July 1985

Report No. STAN-CS-85-1058
Also numbered CSL-85-280

Host Groups: A Multicast Extension for Datagram Internetworks

by

David R. Cheriton and Stephen E. Dewing

Department of Computer Science

Stanford University
Stanford, CA 94305

Host Groups: A Multicast Extension for Datagram Internetworks

David R. Cheriton
Stephen E. Deering

Computer Systems Laboratory
Stanford University

Abstract

The cxtcnsivc usc of local networks is beginning to drive requirements for internetwork facilities that connect these local networks. In particular. the availability of multicast addressing in many local net works nnd its usc by sophisticated distributed applications motivates providing multicast across internetworks.

In this paper. we propose a model of scrvicc for multicast in an internetwork, describe how this service can bc used, and describe aspects of its implementation, including how it would fit into one existing internetwork architecture, namely the US DoD Internet Architecture.[1] [2]

1. Introduction

Multicast is the transmission of a datagram packet to a set of zero or more destination hosts in a network or intcrnctwork, wilh a single address specifying the set of destination hosls. For example, hosts A, B, C and D may bc associated with multicast address X. On transmission, a packet with destination address X is delivered with datagram reliability to hosts A, B, C and D.

Multicast has two primary uses, namely distributed binding and multi-destination delivery. It is useful for binding when one or more of a sc of hosts contain lhc dcsirccl object but particular host addresses arc not known, only a multicast address. For example, in a distributed file system, all the file scrvcrs may bc associated with one multicast address. To bind a file name to a particular server, a client sends a query packet containing lhc file name to the file server multicast address, which is delivered lo all the file servers. The server that recognizes the file name then responds to tne client, allowing subsequent interaction directly with that server host. This also illustrates the use of multicast for *logical addressing*. The multicast address for a group of hosts can denote *function* rather than location. One can similarly associate the group of time scrvcrs, name scrvcrs, computation servers and so on each with their own multicast address.

Multi-destination delivery is useful to several applications, including:

- edislributcclreplicateddatabases[1],
- conferencing[3].
- distributed parallel computation, including distributed gaming[4].

Ideally, multicast transmission lo a set of hosts is not more complicated or cxpcnsivc for the scndcr than transmission lo a single host. Similarly. multicast transmission should not be more cxpcnsivc for the network than traversing the shortest path tree that connects lhc sending host to the hosts identified by lhc multicast address.

Multicast, transmission lo a set of hosts. is properly distinguished from from *broadcast*, transmission lo *all* hosts on a network or internetwork. Broadcast is not a generally useful facility since lhcrc are few reasons for communicating with all hosts. In fact, it is best viewed as an "accident of the technology" for broadcast networks in the same way that self-modifying programs are an accident of lhc technology for stored program machines: just because the technology prol ides it does not mean it is efficient or safe to use. A proper multicast facility allows efficient transmission to multiple hosts while avoiding unnecessary loading of the netw ork and receiv ing hosts that arises with broadcast.

Multicast is now available in standard local networks[5]. For example, the Ethernet[6] prov ides 2^{47} multicast addresses. Sending a packet to an Ethernet multicast address delivers it (with datagram reliability) to the set of hosts listening to that multicast address. A variety of local network applications and systems make use of this facility. For instance, the V distributed system[7] uses network-level multicast for implementing efficient operations on groups of processes spanning multiple machines. Similar use is being made for replicated databases[1] and other distributed applications[8]. Prov iding multicast in the internetwork env ironment would allow porting such local network distributed applications lo the internetwork, as well as making some existing internetwork applications more robust and portable (by. for example, removing wired-in lists of addresses, such as gateway add rcsscs).

In current internetwork environments, an application logically requiring multicast must send indiv idually addressed packets to each recipient. There are two problems with lbis approach. Firstly, requiring the sending host lo know the specific addresses of all the recipients defeats its use as a binding mechanism. For example, a diskless workstation needs on boot to determine the network address of a disk server and it is undesirable lo "wire in" specific network addresses. With a multicast facility, the multicast address of the disk servers (or name servers that holds the nddrcss of the disk serv er) can be *well k now* n, allowing the workstation to transmit its initial queries lo this address.

Secondly, transmitting multiple copies of the same packet makes inefficient use of networ k bandwidth, gateway resources and sender resources. For instance, the same packet may repeatedly traverse the same network links and pass through the same gateways. Furthermore, the network level cannot recognize multi-destination delivery to take advantage of multicast facilities that the underlying networ k technologies may provide. For example, local-area bus, ring, or radio networ ks and even satellite-based wide-area networks can provide efficient multicast delivery directly. Besides using excessive communication resources, the use of multiple transmissions to effect multicast severely limits lhc amount of parallelism in transmission and

[1]This work was sponsored in part by the Defense Advanced Research Projects Agency under contract N00039-83-K-0431 and National Science Foundation Grant DCR-83-52048.

processing that can be achieved compared to an integrated multicast facility.

