid Multicast Architecture
  - Combine ALM and native multicast to provide end-to-end service

IP Multicast, ALM, and OM



IP Multicast **ALM** **OM**

L. Lao, J.-H. Cui, M. Gerla and D. Maggiorini. A Comparative Study of Multicast Protocols: Top, Bottom, or In the Middle? in *Proceedings of 8th IEEE Global Internet Symposium (GI'05)* in conjunction with IEEE INFOCOM'05, Miami, Florida, March 2005.

Terminology

- Tree creation types
  - Centralized
    - Tree manager node collects RTT measurements from group members and computes minimum spanning tree
    - Limited scalability but good tree quality
    - Example: ALMI
  - Mesh-based
    - Members of a group are connected in a mesh
    - Tree is formed using conventional routing algorithm over mesh
    - Need a link evaluation to select links from mesh
    - Scalability: O(n) state for mesh, but see one-hop DHT designs
    - Examples: ESM, Scattercast

Terminology

- Tree creation types
  - Tree-based
    - Group members self-organize
    - Explicitly pick a parent for each new member
    - Needs loop detection and tree-reconnection
    - State is O(E) so scalability to large groups
    - Examples: Yoid, HMTP
  - Implicit
    - Creates a control topology with specific properties such as hierarchy or locality
    - Inherit packet forwarding rule implicitly defines data tree
    - Examples: NICE, CAN, Scribe/Pastry, Bayeux/Tapestry

Terminology

- Rendezvous Point (RP)
  - A designated node for a multicast group which is contacted when a node wants to join that group
  - Used in centralized and tree-based tree construction types.

Terminology - Metrics

- Stress: counts the number of identical packets sent by the protocol over that link or node.
- Stretch: the ratio of the path length along the overlay from the source to the member to the length of the direct unicast path
- Degree: number of edges connecting this node to adjacent nodes in tree
- End-to-end delay
- Control message overhead
- Robustness

Application Layer Multicast

- Many research systems in the past few years have demonstrated the possibility of multicast using end-systems as the routing agent
  - "Application Layer Multicast" (ALM) or "End System Multicast" (ESM)
- Some peer-to-peer overlays have also included support for application layer multicasting
- By moving to the application layer we avoid infrastructure deployment issues
  - But there is a performance penalty
  - And we don't leverage native multicast where it exists

