IP over Huge Scale Broadcast Media

Masataka OHTA (mohta@necom830.hpcl.titech.ac.jp)
Computer Center, Tokyo Institute of Technology, Japan

Manolo SOLA (sola@jet.es)
Graduate School of Science and Engineering, Waseda University, Japan

Abstract

In order to integrate telecommunication and broadcasting in the Internet, the integration of Huge Scale Broadcast Media (HSBM) and the Internet is of crucial importance. Considering the severe scalability requirements imposed by the HSBMs and comparing the result with the real-world TV/radio broadcasting, we show in this paper that IP over HSBM is equivalent to IP multicast over HSBM, and that a modified version of PIM can work as the scalable multicast routing protocol that will provide IP multicast over HSBM.

1. Introduction

A Huge Scale Broadcast Media (HSBM) is a broadcast media (link layer) with a huge number of hosts attached to it. Classical TV/radio broadcasting is an example in which HSBM has been widely used. A typical example of HSBM, and perhaps the only real, is radio wave broadcasting. With satellite broadcasting, there can be hundreds of millions of hosts in an HSBM.

On the other hand, the Internet is built following the CATENET model [CATENET] in which small scale media are interconnected by routers. For the scalability of many Internet protocols, it is important that each media does not contain a lot of hosts. Protocols depending on link broadcast or all-host-multicast will work inefficiently over a link layer with a lot of hosts and eventually stop working over an HSBM.

This paper proposes a way to integrate HSBMs and the Internet. It will help the integration of telecommunication and broadcasting in the Internet.

In this paper, the terminology in [UDLR] is used: "feed" means a host which can both send and receive packets to/from an HSBM, "receiver" means a host which can only receive packets from an HSBM. On the other hand, initiation and termination points of a end-to-end data stream are called "source" and "destination", respectively.

2. What is Huge? How much is Huge?

Huge is a rather relative concept. In this paper, "huge scale" means the scale at which protocols with certain scalability problems do not work. For example, it is impossible to use most broadcast-based protocols, such as RIP, over an HSBM.

Although it depends on applications and protocols, a scale of 100, obviously, is not huge. 1,000,000, perhaps, is huge. If we improve some protocol having scalability difficulties at the scale of 100 in order to work at a scale of 1,000, the improved protocol still can not be considered to be scalable at a huge scale.

Protocols that can not scale can not be used by an HSBM regardless of their performance be improved by small constant factors.

On the other hand, in this paper, today's Internet is considered to be huge enough. For example, the address space of IPv6 or IPv4 is, though finite, huge enough. However, in this context "huge" means that the number is so large that it can safely be assumed that the square of it should be larger than the number of hosts of the Internet or the population of the world.

Using the square inequality given above, it is possible to proof that the number of HSBM-capable media in the world is limited and is not huge. Otherwise, we can regulate the HSBM-capable media used only by the less-than-huge number of hosts, to entirely cover the whole Internet, which makes the media not HSBM.

HSBMs are there, because HSBM-capable media are limited and precious resource.

For example, there is a limited and not so huge number of radio wave broadcast channels in the world.

This makes the use of radio wave frequencies to be strongly regulated by governments and an international organization of ITU-R. The radio wave frequencies are divided into individual bands, within which certain kinds of users can send or receive information. Users are further assigned strict bandwidth and power limitation. Some countries may, then, allow unregulated use of certain bands for in house or regional communication as long as the power is under certain limit covering only a small geographical area and a small number of receivers. Radio amateurs, who can use transmission power large enough to cover the world merely for fun, are admitted partly because of historical reasons and they are strictly licensed with multiple classes. Their communications are restricted for non-commercial purposes and are not protected by the other amateurs.

In such an environment, it costs a lot to gain access to HSBM-capable media for serious, commercial use.

For example, a transponder of a satellite, which has tens of Mbps capacity, costs several million USDs annually.

Although it may not cost so much to share a satellite transponder by a lot of people for unicast phone communication at 4Kbps, it will have a prohibitive cost to use the transponder for serious unicast Internet communications of the coming age at the speed of the coming age.

Of course, there are exceptions for non-commercial emergency use, such as fire fighters and ambulance cars, which may use HSBMs freely, but the cost for the society is the same.

However, if the transponder is used for high (or low of the coming age) bandwidth one-to-many communications, the cost may be divided by the number of the receivers. If the expected number of receivers is huge, the cost is lowered to be even lower than the costs of other media. This is a realistic use of HSBM-capable media, as is exemplified by surface wave or satellite broadcasted TV/radio today.