In this paper, we describe a model of multicast service we call *host groups* and discuss aspects of implementing this service in a datagram internetwork environment. We argue that it is feasible to implement this facility in an internetwork as an extension of the existing "unicast" internetwork datagram model and mechanism.

We restrict ourselves to the communication environment of a datagram-based internetwork, like the IP[9] or XNS[10] internetwork architectures. In these architectures, all hosts employ a common internetwork datagram format and a common internetwork addressing convention to identify the sources and destinations of datagrams. On transmission, an internetwork datagram is delivered to its destination address with "best efforts" reliability, via the transmission services of the underlying networks and the relaying services of the gateways. This service best corresponds to OSI layer 3 or the network level in providing host-to-host delivery. Reliable delivery, including error handling and flow control, is handled by higher-level protocols that operate in terms of internetwork datagrams.

Figure 1 illustrates a heterogeneous collection of independent networks interconnected by hosts that serve as store-and-forward *gateways* typical of datagram internetworks.

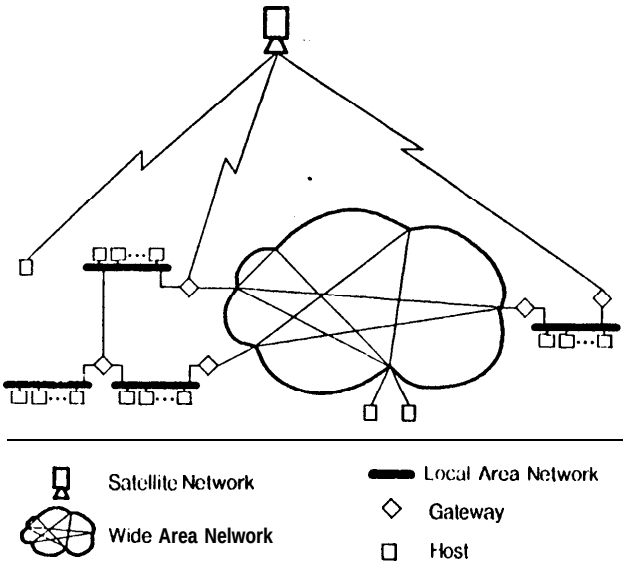

Figure 1 A Typical Internetwork

In Figure 1, a satellite network and a wide area, store-and-forward network connect several local area networks as well as individual hosts. The combination of broadcast and point-to-point technology plus the usual complications of different speeds, delay and maximum transmission unit make an efficient implementation of multicast a challenge

The next section describes the host group model of multicast service. Section 3 describes the implementation strategy we propose. Section 4 describes how this extension fits into the current US DoD Internet architecture and briefly touches on other internetwork architectures. Section 5 illustrates how this facility can be used by a variety of applications. Section 6 relates this model to other proposals. Finally, we conclude with remarks on the status of our experimental prototype implementation of host groups and our future directions for investigation.

2. The Host Group Model

In an internetwork designed in the host group model, each internetwork address identifies a host group. A *host group* is a set of zero or more hosts in one internetwork.[3] When an internetwork packet is sent, it is delivered with "best efforts" datagram reliability to all members of the host group identified by the internetwork address in the packet destination field.

The sender need not be a member of the destination group. We refer to such a group as *open*, in contrast to a *closed* group where only members are allowed to send to the group. We chose to provide open groups because they are more flexible and more consistent as an extension of conventional unicasts models (even though they are harder to implement).

Dynamic management of group membership provides flexible binding of internetwork addresses to hosts. Hosts may join and leave groups over time. A host may also belong to more than one group at a time. Finally, a host may belong to no groups at times, during which that host is unreachable within the internetwork architecture. In fact, an internetwork host need not have an individual internetwork address at all. Some hosts may only be associated with multi-host group addresses. For instance, there may be no reason to contact an individual time server in the internetwork, so time servers would not require individual addresses. Similarly, a bank of shared processors may be identical from the standpoint of clients and only acquire individual internetwork addresses while they are serving individual clients.

Internetwork addresses are dynamically allocated for *transient* groups. groups that often last only as long as the execution of a single distributed program. A range of host group identifiers is reserved for identifying *permanent* groups. One use of permanent host groups identifiers is for host groups with standard logical meanings such as "name server group", "boot server group", "internetwork monitor group", etc. Permanently assigned addresses are also used for conventional single-host addresses.

