

Inter-Domain Multicast Routing (IDMR)
INTERNET-DRAFT

A. Ballardie
Consultant
B. Cain
Bay Networks
Z. Zhang
Bay Networks

March 1998.

Core Based Tree (CBT) Multicast Border Router Specification
<[draft-ietf-idmr-cbt-br-spec-02.txt](#)>

Status of this Memo

This document is an Internet Draft. Internet Drafts are working documents of the Internet Engineering Task Force (IETF), its Areas, and its Working Groups. Note that other groups may also distribute working documents as Internet Drafts).

Internet Drafts are draft documents valid for a maximum of six months. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress."

Please check the I-D abstract listing contained in each Internet Draft directory to learn the current status of this or any other Internet Draft.

Abstract

This draft specifies the behaviour of a CBT multicast border router (BR). This specification assumes the use of CBTv3 - the latest CBT protocol version [3].

CBTv3 has capabilities which make CBT equally well suited for use in stub- or transit- domains; this draft describes mechanisms which enable a CBT distribution tree to span only those routers and links leading to interested receivers or receiver-domains.

1. Changes from Previous Revision

This draft differs significantly from previous revisions, and incorporates mostly new procedures and mechanisms.

2. Interoperability Model

The interoperability model follows that described in [2]. Particular attention is drawn to sections 1 and 2 of that document. For brevity, some of the more fundamental aspects of interoperability are listed below:

- +o logically, a BR has at least two "components", each component being associated with a particular multicast routing protocol. Each component may have more than one associated interface which is running the particular multicast protocol associated with the component. At least one of these components is a CBT component. Figure 1 provides an example (logical) representation of a border router.
- +o besides a CBT component owning its own (private) forwarding cache (hereafter referred to as the PFC), all components share a common, protocol independent, multicast forwarding cache (hereafter referred to as the SFC) which supports source specific (i.e (S, G)), and source independent (i.e. (*, G)) entries. The latest CBT specification recommends that all CBT router implementations include support for an SFC, allowing any CBT router to assume the role of Border Router if necessary.

A CBT component's PFC must support (*, G), (S, G), and (*, Core) entries; (*, Core) entries are not relevant to the SFC.

To ensure that all components have a consistent view of the SFC a BR's components must be able to communicate with each other; how is implementation dependent (guidelines provided in [2]).

- +o the parent for all PFC entries shall point towards the local domain core for G. There is no notion of "incoming" interface wrt any PFC state.

The semantics of the SFC cannot be stated until such time as the inter-domain multicast routing architecture is fully understood.

INTERNET-DRAFT

CBT Border Router Specification

March 1998

- +o It is suggested the SFC is only used on active CBT Border Routers; the PFC is used on all other CBT routers.
- +o mixed multicast protocol LANs are not permitted.