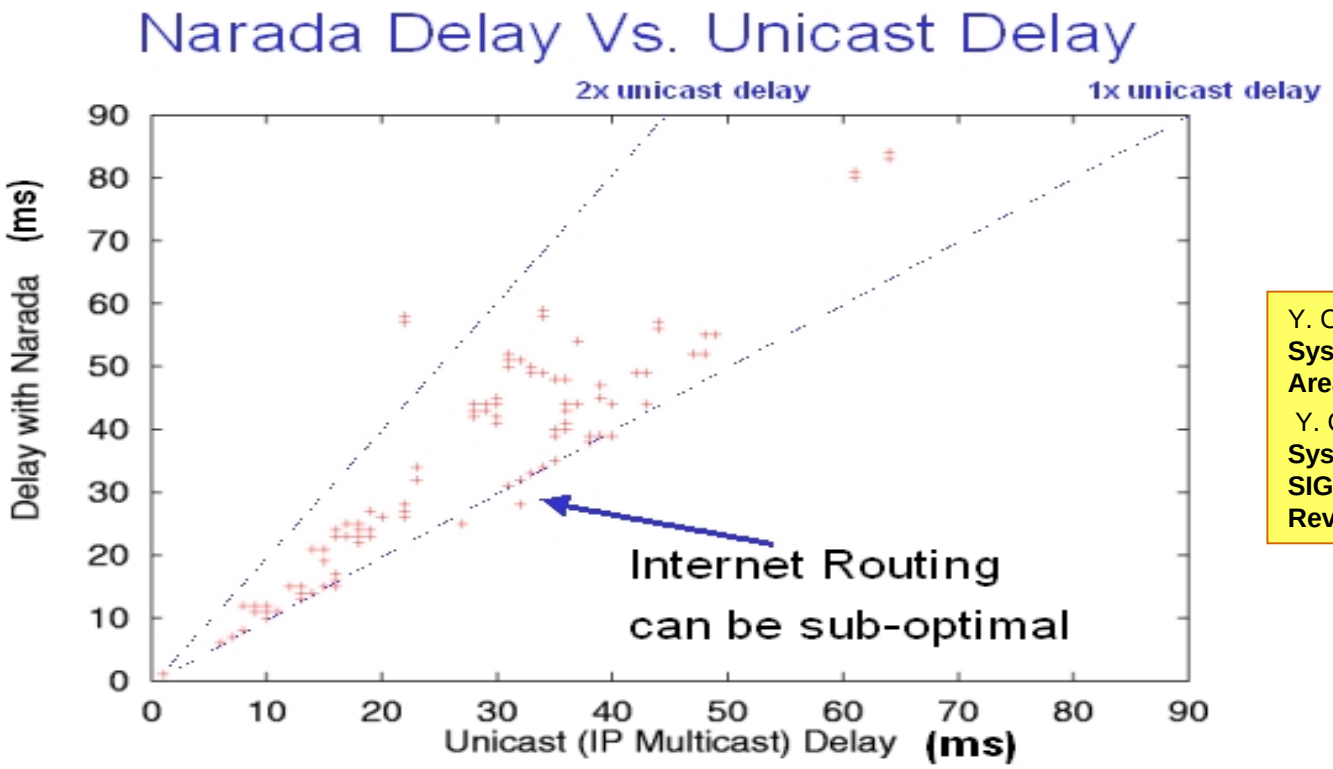
Application Layer Multicast

- Basic idea
  - Multicast is controlled only by participating end-hosts, including group membership, multicast delivery path, and data forwarding without explicit support of intermediate routers or proxies
- How it works
  - A rendezvous point (RP) is registered in a public directory
  - Each node has application software for connecting to multicast sessions
  - Various ways to join the multicast tree, such as:
    - RP sends root node to joining node, and node sends join request to root node.
    - Root propagates request through the tree.
    - Node selects response from possible join points and accepts the best one.

Application Layer Multicast

- Advantages
  - No infrastructure upgrade required
  - Scalability
    - Routers do not need to maintain per-group state
    - End systems do, but they participate in very few groups
  - Leverage solutions for unicast congestion control and reliability
  - No special addresses needed
  - Deployment in hands of user, software download
- Disadvantages
  - Inefficient trees lead to longer latency
  - Dependent on host resources and availability
  - Departing host effects downstream hosts
  - Doesn’t leverage native infrastructure support where it exists

Application Layer Multicast



Y. Chu, S. Rao and H. Zhang**. A Case for End System Multicast. IEEE Journal on Selected Areas in Communications, 2002**

Y. Chu, S. Rao and H. Zhang**. A Case for End System Multicast. (Keynote) ACM SIGMETRICS Performance Evaluation Review, 2000**

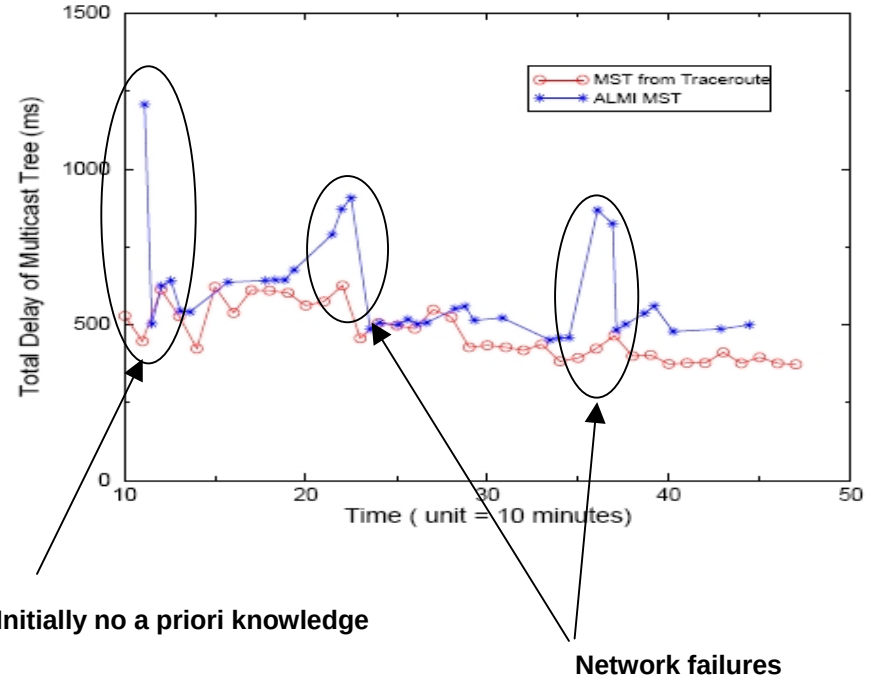
ALMI

- Application Layer Multicast Infrastructure (ALMI)
  - Support of multicast groups of relatively small size (several 10s of members) with many to many semantics
- Centralized session controller node manages a tree
  - Join/leave messages go to session controller
  - Tree is formed as degree-bounded minimum spanning tree according to desired cost metric (e.g., RTT time)
    - Nodes send background probe messages and report results to session controller
- Reliability mechanisms
  - Downstream nodes buffer packets until leaf ACKs are propagated back to the root node, which sends confirmation to all nodes
  - Branches in tree are unicast TCP connections
    - If retranmission fails, then receiver can form separate temporary connection to source to receive missing packets

Dimitrios Pendarakis,Sherlia Shi,Dinesh Verma, and Marcel Waldvogel. ALMI: An Application Level Multicast Infrastructure *Proc. of the 3rd USNIX Symposium on Internet Technologies and Systems,* March 2001.