The host group model of internetwork generalizes the binding of internetwork addresses to internetwork hosts by allowing one address to bind to multiple hosts on multiple networks, more than one address to be bound (in part) to one host, and the binding of an address to host to be *dynamic*, i.e. possible to modify under application control. For performance reasons, the conventional case of single-member groups is handled specially as an optimization. A range of internetwork addresses are reserved for designating groups of at most one internetwork host, allowing the delivery mechanism to make appropriate optimizations. Moreover, if the internetwork address is statically bound to a host permanently attached through one network, a network identifier can be embedded as a subfield of its internetwork address in order to simplify gateway routing. As should be apparent, this special case corresponds to the unicast facility provided by several current datagram-based internetwork architectures, including IP and XNS. Thus, the host group model is a compatible extension of these architectures.

The following subsections provide further details of the model.

2.1 Host Group Management

Dynamic binding of internetwork addresses to hosts is managed by the following three operations available to higher-level protocols or applications:[4]

[3] In reality, the internetwork address is bound to network interfaces or host access ports, not the host machine per se.

[4] In this procedure call notation, the arguments for an operation are listed in parentheses after the operation name, and the returned values, if any, are listed after a --> symbol.

```
CreateGroup ( type )
  --> outcome, group-address, access-key
```

rcqucsts the creation of a new transient host group with the invoking host as its only member. The type argument specifies either a general group or a one-member-only group plus whether the group is rcstriclcd or unrestricted. A restricted group restricts membership based on the access-key. Only hosts presenting a valid host access key arc allowed IO join. All unrestricted host groups have a null access-key. outcome indicates whether the request is approved or dcnicd. If it is approved, a new transient group address is returned in group-address. access-key is the protection key (or password) associated with the new group. This should fail only if thcrc arc no free transient group ad drcsscs.

```
JoinGroup ( group-address, access- key )
  --> Ou tcome
```

rcqucsts that the invoking host become a mcmbcr of the identified host group (pcrmancnt or transient). outcome indicates whether the request is approved or dcnicd. A rcqucst may bc dcnicd ifthc access key is invalid.

```
LeaveGroup ( group-address )
  --> ou tcome
```

rcqucsts that the invoking host be dropped from mcmbcrship in the identified group (pcrmancnt or transient). outcome indicates whether the rcqucst is approved or dcnicd.

There is no operation to destroy a transient host group bccausc a transient host group is dccmcd to no longer exist when its mcmbcrship gccs to zcro.

Note that in conventional intcrnctworks allocation and binding of intcrnctwork addresses is typically performed statically by intcrnctworb administrators

2.2 Packet Transmission

Transmission of a packet in the host group model is controlled b y two parameters o f scope, one being the destination internetwork address and the other being the "distance" IO the members in the group. In particular,

```
Send ( dest-address, source-address,
                              data, distance )
```

transmits the specified data in an internctwork datagram to the hosts in the host group specified by dest-address that arc wilhin Ihc specified distance. The destination address is thus similar lo conventional networks except that delivery may bc to multiple hosts; the distance parameter requires further discussion.

I Distance may be measured in several ways. including number of network hops, time IO deliver and what m i g h t be called administrative distance. Administrative distance refers I o the distance between the administrations of two different networks. For example, in a company the networks of the research group and advanced development group might bc considered quite close to each other, networks of the corporate management more distant, and net works of other companies much more distant. One may wish lo restrict a query lo members within one's own administrative domain because servers outside that domain may n o t be trusted. Similarly. error reporting outside of an administrative domain may not be productive and may in fact be confusing.

Besides limiting the scope of transmission, the distance parameter can bc used to control the scope of multicast as a binding mechanism and to implement an expanding scope of search for a desired scrvicc. For instance, to locate a name server familiar with a given name, one might check with nearby name servers and expand the distance (by incrementing the distance on retransmission) lo include more distant name servers until the name is found.

To reach all members of a group. a sender specifics the maximum value for the distance parameter. This maximum must exceed the "diameter" of the internetwork.

The distnncc parameter can bc viewed as an extension of the time-to-live or hop count parameters that arc used in several intcrnctwork architectures to prevent infinite routing cycles. In those cases, the distance parameter basically ensures that the delivery mcchnnism only expends a finite amount of work in delivery and rhcrcforc discards a packet caught in a routing loop. The distance parameter in the host group model refines this finite bound into further gradations.

Rather than define specific semantics of the distance parameter in the model. wc see it having a refinement of the semantics of the time-to-live or hop count parameters specific to each intcrnctwork architccturc. However, in all cases, thcrc is a need f o r well-known boundaries values t h a t coincide with administrative domains. For instance. there is a need for a distance value that corresponds to "not outside this local network".

Packet rcccplion is the same as conventional architectures. That is,

```
Receive ()
  --> dest-address, source-address, data
```

returns Lhc next internetwork datagram that is, or has been, rcccivcd.

2.3 Delivery Requirements

W e identify several requirements f o r the packet delivery mcchnnism that arc essential to host groups being a useful and used facility.

Firstly, given Ihc predominance of broadcast local-area networks and the locality of communication IO individual netwoi ks. the delivery mechanism must be nblc to exploit the hardware's capability for very efficient multicast within a single local-nrca net work.

