Vulnerability Assessment of InfiniBand Networking

Daryl Schmitt, Scott Graham, Patrick Sweeney, Robert Mills

Chapter 10

VULNERABILITY ASSESSMENT OF INFINIBAND NETWORKING

Daryl Schmitt, Scott Graham, Patrick Sweeney and Robert Mills

Abstract InfiniBand is an input/output interconnect technology for high performance computing clusters – it is employed in more than one-quarter of the world's 500 fastest computer systems. Although InfiniBand was created to provide extremely low network latency with high quality of service, the cyber security aspects of InfiniBand have yet to be investigated thoroughly. The InfiniBand architecture was designed as a data center technology that is logically separated from the Internet, so defensive mechanisms such as packet encryption were not implemented. The security community does not appear to have taken an interest in InfiniBand, but this is likely to change as attackers branch out from traditional computing devices. This chapter discusses the security implications of InfiniBand features and presents a technical cyber vulnerability assessment.

Keywords: InfiniBand, networking, vulnerability assessment

1. Introduction

The cyber threat landscape is becoming more diverse as attackers target new types of networks, devices and applications. According to Symantec's 2018 Internet Security Threat Report, the number of new mobile malware variants in 2017 increased by 54% over the number in 2016 [27]. Much more alarming was the 600% increase in attacks against Internet of Things (IoT) devices. It is safe to assume that state-sponsored cyber groups are building capabilities against networks designated by the U.S. Department of Homeland Security as part of the national critical infrastructure.

Information technology (IT) professionals and cyber defenders alike rely on signature-based detection methods to provide alerts about anomalous and potentially malicious activities in networks, but this approach cedes the initiative to the attacker and relegates the defender to a reactive position. Symantec's findings suggest the need for more proactive measures throughout the com-

puting industry and security community. Cyber security experts must explore and evaluate computing equipment in novel ways, especially from the outsider's perspective.

It is unreasonable to expect system engineers and programmers to compete with elite computer hackers, especially those with the backing of nation-states or large criminal organizations. As a result, much of the onus falls on the research community to investigate the cyber hardening and resilience of computing systems that have not been evaluated. Examples include mobile and Internet of Things devices, industrial control systems and networks, and embedded devices that communicate over vehicular networks. Such non-traditional computing devices typically do not have active traffic monitoring in place, much less security professionals to examine logs and alerts. These systems were not designed with cyber security in mind; instead, they were built for user convenience, durability (availability of services) and profitability. Despite these challenges, even small amounts of cyber hardening can greatly increase the costs to attackers and reduce the threats of cyber attacks to critical infrastructure assets.

This chapter focuses on InfiniBand, an advanced input/output interconnect technology used in high-performance computing (HPC). InfiniBand equipment has not been subjected to thorough external security testing because it is not considered to be a likely target for hackers. However, the creators of InfiniBand did realize the need for hardening and resilience. Indeed, they created the technology to address some of the fundamental weaknesses of Ethernet.

High-speed networking hardware is very expensive and InfiniBand customers rightfully expect that their equipment will not be easily compromised by cyber attacks. According to the November 2018 update to the TOP500 list, InfiniBand equipment powers 27% of the 500 most powerful computer systems in the world, but accounts for 37.4% of the total computing performance [28]. As a result, InfiniBand manufacturers have a lot to lose should a newsworthy cyber attack occur on an InfiniBand network.

The desire to put such concerns to rest is evident in a Mellanox white paper titled "Security in Mellanox Technologies InfiniBand Fabrics" [12]. The paper discusses a security review of InfiniBand protocols and highlights certain Mellanox product offerings. A vendor white paper is likely biased; therefore, the vulnerability assessment described here provides an independent and alternative viewpoint. Furthermore, the assessment has a wider scope than the Mellanox effort, which mainly focuses on the protocol, but not much on other aspects of InfiniBand networking. This chapter describes a technical cyber vulnerability assessment, an apparatus for determining the vulnerabilities that are present in a generic InfiniBand network.

2. Background

This section describes the InfiniBand architecture and the interactions between the architectural components.

Table 1. InfiniBand bandwidth specifications.

InfiniBand Standard	Line Rate	Lines	Total
Quad Data Rate (QDR)	10 Gb/s	4	40 Gb/s
Fourteen Data Rate (FDR)	14 Gb/s	4	56 Gb/s
Enhanced Data Rate (EDR)	25 Gb/s	4	100 Gb/s
High Data Rate (HDR)	50 Gb/s	4	200 Gb/s

2.1 InfiniBand

InfiniBand is a network protocol comparable to Ethernet. It is extremely lightweight and is designed to minimize latency. In the late 1990s, the computing industry recognized that it was facing a tremendous hurdle. Processor speeds were increasing according to Moore's Law, but memory latency and network bandwidth limitations were nullifying processor performance gains. This was not much of a problem for personal computers. However, high-end servers, especially those operating in clusters, needed a solution. In particular, networking (gigabit Ethernet) and storage (fibre channel) cards were pushing the bandwidth limits of motherboard buses and networking cables [11]. The InfiniBand Trade Association (IBTA) was created to come up with a solution.

More than 180 companies assembled in August 1999 to develop the InfiniBand architecture. Individuals from IBM and Intel served as co-chairs of the InfiniBand Trade Association and the steering committee members came from influential companies such as Dell, Compaq, Hewlett-Packard, Microsoft and Sun. With numerous contributors presenting differing needs, the association had to design a flexible system. The specification had to "scale down to cost-effective small server systems as well as scaling up to large, highly robust, enterprise-class facilities" and had to accommodate "new inventions and vendor differentiation" [23]. The InfiniBand Trade Association was striving to design the most secure networks while ensuring the lowest latency and highest application performance [7].

Modern InfiniBand uses individual copper or fiber cables capable of up to 200 Gb/s full bi-directional bandwidth, but the first release was primarily based on 2.5 Gb/s copper [25]. The basic copper link had four wires, a differential signaling pair for each direction. The original specifications called for several speeds: 1x, 4x or 12x copper, and 1x fiber. Table 1 shows the current InfiniBand standards, highlighting the two decades of growth.