ALMI: Experimental Evaluation

- Experiment
  - Single tree, 9 trans-atlantic sites, 6 hours
  - Controller recalculates tree every 5 minutes
- Results (see graph)

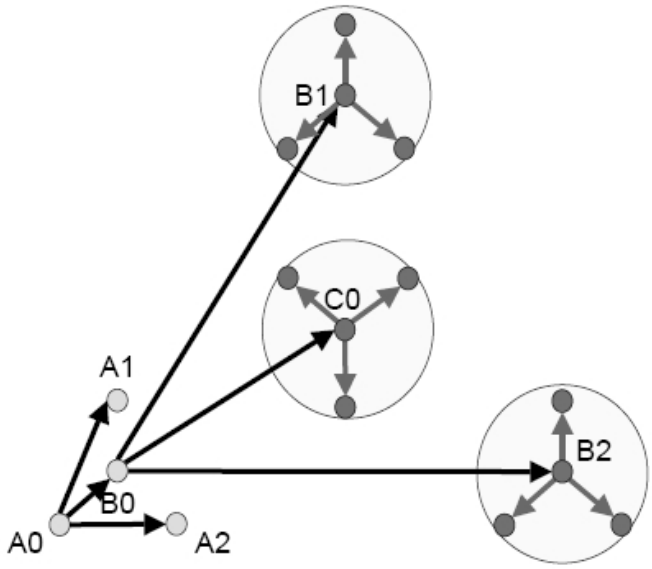


- Analysis vs ESM and Yallcast:
  - "Yallcast and Endsystem Multicast have their end goals align with those of ALMI, the tree construction algorithms are very different in all three protocols. Both Yallcast and Endsystem multicast try to leverage the existing multicast routing protocols and re-apply them at the application level. However, we argue that one of the fundamental complexities comes with IP multicast is its complication in routing protocols."

Dimitrios Pendarakis,Sherlia Shi,Dinesh Verma, and Marcel Waldvogel. ALMI: An Application Level Multicast Infrastructure *Proc. of the 3rd USNIX Symposium on Internet Technologies and Systems,* March 2001.

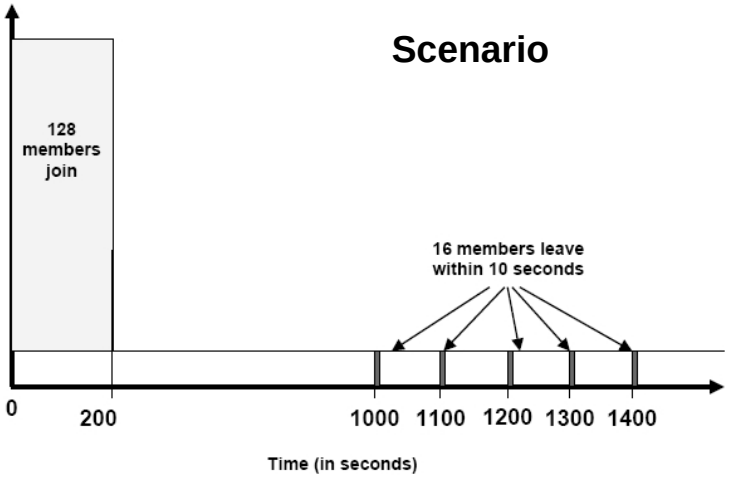
NICE

- Hierarchical
  - Proximity-based clusters
  - Log N layers
  - Clusters managed by soft state, maintained by heart-beats (cluster size of k nodes)
- Two topologies: control (for restructuring overlay), data
- Messages
  - Join: O(log N) RTTs and O(k log N) messages
  - Leave
  - Cluster Split, Merge, Refine