Secondly, the delivery mechanism must scale in sophistication IO efficient delivery across the intcrnctwork as intcrnctworks acquire high-speed wide-area communication links and high per formancc gateways. The former arc being provided by the introduction of high-speed satellite channels and long-haul fiber optic links. The latter arc made feasible by the falling cost of memory nnd processing power plus the increasing importance in controlling access to relatively unprotected l o c a l n e t w o r k environments. A host group delivery mechanism must bc able to take advantage of these trends as they materialize.

Finally, the delivery mechanism must a v o i d "systematic errors" in delivery to members of the host group. That is. a small number of repeated transmissions must result in delivery to all group members within the specified distance, unless a mcmbcr is disconnected or has failed. We refer IO this property as *coverage*. In gcncral. most reliable protocols make this basic assumption for unicast delivery. Il is important lo guarantee this assumption for multicast 3s well or else applications using multicast may fail in unexpected ways when coverage is not provided. For efficiency, the multicast delivery mechanism s h o u l d also avoid regularly delivering multiple copies of a packet IO individual hosts.

Failure notification is riot viewed as an essential requirement given the datagram semantics of delivery. I lowever, a host group cstcnsion of internetwork architectures such as IP and XNS

should provide "hint"-level failure notification as the natural extension of their failure notification for unicasl.

3. Implementation

In this section, wc sketch a design for implementing the host group model in a datagram internetwork. This description of the design is given to further support the feasibility of the host group model as well as point out some of the problems yet to bc addressed.

Implementation of host groups involves implcmcnting a binding mechanism (binding intcrnctwork addresses to zero or more hosts) and a packet delivery mechanism (delivering a packet to each host to which its destination address binds). This facility fits most naturally into the gateways of the intcrnctwork and the switching nodes of the constituent point-to-point networks (as opposed to scpararc machines) because multicast binding and delivery is a natural ex tension of the unicast binding and delivery (i e. routing plus srorc-and-forward). That is. a multicast packet is routed and transmitted to multiple destinations, rather than to a single destination.

A gateway in a host group intcrnctwork is thus viewed as a "communication server", providing multicast dclivcry and host group management. The multicast dclivcry service is invoked implicitly by sending packets addressed to host groups. with unicast deliv ery as a special case. The group management scrvicc is invoked explicitly using a request-response transaction protocol between the client hosts and the server gateways. In addition to the operations for creating transient host groups and adding and deleting host memberships in groups (Section 2.1). the gateway supports operations for administrativ e allocation o f permanent group addrcsscs. including static, single-host group addresses (i.e. unicast addrcsscs).

In the following description, wc start with a basic, simple implementation that provides coverage a n d then refine t h i s mechanism with various optimizations to improve efficiency of deliv ery and group management.

3.1 Basic Implementation

A host group defines a *network group*, which is the scl of networks containing current mcmbcrs of the host group. When a packet is sent to a host group, a copy is delivered to each network in the corresponding network group. Then, within each network. a copy is delivered to each host belonging to the group.

To support such multicast dclivcry, cvcry internet gateway maintains the following data structures:

- *routing table*: conventional intcrnctwork routing information. including the distance and direction to the nearest gateway on cvcry net work.
- *network membership table*: A set of records, one for cvcry currently existing host group. The *network membership record* for a group lists the network group, i.e. the networks that contain members of the group.
- *local host membership table*: A set of records. one for each host group that has members on directly attached networks. Each *local host membership record* indicates the local hosts that arc members of the associated host group. For networks that support multicast or broadcast. the record may contain only the local *network-specific multicast address* used by the group plus a count of local mcmbcrs. Othcrwisc. local group mcmbcrs may be identified by a list of unicast addrcsscs to be used in the software implementation of multicast within the network.

A host invokes the multicast dclivcry scrvicc by sending an internetwork datagram to an immediate neighbour gateway (i.e. a gateway that is directly attached l o the same network as the sending host). Upon receiving a datagram from a directly attached network. a gateway looks up the network mcmbcrship record corresponding to the destination address of the datagram. For each of the networks listed in the mcmbcrship record, the gateway consults its routing table. If, according to the routing table, a member network is directly attached, the gateway transmits a copy of the datagram on that network, using the network-specific multicast address allocated for the group on that network. For a member network that is not directly attached and is within the distance constraint specified in the datagram, the gateway creates a copy of the datagram with an additional inter-gateway header identifying the destination network. This inter-gateway datagram is forwarded to the nearest gateway on the destination network, using conventional store-and-forward routing techniques. At the gateway on the destination network, the datagram is stripped o f its inter-gateway header a n d transmitted to the group's multicast address on that network. Member networks that arc beyond the datagram's distance constraint are ignored.

