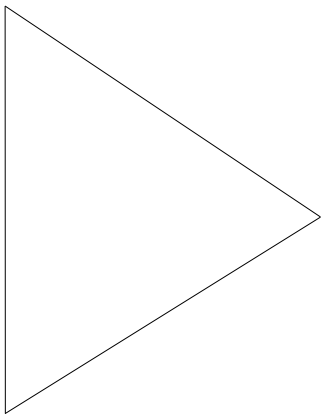
LS2020 LAN Switching Deployment Constraints



Design Implementation Guide

The following set of criteria are designed to ensure that near term LS2020 LAN (FDDI and Ethernet) deployments are successful. They are the result of field experience and considerable internal testing, which has informed us about areas of known deficiencies in the product as of Release 2.1(1.93).

This document consists of two sections. The first contains recommendations concerning feature usage. The second contains information about known capacity limitations.

Recommendations

ARP Spoofing

ARP spoofing is a feature which, if enabled, learns the mapping between MAC and IP addresses, and then issues local replies to ARP messages.

If there are any routers on LAN segments attached to LS2020 ports, **we recommend that ARP spoofing be turned off**. The reason for this is that we have experienced problems with the interaction of this feature and the routers' proxy ARP mechanism. Release 2.1(1.93) allows this feature to be configured on/off on each LAN port. The default configuration is on.

Custom Filters

Custom filters allow the selection of specific LAN to LAN streams based on the contents of the packet header. Once selected, a stream can be blocked, forwarded with the LAN default type of service, or forwarded with a prespecified type of service based on the custom filter that was matched. If forwarded, the stream can be multicast to a prespecified multicast group based on the custom filter that was matched. Custom filters may be specified at the MAC (bridging), IP, and IPX layers.

A custom filter, if matched, creates a pattern that must be added to the high-speed classifier ASIC known as the recognizer. This pattern is called a flow, and there is a limited capacity (described below) to store these flows in each LS2020 chassis' NP memory. Custom filters above the MAC layer may cause an unexpected increase in flow requirements. For example, a pair of routers exchanging IP packets requires one flow. If custom filters are created to differentiate UDP from TCP, then a flow is generated for every LAN stream with a unique combination of Source IP address, Destination IP address, and choice of Layer 4 protocol.

Because of the potential to require massive amounts of flows, **we recommend against casual use of custom filtering above the MAC layer**. This recommendation may be relaxed in circumstances where it has been assured by careful analysis that IP and/or IPX layer custom filters will not generate excessive flows.

Cisco Confidential
Copyright © 1996 Cisco Systems, Inc. All Rights Reserved.

CISCO SYSTEMS

Broadcast and Multicast

Under default circumstances, packets whose destination address indicates they must be delivered to multiple LAN ports do not use the ATM virtual connection mechanism. Instead, they are “flooded” to all appropriate LAN ports in the network. Each packet is multicast by the local NP to every other NP in the network (via an ATM point to multipoint circuit), and each destination NP makes copies for each of it’s chassis’ LAN ports. The broadcast/multicast is delivered only to those ports that share workgroup membership (see below) with the source port.

The High Performance Multicast Service (HPMS) permits the configuration of multicast groups that specify the destination ports, and association of custom filters with multicast groups to specify the packet headers of traffic that should be sent to that multicast group. Traffic that matches these custom filters is delivered using an ATM point to multipoint virtual connection, which operates in hardware without any processor intervention.

The “flooding” mechanism is relatively slow (about 75-100 packets/second), and intended to handle low-frequency broadcast needs only. **We recommend always using HPMS for broadcast and multicast traffic**. In addition to any application-specific HPMS configuration, there should be a “general” HPMS configuration as follows:

1. Multicast Group Definition

   All the LAN ports in the network, or, if there are multiple workgroups, then multiple multicast groups corresponding to each possible combination of port workgroup membership.

2. Custom Filter Definition

   A MAC destination address filter that matches any DA with the “multicast” bit on.