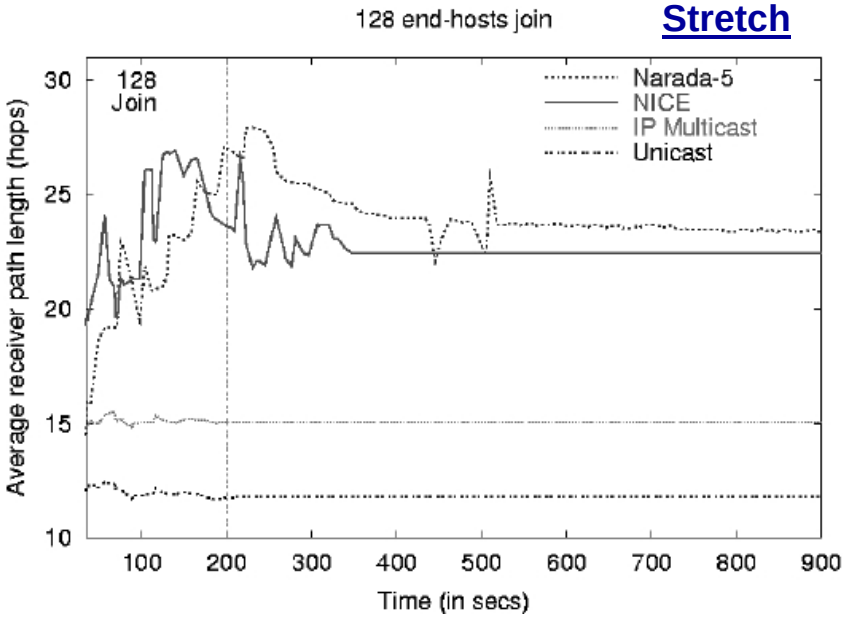


S. Banerjee, B. Bhattacharjee, and C. Kommreddy Scalable Application Layer Multicast. *Proceedings of ACM SIGCOMM* 2002.

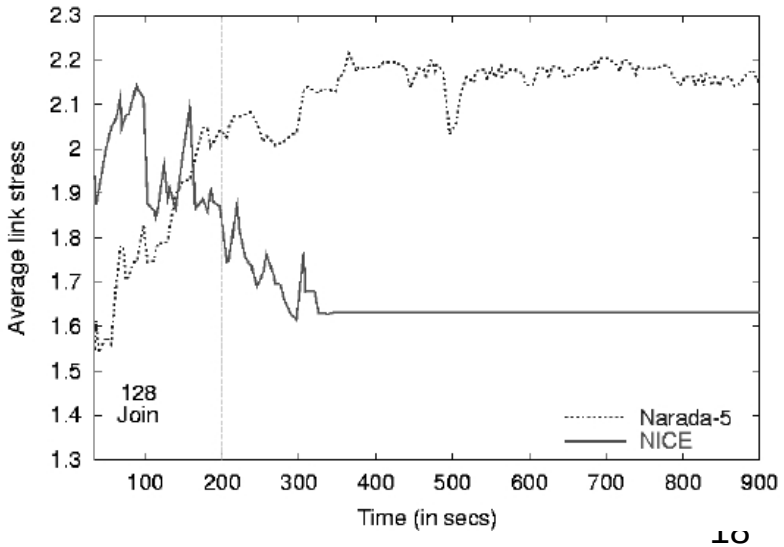
Nice vs Narada - Stress & Stretch



Stretch



Stress



S. Banerjee, B. Bhattacharjee, and C. Kommreddy Scalable Application Layer Multicast. *Proceedings of ACM SIGCOMM* 2002.

Beyond ALM

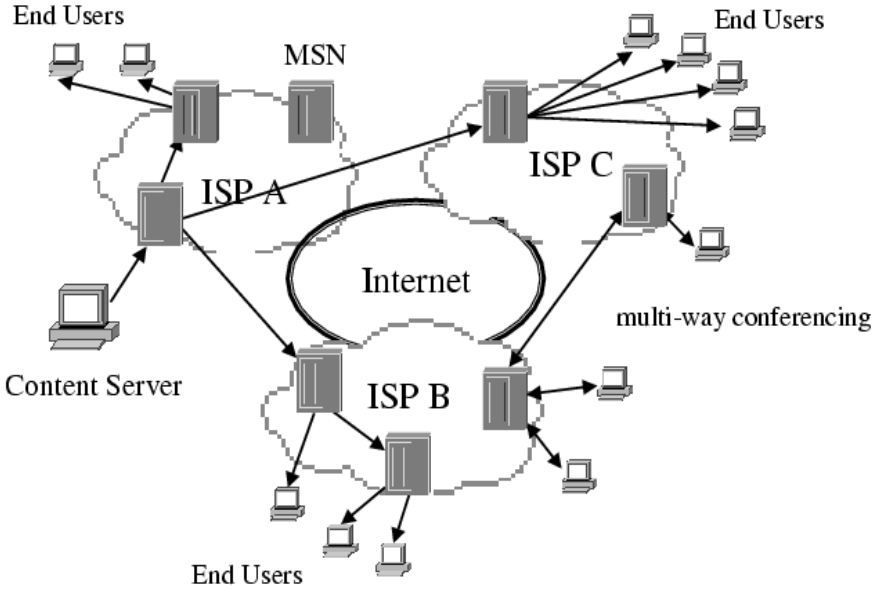
- Because of ALM's performance issues, what if special nodes are placed in the network ?
  - This is called Overlay Multicast (OM)