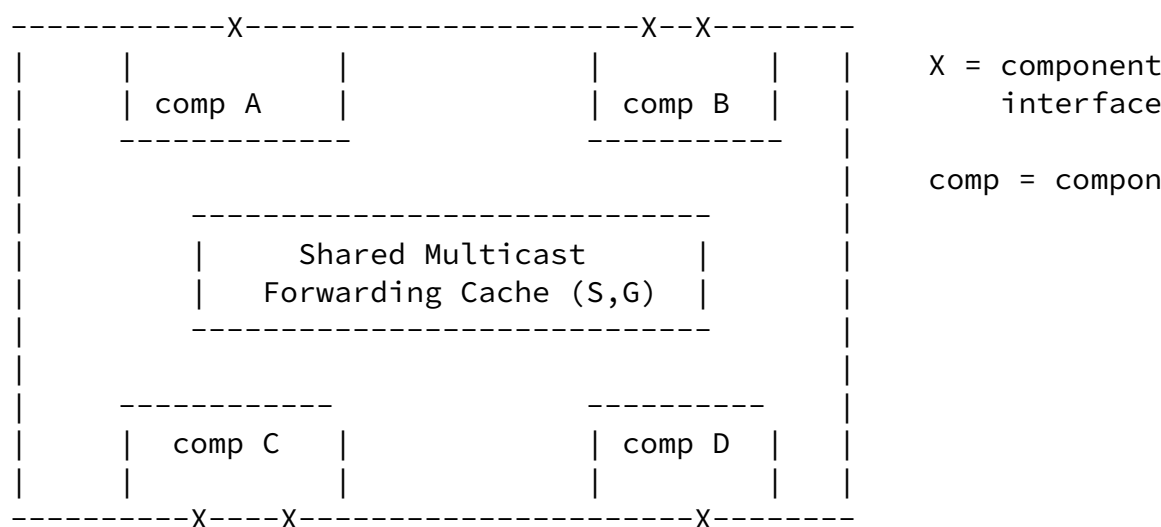


Figure 1: Example Representation of a Border Router

3. Multicast ASs vs. Multicast Domains/"Regions"

It is important to distinguish between a multicast Autonomous System (AS) - an AS whose unicast and multicast routing boundaries are aligned, and a multicast domain (or "region"), whose unicast and multicast routing boundaries are not aligned.

In the former case BGP-4 [6] is often deployed as a separator between interior and exterior routing. For multicast ASs BGP-4+ [1] is assumed, which allows a domain to express multicast policy, i.e. "come from" paths, as well as unicast policy, i.e. "go to" paths, for particular network addresses (or prefixes). The advantage of BGP-4+ is highlighted in the case where a multicast AS is multi-homed (multi-homed stub AS, or transit AS) - BGP-4+ has the ability to select a single ingress border router (BR) per external multicast source (network or prefix), thereby avoiding the potential for the

very damaging effect of multicast packet duplicates being injected into a domain (or AS).

Expires October 1998

[Page 3]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

Note that, if BGP-4+ is assumed, it must be deployed on all of an AS's BRs.

In circumstances where BGP-4+ cannot be assumed, a single ingress BR for a particular multicast source (network or prefix) must be selected by alternate means. One alternative would be to manually configure a CBT domain's multicast BRs, but this does not scale for large numbers of BRs. We therefore recommend that each of a CBT domain's BRs implement an "arbitration process" on each BR, responsible for dynamically selecting a single ingress BR per multicast source (network or prefix). This scheme arbitrates using (multicast) routing metrics as its BR selection criterium; its goal is to select a single ingress BR per external source, not replace BGP-4+ with its fine-grained policy expression capabilities.

The resulting effect of the arbitration process is to allow a CBT component to potentially create/modify/delete a BR shared forwarding cache entry as necessary to prevent the injection of multicast duplicates inside the CBT domain.

A description of one possible implementation of an "arbitration process" is provided in the Appendix.

Hereafter, the terms "multicast domain" and "multicast AS" will be simply referred to as multicast "domain".

[4.](#) The Architectural Model

This section explains the overall architecture of a CBT multicast domain that attaches to other multicast domain(s).

- +o a CBT Border Router (BR) which is used to forward traffic towards an external receiver domain can be thought of as a group member wrt the CBT domain.
- +o Domain Wide Reports (DWRs) (see [section 5](#)) are used in a CBT domain

so that BRs learn of internal group membership, and downstream domain group membership.

DWRs need not necessarily be implemented in singly-homed stub CBT domains, at the potential cost of traffic flowing unnecessarily between the ingress BR and core router(s).

Expires October 1998

[Page 4]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

- +o DWRs are only sent by core routers. They are sent whenever a core router gets its first child, or when a core loses its last child.

DWRs are distributed via the "all-cbt-border-routers" (ABR) multi-cast group, administratively scoped as 239.X.X.X. All CBT BRs must join this group at initialization time.

- +o CBT core routers have authoritative group membership information for the CBT domain for those groups for which they are the core. Hence, DWRs sent by core routers are authoritative - BRs use DWRs to decide whether or not to inject traffic into the CBT domain.
- +o CBT core routers also have authoritative group membership information wrt BRs' attached domain(s) - this is reflected in a core router's forwarding cache; BRs not interested in receiving traffic on behalf of a neighbouring domain send a QUIT_NOTIFICATION (prune) of the corresponding granularity towards the core router. Hence, a core router knows whether or not there are interested receivers downstream of it (internal or external).
- +o CBT BRs may issue (*, Core), (*, G), or (S, G) quits (prunes). In CBT, quits always instantiate uni-directional prune state; by sending a quit the BR is electing not to receive traffic via the CBT domain, but may inject externally sourced traffic into the CBT domain. A quit always follows the state that it is pruning - towards the core; if a quit reaches a core router, it is never forwarded beyond the core router.
- +o A BR may instantiate a priori state between itself and a core router, or that state may be explicitly invoked (see [section 4.1](#) below). For the case where no a priori state exists between a BR and core router, if the BR receives externally sourced data and is the ingress BR for that data, if the BR has a cached DWR Join (received recently from a core router) the ingress BR instantiates

(*, Core) (uni-directional) state between itself and the core, UNLESS the BR has appropriate pre-existing (*, G) or (S, G) state. This way, externally sourced data traffic can always be injected natively into the CBT domain.