3. Port Configuration

   Every LAN port in the network is assigned the filter (2) and associated multicast group (1). In the case of multiple workgroups, the assigned multicast group is the one that corresponds to the combined workgroup membership of the port.

Workgroups

LS2020 workgroups are similar in concept to VLANs. They provide privacy domains to which LAN ports may belong, and which limit the distribution of broadcast and multicast packets. Unlike VLANs, however, a port may have multiple workgroup membership.

There is no specified or automatic relationship between workgroups and multicast groups. If both are configured, the multicast group definition dictates the destinations for HPMS traffic, even if the source and destination ports do not have compatible workgroup configurations. We recommend using multicast groups for containment of multicast traffic, and workgroups for securing unicast traffic across administrative domains.

There are a few interactions between the workgroup and spanning tree mechanisms which may affect a particular network design.

1. BPDU Propagation

   The workgroup configuration does not affect the delivery of bridge protocol (spanning tree) data units. There is only one spanning tree for the entire network and its operation is independent of workgroups. You need to be aware that unusual **spanning tree activity (e.g. root wars) in one workgroup will propagate to other workgroups**, so that workgroups do not provide this kind of broadcast protection.

2. Workgroup partitioning

   The transport resources internal to the LS2020 network are open to traffic from any workgroup. **External LAN segments represent workgroup restricted transport resources**. If the ATM network partitions such that chassis cannot communicate over their ATM trunks, but there is a path between them via a redundantly connected LAN segment, then communication between the two chassis can happen only in workgroups to which the common LAN segment/redundant port pair belongs.

Cisco Confidential
Copyright © 1996 Cisco Systems, Inc. All Rights Reserved.

Redundant Attachments

The LS2020 LAN service includes a conformant implementation of the IEEE 802.1 spanning tree algorithm. This means that external LAN segments may be connected to the network over multiple ports (on the same or different chassis), in order to have redundant connections. The spanning tree algorithm will forward traffic to and from the segment via one port at time, switching to alternate ports only when the primary is not able to perform that function.

It has been observed that cutover from one port to another can create a problem if the change causes the rehoming of large amounts of MAC addresses and tearing down/rebuilding of large numbers of flows. Capacity limitations that apply in these circumstances are described in the next section. **We recommend that if redundant attachments are used in a network design, there needs to be careful analysis of each failure scenario to insure capacity limitations are not exceeded**.

Responsiveness

We have also measured the "settling time" for massive LAN service reconfiguration such as MAC address learning/removal and flow creation/deletion. In all cases, the transient was essentially over within 8 minutes. Since the console or other network management interface might remain unresponsive during this period of intense CPU utilization, **we recommend waiting for at least 8 minutes** after any significant event before taking extreme action based on lack of response at the management interface.

Capacity Limitations

LS2020s

A network of LS2020 chassis connected by trunks implements a variety of distributed algorithms, including congestion avoidance (CA), ATM routing, and MAC database synchronization. As the number of chassis increases, there is additional NP loading and trunk control bandwidth utilization. We recommend deploying **no more than 20 chassis per network** without detailed analysis of scalability issues.

MAC Addresses

The deployment of FDDI and Ethernet switched networks based on LS2020s ran into problems that appeared to be triggered by large MAC address complements. The maximum complement the network is designed to support is 10,000 MAC addresses.

We have done analysis and testing and have a better understanding of this issue. At the root of the problem is the occurrence of a buffer shortage under load that crashes the system. While we will not be able to get a fix for this system problem into Release 2.1(1.93), we have tested the effect of MAC address complements in provoking this behavior. It turns out that the biggest stress in the network is on the chassis with the most number of neighbors, and we can state the MAC address complement limit in terms of the largest number of neighbors of any 2020 chassis in the network.

If the chassis with the most neighbors in the network has four neighbors (i.e. it has four trunks going to four other chassis), **then the maximum number of MAC addresses the network can support is 3,000**. These can be distributed in any way over all the ports and chassis in the net.

If the chassis with the most neighbors in the network has three neighbors, then the maximum number of MAC addresses the network can support is 6,000. These can be distributed in any way over all the ports and chassis in the net.