The network mcmbcrship records and the network-specific multicast structures arc updated in response to group management requests from hosts. A host sends a request to create, join. or leave a group to an immediate neighbour gateway. If the host requests creation of a group. a new network mcmbcrship record is created by the serving gateway and distributed lo all other gateways. If the host is the first on its network to join a group. or if the host is the last on its network to leave a group, the group's network membership record is updated in all gateways. The updates need not be performed atomically at all gateways, due to the datagram delivery semantics: hosts can tolerate misrouted and lost packets caused by temporary gateway inconsistencies, as long as the inconsistencies are resolved within normal host retransmission periods. In this respect, the network membership data is similar to the network reachability data maintained by conventional routing algorithms, and can be handled by similar mechanisms.

In many cases, a host joins a group that already has mcmbcrs on the same network. or leaves a group that has remaining members o n the same network. This i s then a local matter between the hosts and gateways on a single network: only the local host mcmbcrship table needs to be updated to include or exclude the host:

This basic implementation strategy meets the dclivcry requirements stated at the end of Section 2. I lowcvcr. it is far from optimal, in terms of either dclivcry efficiency or group management overhead. One simple improvement is lo recognize the important special case of static, one-member-only groups. This again corresponds to the conventional unicast provided in (for example) IP and NNS. In this case, the internetwork address for the sin&host group encodes within it the network of the one host so there is no need to maintain a separate group mcmbcrship record for that group. Consequently, the number of group mcmbcrship records in the gateways is greatly rcduccd. Also, dclivcry to these groups degenerates to convent ional unicast techniques s u c h a s currently used in II' and XNS implementations. Below, wc discuss some further refinements to the basic implcmcntation.

3.2 Multicast Routing Between Networks

Multicast routing among the internetwork gateways is similar to store-and-forward routing in a point-to-point network. The main difference is that the links between the nodes (gateways) can bc a mixture of broadcast and unicast-type networks with widely di fferent throughput and delay characterist ics. In addition, packets arc addressed to networks rather than hosts (al the gateway level).

We usc the extended reverse path forwarding algorithm of Dalal and Metcalfe[11]. Although originally designed f o r broadcast, it is a simple and efficient technique that can serve well for multicast delivery if network membership records in each gateway are augmented w i t h information f r o m neighbouring

gateways. Th is algorithm uses the *source* net work identi fier, rather than a *destination* network identifier to make routing decisions. Since the source address of a datagram is a general group address, it cannot bc used to identify the source network of the datagram; the first gateway must add a hcadcr specifying the SOUTCC network This approach minimizes redundant Iransmissions when multiple destination networks arc reachable across a common intergateway link, a problem with the basic implementation described earlier.

Note that we eliminated from consideration techniques that fail lo deliver along the branches of the shortest dctay tree rooted at the source, such as Wall's center-based forwarding[12] because this compromises lhc meaning of the multicast distance parameter and detracts from multicast performance in gcncral. Wc also rejected the approach of having a multicast packet Carry more than one network identifier in its inter-gateway hcndcr to indicate multiple destination networks because the resulting variable length headers would cause buffering and fragmentation problems in the gateways.

3.3 Multicasting Within Networks

A simple optimization within a nclwork is to have the scndcr use the local multicast address of a host group for its initial Iransmission. This allows the local host group mcmbcrs lo rcccivc the transmission immediately along with lhc gateways (which must now "eavesdrop" on all multicast transmissions). A gateway only forwards the datagram if lhc destination host group includes members on other nciworks. This scheme reduces the cost l o reach local group mcmbcrs lo one packet transmission from two required in the basic implementation[5] so Iransmission to local mcmbcrs is basically as efficient as the local multicast support provided by the nclwork.

A similar opportunity for reducing packet traffic arises when a datagram mus' traverse a network to get from one gateway to another, and that network also holds mcmbcrs of the destination group. Again. use of a network-specific multicast address which includes member hosts plus gateways can achieve the desired effect. I lowever, in this case, hosts must be prepared lo accept datagrams that include an inter-gateway header or, alternatively, every datagram must include a spare field in its header for use by gateways lll lieu of an additional inter-gateway hcadcr.

3.4 Distributing McmbershipInformation

A refinement lo host group membership maintenance is lo store the host group membership record for a group *only* in those gateways that are directly connected l o member networks. Information about other groups is cached in the gateway only while it is required lo route lo those other groups When a gateway receives a datagram to be forwarded lo a group for which it has no network membership record (which can only happen if the gateway is nol directly connccicd to a member nclwork). it takes the following action. The gateway assumes temporarily that the destinat ion group has mcmbcrs on *every* nclwork in the inter nclwork. except those directly attached lo the sending gateway, and routes the datagram accordingly. In the inter-gateway header o f the outgoing packet, the gateway sets ; l bit indicating that it wishes l o receive a copy o f the network membership record for the destination host group. When such a datagram reaches a gateway on a member network, that gateway sends a copy of the membership record back to the requesting gateway and clears the copy request bit in the datagram.

