                   Issues affecting MARS Cluster Size

Status of this Memo

Abstract

   IP multicast over ATM currently uses the MARS model [1] to manage the
   use of ATM pt-mpt SVCs for IP multicast packet forwarding. The scope
   of any given MARS services is the MARS Cluster - typically the same
   as an IPv4 Logical IP Subnet (LIS). Current IP/ATM networks are
   usually architected with unicast routing and forwarding issues
   dictating the sizes of individual LISes. However, as IP multicast is
   deployed as a service, the size of a LIS will only be as big as a
   MARS Cluster can be. This document provides a qualitative look at the
   issues constraining a MARS Cluster's size, including the impact of VC
   limits in switches and NICs, geographical distribution of cluster
   members, and the use of VC Mesh or MCS modes to support multicast
   groups.

1. Introduction

   A MARS Cluster is the set of IP/ATM interfaces that are willing to
   engage in direct, ATM level pt-mpt SVCs to perform IP multicast
   packet forwarding [1]. Each IP/ATM interface (a MARS Client) must
   keep state information regarding the ATM addresses of each leaf node
   (recipient) of each pt-mpt SVC it has open. In addition, each MARS
   Client receives MARS_JOIN and MARS_LEAVE messages from the MARS
   whenever there is a requirement that Clients around the Cluster need
   to update their pt-mpt SVCs for a given IP multicast group.

   The definition of Cluster 'size' can mean two things - the number of
   MARS Clients using a given MARS, and the geographic distribution of
   MARS Clients. The number of MARS Clients in a Cluster impacts on the
   amount of state information any given client may need to store while
   managing outgoing pt- mpt SVCs. It also impacts on the average rate
   of JOIN/LEAVE traffic that is propagated by the MARS on
   ClusterControlVC, and the number of pt-mpt VCs that may need
   modification each time a MARS_JOIN or MARS_LEAVE appears on
   ClusterControlVC.

The geographic distribution of clients affects the latency between a
client issuing a MARS_JOIN, and it finally being added onto the pt-
mpt VCs of the other MARS Clients transmitting to the specified
multicast group. (This latency is made up of both the time to
propagate the MARS_JOIN, and the delay in the underlying ATM cloud's
reaction to the subsequent ADD_PARTY messages.)

When architecting an IP/ATM network it is important to understand the
worst case scaling limits applicable to your Clusters. This document
provides a primarily qualitative look at the design choices that
impose the most dramatic constraints on Cluster size. Since the focus
is on worst-case scenarios, most of the analysis will assume
multicast groups that are VC Mesh based and have all cluster members
as sources and receivers. Engineering using the worst-case boundary
conditions, then applying optimisations such as Multicast Servers
(MCS), provides the Cluster with a margin of safety. It is hoped
that more detailed quantitative analysis of Cluster sizing limits
will be prompted by this document.

Section 2 comments on the VC state requirements of the MARS model,
while Sections 3 and 4 identify the group change processing load and
latency characteristics of a cluster as a function of its size.
Section 5 looks at how Multicast Routers (both conventional and
combination router/switch architectures) increase the scale of a
multicast capable IP/ATM network. Finally, Section 6 discusses how
the use of Multicast Servers (MCS) might impact on the worst case
Cluster size limits.

2. VC state limitations.

   Two characteristics of ATM NICs and switches will limit the number of
   members a Cluster may contain. They are:

      The maximum number of VCs that can be originated from, or
      terminate on, a port (VCmax).

      The maximum number of leaf nodes supportable by a root node
      (LEAFmax).

   We'll assume that the MARS node has similar VCmax and LEAFmax values
   as Cluster members. VCmax affects the Cluster size because of the
   following:

      The MARS terminates a pt-pt control VC from each cluster member,
      and originates a VC for ClusterControlVC and ServerControlVC.

When a multicast group is VC Mesh based, a group member terminates
a VC from every sender to the group, per group.

When a multicast group is MCS based, the MCS terminates a VC from
every sender to the group.

LEAFmax affects the Cluster size because of the following:

ClusterControlVC from the MARS. It has a leaf node per cluster
member (MARS Client).

Packet forwarding SVCs out of each MARS Client for each IP
multicast group being sent to. It has a leaf node for each group
member when a group is VC Mesh based.