If the chassis with the most neighbors in the network has two or fewer neighbors (including the case of 0 neighbors, i.e. a single chassis network), **then the maximum number of MAC addresses the network can support is 10,000**. These can be distributed in any way over all the ports and chassis in the net.

Flows

There is a limit of **10,000 flows per LS2020** chassis. By default, each LAN card gets 1200 flows. The flows can be re-apportioned by configuration, provided that the sum of flows from each LAN card does not exceed the 10,000 limit.

Cisco Confidential
Copyright © 1996 Cisco Systems, Inc. All Rights Reserved.

Flows are the mechanism by which the NP keeps track of what patterns have been programmed into the recognizer. If the NP has no remaining capacity for a new flow, that pattern cannot be programmed into the recognizer. If a pattern cannot be programmed into the recognizer, a packet containing that pattern will be sent to the NP as an “unrecognized packet” for processing by the NP.

Therefore, the effect of exceeding a card’s flow capacity is to potentially impact the forwarding performance of some LAN streams. Up to a certain point, there is no packet loss, because the NP will drop or forward the packet anyway, even if it is unable to create a flow. It means that the next time that pattern shows up, the packet will go back to the NP instead of being handled automatically by the high-speed recognizer. If this happens a lot, the forwarding rate for those LAN streams caught outside the cache will be very low. Furthermore, once the 50 packet/second NP packet processing rate is exceeded, packets will be lost and the whole system rendered unresponsive to monitoring or control activities.

This “flow overflow” is only serious if on a steady state basis there is a greater number of “active” LAN streams than flows. This is because infrequent patterns will automatically migrate out of the flow cache. In fact, it is more efficient, at a low frequency, to not create the flow anyway, since it will time out before it can be used again. The normal **flow time-out is 5 minutes**.

Some analysis of flow utilization is highly recommended. Flow requirements should be estimated based on simultaneous usage over a 7.5 minute period. Pick the worst period of the day. Note that the worst period for flows may not be the same as the worst period for traffic. Think about what happens at 9 am and right after lunch (many logins, lots of E-mail) as well as peak traffic times. Below, when estimating the numbers and port locations of traffic sources and sinks, consider all cases of single failure. Remember that the flows are shared across all the ports of a LAN card. Thus the estimate must be for the total requirements of all the ports.

A separate flow is created on each port for each unique combination of MAC source address, MAC destination address, and protocol type (three values- IP, IPX, and “other”). This is true whether or not MAC layer custom filters are configured. If IP or IPX layer custom filters are used, there is a further set of flows created for each unique combination of source and destination L3 address and L4 protocol type. Here are some rules of thumb that can be used based on node (PC/server) population, protocol usage, and client/server configuration:

- Client: 10 Flows.
- Server: Estimate the number of clients directly served by and bridged to the server. Add one flow per protocol per router.
- Router: Estimate the number of clients directly served by the router (within the bridged domain which includes the LS2020s). Add one flow per protocol per peer router across that same domain.

There is a tool (*smstats* in*/usr/fldsup/bin*) that will display statistics about currently active flows and can be used to verify the design.

Example Scenario:

Card 4 Port 0: FDDI ring has 4 Cat5000s with 20 clients behind each. The Cat5000s are dual homed to two rings with the spanning tree port costs set such that two Cat5000s are served by each ring.

Card 4 Port 1: Has four file servers each of which serves 100 clients. Each server must take up the load of an equivalent server if that server fails. The network topology is such that a single failure can cause the outage of two of the servers which are load sharing with these.

Flow Requirements	Port 0	Port 1	Card Flows
Normal Operation	2 Cat5000s X 20 clients X 10 Flows = 400	4 servers X 100 Flows = 400 Flows	800
Server Redundancy	2 Cat5000s X 20 clients X 10 Flows = 400	2 servers X 100 Flows + 2 servers X 200 Flows = 600 Flows	1000
Terminal Redundancy	4 Cat5000s X 20 clients X 10 Flows = 800	4 servers X 100 Flows = 400 Flows	1200

Cisco Confidential
Copyright © 1996 Cisco Systems, Inc. All Rights Reserved.

