WHITE PAPER

Security in Mellanox Technologies InfiniBand Fabrics

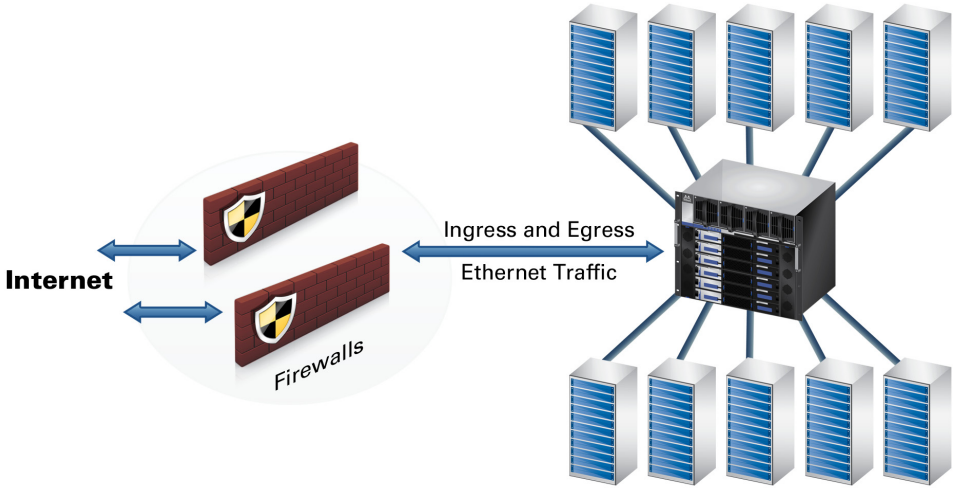
Technical Overview

Overview

InfiniBand is a new systems interconnect designed for Data Center Networks, and Clustering environments. Already, it is the fabric of choice for high-performance computing, education, life sciences, oil and gas, auto manufacturing and increasingly financial services applications.

As a local area network that will be deployed in data centers, InfiniBand architecture targets one of the main concerns in such environments which is Security, and has many built in mandatory features that enable much better Isolation and Security than current networks and other cluster interconnects.

InfiniBand is a data center technology and is not typically deployed in DMZ's unless firewalls or other similar access controlled layers are placed in front of an InfiniBand fabric as depicted in below figure:



As a non-DMZ layer-2 networking technology, InfiniBand is typically more secure than Ethernet for architectural and Mellanox Technologies-implementation-specific reasons. This document is meant as an intermediate level Mellanox Technologies InfiniBand security review aimed at a security professional or a technical reader.

This document is divided into multiple sections.

- First, overall Mellanox Technologies design of its equipment and software is described as it relates to security – The Big Picture.
- Second, the document delves into specific Mellanox Technologies products offerings as they relate to security.
- Third, the document discusses InfiniBand protocols as they are implemented by Mellanox Technologies end-to-end to give the reader a more complete picture.

The Big Picture

InfiniBand is a layer 2 protocol much like Ethernet. This means that almost all Layer 3-7 application security mechanisms that are built on top of Ethernet in typical applications will work the same way with InfiniBand. As an example: SSH sessions that run over InfiniBand (IPoIB or SDP) will have all the inherent security that are built into SSH on Ethernet. Namely: high grade encryption, inability to perform replay attacks, protection mechanisms against "man-in-the-middle" attacks, etc. This is very important to keep in mind as applications designed for InfiniBand from scratch (from a security perspective) need to be built on the same premises and principles as the same application would have been on Ethernet. Application encryption, authentication, authorization and audit is still the responsibility of the application developer regardless of the underlying network type.

Because of the above, it is important to approach InfiniBand security review the same way as one would approach Ethernet or FC security review. This document will cover some major known attacks against Ethernet devices and how InfiniBand deals with protecting against those vulnerabilities. As well as enhanced mechanisms for preventing new attacks that may be discovered in Ethernet; which are impossible to perform in a Mellanox Technologies InfiniBand fabric due to inherent security of the underlying layer 2 protocol, and securing physical InfiniBand fabric devices will be covered.

Mellanox Technologies Product Security

Mellanox Technologies designs and sells various size director InfiniBand switches as well as gateway devices to connect into Ethernet networks. Each class of Mellanox Technologies products will be covered separately in this document.

Few security-related features are common across all Mellanox Technologies devices.