Packet forwarding SVCs out of each MCS for each IP multicast group
being sent to. It has a leaf node for each group member when a
group is MCS based.

If we have N cluster members, and M multicast groups active (using VC
Mesh mode, and densely populated - all receivers are senders), the
following observations may be made:

ClusterControlVC has N leaf nodes, so
        $N <= LEAFmax$.

The MARS terminates a pt-pt VC from each cluster member, and
originates ClusterControlVC and ServerControlVC, so
        $(N+2) <= VCmax$.

Each Cluster Member sources 1 VC per group, terminates (N-1) VC
per group, originates a pt-pt VC to the MARS, and terminates 1 VC
as a leaf on ClusterControlVC, so
        $(M*N) + 2 <= VCmax$.

The VC sourced by each Cluster member per group goes to all other
cluster members, so
        $(N-1) <= LEAFmax$.

Since all the above conditions must be simultaneously true, we can
see that the most constraining requirement is either:

    (M*N) + 2 <= VCmax.

or

    N <= LEAFmax.

The limit involving VCmax is fundamentally controlled by the VC
consumption of group members using a VC Mesh for data forwarding,
rather than the termination of pt-pt control VCs on the MARS. (It is
in practice going to be very dependent on the multicast group
membership distributions within the cluster.)

The LEAFmax limit comes from ClusterControlVC, and is independent of
the density of group members (or the ratios of senders to receivers)
for active multicast groups within the cluster.

Under UNI 3.0/3.1 the most obvious limit on LEAFmax is 2^15 (the leaf
node ID is 15 bits wide). However, the signaling driver software for
most ATM NICs may impose a limit much lower than this - a function of
how much per-leaf node state information they need to store (and are
capable of storing) for pt-mpt SVCs.

VCmax is constrained by the ATM NIC hardware (for available
segmentation or reassembly instances), or by the VC capacity of the
switch port that the NIC is attached to. VCmax will be the smaller
of the two.

A MARS Client may impose its own state storage limitations, such that
the combined memory consumption of a MARS Client and the ATM NIC's
driver in a given host limits both LEAFmax and VCmax to values lower
than the ATM NIC alone might have been able to support.

It may be possible to work around LEAFmax limits by distributing the
leaf nodes across multiple pt-mpt SVCs operating in parallel.
However, such an approach requires further study, and doesn't solve
the VCmax limitation associated with a node terminating too many VCs.

A related observation can also be made that the number of MARS
Clients in a Cluster may be limited by the memory constraints of the
MARS itself. It is required to keep state on all the groups that
every one of its MARS Clients have joined. For a given memory limit,
the maximum number of MARS Clients must drop if the average number of
groups joined per Client rises. Depending on the level of group
memberships, this limitation may be more severe than LEAFmax.

3. Signaling load.

   In any given cluster there will be an 'ambient' level of
   MARS_JOIN/LEAVE activity. The dynamic characteristics of this
   activity will depend on the types of multicast applications running
   within the cluster. For a constant relative distribution of multicast
   applications we can assume that, as the number of MARS Clients in a
   given cluster rises, so does the ambient level of MARS_JOIN/LEAVE
   activity. This increases the average frequency with which the MARS
   processes and propagates MARS_JOIN/LEAVE messages.

   The existence of MARS_JOIN/LEAVE traffic also has a consequential
   impact on signaling activity at the ATM level (across the UNI and
   {P}NNI boundaries). For groups that are VC Mesh supported, each
   MARS_JOIN or MARS_LEAVE propagated on ClusterControlVC will result in
   an ADD_PARTY or DROP_PARTY message sent across the UNIs of all MARS
   Clients that are transmitting to a given group. As a cluster's
   membership increases, so does the average number of MARS Clients that
   trigger ATM signaling activity in response to MARS_JOIN/LEAVEs.

   The size of a cluster needs to be chosen to provide some level of
   containment to this ambient level of MARS and UNI/NNI signaling.

   Some refinements to the MARS Client behaviour may also be explored to
   smooth out UNI signaling transients. MARS Clients are currently
   required to initiate revalidation of group memberships only when the
   Client next sends a packet to an invalidated group SVC. A Client
   could apply a similar algorithm to decide when it should issue
   ADD_PARTYs. For example, after seeing a MARS_JOIN, wait until it
   actually has a packet to send, send the packet, then initiate the
   ADD_PARTY. As a result actively transmitting Clients would update
   their SVCs sooner than intermittently transmitting Clients.