InfiniBand was built primarily for high-performing computing clusters that are logically isolated from the open Internet. InfiniBand nodes are capable of communicating across the web, but the Internet backbone could not likely run on InfiniBand due to features such as predetermined static routing. The InfiniBand Trade Association was aware of this, when it said that the "present [router] specification does not cover the routing protocol nor the messages exchanged between routers." True routers are optional in InfiniBand networks;

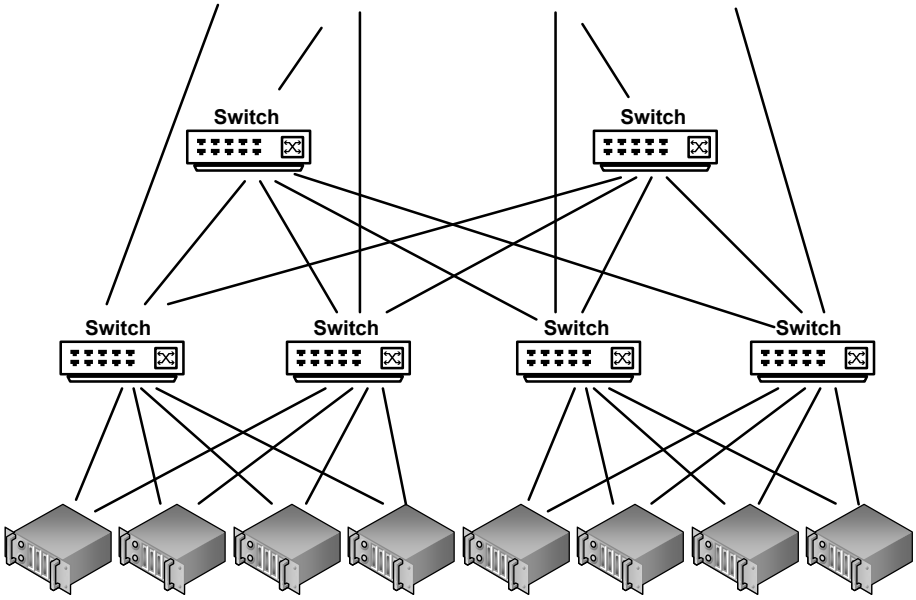

Figure 1. Switched fabric topology.

InfiniBand has been successfully utilized without routers in a production environment between two distant clusters [10]. Nonetheless, InfiniBand is fundamentally a data center technology that is not typically deployed in a network demilitarized zone unless firewalls or other similar access-controlled layers are placed in front of the InfiniBand fabric [12]. The logical segregation provides a fair amount of security, but motivated, well-resourced cyber actors would want to access the valuable data stored inside InfiniBand networks.

2.2 InfiniBand Terminology

InfiniBand is more than just a protocol – the InfiniBand Trade Association envisioned a network infrastructure around it. In fact, the association wanted to improve on the typical hierarchical structure of switches and routers used by Ethernet-based networks.

Figure 1 shows InfiniBand's switched fabric topology, which is a partial mesh that provides connection reliability to interprocessor-communications-based systems by allowing multiple paths between systems. Scalability is supported via fully hot swappable connections managed by a single unit called the subnet manager [11]. InfiniBand hosts have network cards, called host channel adapters (HCAs), which are equivalent to Ethernet network interface cards (NICs). Host channel adapters usually have at least two physical ports so that a node can be connected to two or more InfiniBand switches simultaneously. It would be impractical to create a full mesh by establishing direct

links between all devices, but switched fabrics provide a good compromise by enhancing redundancy, load balancing and routing speeds.

Ethernet networks use the dynamic Address Resolution Protocol (ARP) and routing tables to determine how and where to send traffic based on link speeds and congestion. InfiniBand switches do not make routing decisions. Instead, all the shortest paths are calculated by the subnet manager during network initialization and after configuration changes. The subnet manager then pushes forwarding tables to every device in the subnet, including all the compute nodes. Multiple subnet managers may exist, but only one acts as the master. Each host channel adapter and switch have a subnet management agent that enables communications with the subnet manager. The subnet manager sets up and maintains every link in the subnet. Network discovery is performed periodically in InfiniBand, but nodes tend not to be added or removed as often as in Ethernet networks.

Like the Transmission Control Protocol/Internet Protocol (TCP/IP) stack, the InfiniBand protocol stack is based on the seven-layer Open Systems Interconnection (OSI) model. Layer-2 addressing is done via a local identifier (LID), which is dynamically assigned by the subnet manager [10]. The local identifier is a 16-bit value, so a single subnet can support up to 65 K hosts. In contrast, media access control (MAC) addresses used in Ethernet are burned into the network cards by manufacturers. However, these addresses are easily changed via software – a simple command such as `ifconfig eth0 hw ether 02:01:02:03:04:08` accomplishes this in many Linux distributions. This is significant because InfiniBand addresses cannot be easily modified in such a manner.

The layer-3 InfiniBand locators are called global identifiers (GIDs). These are valid IPv6 addresses for the most part. The first half (i.e., 64 of the 128 bits) of each GID is called the global unique identifier (GUID). GUIDs are embedded in the host channel adapter, although there is not just one of them per network card like MAC addresses. Each host channel adapter port has its own GUID. Distinct port addressing helps enforce the static routing, as discussed above.

InfiniBand does not use sockets or virtual ports like Ethernet networks. Instead, InfiniBand connections are established between two endpoints by queue pairs (QPs). Each queue pair consists of a send queue and a receive queue, and each queue pair represents one end of a channel. If an application requires more than one connection, additional queue pairs are created. A send queue and receive queue are collectively referred to as a work queue (WQ).

Work queues put the results of completed work requests (WRs) in an associated completion queue. This includes successfully-completed work requests and unsuccessfully-completed work requests. Completion queues notify applications about ended work requests (status, opcode, size and source) [13]. The user may insert a completion notification routine to be invoked when a new entry is added to a completion queue. This nomenclature lends itself to viewing InfiniBand as a messaging service.