Overlay Multicast

- Basic idea
  - Construct a backbone overlay by deploying special intermediate proxies
  - Proxies create multicast trees among themselves
  - End hosts communicate with proxies via unicast or native multicast
- Examples
  - Overcast, RMX, OMNI, Scattercast, Amcast

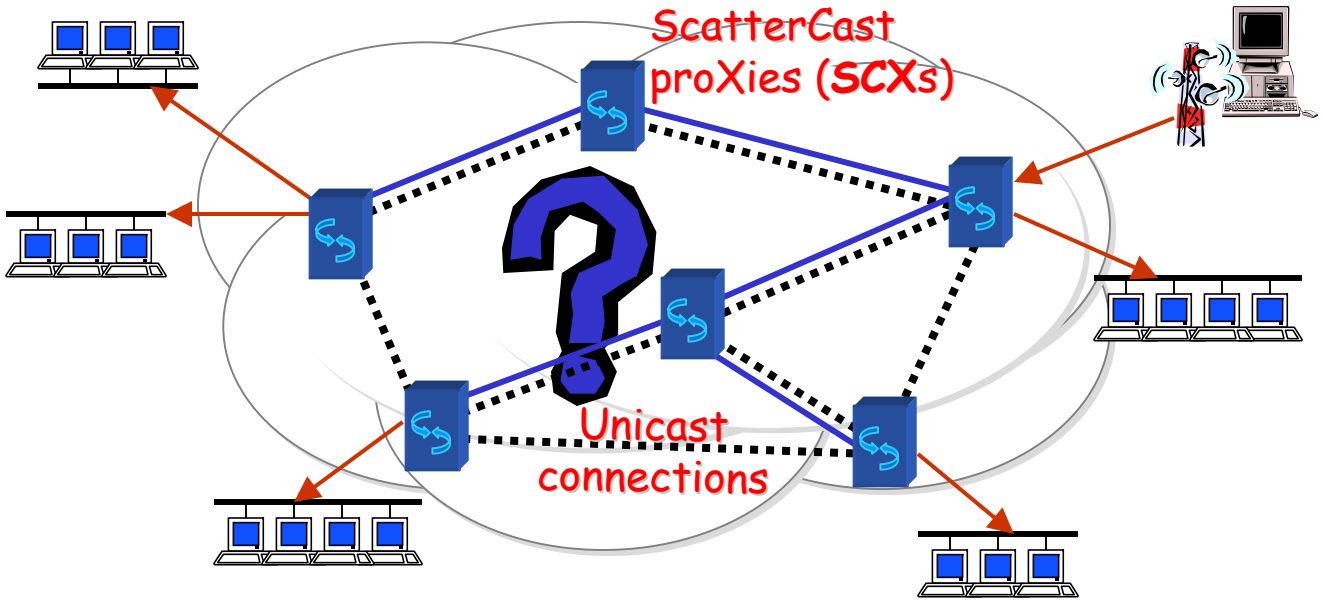
OM Example: AMcast

- Design problems
  - Where to place Multicast Service Nodes (MSNs)
  - How much bandwidth capacity should each MSN have, and how is that related to its geographic position
  - Balancing delay with bandwidth usage



Sherlia Y. Shi and Jonathan S. Turner, Multicast Routing and Bandwidth Dimensioning in Overlay Networks *IEEE Journal on Selected Areas in Communications*, Vol.20, No.8. October 2002.