4. Group change latencies

   The group change latency can be defined as the time it takes for all
   the senders to a group to have correctly updated their forwarding
   SVCs after a MARS_JOIN or MARS_LEAVE is received from the MARS. This
   is affected by both the number of Cluster members and the
   geographical distribution of Cluster members. (Groups that are MCS
   based create the lowest impact when new members join or leave, since
   only the MCS needs to update its forwarding SVC.) Under some
   circumstances, especially modelling or simulation environments, group
   change latencies within a cluster may be an important characteristic
   to control.

As noted in the previous section, the ADD_PARTY/DROP_PARTY signaling
load created by membership changes in VC Mesh based groups goes up as
the number of cluster members rises (assuming worst case scenario of
each cluster member being a sender to the group). As the UNI load
rises, the ATM network itself may start delivering slower processing
of the requested events.

Wide geographic distribution of Cluster members also delays the
propagation of MARS_JOIN/LEAVE and ATM UNI/NNI messages. The further
apart various members are, the longer it takes for them to receive
MARS_JOIN/LEAVE traffic on ClusterControlVC, and the longer it takes
for the ATM network to react to ADD_PARTY and DROP_PARTY requests. If
the long distance paths are populated by many ATM switches,
propagation delays due to per-switch processing will add
substantially to delays due to the speed of light.

(Unfortunately, mechanisms for smoothing out the transient ATM
signaling load described in section 3 have a consequence of
increasing the group change latency, since the goal is for some of
the senders to deliberately delay updating their forwarding SVCs.
This is an area where the system architect needs to make a
situation-specific trade-off.)

It is not clear what affect the internal processing of the MARS
itself has on group change latency, and how this might be impacted by
cluster size. A component of the MARS processing latency will depend
on the specific database implementation and search algorithms as much
as on the number of group members for the group being modified at any
instant. Since the maximum number of group members for a given group
is equal to the number of cluster members, there will be an indirect
(even if small) relationship between worst case MARS processing
latencies and cluster size.

5. Large IP/ATM networks using Mrouters

   Building a large scale, multicast capable IP over ATM network is a
   tradeoff between Cluster sizes and numbers of Mrouters. For a given
   number of hosts, the number of clusters goes up as individual
   clusters shrink. Since Mrouters are the topological intersections
   between clusters, the number of Mrouters rises as the size of
   individual clusters shrinks. (The actual number of Mrouters depends
   largely on the logical IP topology you choose to implement, since a
   single physical Mrouter may interconnect more than two Clusters at
   once.) It is a local deployment question as to what the optimal mix
   of Clusters and Mrouters will be.

Currently two broad classes of Mrouters may be identified:

   Those that originate unique VCs into target Clusters, and
   forward/interleave data at the IP packet level (the Conventional
   Mrouter).

   Those that originate unique VCs into target Clusters, but create
   internal, cell level 'cut through' paths between VCs from
   different Clusters (e.g. the Cell Switch Router).

How these Mrouters establish and manage the associations of VCs to IP
traffic flows is beyond the scope of this document. However, it is
worth looking briefly at their impact on VC consumption and ATM
signaling load.

5.1 Impact of the Conventional Mrouter

A conventional Mrouter acts as an aggregation point for both
signaling and data plane loads. It hides host specific group
membership changes in one cluster from senders within other clusters,
and protects group members (receivers) in one cluster from having to
be leaf nodes on SVCs from senders in other Clusters.

When acting as an ingress point into a cluster, a conventional
Mrouter establishes a single forwarding SVC for IP packets. This
single SVC carries data from other clusters interleaved at the IP
packet level. Only this single SVC needs to be modified in response
to group memberships changes within the target cluster. As a
consequence, there is no need for sources in other clusters to be
aware of, or react to, MARS_JOIN/LEAVE traffic in the target cluster.
(The consequential UNI signaling load identified in section 3 is also
localized within the target Cluster.)

MARS Clients within the target cluster also benefit from this data
path aggregation because they terminate only one SVC from the Mrouter
(per group), rather than multiple SVCs originating from actual
senders in other Clusters.