Table 2. InfiniBand transport services.

Class of Service	State	Response Sent
Reliable Connection (RC)	Connection-oriented	Acknowledged
Reliable Datagram (RD)	Multiplexed	Acknowledged
Unreliable Connection (UC)	Connection-oriented	Unacknowledged
Unreliable Datagram (UD)	Connectionless	Unacknowledged
Raw Datagram	Connectionless	Unacknowledged

Keeping applications informed of network activity is vital considering that InfiniBand has a major speed enhancing feature called remote direct memory access (RDMA). Remote direct memory access permits data transfers without interrupting either processor. In order to avoid involving the operating system, applications at each end of a channel must have instant access to queue pairs. This is accomplished by mapping the queue pairs directly to the virtual address space of each application. Thus, the application at each end of the connection has direct, virtual access to the channel connecting it to the application (or storage) at the other end of the channel. This concept is referred to as channel input/output [5]. Because there is no extra copying of data (e.g., to various levels of cache), InfiniBand is referred to as "zero copy" networking.

Reliable connection types send acknowledgements after every transmission. InfiniBand offers stateful and stateless connection types similar to the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP), but these are not as critical in InfiniBand. The reliable connection types in InfiniBand are called reliable connection (RC) and unreliable datagram (UD). The unreliable connection (UC), reliable datagram (RD), raw IPv6 datagram and raw Ethertype datagram transport media exist as well, although these are not as mainstream. Table 2 summarizes the transport options.

In Ethernet networks, most higher-level protocols run over TCP to guarantee 100% packet delivery. Only trivial traffic that can be resent easily (e.g., domain name queries) or traffic that requires high speeds (e.g., streaming video) runs over UDP. Conversely, unreliable datagram is extremely common in InfiniBand. InfiniBand is more efficient at avoiding congestion due to its priority-based flow control. The Ethernet pause frame "stops all traffic indiscriminately" whereas InfiniBand "strictly avoids packet loss by employing link-by-link flow control, which prevents a data packet from being sent from one end of a link if there is insufficient space to receive the packet at the other end of [the] link" [10]. More importantly, the InfiniBand receiver host channel adapter drops all out-of-order packets because it is an error condition as far as the InfiniBand receiver is concerned. This setting can be changed, but the default is to disallow out-of-order delivery.

Possible responses are either a positive acknowledge (Ack) or a negative acknowledge (Nak). A negative acknowledge is triggered under three conditions:

Table 3. Queue pair operations.

Operation	**UD**	**UC**	**RD**	**RC**
Send (with immediate)	X	X	X	X
Receive	X	X	X	X
RDMA Write (with immediate)		X	X	X
RDMA Read			X	X
Atomic: Fetch and Add			X	X
Atomic: Compare and Swap			X	X
Maximum Message Size	MTU	1 GB	1 GB	1 GB

(i) temporary receiver not ready (RNR Nak); (ii) packet sequence number error (PSN error Nak); and (iii) fatal Nak error code. The reliable connection, unreliable datagram and reliable datagram classes support remote direct memory access and require unique queue pair numbers (QPNs), meaning that no two connections can share the same queue pair numbers simultaneously. Since virtual ports do not exist, this is the primary way that connections can be distinguished from each other. Alternatively, unreliable datagram queue pairs use the same queue pair number, because they can send and receive messages to and from any other unreliable datagram queue pair using the unicast (one-to-one) or multicast (one-to-many) modes; however, only send operations are supported. In addition to remote direct memory access reads and writes, atomic extensions to the remote direct memory access operations also exist. These are essentially a combined write and read remote direct memory access, carrying the data involved as immediate data [23]. Table 3 shows a detailed listing of the available queue pair operations by connection type.

Remote direct memory access has become popular as a result of InfiniBand and it is now used in other I/O interconnects. The reason is that remote direct memory access enables high-throughput, low-latency networking with low CPU utilization. These advantages make it especially useful in massively parallel compute clusters. The Internet Wide-Area RDMA Protocol (iWARP) and RDMA over Converged Ethernet (RoCE) now bring similar capabilities to networks employing Ethernet-based software. The main difference between the two is that iWARP uses a "complex mix of layers, including DDP (direct data placement), a tweak known as MPA (marker PDU aligned) framing, and a separate RDMA Protocol (RDMAP) to deliver RDMA services over TCP/IP" whereas RoCE operates "over standard layer-2 and layer-3 Ethernet switches" [14]. RoCE's superior performance metrics compared with iWARP have made it the market frontrunner.

InfiniBand products (e.g., by Mellanox) support Ethernet by offering Internet Protocol over InfiniBand (IPoIB) and Ethernet over InfiniBand (EoIB) services. IPoIB uses an upper layer protocol (i.e., application layer) driver that enables it to encapsulate IP datagrams over an InfiniBand connected or

datagram transport service. EoIB is akin to IPoIB except that it includes the (layer-2) Ethernet header and only runs on UD. EoIB performs an "address translation from Ethernet layer-2 MAC addresses (48-bits long) to InfiniBand layer-2 addresses made of LID/GID and QPN" whereas IPoIB "exposes a 20-byte [hardware] address to the [operating system]" [15]. As a result, EoIB requires additional equipment, specifically a BridgeX gateway that connects an InfiniBand fabric to its external side (i.e., an Ethernet network segment). This may be a reason why EoIB is being phased out; this is evidenced by the fact that it is not mentioned in the latest version of the Mellanox OpenFabrics Enterprise Distribution (OFED) Linux User's Manual. The Ethernet Tunneling over IPoIB (eIPoIB) driver appears to have replaced this functionality.

2.3 InfiniBand Security Features

The InfiniBand architecture provides isolation and protection services using keys. Keys are "values assigned by an administrative entity that are used in messages in order to authenticate that the initiator of a request is an authorized requester and that the initiator has the appropriate privileges for the request being made" [22]. InfiniBand has five types of keys: (i) partition keys (P_Keys); (ii) memory keys (L_Keys and R_Keys); (iii) queue keys (Q_Keys); (v) management keys (M_Keys); and (vi) baseboard management keys (B_Keys).