On the other hand, the protocol in the Internet that is expected to replace TV/radio broadcast is IP multicast, and that is why in this paper we concentrate on "IP multicast over HSBM".

3. HSBM and UDL

In this section, the relationship between HSBM and Uni-Directional Link Layer (UDL) is discussed.

In the IETF, the UDLR WG is working on unicast/multicast communication over unidirectional link layers using satellite communication as a real world example. [UDLR] recognizes a satellite network as:

> The advantage of a satellite network is to provide high bandwidth services independent of the user's location over a large geographical area.

It can be shown that HSBM is likely to be UDL.

An HSBM is mostly unidirectional. This is because if all the hosts can send some data or control traffic, however infrequently, the huge number of hosts in the HSBM can saturate the total bandwidth of the HSBM. For example, if one million hosts transmit 4Kbps of low-speed voice stream, the total bandwidth would be 4Gbps, much larger than that of typical satellite transponders. Or, if the same number of hosts transmit a 40 byte control packet once in a 30 seconds, the total average bandwidth used would be 10.7 Mbps, more than that for NTSC-quality digital TV broadcasting. If the number of hosts of an HSBM is one hundred million, 1.07 Gbps would be consumed.
That is, considering the scalability and the finite bandwidth, it is impossible that many of the hosts in an HSBM send packets at some interval. This fact motivates the use of the media as UDL, because the extra cost to make the receivers feeds can be cut off.

The UDLR WG attempts to support small number of users, some of which may be "Receivers", over a large geographical area, by emulating bi-directional communication with tunneling. However, the problem of unicast communication over HSBM is that it is impossible that receivers periodically advertise routes beyond them to the feeds, and that means that feeds can not have dynamically updated routing table entries for routers over an HSBM.

That is, unicast communication over an HSBM, in general, does not scale and is impossible.

The lack of the scalability is so fundamental that tunneling techniques in [UDLRRIP] does not scale either. That is, even with the techniques in [UDLRRIP], unicast communication over an HSBM is still impossible.

This is another reason why we concentrate on "IP multicast over HSBM".

It should be noted that less-than-huge number of hosts can still use an HSBM-capable media, and at that scale the techniques of UDLR WG can work. But, then, the cost for each receiver is larger. If receivers can afford the cost, they may not be motivated to cut the cost off to be the feeds and the media will become fully bi-directional.

4. Models of IP over HSBM

Figure 1 illustrates three models to integrate HSBM and the Internet. They are also the models for providing IP over HSBM.

It is not difficult to use an HSBM as the last hop of the communication, like in model (a). That is, if almost all the hosts are directly attached to an HSBM or through bridges but without intermediate routers, packets to such hosts of the HSBM can be just sent to the HSBM. But, it is an uninteresting case and is not a full integration of HSBMs into the Internet.

The Internet is an integration of heterogeneous media by routers and each media should operate independent of the other media. They interoperate only through Layer 3 (or above) protocols. Any other attempts not doing this, such as NHRP or MPLS, are known to have complex efficiency, scalability or loop problems [NHRP, MPLS], some of them proved to be unsolvable and some so complex that real world operatability (or operational scalability) is yet unknown.

So, in this paper, only the fully integrated model of (c) is considered as the model of IP Multicast over HSBM.