Scattercast

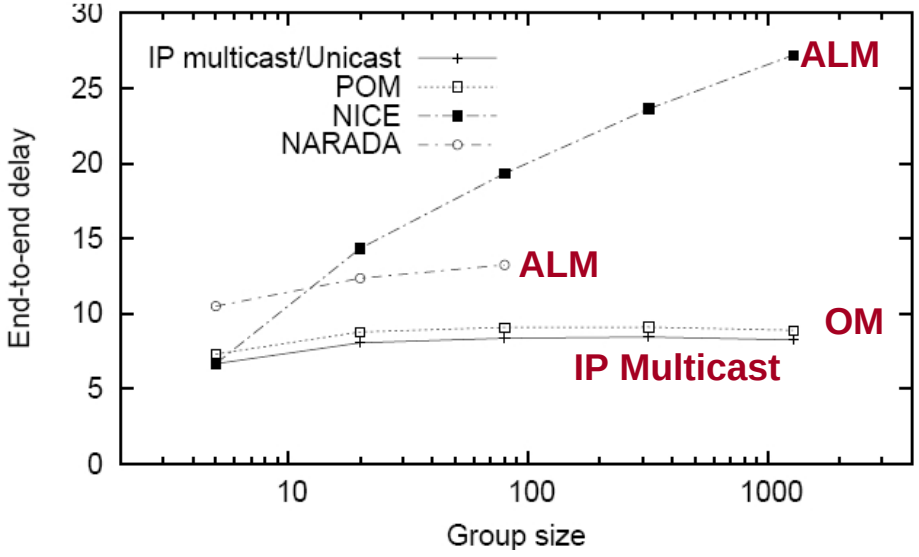


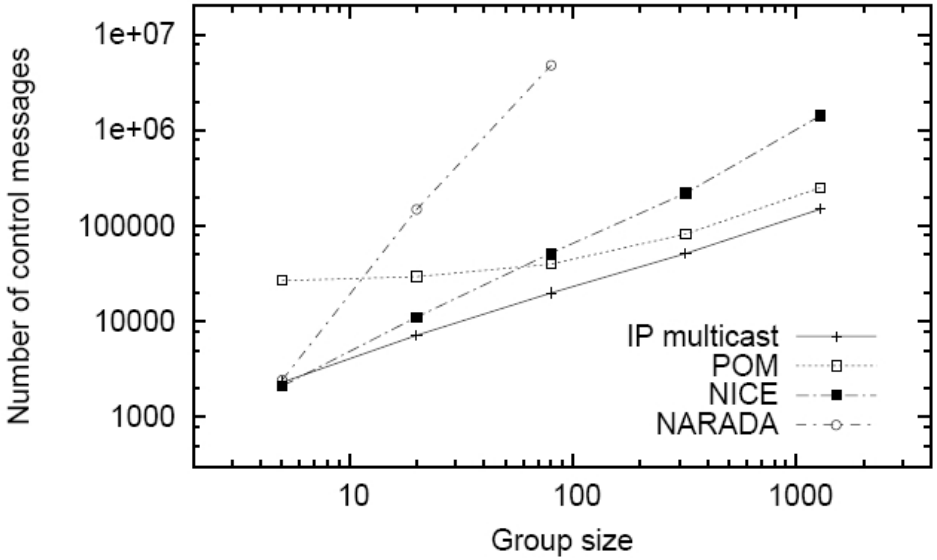
- Source injects data into a session via its local SCX
- SCXs dynamically construct overlay network of unicast connections: the *mesh*
- Run DVMRP-style routing on top of this network to construct distribution trees
- Restrict degree of each SCX based on its bandwidth capabilities

Yatin Chawathe. Scattercast: An Adaptable Broadcast Distribution Framework. In a special issue of the *ACM Multimedia Systems Journal* on Multimedia Distribution, 2002.

Overlay Multicast

- Advantages
  - Doesn't require router upgrade
  - Performance can approach native multicast





L. Lao, J.-H. Cui, M. Gerla and D. Maggiorini. A Comparative Study of Multicast Protocols: Top, Bottom, or In the Middle? in *Proceedings of 8th IEEE Global Internet Symposium (GI'05)* in conjunction with IEEE INFOCOM'05, Miami, Florida, March 2005.

Overlay Multicast

- Disadvantages
  - Requires infrastructure deployment
    - Host level rather than router level
  - Requires provisioning decisions
    - Where to place multicast service nodes (MSNs)
    - How much bandwidth capacity should each MSN have, and how is that related to its geographic position
  - Faces inter-domain interoperability issues

Beyond OM

- OM seems to offer a middle ground between ALM and native multicast
  - Better performance than ALM
  - Simpler deployment than native multicast
- But
  - Requires wide deployment to provide service through out network

Hybrid Approaches

- Combine islands of IP multicast deployment with application level multicast
- Dynamically map ALM path to underlying IP multicast path where available to optimize performance
- Within a region, dynamically transition multicast groups and flows between multicast protocols/mechanisms in response to changes in traffic characteristics, group properties, and network topology