- +o A join explicitly invoked by another BR component (as opposed to a DWR) - signalling that a neighbouring domain is interested in a group - instantiates bi-directional state between the BR and core (otherwise data would not be "pulled down" to the BR). This state may be (*, Core), (*, G), or (S, G).

Expires October 1998

[Page 5]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

In each of the following two sections we look at the procedures followed by two different circumstances: firstly, when a BR's neighbouring domain is able to explicitly signal its group membership, and secondly, when a BR's neighbouring domain cannot (or cannot to the same degree compared to the first case) explicitly signal group membership.

[4.1.](#) A Neighbouring Domain can Explicitly Signal Group Membership

- +o CBT BR's send (*,G), (S,G), or (*, Core) JOIN_REQUESTs towards the relevant core router when a neighbouring domain has group members, i.e. a join-alert is received by the CBT BR component from another BR component. The resulting state is bi-directional.
- +o CBT BR's send (*,G), (S,G), or (*, Core) QUIT_NOTIFICATIONs towards the relevant core router when the neighbouring domain no longer has group members, i.e. a prune-alert is received by the CBT component from another BR component.
- +o When a core router gets its first child or loses its last child it issues a DWR of the corresponding granularity. This is received by all BR's; as a result, the BRs know whether or not to inject externally sourced traffic.

[4.2.](#) A Neighbouring Domain cannot Explicitly Signal Group Membership

- +o CBT BR's instantiate (*,Core) state at initialization time to all core routers in a CBT domain that are associated with inter-domain scoped groups. The resulting state is bi-directional.
- +o Though a neighbouring BR component might not be explicitly informed of group membership inside its domain, it may still send prune-alerts (e.g. DVMRP) to the CBT component. Upon receiving a prune-alert from another component, the CBT component sends a (*, G), (S, G), or (*, Core) QUIT_NOTIFICATION (prune) towards the relevant core. Since a quit (prune) is always uni-directional, the BR - if ingress for some external sources - is still able to inject externally sourced data into the CBT domain.

Expires October 1998

[Page 6]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

- +o Though a neighbouring BR component might not be explicitly informed of group membership inside its domain, it may still send join-alerts (e.g. DVMRP) to the CBT component. Upon receiving a join-alert from another component, the CBT component sends a (*, G) (or (S, G)) JOIN_REQUEST toward the relevant core, UNLESS there exists appropriate non-pruned less specific state, i.e. (*, Core). The resulting state is bi-directional.

[4.3.](#) Architectural Summary

In the context of multicast domain interconnection, a CBT domain exhibits the following attributes:

- +o if at least one of a BR's neighbouring domains cannot explicitly signal group membership, the BR must instantiate a priori (*, Core) state (bi-directional) between itself and each domain core.
- +o if all of a BR's neighbouring domains can explicitly signal group membership, the BR need not instantiate any state between itself and domain cores until group membership is signalled.
- +o if a BR receives a DWR Join from a domain core, the DWR is cached. If, during the DWR cache lifetime data arrives for a member group and the BR is the ingress BR for that data, the BR instantiates

uni-directional (*, Core) state between itself and the core so the data can be injected into the CBT domain natively, UNLESS there exists appropriate (*, G) or (S, G) state. A BR may explicitly tear down (using a quit message) uni-directional (*, Core) state after a suitable data flow idle period, or the state may remain.

- +o if a BR receives a DWR Leave from a domain core, the DWR is cached. A DWR Leave results in BRs with any (*, G) or (S,G) PFC (bi-directional) states pruning the relevant state by sending a quit message. The BR also removes the appropriate interface from its SFC entry. An ingress BR which is receiving traffic for the now pruned group (injecting it using (*,Core) state) either tears down the one-way (*, Core) state, or marks its parent PFC as pruned; this is the ONLY instance of a parent interface being pruned.