A partition key designates a network partition for a channel adapter port. Each port is assigned at least one partition key by the subnet manager; these values point to entries in the port's partition key table. InfiniBand partitions are equivalent to Ethernet virtual local area networks (VLANs), so partition keys are like VLAN tags.

Memory keys are needed for remote direct memory access operations and come in the form of local keys (L_Keys) and remote keys (R_Keys). System memory is registered to provide access to local and remote channel adapters. Registration returns the keys, each of which has the associated access permission (i.e., read-only versus read/write) [3]. The same memory buffer can be registered several times, even with different permissions, and every registration results in a different set of keys.

Queue keys are preshared keys that are used in the datagram connection types (reliable datagram and unreliable datagram). During communications setup, channel adapters exchange queue keys between queue pairs. Receipt of a packet with a different queue key than the one provided to the remote queue pair indicates that the packet is not valid and is, therefore, rejected.

Management and baseboard management keys enforce control of the master subnet manager and subnet baseboard manager, respectively. The baseboard manager component communicates with nodes to provide an in-band mechanism for managing each baseboard configuration [22]. The baseboard manager's purview covers topics such as the retrieval of vital product data (e.g., serial number and manufacturing information), environmental data and adjusting power and cooling resources [24]. The baseboard manager communicates with a baseboard management agent (BMA) on each node, just as the subnet

manager does with every subnet manager agent. Every channel adapter port and every switch have a management key and a baseboard management key. These do not need to be identical across all devices, but they must match what the destination is expecting in order to verify that the source of a management packet is correct.

InfiniBand also provides integrity and quality of service (QoS). Integrity is ensured by two cyclic redundancy checksums (CRCs). As the name implies, the 16-bit variant CRC (VCRC) is recalculated at each hop. The 32-bit invariant CRC (ICRC) complements the VCRC by protecting the fields that do not change along the communications pathway. Each packet has a VCRC and an ICRC; a per-block CRC exists as well for each memory block sent in the payload. As for quality of service, packets are assigned a priority between 0 (lowest) to 15 (highest). This priority translates to a virtual lane (VL) through which the packet can transit. Each physical link can support up to sixteen virtual lanes, with VL 15 reserved for management packets.

InfiniBand management is performed in-band, using management datagrams, which are unreliable datagrams with maximum transmission units (MTUs) as low as 256 bytes. Some management datagrams are called subnet management packets, which are unique in several ways. In addition to transiting in VL 15, they are always sent and received on queue pair 0 of each port, and they can use directed routing [23]. Directed routing occurs when a subnet management packet tells a switch which ports to send it on. This is necessary when the forwarding tables have not been initialized.

InfiniBand software derives from the OFED suite from the OpenFabrics Alliance, a collaboration involving major high performance I/O vendors. Mellanox has augmented this package to create its own version of OFED. It supports both InfiniBand and Ethernet (technically, RoCE), although many network cards cannot process both interconnect types [18]. OFED includes custom diagnostic tools for ascertaining the status of the fabric. Two such utilities are `ibstat` and `ibdump`, which are analogous to the traditional Linux `ifconfig` and `tcpdump` utilities, respectively. OFED also includes open-source software called OpenSM, which provides subnet manager functionality.

InfiniBand-supported applications are written using a series of functions called "verbs." The InfiniBand architecture "contains no APIs, defined registers, etc. Instead it is specified as a collection of verbs – abstract representations of the functions that must be present, but may be implemented with any combination and organization of hardware, firmware and software" [23].

For example, the InfiniBand standard does not specify how a queue should be implemented internally in the host channel adapter hardware. Each manufacturer must provide a driver in the OFA Verbs API, whose inputs are function calls and data structures defined in detail by the API. Due to latency requirements, Mellanox programming is done in the C language according to its "RDMA-Aware Networks Programming User Manual" [13]. Example verbs/-functions are `ibv_get_device_list()`, `ibv_reg_mr()` for registering a memory region and `ibv_create_qp()` for creating a queue pair.

Table 4. Ethernet versus InfiniBand features.

Feature	**Ethernet**	**InfiniBand**
Network Card	Network interface card (NIC)	Host channel adapter (HCA)
Programming Model	Sockets	Verbs
Layer-2 Addressing	Media access control (MAC) address is statically assigned by the NIC manufacturer	Local identifier (LID) is dynamically assigned by the subnet manager
Layer-3 Addressing	Internet Protocol (IP) address	Global identifier (GID) is a 64-bit subnet ID assigned by the subnet manager plus a 64-bit global unique identifier (GUID) assigned by the HCA manufacturer
Forwarding Tables	Distributed control; each switch discovers neighbors independently	Centralized control by the subnet manager
Packet Capture	Standard operating system tools (e.g., Wireshark and `tcpdump`)	Vendor-specific tools (e.g., `ibdump` from Mellanox)

Table 4 juxtaposes some relevant Ethernet and InfiniBand features.

2.4 Cyber Vulnerability Assessment

A cyber vulnerability assessment (CVA) is an integral part of a good security program. It is the process of identifying and analyzing security vulnerabilities that might exist in a computer system. The term system usually refers to a network or enterprise, but it can be an individual device or component. Vulnerability assessments are typically conducted through "network-based or host-based methods, using automated scanning tools to conduct discovery, testing, analysis and reporting of systems and vulnerabilities. Manual techniques can also be used to identify technical, physical and governance-based vulnerabilities" [8].