Copies of network membership records sent lo gatewavs outside of a group's mcmhcr networks arc cached for use in subsequent transmissions b y those gateways. That raises the danger of a stale cache entry leading to systematic dctivcry failures. To counter that problem, the inter-gateway hcndcr contains a field which is a hash value or checksum on the network membership record used lo route lhc datagram. Gateways on member networks compare the checksum on incoming datagrams with their up-to-date records. If the checksums don't match, an up-to-date copy of the record is returned to the gateway with the bad record.

This caching strategy minimizes intergateway traffic for groups that arc only used within one network or within the set of networks on which mcmbcrs reside, the expected common cases. Partial replication with caching also reduces the overhead for nclwork traffic to disseminate updates and keep all copies consistent. Finally, it also reduces the space cost for data in large internetworks with large numbers of multiple host groups.

We have not addressed here the problem of maintaining up-to-date, consistent network membership records within tic set of gateways connccictl lo mcmbcrs of a group. This can bc viewed as a distributed database problem which has been welt studied in other contexts. The loose consistency requirements on network membership records suggest that the techniques used in Grapevine[13] might be useful for this application.

4. Integration into the DoD Internet

To show how the host group model can bc supported by straight forward ex lcnsion ol an existing internetwork architecture, wc outline how it might fit into lhc US DoD Internet.

The current Internet provides unicast datagram delivery between hosts on a wide variety of networks, both local-area and wide-area, broadcast and point-to-point. An Internet address is a 32-bit value consisting of lwo subfields: a network number and a host-within-network number. Every Internet gateway maintains a routing table that specifies the distance and direction to every network in the Internet, relative lo the gateway. Thus, given a datagram, a gateway can determine from the network number subfield of its destination address, where lo send it next on the path towards its destination. When the datagram reaches a gateway i n t o its destination network, that gateway maps the host-within-network number to a local network address for final delivery.

The existing architecture supports our model of static, one-member-only groups Wc extend this architecture Lo support multiple host groups by reserving a single network number to identify all such g r o u p s . Each multiple host group is distinguished b y a unique value in the host-within-network subfield o f its internet address. The Internet gateways a r c augmented with lhc data structures and procedures discussed in Section 3 lo support internet multicast.

An I I ' datagram contains a "time to live" field which is decremented b y the gateways once a second a n d o n every nclwork hop. If the time lo live goes to zero before the datagram reaches its destination, the datagram is discarded. In the host group implementation, this field is used lo limit the delivery distance 0 f multicasts.

Other datagram internetwork architectures yield lo similar extensions. For example, the Xerox Network Systems architecture i s essentially identical to the DoD Internet w i l h regards lo nddrcss encoding (network, host-within-network) and contents of routing tables. XNS datagrams contain a hop count field that can bc used for multicast scope control.

The proposed ISO internetwork protocol[14] provides the same style of internetwork datagram service as I I ' or XNS. The draft proposal for ISO internetwork addresses[15] specifies a much more complex structure than the fixed-length, lwo-tcvcl hierarchical addresses of IP and XNS. A more sophisticated, possibly hierarchical, distribution of the nclwork membership records would be appropriate for the cnormious potential size of the ISO

[5] One unicast transmission from sender l o gateway and one multicast transmission from gateway to local group members

"world network".

5. Use of Multicast

A number of applications that can usc multicast have been cited earlier in the paper, including distributed databases, confcrcncing. distributed computation and locating intcrnctwork services. Rather than describe these applications in greater detail, wc focus on some general issues that were identified in previous work[7]. (This work dealt with the use of local network multicast in a distributed operating system to support the concept of interprocess group communication where process groups are distributed across host groups.)

A key issue is providing reliable communication as required by the application. Firstly, some applications, such as real-time confcrcncing. do not need reliable delivery, assuming the periodic updates arc generally reccivcd. Secondly, binding applications. such as locating a name server, do not require delivery to al! but simply a positive response from al least one host. Retransmission with possibly expanding scope of search until a response is received provides the rcquircd semantics.

As an aside, one might argue that the binding usc is only really required to locate a name scrvcr. While true in theory. it may be simpler for some applications to locate other servers directly using this simple search protocol Then they do not need to implement the protocol to lookup a name in the name server as well as this simple search protocol to locate the name server in the first place. For example, the PROM network loader for diskless workstations might be simpler if it can locate a boot server using a boot server group address directly rather than going through a name server.

For applications requiring reliable delivery, there arc basically two approaches. The most common approach is to place the onus for reliable delivery on the sender. Ilcrc, the scndcr knows the membership of a group and retransmits to the group until it has receivcd acknowledgements from each group member. As an optimization, the sender can usc unicast to retransmit to particular group members if the number of missing acknowledgements is relatively small compared to the cardinality of the host group.