Expires October 1998

[Page 7]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

5. Domain Wide Reports (DWRs)

Domain Wide Reports (DWRs) are used in a CBT domain to enable BRs to learn - in a dynamic and timely fashion - of internal group membership, and downstream domain group membership. Group membership/absence is indicated by means of DWR Join and Leave messages, respectively.

It is assumed DWRs are refreshed periodically, and cached by receiving BRs for a lifetime of X seconds, after which time they expire in the BR's DWR cache.

If DWRs are in use in a CBT domain, they are only ever issued by core routers. DWRs issued by core routers are authoritative.

DWRs can be source-group specific (S, G), or source independent (*, G).

A DWR represents aggregated state where possible. For example, if a core has only one child for each of its (*, G) and (S, G) states, it generates a (*, G) DWR to cover (*, G) and (S, G). Finer grained aggregates may be represented if DWRs support mask information.

The DWR processing rules are as follows:

- +o whenever a CBT component receives an (S,G) DWR Join message it is only processed by the ingress BR for S. If there exists no (S, G) SFC entry at the ingress BR, an (S, G) SFC entry is created by the CBT component, and an (S, G) Creation-Alert is generated. Then (or if an (S, G) entry already exists) the interface via which the DWR originating core router is reachable is added (or un-pruned) in the entry's child list. If this interface is the only interface in the child list of the entry, the CBT component generates an (S, G) Join-Alert.

If the CBT component's PFC does not have equal- or less specific state that includes the same interface, a (*, G) JOIN_REQUEST (including the "uni-directional" join option) is sent over the interface towards the domain core for G. [An (S,G) join is not sent because (S,G) state does not exist between the ingress BR and core router].

- +o whenever a CBT component receives an (S,G) DWR Leave message it is processed only by the ingress BR for S. If the ingress BR for S has an equal- or less specific SFC entry that includes a pruned

Expires October 1998

[Page 8]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

outgoing interface corresponding to that leading to the relevant domain core router, no further action need be taken.

Otherwise, an (S, G) SFC entry is created (causing an (S, G) Creation-Alert), and the interface leading to the relevant domain core router is included in the outgoing interface list and marked as pruned.

If no further non-pruned children remain in the (S, G) SFC child list, the CBT component sends an (S, G) Prune-Alert to the entry's owner.

- +o whenever a CBT component receives a (*, G) DWR Join, a (*, G) SFC entry is created (unless it, or a less specific entry, already exists) and the interface leading to the domain core for G is added as a child in the entry. The CBT component's PFC is checked to ensure the same interface belongs to an equal- or less specific entry. If no such entry exists, the CBT component sends a (*, G)

JOIN_REQUEST (uni-directional) towards the domain core for G, instantiating (*, G) PFC state.

- +o When a DWR (*, G) Leave message is received by a CBT component if no (*, G) SFC entry exists one is created, and the interface leading to the relevant domain core router is added as a child and marked as pruned.

If no more non-pruned (*, G) SFC children remain, the CBT component sends an (*, G) Prune-Alert to the entry's owner.

6. More BR Component Interactions

- +o upon receipt of an (S,G) Join-Alert (see [2]) by a CBT component, if the interface towards the domain core for G is owned by the CBT component, it adds the interface as the (S, G) SFC entry's incoming interface.

If the CBT component's PFC has no equal- or less specific state that includes the same interface, an (S, G) JOIN_REQUEST is sent over the interface towards G.

- +o the receipt of a (*,G) Join-Alert (see [2]) by a CBT component results in the CBT component including the interface leading to the

relevant domain core as the parent in the (*, G) SFC entry. The CBT component's PFC is checked to ensure the same interface belongs to an equal- or less specific entry. If no such entry exists the CBT component sends a (*, G) JOIN_REQUEST towards the domain core for G, instantiating (*, G) PFC state.

- +o upon receipt of an (S,G) Prune-Alert (see [2]) by a CBT component, if the next-hop interface towards the domain core for G is owned by the CBT component, the CBT component removes the interface from the (S, G) SFC entry, and an (S, G) QUIT_NOTIFICATION is sent towards the domain core for G, instantiating (S, G) PFC prune state.
- +o the receipt of an (*,G) Prune-Alert (see [2]) by a CBT component causes the CBT component to remove the interface leading to the

relevant domain core from the (*, G) SFC entry, then send a (*,G) QUIT_NOTIFICATION over that interface.