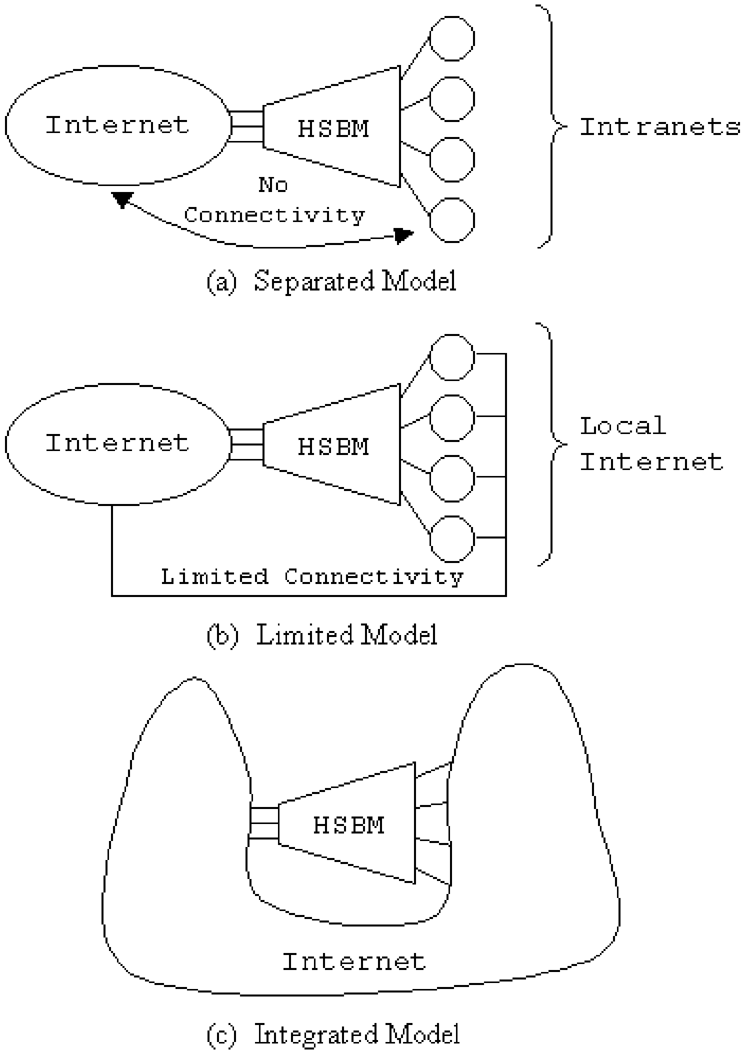

Figure 1. Models to use HSBM with the Internet

It should be noted that [UDLR] has the same strategy. Though it is not obvious to claim limited connectivity and to have the receiver function collocated to the router in the subnet (see Figure 2 in [UDLR]), its conclusion says:

> Fortunately these changes should not lead to new versions of routing protocols (RIP and DVMRP) and should be transparent for routers not connected to satellite networks.

By the way, full integration does not mean that HSBM is completely invisible, at least not at the Layer 3. For example, Internet addresses, the Layer 3 identities, of DNS servers of some DNS zone may be visible to someone trying to resolve the host addresses of the zone. Just like that, some Layer 3 entity related to an HSBM may be visible as a favorite Layer 3 entity to someone trying to do something if it is beneficial for the person to use the HSBM, of course.

On the other hand, it is not necessary to consider the case of using HSBMs more than twice in a single communication between the source and the destination. As is discussed in section 2, HSBMs are so huge that to cover the entire Internet it is enough to have an HSBM at the first stage and a huge number of none-huge broadcast media at a second stage. Other approach is to use a none-huge number of HSBMs in parallel.

This is the reason why in this paper protocols' traffic is considered to use the HSBM only once.

5. Multicast Protocols over HSBM

IP over HSBM means IP multicast over HSBM.

Currently, there are several multicast protocol proposals. In this section, we consider how to avoid the use of IGMP in HSBMs and how to modify DVMRP, MOSPF, CBT and PIM-SM in order to be able to use them over HSBMs.

5.1 IGMP and HSBM

IGMP is something like ARP for multicast. It is a mechanism to let routers know that there is at least one destination for a group in a subnet [IGMP]. IGMP relies on feedback messages from receivers. On usual subnets, the implosion of messages due to receivers' feedback can be avoided by randomly delaying the feedback and by suppressing it completely if the receivers listen to other receiver's reply. However, on HSBMs, the number of possible receivers is so huge that the waiting time is also huge and IGMP can not be used. In [IPSAT], there was a proposal to exponentially increase the possibility of reply as the waiting time goes by, which makes the expected waiting time logarithmic to the number of hosts in an HSBM multiplied by the RTT of the HSBM. However, the delay is still considerably large, especially because the estimation of RTT is not easy and RTT of some satellite links is actually large.

However, this is not a problem and we don't need IGMP over HSBM.

IGMP is a mechanism to avoid to waste bandwidth if there are no receivers or indirect destinations. In HSBM, there are so many receivers that it is almost always certain that at least one receiver is connected directly or indirectly to at least one destination. Remember that HSBM is a precious resource and its purpose is regulated or controlled. Only traffic worth using the HSBM is admitted to enter the HSBM. So, even if there are no receivers, in such a case of broadcasting boring discussions in congresses, it is acceptable for the authority (often rooted by the congress, of course) to use the HSBM for the authorized traffic.