Mellanox Technologies has made the decision to have all switch and switch component administration to be done out-of-band via Ethernet ports that are built into Mellanox Technologies switches, Ethernet to InfiniBand bridges and FC to InfiniBand bridges. This means that even if Mellanox Technologies InfiniBand switches are connected to an in-secure network, management of Mellanox Technologies switches or bridges cannot be done from a compromised node on that network. This is very useful as it allows network designers to place management ports of Mellanox Technologies devices on a protected management network that can be secured by firewalls, RSA token access, RADIUS, and TACAC's by forcing a switch administrator to have to traverse a security bastion host and other layered security infrastructure.

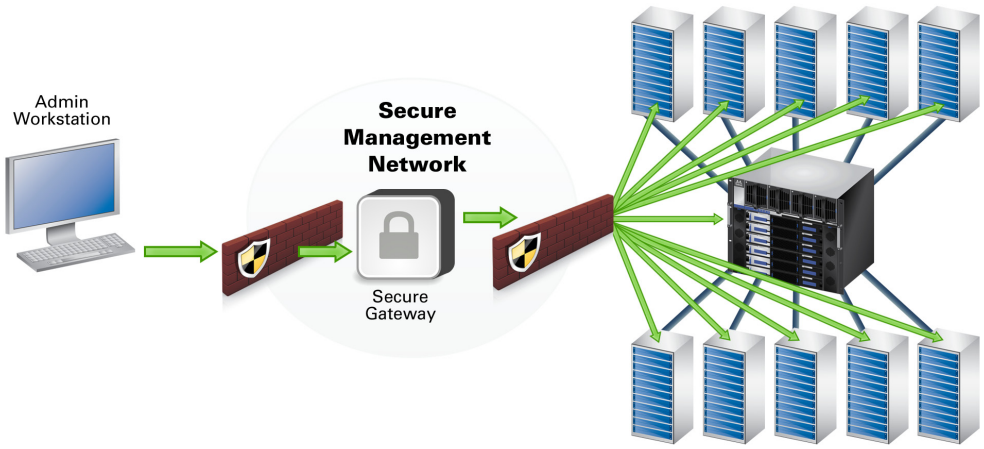
Mellanox Technologies director grade switches also have a number of important security features. Below are some of them:

- RADIUS authentication
- IP filtering for switch administrative (Ethernet) management ports
- Two classes of users: read-only and full rights (a large number of users with individual accounts is supported via RADIUS)
- SSH Support (including scp to be used for upgrades and configuration backup)

- Ability to administratively turn off ports

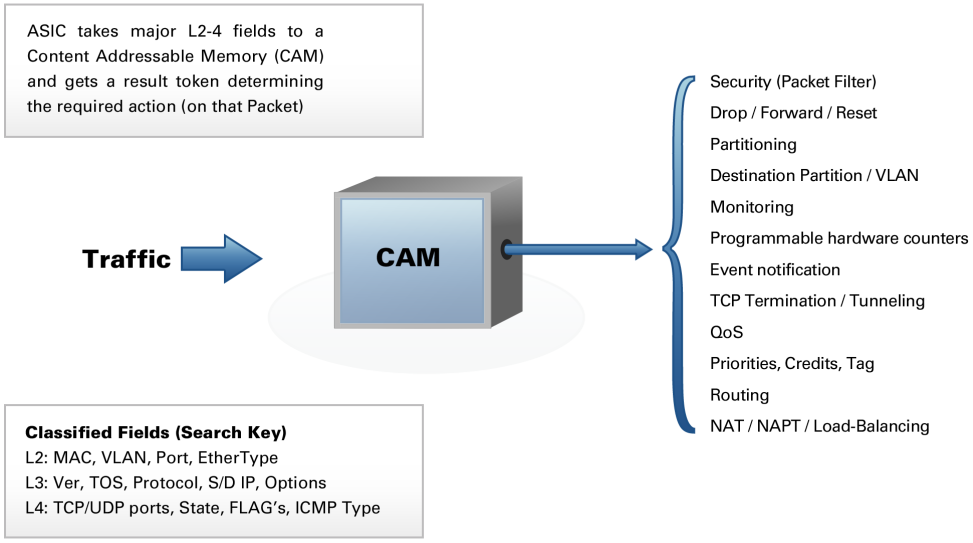
As InfiniBand switch ports are not addressable, compromised hosts do not pose a break-in risk for the InfiniBand switches and bridge, there is no "address" to bind the switch to as a pass-thru device. A compromised node of course poses a risk to other nodes on the same network as insecure services can be exploited on other hosts (same way as Ethernet). Standard layer 2 attacks that are very popular in Ethernet cannot be exploited in InfiniBand. Such as SYN flood denial-of-service attacks, as the handshaking process is very different for InfiniBand and the hardware on the receiver will auto-discard syn-flood packets (instead of passing them to the OS).