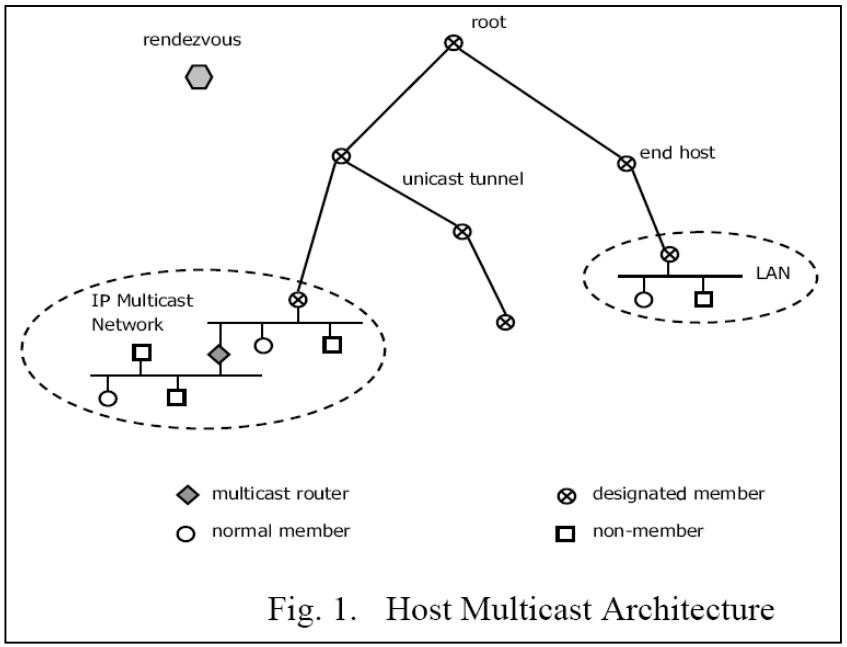


Fig. 1. Host Multicast Architecture

B. Zhang, S. Jamin, and L. Zhang. Universal IP multicast delivery. In *Proc. of the Int'l Workshop on Networked Group Communication (NGC)*, Oct. 2002

- Issues
  - Mapping different join/leave and routing protocols
  - Different group management mechanisms
  - Application sensitivity to performance variations

Hybrid Approaches

- Advantages
  - Enables end-to-end multicast with incremental native multicast roll-out
- Disadvantages
  - Complexity and performance loss due to
    - Mapping different join/leave and routing protocols
    - Brokering different group management mechanisms
  - Application sensitivity to performance variations

HMTP: Host Multicast Tree Protocol

- Combines ALM with Native Multicast (NM)
- Each NM island has a host called designated member (DM)
- IP Multicast islands are connected by bi-directional UDP tunnels between DMs
- RP
  - Initial join point for joining group
  - assocates Group ID with local addresses

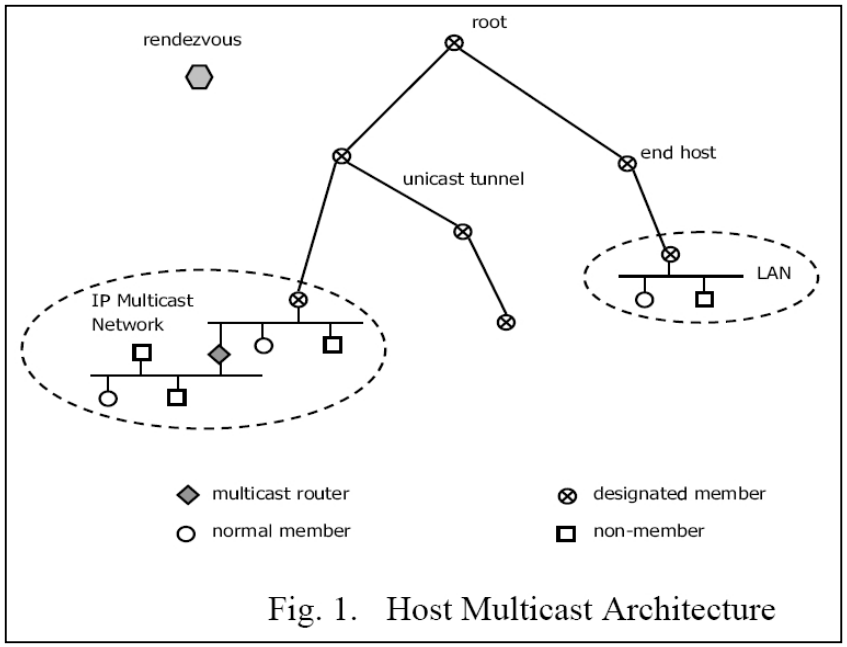


Fig. 1. Host Multicast Architecture

Beichuan Zhang, Sugih Jamin, and Lixia Zhang, Host Multicast: A Framework for Delivering Multicast To End Users *Proc. of IEEE INFOCOM'02* June 2002.

HTMP

- Constructs group-shared tree
- RTT used as distance metric
- Join
  - Newcomer contacts root for list of children
  - Selects next closest one and iteratively descends tree until closest node is found
  - Request to join is determined by existing node depending on load, capacity, etc.
  - If request fails, goes to next closest node

HTMP

- Maintenance
  - Children periodically send refresh message to parent
  - Parent sends its path to root back to children
  - Missing messages are signal of disconnection
- Other
  - Tree improvement
  - Partition recovery
  - Loop detection and recovery