Redundant Attachments

We have observed two failure scenarios associated with redundant attachments. These scenarios were associated with internal testing and are at the extreme of possible operational usage.

1 Spanning Tree Overload

If new MAC addresses and corresponding flows appear, at a sufficiently high rate, on a LAN that is redundantly attached, we have observed a failure scenario whereby the blocked port reverts to forwarding mode despite the fact that the other port is also in forwarding mode. This happens when new MAC addresses (and corresponding packets to be delivered to known destinations) are presented on a redundantly connected LAN at a rate **exceeding 2,230 new addresses/second for over 30 seconds**.

We are in the process of determining a fix for this problem, which is presumed to be caused by insufficient cycles available for BPDU processing. Since this problem will not be fixed in Release 2.1(1.93), we recommend that designs containing redundant attachments consider if the above scenario is possible.

2 Learning Overload

In addition to the network wide MAC address database on 10,000 entries kept on the NP, each line card also has storage for up to 16,000 local entries. Each of these entries corresponds to a particular MAC address learned on a particular port. If a LAN is redundantly connected, then there will be redundant entries for each learned address. Furthermore, all the source addresses in packets delivered over the forwarding port will be learned as local addresses by the blocking port(s). Immediately following a spanning tree topology change the number of frames that are “flooded” by the LS2020 bridge and therefore seen by all blocking ports as local may be surprisingly high. Thus, redundant attachments have the possibility, though the multiplicative effect described above, of exceeding a card’s local address cache.

If the capacity of 16,000 MACs per LAN card is exceeded, the card will crash. We are in the process of determining a fix for this problem, but the fix will not be available in Release 2.1(1.93). The following is a very conservative formula to use in a design containing redundant attachments:

if **M = number of distinct MAC addresses in the network**, and

if **R = maximum number of LAN ports in spanning tree blocking state on any one card**,

then **M * R < 16,000**, or else there is a risk that a spanning tree topology change may cause that card to crash.

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
World Wide Web URL:
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

Cisco Systems has more than 125 sales offices worldwide. To contact your local account representative, call Cisco's corporate headquarters (California, USA) at 408 526-4000 or, in North America, call 800 553-NETS (6387).
0496R

AtmDirector, Catalyst, CD-PAC, CiscoAdvantage, CiscoFusion, Cisco IOS, the Cisco IOS logo, *CiscoLink*, CiscoPro, the CiscoPro logo, CiscoRemote, Cisco Systems, CiscoView, CiscoVision, CiscoWorks, ClickStart, ControlStream, EtherChannel, FastCell, FastForward, FastManager, FastMate, FragmentFree, HubSwitch, Internet Junction, LAN^{2}LAN Enterprise, LAN^{2}LAN Remote Office, LightSwitch, Newport Systems Solutions, *Packet*, PIX, Point and Click Internetworking, RouteStream, SMARTnet, StreamView, SwitchProbe, SwitchVision, SwitchWare, SynchroniCD, *The Cell*, TokenSwitch, TrafficDirector, VirtualStream, VlanDirector, WNIC, Workgroup Director, Workgroup Stack, and XCI are trademarks, Access by Cisco, Bringing the power of internetworking to everyone, and The Network Works. No Excuses. are service marks, and Cisco, the Cisco Systems logo, CollisionFree, Combinet, the Diamond logo, EtherSwitch, FastHub, FastLink, FastNIC, FastSwitch, Grand, Grand Junction, Grand Junction Networks, the Grand Junction Networks logo, the Highway logo, HSSI, IGRP, Kalpana, the Kalpana logo, LightStream, Personal Ethernet, and UniverCD are registered trademarks of Cisco Systems, Inc. All other trademarks, service marks, registered trademarks, or registered service marks mentioned in this document are the property of their respective owners. 0496R
1096R

Copyright © 1996 Cisco Systems, Inc. All Rights Reserved.