Link Aggregation

Link aggregation allows for the bundling of two or more links into one virtual channel. Link aggregation is also known as EtherChannel, Port Channel, Port aggregation or trunking, depending on the vendor involved. The IEEE 802.3ad or LACP specifications is applicable.

Pricing and Availability

MN-IX currently offers link aggregation on any 1G/10G/100G physical connection. Aggregating links acquired from different partners or resellers, however, is not supported. The port prices for aggregated links are identical to the normal port prices. Refer to www.mn-ix.com/fees

Due to technical limitations of the switches used by MN-IX it may be necessary to relocate your existing port. If this turns out to be the case, MN-IX will inform you and advise you of any additional steps necessary for this process.

MN-IX can deliver aggregated links at all co-locations. Although a strict reading of the spec forbids it, we can offer aggregated links over different media types of the same speed.

Load-Balancing Algorithm for the NetIron MLX and SLX platforms

The load-balancing algorithms used in our SLX switches uses a modulo operation, leading to the best distribution over links with the entropy available in source and destination IPv4, TCP or UDP source and destination port number, as applicable.

LACP & MN-IX Topology

LACP is supported at MN-IX for all connection types. When used with 10Gbit/s and 100Gbit/s links (where available) use of LACP does introduce an inconvenient side effect. After a topology failover ports with LACP enabled will stay in blocking mode until the first LACP frame is received. Because this may take several tens of second (depending on vendor implementation) this can cause BGP sessions to flap.

When enabling LACP we advise to configure the LACP timeout to short, to limit the maximum failover time to 3 seconds.

Configuration Hints

We have collected information about link aggregation for several router platforms in our configuration guide.

SEE CONFIG GUIDE(link to previous config guide here...)

Quarantine VLAN

MN-IX has implemented a feature called “Quarantine VLAN” whereby all new ports are placed in a separate VPLS instance, which is used for testing purposes. before the customer connection is moved to the production environment.

What is a Quarantine VLAN?

A quarantine VLAN is a VPLS instance on the MN-IX switch containing the following:

* (new) customer ports

* MN-IX monitoring system

The monitoring system sniffs all broadcast, multicast and unknown unicast in the quarantine VLAN.

Why have Quarantine VLANs?

MN-IX defines a fairly strict set of allowed traffic types on the peering LANs. Not all routers (and intermediate L2 devices) adhere to these guidelines; they typically have various protocols turned on by default such as CDP, EDP, STP, DEC MOP, etc., or they present more than one MAC address to the platform. These misbehaving/misconfigured devices potentially endanger the stability of the peers and/or switching platform. Hence, we cannot allow them on the peering LANs.Rather than act reactively once a customer port is in production, we prefer to detect and fix these issues beforehand. Therefore, we introduced the concept of a quarantine VLAN. Once a customers router is connected and the port is up, we can quickly see if it adheres to the rules. If not, the violating traffic does not harm the rest of the platform.

When do you use Quarantine VLANs?

New ports are always put into a quarantine VLAN first. This also is the case for upgrades, downgrades and relocations.

In addition to the above, existing customer ports may be moved into quarantine VLAN if they violate the allowed traffic types. Please note that this is only done in extreme cases.

How do I get out of a Quarantine VLAN?

If your port is moved into Quarantine the MN-IX NOC will notify you for this. If the reason is because you are sending illegal traffic, your configuration should be updated accordingly. Once you are confident the port adheres to the rules please contact the MN-IX NOC and request the port be put back in production. The NOC will check the port's behaviour again. If all is fine, the port will be moved (back) into production. If not, we will notify you with details of the problem.

sFlow at MN-IX

To analyse and optimise high speed networks, an efficient monitoring system is required. MN-IX uses sFlow for its traffic analysis.

Introduction

sFlow is a standard to capture traffic data in switched or routed networks. It uses a sampling technology to collect statistics from the device and is for that reason applicable to gigabit speeds or higher. Due to it being an open standard, described in RFC 3176, it is implemented on a wide range of devices, like the MN-IX Brocade/Extreme switches.

The sFlow agent (Brocade/Extreme switch) supports two forms of operation:

* time-based sampling of counters

* packet-based sampling of ethernet frames

The counter samples provide exactly the same information MN-IX uses for its traffic statistics now, therefore the sFlow implementation at MN-IX makes use of the packet-based samples (called flowsamples) to provide additional analysis of the exchanged traffic.

Packet Based Sampling

Based on a defined sampling rate, one out of N frames from the incoming traffic for each interface gets sampled and sent to a central server which is statistically analyzing the traffic. If we see one packet out of N, we assume that all the N-1 packets we haven't seen are the same type and size.

Note: this type of sampling does not provide 100% accurate results.

Without sampling technology, packet analysis on a network with a throughput like MN-IX would not be possible. For more detailed information about the accuracy of packed based sampling see the documents on the official sFlow website.

Software

The sFlow samples on the server get analysed by software developed at MN-IX. The software package is written in PERL and based on the sFlow decoding module Net::sFlow.

Note: While the sFlow packet format supports sampling of IP and TCP/UDP flows, our software only looks at Layer-2 (Ethernet) fields. We neither process nor store flow information from higher layer protocols.