It is important to note that each IPR has its own management port that is independent from the switch management port allowing different access rights to each device. A person can set firewall rules stating that user "joe-smith" from node IP x.x.x.x can access management port of IPR1 with IP of x.x.y.y only and not the adjacent IPR with IP x.x.y.z or switch management port with IP x.x.z.z. Layered access to fabric management is depicted below.

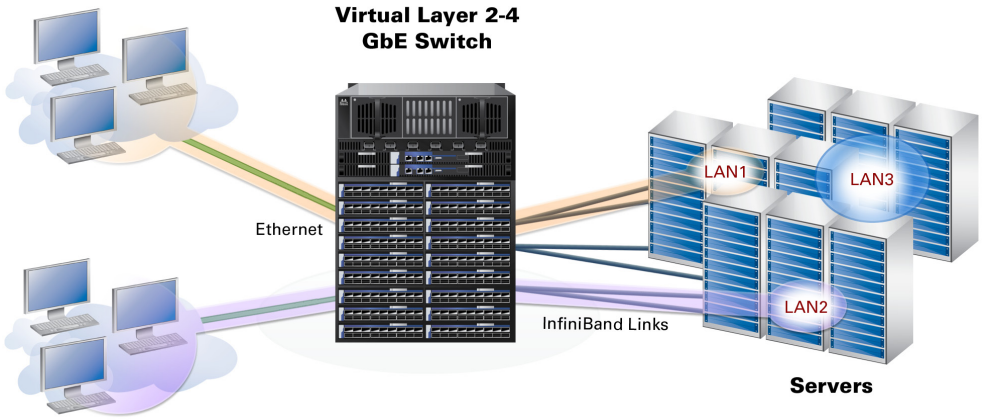


Current and Future Mellanox Technologies Ethernet Gateway Products

It is important to provide a brief description of Mellanox Technologies InfiniBand to Ethernet gateway and its filtering mechanisms. Mellanox Technologies' custom ASIC takes each packet coming into the ASIC and places all of the L2-4 clasifyer fields into CAM memory. Based on explicit rules set on the IPR, the IPR can drop or forward packet that matches or does not match L2-4 criteria set forth to allow the packet to pass. Please refer to below figure for further detail:



IPR transparently extends Ethernet VLANs to InfiniBand fabric and enforces isolation on both sides across both Fabrics.



Central Resource Management/Hardware Based Protocols

Unlike traditional networks InfiniBand (IB) standards also defines the Fabric Management services that include standard mechanisms to configure Security, QoS, Services, Routing, and collect status/statistics on various Fabric elements.

The Fabric Management is a central entity that can also have persistent backups for redundancy, which configures and initializes all the ports and switches in the fabric.

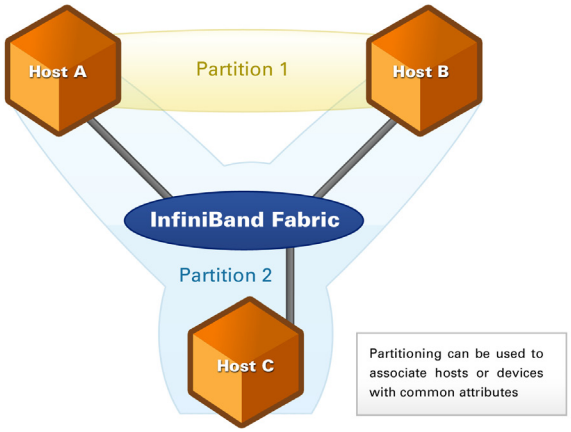
The fabric configuration is done using special keys such as the Management key (M_Key), and has several mechanisms to protect against malicious acts. Such as standard traps that are sent automatically on any violation attempt.

InfiniBand protocols are implemented in hardware and the communication attributes such as addressing and partitioning. Multicast are configured centrally in a way that doesn not enable software applications to gain control over them and maliciously change those attributes. Any such attempt is recorded and sent to the administrator via SNMP and switch Device and Fabric management GUIs.

In addition, each node and port is identified by the Fabric Management through a 64-bit Global unique identifier (GUID) maintained by hardware, helping to avoid impersonation. Unlike Ethernet (where MAC can be changed via software to impersonate another machine or a packet can be pre-constructed in software), crafting malicious packets in software or changing hardware addressing is basically impossible – such packets will be discarded either by the switch or HCA's.