A cyber vulnerability assessment has two main phases: (i) planning the vulnerability assessment; and (ii) performing the vulnerability assessment. The planning phase is extremely important, because it entails "gathering all relevant information, defining the scope of activities, defining roles and responsibilities," and more [1]. A cyber vulnerability assessment of a production network entails interviewing system administrators and reviewing appropriate policies and procedures relating to the systems being assessed. However, the experi-

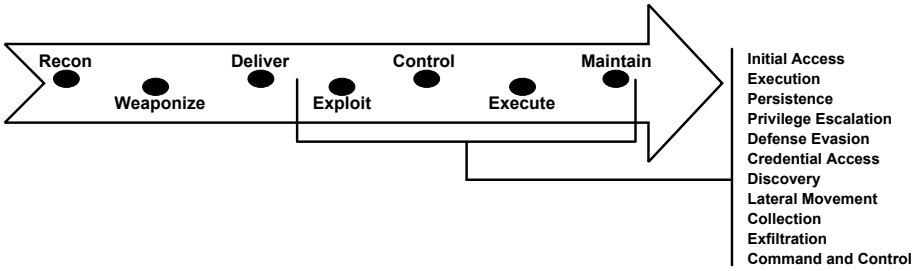

Figure 2. MITRE enterprise tactics.

mental setup comprised just a few Linux hosts connected to a single switch, so the effort is called a technical cyber vulnerability assessment.

The process of defining the scope is almost always up to the customer or network owner. This determines what entities are in play, but the execution strategy is usually up to the assessor. Many cyber experts believe in adopting an attacker's perspective by employing the "hacker methodology." This progression lists the stages of a cyber attack from reconnaissance and enumeration to exfiltrating data and covering tracks. Physical attacks are carried out in much the same manner, but each type of attack does not necessarily incorporate every stage in the progression. Some of the codified models include Lockheed Martin's Cyber Kill Chain [6], MITRE's ATT&CK Matrix [19] and the STRIDE model of Garg and Kohnfelder from Microsoft [26]. This assessment has adopted MITRE's framework because it is widely accepted by the U.S. Government cyber community.

The ATT&CK Matrix begins with gaining initial access to a device. All the adversarial actions taken prior to establishing a foothold in the network are covered by the PRE-ATT&CK Matrix. These steps are extremely important for an actual attacker, but are not relevant here. Indeed, the assumption here is that a host or other device on a generic InfiniBand network could be compromised somehow, but the method or means by which the unintended access might be acquired is tangential to the effort. The full ATT&CK Matrix covers techniques spanning Windows, Macintosh and Linux platforms. Many techniques are operating system dependent. InfiniBand is supported by newer Windows distributions, but the focus here is on Linux-style attacks. No cyber attack model explicitly covers InfiniBand, so the effort sought to discover and document specific techniques using the ATT&CK tactics as guidelines.

The eleven ATT&CK Matrix tactic categories are: (i) initial access; (ii) execution; (iii) persistence; (iv) privilege escalation; (v) defense evasion; (vi) credential access; (vii) discovery; (viii) lateral movement; (ix) collection; (x) exfiltration; and (xi) command and control [19] (Figure 2).

These functions are typically performed in the specified order, although attackers have their own tradecraft and preferences. The available time on target and required stealth also influence the sequence of events. Execution is the

means by which cyber effects are produced. Common options include a command line interface, a graphical user interface, a script or a compiled binary. Persistence enables an attacker to quickly and/or easily regain access to a system should the connection be severed. Privilege escalation involves increasing the levels of access to files, directories and programs. Ideally, an attacker would have full administrative rights such as being able to modify, add or delete anything on the filesystem. In Linux systems, the default administrator is the `root` account.

Defense evasion involves bypassing security measures (e.g., anti-virus software and firewalls) and avoiding detection. Credential access is the process of harvesting usernames, passwords, personal identification numbers, and even cryptographic keys. Discovery involves gaining information about the other systems in the internal network, after which a decision may be made to pivot to another system (lateral movement). Collection is the assembling and staging of the victim's data so that it can be exfiltrated to a location of the attacker's choosing. Lastly, command and control is how an attacker communicates with his or her malicious beacons and implants.

2.5 InfiniBand Security Research

Warren [29] has presented a GUID spoofing attack that altered the values in firmware. This is a significant contribution, but with limited realism because the victim machine (to which the GUID belonged) was taken offline prior to the attack. An attacker who compromises a single host would not be able to shut down another host without first gaining access to it, which negates the benefit of spoofing its address. (The exception might be launching a successful denial-of-service attack.) Nonetheless, this precaution led to a more straightforward proof-of-concept experiment. In Ethernet networks, duplicate MAC addresses in the same subnet can cause network instability. It is unclear how the subnet manager would react to a GUID change in a live network because GUIDs are not supposed to change, unlike LIDs or even GIDs [29].

In a white paper, Mellanox [12] asserts that the InfiniBand architecture "targets one of the main concerns in such environments [a data center LAN] which is security, and has many built-in mandatory features that enable much better isolation and security than current networks and other cluster interconnects." The paper emphasizes that InfiniBand is a layer-2 protocol much like Ethernet, so almost all layer-3 through layer-7 security mechanisms work the same way with InfiniBand. Switch administration is done out-of-band via management ports as opposed to many switches that can be configured remotely. The switches support RADIUS authentication, although it uses MD5 hashes that are now considered to be insecure. Mellanox [12] also contends that hardware-based features such as packet construction and GUID addressing significantly improve security by preventing software applications from gaining control over them and maliciously changing the attributes. It also claims that standard layer-2 attacks such as MAC floods, gratuitous ARP and VLAN hopping are not possible in InfiniBand.

In contrast, Lee et al. [9] believe that the InfiniBand architecture specification omits security, resulting in security vulnerabilities that could be exploited with moderate effort. Lee and colleagues did not orchestrate attacks, but base their arguments on the lack of encryption in InfiniBand. They are concerned that the keys used for authentication and management are sent over the wire in plaintext. The keys could be easily captured by a traffic sniffer and then spoofed to achieve powerful effects. They infer that, having infiltrated an InfiniBand network, a hacker could abuse its extensive computational power and massive storage capacity of the cluster in "another attack and as a repository for illegal content." Although their focus was on data confidentiality, Lee and colleagues proposed a security enhancement to protect against denial-of-service attacks. They constructed a simulation testbed with a stateful partition enforcement mechanism in switches using trap messages; the security mechanism filtered packets with invalid partition keys.