- +o whenever a more specific PFC entry is created and there exists a less specific entry/entries, the child list of the new entry is the union of the less specific entry/entries child list(s). The child list of the new entry must also include the interface over which the triggering control message was received.

7. Tunnel Issues

IP multicast deployment in the Internet is a slow process. A unicast AS may not have the resources, or it may not be practical, to migrate a complete AS into one that is completely multicast capable (i.e. multicast AS), so multicast "islands" (i.e. multicast domains - see [section 3](#)) may be created within the unicast AS infrastructure as part of a longer term migration strategy.

A shortage of resources may mean that core routers must be shared between any multicast domains, implying that, from the perspective of the Bootstrap Mechanism [[5](#)], multiple multicast domains may be seen as one single domain. Configured (IP-in-IP) tunnels provide multi-cast connectivity between the multicast domain "islands".

Under these circumstances the Bootstrap Mechanism operating within each multicast domain must be modified such that, if the core router for some set of groups belongs to another domain, the local domain tunnel end-point advertises ("proxies") itself as the core router for that set of groups. The tunnel end-point (router) simply acts as a

"relay" agent for CBT joins, forwarding them to the remote tunnel end-point for onward forwarding.

Acknowledgements

Special thanks goes to Paul Francis, NTT Japan, for the original brainstorming sessions that led to the development of CBT.

Others that have contributed to the progress of CBT include Ken Carlberg, Eric Crawley, Jon Crowcroft, Bill Fenner, Mark Handley, Ahmed Helmy, Nitin Jain, Alan O'Neill, Steven Ostrowski, Radia Perlman, Scott Reeve, Benny Rodrig, Clay Shields, Martin Tatham, Dave Thaler, Sue Thompson, Paul White, and other participants of the IETF IDMR working group.

Thanks also to 3Com Corporation and British Telecom Plc for assisting with funding this work.

References

- [1] Multiprotocol Extensions to BGP-4; T. Bates et al.
<ftp://ds.internic.net/internet-drafts/draft-ietf-idr-bgp4-multiprocol-02.txt>
- [2] Interoperability Rules for Multicast Routing Protocols; D. Thaler;
<ftp://ds.internic.net/internet-drafts/draft-thaler-multicast-interop-01.txt>; Working Draft, March 1997.
- [3] Core Based Trees (CBT_{v3}) Multicast Routing: Protocol Specification; A. Ballardie, B. Cain, Z. Zhang; ftp://ds.internic.net/internet-drafts/draft-ietf-idmr-cbt-spec-*.txt Working Draft, March 1998.
- [4] Domain Wide Multicast Group Membership Reports; W. Fenner; <draft-ietf-idmr-membership-reports-00.txt>; Working Draft, November 1997.
- [5] A Dynamic Bootstrap Mechanism for Rendezvous-based Multicast Routing; D. Estrin et al.; Technical Report, available from:
<http://netweb.usc.edu/pim>
- [6] A Border Gateway Protocol 4 (BGP-4); Y. Rekhter and T. Li; [RFC 1771](RFC1771), March 1995. <ftp://ds.internic.net/rfc/rfc1771.txt>

Expires October 1998

[Page 11]

INTERNET-DRAFT

CBT Border Router Specification

March 1998

APPENDIX

The BR "Arbitration Process"

The specific details of the arbitration process (AP) are implementa-

tion dependent, but we provide an outline of its possible operation for reference.

The AP is applicable to any multi-homed (stub or transit) CBT domain whose BRs have not deployed BGP-4+ [1]. The goal of the arbitration process is to allow CBT BRs attached to a CBT domain to select a single ingress BR per external multicast source in a timely fashion. A diagram showing the arbitration process in a CBT BR component is shown in figure 2.