This is rather different than the current situation in Ethernet networks that have a very complex protocol stack implemented in software (especially on servers where the NIC almost always will pass the packet to hardware driver/TCP stack and will not discard the packets itself), which can be compromised by software agents.

InfiniBand Partitioning



One of the security and isolation mechanisms provided by InfiniBand is partitioning. Partitioning enables definition to the resource group membership for each node or port. This way a port in one partition can be denied access to another partition through hardware mechanisms, similar to Network VLAN mechanisms.

The uniqueness of the InfiniBand partitioning mechanism is that it is well defined. It is centrally controlled by the Subnet Management (SM, part of the Fabric Management) entity, and in addition a node cannot determine its own partition (it is set and enforced by subnet manager software on the switches). This enables organizations that utilize InfiniBand to eliminate potential hacking and security holes possible in standard networks today via compromised servers.

The InfiniBand partitioning mechanism works in the following way:

1. The administrator determines the group membership using one of the Subnet Management remote interfaces, over a secured link (SSH, etc)
2. For any new node in the fabric
    - **A.** The SM initializes administrator configured port attributes on switch
    - **B.** The SM initializes administrator configured port attributes on the server
    - **C.** The SM assigns a local address (LID) to the HCA port
    - **D.** The SM assigns partition attributes to the HCA LIDs in the port partition table on the switch
3. The SM configures the partitions in the switch tables –- compromising server nodes is useless as servers have no say about what partition they are members of.
4. The host port attributes are maintained in the switch and HCA hardware and can only be accessed through special Management Key (M_Key) used and known by SM and hardware (IB silicon).
5. When an application chooses to connect to a remote node, it first resolves the destination location and the partition using standard name resolution functions (similar to IP ones), and it opens a connection to the destination.
6. The application CANNOT specify the partition, but only the existing index in the local partition table, this way an application cannot choose to connect to a partition which the node is not a member of (even if the switch were to theoretically allow this – which it does not).

The mentioned process enables better isolation in InfiniBand based partitions then in the case of Ethernet VLAN mechanism that is only enforced by the switches.

Network and Transport Layer Security

A potential attacker will try to access unauthorized destinations, listen on someone else's traffic, or impersonate another entity. The following InfiniBand mechanisms minimize these threats.

Privacy

In the basis, InfiniBand is a strict switched fabric. Traffic cannot just arrive to an unwanted node. The switching tables are explicitly defined by the fabric management (LID granting process as discussed above), unlike Ethernet which may forward traffic to multiple ports when it is not aware of the destination port for a given MAC address (standard CAM-table attacks).

In the case of InfiniBand Multicast traffic, the node first requests to join a Multicast group from the Fabric Manager, which should grant the access, and then configure the Multicast forwarding tables in the switches. The node cannot just send or listen on a broadcast or multicast traffic like in Ethernet, or issue IGMP request that will operate the IGMP snooping mechanisms in the switches.

The explicit Unicast and Multicast forwarding in InfiniBand eliminates the need for encryption mechanisms within the fabric – A host cannot listen to traffic not explicitly destined for it.

InfiniBand defines two transport types Reliable (RC, RD) and Unreliable (UD). These have several mechanisms that enable to secure against unauthorized access and session hijacking.

Security in Unreliable Transport

When using Unreliable Datagram (UD), traffic the receiver opens a Queue Pair (QP) and listens on incoming traffic from multiple sources, the senders opens their own QP's and transmits Unicast or Multicast traffic over the fabric.

Since UD traffic is stateless and more vulnerable to attacks, the two nodes that participate in a certain service communication must use a key (Q_Key) that travels with the packets. If the receiver sees that the key doesn't correspond to its own Q_Key, than it will drop the packets. This can help isolating traffic of a specific service, even between two members of the same partition.

If ever a synthetically generated packet with a wrong Q_Key arrives to a receiver, the receiver drops the packet and sends a trap automatically to the Fabric Manager. This event is recorded by the SM and can be forwarded as a SNMP trap to a monitoring system, such as Tivoli or Unicenter, etc.

Security in Reliable Transport

InfiniBand Reliable transports, which is what SDP uses, are implemented in hardware, with no software access to the protocol. The two peers first use an entity called Communication Management (CM) to initiate the communication where every node sends its credentials to the other, over management datagrams (MAD's). Followed by requesting the hardware to move the Queue Pair (QP) to an operational state with all the connection attributes.

After a QP is functioning, every message includes: addresses, Sequential numbers and two CRC numbers that are all generated by hardware. If for some reason one of the values is unexpected or wrong, it is dropped and registered. InfiniBand uses strict ordering, no windows or out of order packet delivery is allowed, unlike TCP which also assists in getting the high-performance.