Controlling ARP Traffic on MN-IX platform

Controlling ARP Traffic on MN-IX platform.

1. ARP (Address Resolution Protocol)

ARP (Address Resolution Protocol) is the Layer-2 protocol used by MN-IX member's router to associate IPv4 address with the MAC address of peers interfaces. Learn more about ARP here.

2. Problems caused by too much ARP traffic

On Ethernet networks, the Address Resolution Protocol (ARP) is used to find the MAC-address for a given IPv4 address. ARP uses Ethertype 0x0806 together with Ethernet broadcasting. A node will broadcast an ARP Request packet to ask for the MAC address of an unknown IPv4 address. The node using the requested IP address replies (using regular unicast) with an ARP Reply packet, which includes its MAC address. In order to work, it is important that all nodes using IPv4 listen for ARP packets and reply to them if necessary.

The nodes therefore need to process all Ethernet broadcast messages with Ethertype 0x0806. For each ARP packet, they must decide whether or not to reply. Processing ARP packets can take a lot of processing power. Because all ARP packets need to be examined in order for ARP to work, processing ARP packets may take precedence over other activities, depending on the Operating System. As such, when there is a lot of ARP traffic, routers may be unable to do other processing tasks like maintaining BGP sessions.

This problem was noticed on MN-IX when the ISP peering LAN was renumbered to new IPv4 addresses. Members in the new IPv4 range were trying to reach members in the old IPv4range and vice versa. Larger amounts of ARP packets than usual crossed the network, consuming all available processing power on some customer routers, not leaving enough resources to process BGP in a timely manner, resulting in lost BGP sessions. Also, routers trying to re-establish old BGP sessions started sending ARP packets, resulting in an ARP storm that caused even more problems on customer equipment.

3. ARP Sponge – the MN-IX solution

To help routers survive heavy ARP traffic, MN-IX tries to keep the amount of ARP traffic to a minimum. For this purpose, MN-IX developed a daemon, written in Perl, called ARP Sponge. The ARP Sponge daemon listens on the ISP peering LAN for ARP traffic. When the number of ARP Requests for a certain IP address exceeds a threshold, the ARP Sponge sends out an ARP Reply for that IP address using its own MAC address. From that moment, the IP address is sponged: all traffic to that node is sent to the ARP Sponge. This prevents ARP storms because it keeps the amount of ARP traffic limited.

When the interface of a sponged IP address comes up again, it generally sends out a gratuitous ARP request packet. This is an ARP packet with both source and destination IP address set to the IP address of the node sending the packet. It is used mostly in case the MAC-address changed, so that other nodes can update their ARP caches. When the ARP Sponge receives any traffic from a sponged IP address (including but not limited to gratuitous ARP requests, ARP requests for other nodes, BGP peering initiations, etc.), it ceases sponging the IP address, thus no longer sending out ARP replies for that IP address.

4. Common Issue with IPv4 addresses after being sponged by ARP Sponge

* Unable to exchange traffic with MN-IX peers when IPv4 address comes up after being inactive for a period of time

If a IPv4 is sponged, it means that in the members ARP tables, the ARP entry for this IP is registered with the ARP sponge MAC address. After the IPv4 is again reachable again and being "un-sponged", the ARP table of peers might not be updated fast enough with the customer’s MAC address, result in traffic from these peers toward the recovered IP still being forward to the sponged MAC address.

For instance, if the IP 77.69.248.x of member A with MAC address AAAA.AAAA.AAA is sponged with the sponged MAC address EEEE.EEEE.EEEE, then member B ARP entry for the address will be 77.69.248.x - EEEE.EEEE.EEE. After the member A recovers, it send traffics toward member B, but member B ARP entry is not yet updated with the original address AAAA.AAAA.AAAA, then traffic will be ended up sending to EEEE.EEEE.EEEE, until member B updates the ARP entry.

This issue should be automatically resolved after a certain period of time, after the daemon stop replying to ARP reply for this IP and let the un-sponged IP and peers update ARP entries themselves.

The issue is more noticeable with members that only have peering sessions with the MN-IX route-servers. If members do not have peering sessions with route-servers, BGP sessions with peers must be brought up one-by-one and ARP entries are sure to be updated through the BGP initialisation process. Subsequently, traffic will be properly forwarded and received from each peers. However, if the newly "un-sponged" member only has peering sessions with route-servers, and after recovery establishes BGP sessions and receives MN-IX peers prefixes from there, there could be a case that traffic is forwarded to the next hop IP of peers that still have the spoofed ARP entries.

Therefore, MN-IX NOC recommend members that have their IPv4 address being unreachable for prolonged period of time (so it is certainly sponged), to temporary shutdown peering with route-servers and send gratuitous ARP request to update peer's ARP tables.

MN-IX ROUTE SERVERS

Acknowledgement

The ARP sponge explanation section is an excerpt from the report of Marco Wessel and Niels Sijm from Universiteit van Amsterdam in 2009, after they did the research about effect of IPv4 and IPv6 address solution on AMS-IX platform and the ARP sponge during their course for Master in System and Network Engineering.