3. Methodology

A test network was created to perform the cyber vulnerability assessment of InfiniBand. Generic equipment and software were employed because the intent was to investigate potential vulnerabilities in core InfiniBand equipment and software, not custom InfiniBand-supported applications. Mellanox products were chosen because it is the largest InfiniBand vendor.

3.1 Equipmental Setup

A minimal network was constructed for the technical cyber vulnerability assessment. It comprised three desktop computers and a Mellanox SX6012 switch. The switch had twelve ports, each capable of 56 Gb/s full bi-directional bandwidth. It also had Ethernet, RS-232 and mini-USB management ports for out-of-band maintenance [16]. The computers were high-performance machines that were built to handle the requirements of the assessment. They had identical hardware and software.

One of the computers was assigned the role of subnet manager. As a result, it ran the OpenSM program in the master mode. In addition, this computer was connected to the switch via an RJ-45-to-DB9 serial cable. This did not affect the normal in-band traffic, although it could provide an attacker with the means to access the switch.

3.2 Approach

A cyber vulnerability assessment of InfiniBand must take a holistic approach, looking at its protocol, physical equipment (hardware), supporting software and network architecture. Hardware vulnerabilities present the most challenges to defenders because they are by far the most difficult to detect. InfiniBand switches and host channel adapters use custom application-specific integrated circuits (ASICs) that are capable of sending and receiving data at rates up to

200 Gb/s per port. The newest Mellanox switch and host channel adapter product lines are Quantum and ConnectX-5, respectively [17]. InfiniBand hardware is cost prohibitive for most businesses and individuals, so these chipsets have not been externally tested or brute forced like, for example, Intel i7 processors. In addition, assembly language or microcode is not available, requiring programmers and users to use vendor-specific tools and APIs. Thus, if hardware vulnerabilities or backdoors were to exist, they would be nearly impossible to discover without insider knowledge. Reverse engineering a microchip begins with an expensive and time-consuming process called delidding, which progressively strips layers off the chip. Images are taken of each chip cross-section using a scanning electron microscope [4]. With more than a billion transistors on a modern ASIC chip, performing these tasks and then analyzing the images is impractical. In any case, a more likely scenario is supply chain tampering rather than a manufacturing or design flaw.

The National Institute of Standards and Technology lists cyber supply chain risks as the "insertion of counterfeits, unauthorized production, tampering, theft, insertion of malicious software and hardware, as well as poor manufacturing and development practices in the cyber supply chain" [20]. In the case of InfiniBand products, this would be an extremely sophisticated attack, probably requiring nation-state support. Malicious variants of integrated circuits could be produced using an embedded rootkit or logic bomb. These could then be substituted for the original chips while the devices are in transit from the manufacturer to the customer. The hope would be that these compromised devices would find their way into facilities or in networks that would otherwise be out of reach of the attacker. Testing for hardware vulnerabilities is such a complex operation that it will not be discussed further in this chapter.

The key component of InfiniBand networking to be evaluated is its architecture. The switched fabric topology is not impervious to attack, but it does have advantages over traditional switched networks. Having more than one physical connection to the rest of the subnet provides redundancy and resilience. If one link were to go down, the endpoint should still be able to communicate through its other channel adapter port. Furthermore, man-in-the-middle attacks are more difficult and generally less successful when potential routes between two endpoints do not share at least one common intermediate node. Valuable intelligence can be gathered when a cyber actor sniffs traffic from a switched port analyzer (SPAN) port on an Ethernet switch; all (or selected) traffic traversing the switch is mirrored on a different port to the host with the listener. InfiniBand nodes are not always connected to a single switch, so in theory only a portion of the packets headed to or from a specific node would transit through the compromised switch.

Switched fabrics also cut down the attack surface by taking away the ability of switches to perform dynamic routing. This is not really a security enhancement of the architecture, but the instantiation by InfiniBand. ARP cache poisoning and routing table overflows are examples of attacks that are not pos-

Table 5. Switched fabric versus shared bus architecture.

Feature	**Switched Fabric**	**Shared Bus**
Topology	Switched	Shared Bus
Pin Count	Low	High
Number of End Points	Many	Few
Maximum Signal Length	Kilometers	Inches
Reliability	Yes	No
Scalable	Yes	No
Fault Tolerant	Yes	No

sible because of InfiniBand's predetermined routes. Table 5 summarizes some of the advantages of switched fabrics [11].

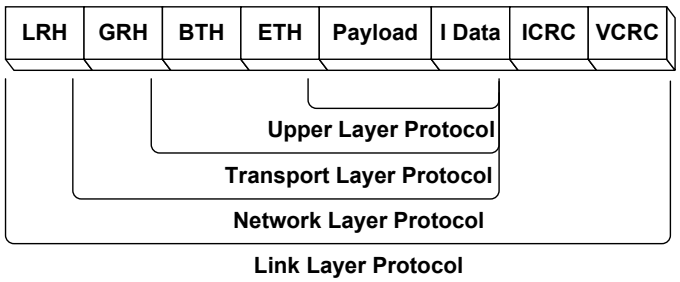

Figure 3. Data packet format.

The InfiniBand protocol follows the OSI model as seen in Figure 3. The local route header (LRH) corresponds to layer 2, the global route header (GRH) layer 3, and the base transport header (BTH) and extended transport header (ETH) comprise layer 4. The LIDs, and service level and virtual lane information are in the local route header. The global route header has the IP version and GIDs, but is omitted entirely during local (within subnet) transmissions. The packet sequence number as well as the queue, partition and memory keys are in the layer-4 headers. The fields in the extended transport header differ based on the base transport header operation or the next header of the local route header. The invariant and variant cyclic redundancy checks (ICRC and VCRC) are the checksums for bits that do not change during the transmission and that are recalculated at each hop, respectively. Breaking the checksum up into two parts makes the packets slightly harder to spoof. However, this design feature was intended to decrease the transmission delay time by limiting the work done by switches.