The second approach places the onus on the receivers to implement reliable delivery, what wc call *publishing*. It is so named because it mimics real world publishing. That is, information lo be sent to a group, the *subscribers*, is filtered though the *publisher*, which collates and numbers the information before issuing it lo lhc subscribers. A subscriber noticing a missing issue by 3 gap in the issue numbers or a new issue not being received in the expected time interval requests the *back issue* from the publisher. Thus, instead of automatic retransmission until lhc rcccivcr acknowledges the message, the receiver must request retransmission if it is rcquircd.

A family of reliable multicast protocols is specified by Chang and Maxemchuk[16] that combines both techniques built on top of an unreliable broadcast or multicast network. They describe a protocol that guarantees not only that all group members receive all messages, but also that they all receive the messages in the same order, regardless of lhc number of senders Furthermore, this strong level of reliability is achieved with only one acknowledgement per message in the normal case, no single point of failure, and survival in the face of multiple host failures and recoveries. In another paper[1], Chang describes the usc of this protocol to support a distributed, replicated database.

In general. the problem is not implementing reliable delivery for multicast dclivcry hut choosing the right trade-off between cost, performance and reliability as required by the application. Wc have briefly described some basic techniques. I lowever, further study is required lo understand these trade-offs with various applications and internetworking parameters.

6. Related Work

Thcrc is relatively little published work on the use or implementation of intcrnctwork multicasting.

Wall's thesis[12] presents several mechanisms for performing efficient broadcast and multicast delivery in point-to-point networks. His results can bc applied to providing multicast within point-to-point networks that arc constituents of an intcrnctwork, and to the problems of multicast routing to "network groups" of gateways.

Boggs, in his thesis[8], describes a number of distributed applications that arc impossible or very awkward to support without the flexible binding nature of broadcast addressing. Although hc recognizes that almost all of his applications would be best serv ed by a multicast mechanism, he advocates the usc of "directed broadcast" because it is easy lo implement within many kinds of networks and can be extended across an intcrnctwork without piacrng any new burden on internetwork gateways. Unfortunately, broadcasting has the undesirable side effect of delivering packets to more hosts than necessary, thus incurring overhead on uninvolv cd parties and possibly creating security problems. Furthermore, directed broadcasting supports simple communication with unknown destinations on directly connected networks only: for destinations on more distant networks, the sender must know their network numbers or perform a search using gateway routing tables.

Recent proposals by Mogul[17] and Aguilar[18] have addressed the issue of multi-destination delivery within the DoD Internet. Mogul proposes an implementation of Bogg's directed broadcast facility. Aguilar suggests allowing an IP datagram to carry additional destination addresses, which arc used by the gateways to route the datagram to each recipient. Such a facility would alleviate some of the inefficiencies of sending individual datagrams to a group, but it would not be able lo take advantage of local net work mu lticast facilities. More seriously, Aguilar's scheme requires the scndcr to know the individual IP addresses of ail members of the destination group and thus lacks the flexible binding nature of true multicast or broadcast.

Blaustein c l al[19] discuss a variety of protocols for reliable multicast delivery based on various (inter)network characteristics (e.g. point-to-point or broadcast or both, clusters of fast networks joined by slower networks, degree of multicast support provided by the networks, etc). As well as making a case for unreliable multicast services at the internetwork level, their w o r k suggests ways of achieving efficient multicast a m o n g gateways i n a heterogeneous internetwork.

7. Concluding Remarks

Wc have described a modct of multicast communication for datagram-based internetworks. As an extension of existing intcrnctwork architectures, it views unicast communication and time-to-live constraints as special cases of the more general form of communication arising with multicast. Wc have argued that this model is implementable in current and future internetworks and that it provides a powerful facility f o r a variety o f applications. In some cases, it provides a facility that is required for certain applications lo work in the intcrnctwork environment. In other cases, it provides a more efficient, robust a n d possibly more elegant way o f implementing e x i s t i n g internetwork applications.

Wc arc currently implementing a prototype host group facility as an extension of IP. For practical reasons. this prototype implements a l l group management functions a n d multicast r o u t i n g outside of Internet gateways, i n special hosts called *multicast agents*. The collection of multicast agents i n effect provides a second gateway system on lop of lhc existing Internet, f o r multicast purposes. The major c o s t s of this separation are redundancy of routing tables between gateways and multicast agents and the increased delay and unreliability of extra hops in

the delivery path. Much of the routing information in the multicast agents must bc "wired-in" because they do not have access to the gateways' routing tables. 1 lowcvcr, this rudimentary implementation provides an environment for evaluating the interface to the multicast service and for investigating group managcmcnl and multicast routing protocols for eventual use in the gateways. It also serves as a testbed for porting multicast-based distributed applications to an intcrnctwork from the V distributed operating system.