That is, it is not a problem if data stream is sent always, regardless of the number of destinations reachable through the HSBM.

On the other hand, HSBM will not be used for usual traffic of individual users.

The phenomenon discussed in this subsection is related to what is happening in TV/radio broadcasting today. Though there is high delay feedback on the content of broadcasting based on the audience rate, the transmission of the content is not affected by whether there are receivers or not. That is, the behavior is already socially accepted.

5.2 DVMRP over HSBM

DVMRP is a multicast routing protocol to construct reverse shortest path trees rooted by the source by globally flooding routes from the sources [DVMRP]. As is analyzed in [MANOLO], it needs routing table entries and forwarding state for each sender and each group on all routers so it can not be used at a large scale in a large network with a lot of groups. However, unlike unicast routing, DVMRP advertise route from the source to the destinations, from the feed to the receivers, so that there is no route implosion problem at the feed. That is why there is a possibility that DVMRP may be used by a huge number of receivers near HSBM's receivers with a small number of groups.

With DVMRP, routers use Reverse Path Forwarding (RPF) checks against the source of the multicast to avoid the possibilities of loop and packet duplication. On a broadcast multiaccess media, there may be routers on a shortest path tree of a group through the media so multicast packets have to be sent to the media. However other routers on the media may be on the shortest path tree through a different link. Thus, RPF check against the source is necessary on routers. That is, routers drop multicast packets if they are not incoming from the interface on the shortest path from the source to the router.

For bandwidth efficiency, DVMRP routers perform pruning, too. That is, if the router is connected to a leaf subnet of some group tree but no member is found in the subnet, the branch of the tree is pruned for a period of time.

Obviously, pruning is impossible in an HSBM.

However, as discussed in subsection 5.1, this is not a problem if the authorized data is always sent over the HSBM.

The remaining possible problem of DVMRP over HSBM is that multiple feeds may transmit the same data to the HSBM, which is a waste of so precious bandwidth in HSBM. In DVMRP specifications [DVMRP], feeds monitor each other so that only the feed with the smallest metric (from the source) transmit packets. If the metric is the same, the feed with the smaller IP address wins.

However, the identity of the feed is not contained in IP packets but, hopefully, in link headers. The problem is that, as was discussed in [IPSAT], some link protocol, including that of DVB, the major protocol for satellite communication, does not have source MAC address in link headers.

That is, a new mechanism is necessary to choose a feed. If link format can be changed, it is fine. Otherwise, there must be some external mechanism to choose the proper feed. As it is as difficult as Core/RP location of CBT/PIM, it is seemingly better to use CBT or PIM, if possible, than to modify DVMRP with inherent scalability limitations.

5.3 MOSPF over HSBM

MOSPF is a multicast extension of OSPF and globally floods locations of receivers [MOSPF]. Obviously, such flooding has scalability problems so it is impossible to use flat MOSPF over HSBMs. To overcome the problem to some extent, MOSPF (and OSPF) have a notion of areas, which aggregates a group of hosts and subnets. However, even if an HSBM is somehow considered to be an area, the number of areas downstream to the HSBM is still huge so that MOSPF still does not scale. As a less serious problem, MOSPF needs RPF checks against sources, too.

5.4 CBT over HSBM

CBT uses a shared bi-directional tree rooted at a Core for communication [CBT]. Sources just send packets to the Core using unicast routing table. Packets reaching the shared tree are forwarded from there to all the branches of the tree.

The problem of CBT over HSBM is obvious, it needs bi-directional communication over the HSBM and this does not scale.

5.5 PIM-SM over HSBM

PIM-SM is like CBT, but uses a shared uni-directional tree rooted at an RP (Rendez-vous Point) for communication [PIM]. Sources (actually, a designated router near the source) encapsulate packets in unicast packets and send it to the RP. Receivers send JOIN messages to the RP to construct the shared tree. The RP receiving encapsulated packets copies them to the tree.

In [PIM], it is allowed to switch from the shared tree to the source rooted tree. This has demerits for HSBM. PIM-SM also uses RPF check, but, as the tree is shared and rooted by the RP, RPF check is against the RP, not to the source.

Thus, using the shared tree with RP as the receiver, it is possible to construct the tree and perform RPF check outside of the HSBM. The problem, however, is that there will be huge number of RPs and shared trees. But, simple routing trick used for private use addresses or administratively scoped multicast address [PRIV, ASM] can solve the problem. That is, the unicast address of a RP may have different identity in different domains/areas/regions as long as communication within a domain/area/region is consistent.