An obvious security concern with the InfiniBand protocol is the omission of encryption at the link level. Encrypting the payload and possibly some of the metadata contained in the higher levels of encapsulation (protocol data unit

headers) can significantly improve data confidentiality. The Ethernet stack offers encryption down to layer 3 via Internet Protocol Security (IPsec), but other common protocols such as Transport Layer Security (TLS) and Secure Shell (SSH) operate at the application layer. The InfiniBand Trade Association chose not to implement encryption because it is computationally expensive and increases latency. Technically, SSH is still available using IP-over-InfiniBand, but even this would not protect the keys.

Another negative, albeit necessary, feature is forced routing. This could enable an attacker to ignore forwarding tables and send a packet along any pathway. Positive security attributes include having virtual lanes, keys (even cleartext ones) and unique queue pair numbers.

Lastly, the InfiniBand supporting software must be examined. The open-source portions of the OFED suite can be modified and recompiled relatively easily to create a new cyber weapon. OpenSM is one such application that is susceptible to exploitation, along with many Linux shell scripts in the filesystem. InfiniBand diagnostic tools comprise the majority of the OFED binaries. Binaries could be overwritten, fuzzed for input validation vulnerabilities and/or brute forced by testing all the command line and graphical options.

3.3 Cyber Attacks

This section discusses the types of cyber attacks that were attempted. Some of the attacks are feasible on Ethernet networks, so the intent was to launch equivalent attacks on InfiniBand. The vectors were selected based on the authors' experience and research, and using the ATT&CK Matrix as a guide. All eleven tactic categories in the matrix do not pertain to InfiniBand. In particular, initial access, persistence, privilege escalation and defense evasion involve methods that are specific to the operating systems being used.

Execution. Security researchers and the hacking community have created many cyber tools for Ethernet networks. Very few, if any, of these could be applied directly to InfiniBand due to hardware packet crafting, lack of virtual ports, etc., without activating IPoIB or EoIB. Using these protocols is a legitimate technique, but this work does not consider InfiniBand as "running in the Ethernet mode."

- **OFED Diagnostic Tools:** The diagnostic utilities in the OFED suite can help debug the connectivity and status of InfiniBand devices in a fabric. Due to the lack of custom cyber security tools in InfiniBand, these utilities could serve as building blocks for cyber weapons, enabling an attacker to manipulate settings and network traffic. Running standard operating system commands and using the available diagnostic tools are much stealthier techniques than transferring and executing non-native files to an InfiniBand environment. OFED tool usage should not set off any alarms nor should it put the attacker's code at risk of being quarantined or captured.

All the OFED diagnostic tools have to be studied and tested. Individual packet captures have to be taken for each tool to understand the network traffic generated during its execution. All possible combinations of parameters cannot be executed and evaluated. Instead, options that appeared to have dangerous ramifications were chosen and tested (e.g., `ibping` with the flood option).

- **RDMA Programming:** RDMA programming for InfiniBand, RoCE and iWARP is accomplished via the Verbs API. Mellanox states that its architecture "permits direct user mode access to the hardware" through a "dynamically loaded library" [13]. Networking experts can program with verbs in order to customize and optimize the RDMA network or generate malicious effects.

- **Malicious Firmware Installation:** Firmware is embedded code on a hardware device. The host channel adapter and possibly the switch firmware would be of particular interest. Reprogramming an InfiniBand host channel adapter could enable an attacker to intercept incoming packets and modify outgoing packets. Firmware is chipset dependent, so a code modification would not be guaranteed to work on all InfiniBand host channel adapters. For this and other reasons, the experiments did not delve into firmware, but instead investigated how malicious firmware could be burned on a device.

Credential Access. InfiniBand does not use usernames and passwords for authentication, nor does it require access tokens or tickets like Kerberos. Usernames and passwords are operating system mechanisms meant for human users. In contrast, high-performance computing clusters usually run automated processes, and only matching source addresses and keys enable communications access between nodes. Additionally, administrative privileges are needed to run most InfiniBand tools.

- **Address Spoofing:** Firewalls and intrusion detection/prevention systems typically block traffic and generate alerts based on the source (MAC and/or IP) addresses. Modifying source addresses can enable an attacker to bypass these middleware devices.

  Spoofing can cause a destination computer to grant elevated permissions if no other authentication/authorization mechanisms are in place. Address spoofing can also enable reflected attacks because responses would be sent to the true owner of the spoofed address. Popular hacking tools such as Nmap and Scapy can be used as follows:

```
nmap -S $IP_Address
ifconfig eth0 hw ether $MAC_Address
nmap -spoof-mac $MAC_Address
```

Address spoofing is an attack on data confidentiality because it can enable an unintended user or computer to gain access to resources that would otherwise be denied. The experiments attempted to duplicate the GUID spoofing accomplished by Warren in 2012 [29]. LID spoofing was investigated as well.

Discovery. A cyber actor can learn the addresses and structure of an internal network in different ways. Typically, discovery is performed passively through traffic analysis and actively through scanning. Traffic analysis utilizes programs such as `tcpdump` and Wireshark whereas scanning uses tools such as `traceroute`, Nmap and Solarwinds.

- **Network Traffic Sniffing:** Traffic sniffing can give an attacker valuable situational awareness about a network. It may not be thought of as an attack in and of itself, but it is an attack on data confidentiality. Monitoring network traffic requires an attacker or attack tool to be positioned between the sender and recipient, unless the interest is in conversations involving one entity. The reason is that messages are not always sent to every device, unless the topology of the network is a shared bus or a network hub is used instead of a switch or router. Broadcast and multicast messages exist in Ethernet and InfiniBand, but these are not private messages. Therefore, switches would tend to be the preferred devices on which to perform sniffing. The experiments explored methods for sniffing InfiniBand traffic.

- **Network Mapping:** The ideal byproduct of the discovery step is a complete and accurate network map. Not every node communicates regularly, so active scanning may be necessary to identify all the connected devices. Even during relatively idle times, Ethernet networks produce a lot of noise in the form of ARP, Network Time Protocol (NTP) and Simple Network Management Protocol (SNMP) traffic. InfiniBand, on the other hand, has very little overhead. A network mapping tool, especially one with a visual display, could provide InfiniBand users and administrators with valuable situational awareness about their networks. A few OFED diagnostic tools deliver this functionality in text-only output form, so efforts focused on augmenting the tools with graphical interfaces.