For now, WC arc restricting group membership to local networks that already have a broadcast or mullicast capability, such as the Ethernet. We feel that. in the future, any network that is to support hosts other than just gateways must have a multicast addressing mode. Efficient implementation of multicast within point-to-point or virtual circuit networks deserves investigation.

A significant issue raised by the host group model is authentication and access control in intcrnctworks. Gateways must control which hosts can create and join host groups, presumably making their decision based on the identity of the rcqucstor {thus requiring authentication) and permissions (access con lrol lists). This issue does not arise in conventional intcrnctwork architectures because host addrcsscs are administratively assigned with no notion of dynamic assignment and binding as provided by host groups. WC bclicvc that access control should bc recognized as a proper and necessary function of gateways so as to protect the hosts of local networks from general internetwork activity. Thus, group access control can be subsumed as part of this more general mechanism, although more investigation of the general issue is called for.

On a philosophical point, there has been considerable reluctance to make open use of multicast on local networks because it was network-specific and not provided across intcrnctworks. WC were originally of that school. However, we recognized that our "hidden" uses of multicast in the V distributed system were essential unless WC resorted to dramatically poorer solutions - wired-in addrcsscs. WC also recognized, as described in this paper, that an adequate multicast facility for internetworks was feasible. As a consequence, we now argue that multicast is an important and basic facility to provide in local networks and intcrnctworks. I lighcr levels of communication, including applications, should feel free lo make use of this powerful facility. Networks and intcrnctworks lacking multicast should bc regarded as deficient relative lo the future (and present) requirements of sophisticated distributed applications and communication systems.

Acknowledgements

The original proposal for host groups was dcvclopcd a a proposal for extending the US DoD Internet architecture. Work is continuing develop a specific Internet extension in conjunction with the DARPA Internet task force on end-to-end protocols. headed by Bob Braden This task force has provided valuable input jn the development of the host group model.

References

1. J-M. Chang. "Simplifying Distributed Database Design by Using a Broadcast Network," *SIGMOD '84*, ACM, June 2984, .

2. F. Cristian et al, "Atomic Broadcast: from simple message diffusion to Byzantine agreement," *15th International Conference on Fault Tolerant Computing*, , Ann Arbor, Michigan, June 1985, .

3. 11. Forsdick, "MMCF: A Multi-Media Conferencing Facility", personal communication

4. E. J. Berglund and D. R. Cheriton, "Amaze: A distributed multi-player game program using the distributed V kernel," *Proceedings of the Fourth International Conference on Distributed Systems*. IEEE, May 1984, .

5. IEEE, "Standards for Local Area Networks: Logical Link Control," The Institute of Electrical and Electronic Engineers, Inc., New York, New York, 1985,.

6. Digital Equipment Corporation, Intel Corporation, and Xerox Corporation, "The Ethernet. A Local Area Network: Data Link Layer and Physical Layer Specifications," Version 1.0, September, 1980

D. R. Cheriton and W. Zwaenepoel. "Distributed Process Groups in the V Kernel," *ACM Transactions on Computer Systems*, Vol. 3, No. 3, May 1985, .

8. D. R. Boggs, *Internet Broadcasting*, PhD dissertation, Stanford University, January 1982.

9. J. Postel, "Internet Protocol," Tech. report RFC 791, SRI Network information Center, September 1981.

10. Xerox Corporation, *Internet Transport Protocols*. XSIS 028 112, Xerox System Integration Standard, Stamford, Connecticut, December, 1981.

21. Y. K. Dalal and R. M. Metcalfe, "Reverse Path Forwarding of Broadcast Packets," *Communications of the ACM*, Vol. 21, No. 2, December 1978, pp. 1040-1047.

12. D. W. Wall, "Mechanisms for Broadcast and Selective Broadcast," Tech. report 190, Computer Systems Laboratory, Stanford University, June 1980.

13. A. D. Birrell et al, "Grapevine: an exercise in distributed computing," *Communications of the ACM*. Vol. 25. No. 4. April 1382, pp. 260-274.

14. ISO TC97 SCO. "Information Processing Systems - Data Communications - Protocol for Providing the Connectionless Network Service," Draft Proposal N297 I,.

15. ISO TC97 SC6, "Addendum to the Network Service Definition Covering Network Layer Addressing," Draft Proposal NW44,.

16. J-M. Chang and N. F. Maxemchuk, "Reliable Broadcast Protocols." *ACM Transactions on Computer Systems*, Vol. 2, No. 3. August 1984, .

17. J. Mogul, "Broadcasting Internet Datagrams," Tech. report RFC 919, SRI Network Information Center, October 1984.

18. L. Aguilar, "Datagram Routing for Internet Multicasting," *ACM SIGCOMM '84 Communications Architectures and Protocols*, ACM, June 1984. pp. 58-63.

19. B. Blaustein et al. "Notes on the Reliable Broadcast Protocol (The Distributor)". Computer Corporation of America. working paper.