The source must transmit data to feed-only RP, attached to an HSBM.

That is, if PIM-SM is extended to disallow source rooted trees and RPs are separated to be receive-only and feed-only RPs, PIM-SM over HSBM scales.

The remaining problem is how to advertise the addresses of receive-only and feed-only RPs. The mechanism in [PIM] can be extended easily. Or, DNS may be used [MANOLO] with separate RR types for feed-only and receive-only RPs. However, if DNS is used, the same IP addresses are used globally. The address can not be aggregated with the addresses of local domain/area/region so that one routing table slot is consumed for the RP. But, the RP can be shared by all the groups over an HSBM and as there are less-than-huge HSBMs in the world, it is not a problem of routing table size. Note that, as it is analyzed in [MANOLO], a flooding of the same order of magnitude is necessary for the original mechanism of RP selection in [PIM], which, again, is not a problem for the less-than-huge number of HSBMs.

If a multicast source wants to use multiple HSBMs, multiple feed-only RPs may be given by DNS [MANOLO].

Finally, the PIM model naturally gives control of the content over the HSBM. Unwanted or unauthorized data can be filtered at RP of usual PIM and at feed-only RP of the modified PIM.

6. Resource Reservation and HSBM

Multicast capable resource reservation protocols such as RSVP [RSVP] should also be usable over HSBMs. However, it is not possible for receivers send some feedback, such as RESV of RSVP, over an HSBM.

However, as discussed in subsection 5.1, this is not a problem because among the huge number of receivers, the highest possible quality will be almost always desired by someone. Thus, the feed should simply use the maximum possible bandwidth allowed by the authority of the HSBM.

7. Conclusions

In this paper we show that HSBM (Huge Scale Broadcast Media) can be fully integrated within the Internet. References to the real world TV/radio broadcasting model help a lot to rationalize some properties of IP over HSBM. We also show how IP over HSBM is equivalent to IP multicast over HSBM, and how a modified version of PIM can work as the scalable multicast routing protocol that will provide IP multicast over HSBM.

References

[ASM] "Administratively Scoped IP Multicast," D. Meyer. July 1998, rfc2365.txt

[CATENET] J. Postel, C. Sunshine, and D. Cohen, "The ARPA Internet Protocol", Computer Networks 5, pp. 261-271, 1983.

[CBT] "Core Based Trees (CBT version 2) Multicast Routing -- Protocol Specification --," A. Ballardie. September 1997, rfc2189.txt

[DVMRP] "Distance Vector Multicast Routing Protocol," D. Waitzman, C. Partridge, S.E. Deering. November 1988, rfc1075.txt

[IGMP] "Internet Group Management Protocol, Version 3", B. Cain, S. Deering, A. Thyagarajan. Work in Progress as <draft-ietf-idmr-igmp-v3-01.txt>, February 1999.

[IPSAT] IPSAT BoF, 42nd IETF meeting, August 24-28, 1998; Chicago, IL, USA

[NHRP] "NBMA Next Hop Resolution Protocol (NHRP)," J. Luciani, D. Katz, D. Piscitello, B. Cole, N. Doraswamy. April 1998, rfc2332.txt

[MANOLO] "Scalability of Internet Multicast Protocols", Manolo Sola, Masataka Ohta, Toshinori Maeno, in Proc. INET'98, Geneve, Switzerland, Jul.1998.

[MOSPF] "MOSPF: Analysis and Experience," J. Moy. March 1994, rfc1585.txt

[MPLS] "Multiprotocol Label Switching Architecture," Eric C. Rosen, Arun Viswanathan, Ross Callon. Work in Progress as <draft-ietf-mpls-arch-04.txt>, February 1999.

[PIM] "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," D. Estrin et. al. . June 1998, rfc2362.txt

[PRIV] "Address Allocation for Private Internets," Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot & E. Lear. February 1996, rfc1918.txt

[RSVP] "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification," R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin. September 1997, rfc2205.txt

[UDLR] "Supporting Unidirectional Paths in the Internet", Emmanuel Duros, Walid Dabbous. Work in Progress as <draft-ietf-udlr-general-00.txt>, November 1997.

[UDLRRIP] "Handling of unidirectional links with RIP", Emmanuel Duros, Christian Huitema. Work in Progress as <draft-ietf-udlr-rip-00.txt>, November 1997.