Conventional Mrouters help control the limiting factors described in
sections 2, 3, and 4. A hypothetical 10000 node Cluster could be
broken into two 5000 node Clusters, or four 2500 node Clusters, etc,
to reduce VC consumption. Or you might have 200 nodes of the overall
10000 that are known to join and leave groups rapidly, whilst the
other 9800 are fairly steady - so you deploy clusters of 200, 2500,
2500, 2500, 2300 hosts respectively.

5.2. Impact of the Cell Switch Router (CSR).

   Another class of Mrouter, the Cell Switch Router (CSR) attempts to
   utilize IP level flow information to dynamically manage the switching
   of data through the device below the IP level. Once the CSR has
   identified a flow of IP traffic, and associated it with an inbound
   and outbound SVC, it begins to function as an ATM cell level device
   rather than a packet level device.

   Even when operating in this mode the CSR isolates attached Clusters
   from each other's MARS_JOIN/LEAVE activities, in the same manner as a
   conventional Mrouter. This occurs because the CSR manages its
   forwarding SVCs just like a normal MARS Client - responding to
   MARS_JOIN/LEAVE messages within the target cluster by updating the
   pt-mpt trees rooted on its own ATM ports.

   However, since AAL5 AAL_SDUs cannot be interleaved at the cell level
   on a single SVC, a CSR cannot simultaneously perform cell level cut-
   through and aggregate the IP packet flows from multiple senders onto
   a single SVC into a target Cluster. As a result, the CSR must
   construct a separate forwarding SVC into a target cluster for each
   SVC it is a leaf of in a source Cluster (to to ensure that cells from
   individual sources are not interleaved prior to reaching the re-
   assembly engines of the group members in the target cluster).

   Interestingly, the UNI signaling load offered within the target
   Cluster by the CSR is potentially greater than that of a conventional
   Mrouter. If there are N senders in the source Cluster, the CSR will
   have built N identical pt-mpt SVCs out to the group members within
   the target Cluster. If a new MARS_JOIN is issued within the target
   Cluster, the CSR must issue N ADD_PARTYs to update the N SVCs into
   the target Cluster. (Under similar circumstances a conventional
   Mrouter would have issued only one ADD_PARTY for its single SVC into
   the target Cluster.)

   Thus, without the ability to provide internal cut-through forwarding
   with AAL_SDU boundaries intact, the CSR only provides for the
   isolation of MARS_JOIN/LEAVE traffic within clusters. It cannot
   provide the data path aggregation of a conventional Mrouter.

6. The impact of Multicast Servers (MCSs)

   Since the focus of this document is on worst-case scenarios, most of
   the analysis has assumed multicast groups that are VC Mesh based and
   have all cluster members as sources and receivers. The impact of
   using an MCS to support a multicast group can be dramatic in the
   context of the group's resource consumption, but less so in the
   over-all context of cluster size limits.

   The intra-cluster, per group impact of an MCS is somewhat analogous
   to the inter-cluster impact of a conventional Mrouter. The MCS
   aggregates the data flows (only 1 SVC terminates on each group
   member, independent of the number of senders), and isolates
   MARS_JOIN/LEAVE traffic (which is shifted to ServerControlVC rather
   than ClusterControlVC). The resulting UNI signaling traffic and load
   is reduced too, as only the forwarding SVC out of the MCS needs to be
   modified for every membership change in the MCS supported group.

   Deploying a mixture of MCS and VC Mesh based groups will certainly
   improve resource utilization. However, the actual extent of the
   improvements (and consequently how large the cluster can be made)
   will depend greatly on the dynamics of your typical applications and
   which characteristics from sections 2, 3, and 4 are your primary
   limitations.

   For example, if VCmax or LEAFmax (section 2) are primary limitations,
   one must keep in mind that each MCS itself suffers the same NIC
   limits as the MARS and MARS Clients. Even though using an MCS
   dramatically reduces the number of VCs per MARS Client per group,
   each MCS still needs to terminate 1 SVC per sender - potentially up
   to 1 SVC from each Cluster member. (This may become 1 SVC per member
   per group if the MCS supports multiple groups simultaneously.)

   Assume we have a Cluster where every group is MCS based, each MCS
   supports only one group, and both VCmax and LEAFmax apply equally to
   MCS nodes as MARS and MARS Clients nodes. If we have N cluster
   members, M groups, and all receivers are senders for a given MCS
   supported group, the following observations may be made:

      Each MCS forwarding SVC has N leaf nodes, so
            N <= LEAFmax.