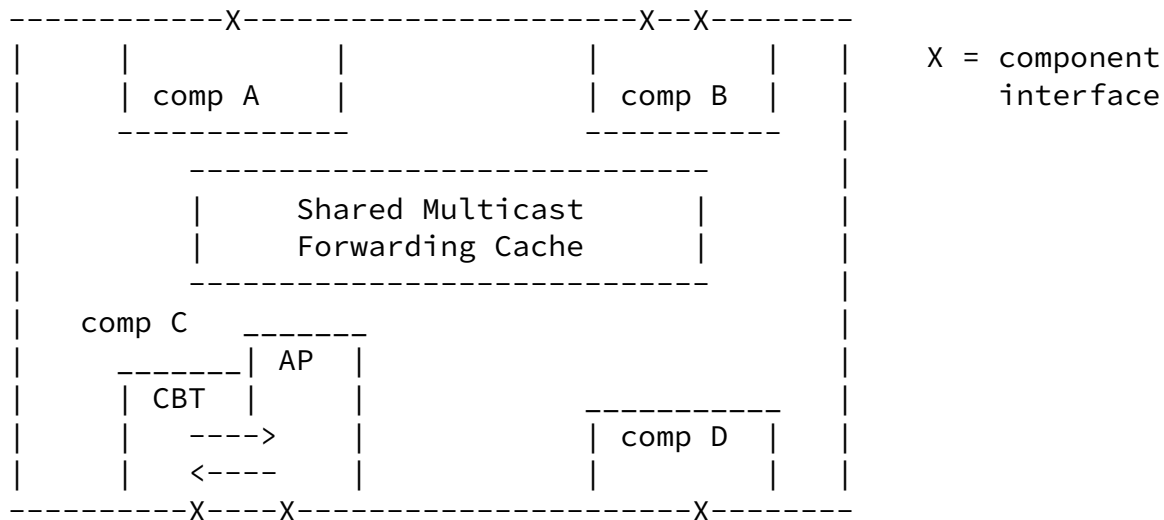


Figure 5: Logical Representation of the Arbitration Process

The arbitration process is only triggered by externally sourced packets, i.e. those passed to a CBT component interface by the BR process that handles SFC forwarding); the AP is not triggered by data packets arriving from inside the CBT domain.

All of a CBT domain's BRs are joined to the "all-CBT-border-routers" (ABR) multicast group (see [section 4](#)), the group address for which is domain-scoped, 239.X.X.X. This group is used both by the arbitration process, and by CBT transit domains for propagating multicast routing

information (in cases where multicast topology discovery is necessary).

Whenever a CBT component receives an externally sourced data packet for the first time (since some time 't-zero') the CBT component arbitration process is invoked.

The arbiter queries the BR component owning the interface via which the externally sourced data packet arrived, i.e. the component owning the interface nearest to S, to find out that component's current (multicast) metric for S. The arbiter multicasts an (S, G, metric, protocol component) tuple - where protocol component is DVMRP, PIM-DM, etc. - to the ABR group, and the message is processed by each receiving CBT component arbitration process. The receiving arbiter(s) cache the received tuple.

The unsolicited arrival of a triple triggers the receiving arbiter to reply with its own corresponding tuple; those CBT components not receiving the (S,G) multicast data simply reply with a NULL tuple, where "metric" and "protocol component" are both NULL. Failure to receive a reply from each other BR belonging to the ABR group results in the tuple being resent (unicast, unless no reply is forthcoming from any BR) after 3 (??) seconds.

The receipt of a non-NULL tuple with a better metric for S than this BR's tuple (for the same protocol), or an equal metric but sent by a lower-addressed BR, causes the arbiter to instigate the removal of the CBT component interface from the relevant SFC entry's child list.

The subsequent arrival of data packets from S are injected into the CBT domain via a single BR, with the same packets arriving at any of the other BRs being filtered. The arrival of subsequent data packets does not result in any exchanges between BR arbiters for the lifetime of the relevant cache entry's timeout period, which must be synchronised across all of the domain's BRs (recommended lifetime, X secs).

Whilst this method does not guarantee against multicast duplicates being injected into the CBT domain, it should ensure that any dupli-

cation is short-lived.

Author Information:

Tony Ballardie,
Research Consultant.

e-mail: ABallardie@acm.org

Brad Cain,
Bay Networks Inc.,
3, Federal Street,
Billerica, MA 01821, USA.
e-mail: bcain@baynetworks.com
voice: +1 978 916 1316

Zhaohui "Jeffrey" Zhang,
Bay Networks Inc.,
600 Technology Park Drive,
Billerica, MA 01821, USA.
Phone: +1 (978) 439 0280
Fax: +1 978 670 8760
e-mail: zzhang@baynetworks.com