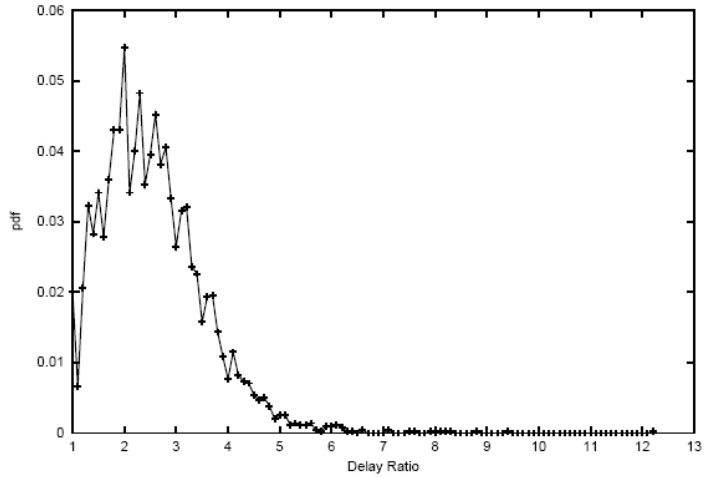


Fig. 11. Distribution of delay ratio in a HMTP tree of 100 members. Last hop delay is 3ms.

Universal Multicast (UM)

- A framework for hybrid multicast
  - Uses HMTP and Native Multicast
- Used at several SIGCOMM conferences
- UM/HMTP demonstrates feasibility of hybrid approach for best effort service

B. Zhang, S. Jamin, and L. Zhang. Universal IP multicast delivery. In *Proc. of the Int'l Workshop on Networked Group Communication (NGC)*, Oct. 2002

Next Steps

- Several ALM summaries exist (see next 2 slides)
- Different design points
  - Small groups
  - Tree metrics
  - Latency in joining/reconfiguring
- ALM analysis could
  - Categorize by adaptativity criteria
  - Incorporate other dimensions of work (e.g., QoS, Mobility)

Summary of Selected ALM Designs

Scheme	Type	Tree-type	Max. Path length	Max. Tree degree	Avg. Control Overheads
Narada	Mesh-first	Source-specific	Unbounded	Approx. bounded	$O(N)$
HMTP/Yoid	Tree-first	Shared	Unbounded	$O(\text{max. degree})$	$O(\text{max degree})$
Bayeux/Scribe	Implicit	Source-specific	$O(\log N)$	$O(\log N)$	$O(\log N)$
CAN-multicast	Implicit	Source-specific	$O(dN^{1/d})$	constant	constant
NICE	Implicit	Source-specific	$O(\log N)$	$O(\log N)$	constant

TABLE I

A COMPARISON OF DIFFERENT APPLICATION LAYER MULTICAST SCHEMES.

"In general, it is difficult to analytically compute either the stretch or stress metrics for most of the protocols. In particular, an analysis of the stress metric significantly depends on the characteristics of the underlying topology."

- Mesh-first protocols are efficient for small multicast groups, while implicit protocols scale well with increasing group sizes.
- Tree-first protocols are less suited for latency sensitive (e.g. real-time) applications but are useful to implement for high-bandwidth data transfers.
- Implicit protocols are particularly beneficial when the size of the multicast group is very large, and can be adapted for both latency-sensitive applications (due to their short path lengths) and high-bandwidth applications (due to low tree degree).

S. Banerjee, B.Bhattacharjee. **A Comparative Study of Application Layer Multicast Protocols. Submitted**

Summary of Design Choices

Table 1. Design choices for each of the application level multicast solutions

	Routing	Multicast tree construction	Tree type	Overlay management	Group size	Closeness metric	Delivery guarantees
Narada	Tree	Mesh first	Source sp.	Distributed	Small	Latency	AMO
ALMI	Tree	Mesh first	Shared	Centralized	Medium	Latency	AMO
Yoid	Tree	Tree first	Shared	Distributed	Medium	Data Loss	ZOM
NICE	Tree	Implicit	Source sp.	Distributed	Large	Latency	AMO
Bayeux	Tree	Implicit/Tree-first	Source sp.	Distributed	Large	none	AMO
CAN	Intelligent flooding	N/A	N/A	Distributed	Large	Latency	ZOM
Scribe	Tree	Similar to RPF	Shared source sp.	Distributed	Large	Latency	AMO
SplitStream	Multiple trees	Any	Source sp.	Distributed	Large	Latency	AMO
Bullet	Mesh	Any	Source Sp.	Distributed	Large	E2E bandwidth	ZOM
Lpbcast	Random flooding	N/A	N/A	Distributed	Large	none	ZOM
BTP	Tree	Tree first	Shared tree	Distributed	Medium	Latency	AMO
Overcast	Tree	Tree first	Shared	Distributed	Not comparable	10KB download time & traceroute distance	AMO
HostCast	Tree	Tree first	Source sp.	Distributed	Medium	Available bandwidth & root-path latency	AMO

Cristina Abad, William Yurcik, and Roy H. Campbell. A Survey and Comparison of End-System Overlay Multicast Solutions Suitable for Network-Centric Warfare, *SPIE Defense and Security Symposium / BattleSpace Digitization and Network-Centric Systems IV*, 2004