      Each MCS terminates an SVC from N senders, originates 1 SVC
      forwarding path, originates a pt-pt control SVC to the MARS, and
      terminates 1 SVC as a leaf on ServerControlVC, so
            N + 3 <= VCmax.

MARS ClusterControlVC has N leaf nodes, so
        N <= LEAFmax.

MARS ServerControlVC has M leaf nodes, so
        M <= LEAFmax.

The MARS terminates a pt-pt VC from each cluster member, a pt-pt
VC from each MCS, originates ClusterControlVC, and originates
ServerControlVC, so
        N + M + 2 <= VCmax.

Each Cluster Member sources 1 VC per group, terminates 1 VC per
group, originates a pt-pt VC to the MARS, and terminates 1 VC as a
leaf on ClusterControlVC, so
        2*M + 2 <= VCmax.

Since all the above conditions must be simultaneously true, we can
see that the most constraining requirements are:

    N + M + 2 <= VCmax (if M <= N)

    2*M + 2 <= VCmax (if M >= N)
or
    N <= LEAFmax.

(Assuming that in general M+2 > 3, so the VCmax constraint at each
MCS is not a limiting factor.)

We can get a feel for the relative impacts of VC Mesh groups vs MCS
based groups by considering a cluster where M1 represents the number
of VC Mesh based groups, and M2 represents the number of MCS based
groups. Again we assume worst case group density (all N cluster
members are group members, all receivers are also senders).

As noted in section 2, the VCmax constraint in VC Mesh mode comes
from each MARS Client, and is:

    N*M1 <= VCmax - 2

For the MCS case we have two scenarios, M2 <= N and M2 >= N.

If M2 <= N we can see the VC consumption by VC Mesh based groups will
become the applicable constraint on cluster size N when:

    N + M2 <= N*M1
i.e.
    M1 >= 1 + (M2/N)

Thus, if there is more than 1 VC Mesh based group, and less MCS based groups than cluster members (M2 < N), the constraint on cluster size is dictated by the VC Mesh characteristics: N*M1 <= VCmax - 2. (If M2 == N, then there may be 2 VC Mesh based groups before the VC Mesh characteristics are the dictating factor.)

Now, if M2 > N (more MCS based groups, and hence MCSes, than cluster members) the calculation is more complex since in this case VCmax at the MARS Client is the limiting parameter for both VC Mesh and MCS cases. The limit becomes:

    N*M1 + 2*M2 <= VCmax - 2

However, on face value this is an odd situation anyway, since it implies more MCS entities than hosts or router interfaces into the cluster (given the assumption of one group per MCS).

The impact of MCS entities that simultaneously support multiple groups is left for future study.

7. Open Issues

   There is a wide range of qualitative analysis that can be extracted from typical MARS deployment scenarios. This document does not attempt to develop any numerical models for VC consumptions, end to end latencies, etc.

8. Conclusion

   This document has provided a high level, qualitative overview of the parameters affecting the size of MARS Clusters. Limitations on the number of leaf nodes a pt-mpt SVC may support, sizes of the MARS database, propagation delays of MARS and UNI messages, and the frequency of MARS and UNI control messages are all identified as issues that will constrain Clusters. Conventional Mrouters are identified as useful aggregators of IP multicast traffic and signaling information. Cell Switch Routers are noted to offer only some of the aggregation attributes of conventional Mrouters. Large scale IP multicasting over ATM requires a combination of Mrouters and appropriately sized MARS Clusters. Finally, it has been shown that in a simple cluster where there are less MCS based groups than cluster members, two or more VC Mesh based groups are sufficient to render the use of Multicast Servers irrelevant to the worst case cluster size limit.

Security Considerations

    Security issues are not discussed in this memo.

Acknowledgments

    Thanks must go to Rajesh Talpade (Georgia Tech) for specific input on
    aspects of the VC Mesh vs MCS tradeoffs, and Joel Halpern (Newbridge)
    for general input on the document's focus.

Author's Address

    Grenville Armitage
    Bellcore, 445 South Street
    Morristown, NJ, 07960
    USA

    EMail: gja@thumper.bellcore.com
    Phone +1 201 829 2635

References

    [1] Armitage, G., "Support for Multicast over UNI 3.0/3.1 based ATM
    Networks.", Bellcore, RFC 2022, November 1996.