Memory Protection

InfiniBand transports implement a mechanism called Remote Direct Memory Access (RDMA) that enables avoiding unnecessary copies and reducing latencies, similar mechanisms are also available in other interconnects, such as Myrinet.

RDMA mechanisms can create a potential security vulnerability. A node may access another nodes memory region that was supposed to be banned.

In order to protect remote memory access to unauthorized memory areas, InfiniBand defines memory protection mechanisms, where a remote memory access requires a special key (R_Key). The R_Key is negotiated between the peers and is validated at the target's system HCA card. In case of illegal key the packet is dropped. The R_Key requirement is built into silicon and driver code and cannot be disabled even when attacker compromises root/admin/superuser account on one or multiple servers.

Standard Layer 2 Attacks and Holes in Ethernet and How InfiniBand handles those

Standard Level 2 Attacks

- MAC Flooding Attacks - The MAC flooding attacks are targeted denial-of-service attacks designed to get the switch to fail open.
- Tagging Attacks - Attempt to get the switch to forward frames from one VLAN to another.
- ARP Attacks - ARP attacks involve using a known MAC and IP address of a host on a remote VLAN to get the switch to forward packets
- Proxy Attack - Hosts configured for standard IP forwarding will forward packets sent from one subnet to the same subnet.
- VLAN Hopping - Using Spanning Tree Protocol (STP)

MAC Flooding is not possible in InfiniBand. Since the forwarding tables are defined explicitly, there is no learn process like in Ethernet, there is a linear forwarding table. In InfiniBand the Tagging is done at the hardware, a node cannot determine its partition (P_Key) but just the index. Also, the partitioning is defined on all the switch ports, from the switch port in the first hop which eliminates the possibility of that port to send a tag which is not allowed for it.

Hopping is also not possible since there is no spanning tree protocol in InfiniBand, but rather a manager configuring the partitioning table uses the M_Key. A trap will be immediately triggered when violation attempt is detected.

Multicast/UDP: Mellanox Technologies' IP emulation over InfiniBand, which also provides the IP addressing, works just like any other Ethernet driver. It sends packets in unreliable datagrams using multicast messages and the layer 2 forwarding is implemented differently (using linear forwarding tables). Just like in Ethernet, the differences are that the MAC is derived from the Port GUID (and channel number).

To resolve IP-MAC relationship we use ARP packets like in Ethernet

So in the layer 3 translation mechanisms are slightly more secure then Ethernet, because the MAC's cannot be compromised (as they are derived from InfiniBand GUID which is burned into the HCA silicon) and InfiniBand provides stronger VLAN partitioning mechanisms.

TCP: When a reliable connection is established, it uses dedicated hardware queues. Each packet sent is signed by the HW address, queue number, relevant sequence numbers, and is summed by two CRC's. If any of that information is incorrect, the packet is simply dropped. It is impossible for an attacker to get a packet out that will match the ingress filtering criteria, and will be a fake for the following reasons:

1. Attacker needs correct sequence numbers (unless attacker already compromised one of the nodes involved in the communication, there is no way to get the sequence number from any other device on the fabric).
2. Even if a attacker managed to compromise one node that is having an authorized unencrypted conversation with another node (using rsh for example for a remote root session), compromising the other node via the InfiniBand fabric is impossible. While the attacker can get at sequence numbers on the sender, they cannot generate the next packet with correct sequence number and CRC's in time before the REAL next packet is transmitted. This is because the InfiniBand is a very fast low-latency fabric and because the InfiniBand receiver HCA drops all out-of-order packets as that is an error condition as far as the InfiniBand receiver is concerned.

The above is trivial to do in Ethernet and would grant an attacker root shell on remote machine, as out of order packet delivery is acceptable.

Also, the famous 'Kevin Mitnick' attack would be impossible in InfiniBand as LID impersonation is impossible because of the static forwarding tables in InfiniBand switches. An attacker cannot launch a denial-of-service against a server and try to assume the identity of that server that has been forced off the fabric by the denial-of-service attack – LID's are assigned by the subnet manager and not the node itself. A node does not determine what the LID should be. Since LID's are dynamically allocated by the fabric when new devices join the fabric, this provides a very nice security layer.

Summary

InfiniBand architecture defines an infrastructure that does not only have superior performance and latency, but provides better data center security than standard high-performance networks.

Security is usually categorized into confidentiality, access control, authentication, and data integrity. InfiniBand addresses those issues in a way that provides a robust security mechanisms to organizations such as ISE.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com