Lateral Movement. Traditional pivoting is not necessary in an InfiniBand network because remote interactive logins are not used (again, excluding SSH via IPoIB). High-performance computing clusters automate work using scripts and nodes share resources. In a sense, hosts are extensions of each other due to remote direct memory access, far more so than file sharing via server message block (SMB) or the File Transfer Protocol (FTP).

- **Malicious Subnet Manager:** As discussed above, the subnet manager is an extremely powerful entity in an InfiniBand network, analogous to a Windows domain controller, albeit much more primitive. There must

be one master subnet manager, but several other nodes can run in the slave or standby mode. The backups perform (vendor-specific) polling to ensure that the master is operational and a failover to one of the backups occurs when the master is not operational [23].

The OpenSM software is open source, so an attacker could download the source code, modify and recompile it [21]. The next step would be to run the weaponized OpenSM on the compromised machine and execute a denial-of-service attack that prevents the master subnet manager from communicating, causing it to be replaced as the master.

A malicious subnet manager could affect the integrity, confidentiality and/or availability of an InfiniBand network. A weaponized version of OpenSM was not created. Instead, the experiments investigated how to cause the master subnet manager to fail remotely.

Collection. Collection is the method used by an attacker to obtain the desired information.

- **Falsified Memory Keys:** Acquiring remote direct memory access memory keys could enable an attacker to read from or write to a remote memory region.

Exfiltration. Exfiltration is loosely interpreted as creating the desired effect instead of merely exfiltrating data from a victim device or network.

- **Denial-of-Service (DoS) Attacks:** Denial-of-service attacks attempt to disable a node or program by various means, including consuming all its resources or shutting it down entirely. This is most often a means to an end, such as enabling an attacker to thwart defenses or migrate to a failover service or situation that may be more advantageous. A common denial-of-service attack is a ping flood, which saturates a victim machine or network link with ping packets to cause legitimate traffic to be dropped or severely stalled. In the experiments, the `ibping` command was executed in a (command line) terminal on one machine, on multiple terminals on one machine and on multiple machines.

4. Experimental Results and Analysis

This section discusses the outcomes of executing the attacks identified in the previous section.

4.1 Malicious Firmware Installation

How this attack is carried out depends on where the malicious firmware is located. Mellanox provides an automatic updater tool named `mlxfwmanager` for Internet-connected devices. The normal syntax is:

```
mlxfwmanager --online -u -d $device
```

Manual firmware installation is accomplished as follows:

```
mlxfwmanager_pci -i fw_file.bin
```

The `mlxfwmanager` binary could be replaced with a malicious version that downloads firmware from a location of the attacker's choosing. This would cause an authorized user to unknowingly install dangerous code. However, the attack would not work if the file hash of the correct firmware was verified from its true source.

Alternatively, the attacker could pull a copy of the malicious firmware via his or her beacon and install it manually. Covering the tracks after the attack could be problematic, although the old version of the firmware could be re-installed after the attack.

4.2 OFED Diagnostic Tools

Table 6 lists the OFED diagnostic tools. A "Yes" in the third column indicates that the tool can be used in a malicious manner, including as a part of a larger cyber weapon. Some of the OFED tools require the user to be running as `root` or as a local administrator.

The tools can run on any node and affect the entire InfiniBand fabric just the same. The result is that every node is critical. Note that Windows-based networks differentiate between local and domain accounts so a local user cannot make changes on a remote node without submitting valid credentials.

The `ibccconfig` is particularly susceptible. The manual page for this command says: "WARNING −− You should understand what you are doing before using this tool. Misuse of this tool could result in a broken fabric." InfiniBand has robust quality of service, but a key point is that every packet is assigned a service level (SL). A table that maps the service level of each port to a virtual lane determines the virtual lane on which a packet will be sent. The InfiniBand architecture specifies a dual priority weighted round robin scheme. In this scheme, each virtual lane is assigned a priority (high or low) and a weight. Within a given priority, data is transmitted from virtual lanes in approximate proportion to their assigned weights (excluding, of course, virtual lanes that have no data to be transmitted).

The `ibccconfig` tool can be abused in several ways. Mapping all the service levels to the same virtual lane will essentially eliminate all priorities. An attacker who wishes to subtly (or not) impede certain traffic could manipulate the virtual lane weights so that the targeted virtual lane is allocated a lower percentage of the total bandwidth. Another attack significantly increments the `HighPriCounter` so that all the low priority lanes rarely get their turn. The impacts of changing the sizes of maximum transmission units are minimal.

4.3 Address Spoofing

As stated above, LIDs and GUIDs/GIDs are the InfiniBand layer-2 and-3 addresses, respectively. Mellanox states that "a node does not determine what

Table 6. OFED diagnostic tools.

Commands(s)	Manual	Exploitable	Function
`ibaddr`	Yes	No	Simple address resolver
`ibdev2netdev`	No	No	Device/port status checker
`ibdiagnet`, `iblinkinfo`, `ibnetdiscover`, `ibnodes`, `ibswitches`	Yes	Yes	Fabric scanners
`ibdiagpath`, `ibtracert`	No, Yes	No	Route tracers
`ibdump`	Yes	Yes	Traffic sniffer
`ibnetsplit`	Yes	No	New subnet creator
`ibping`, `ibsysstat`	Yes	Yes	Connectivity verifiers
`ibportstate`	Yes	Yes	Port state querier/modifier
`ibqueryerrors`, `perfquery`	Yes Yes	No No	Report port errors
`ibroute`, `dump_fts`	Yes, No	Yes	Display forwarding table(s)
`ibstat`, `ibstatus`	Yes	No	`ifconfig`, `ipconfig` equivalent
`ibtopodiff`	Yes	No	Topology difference checker
`mstflint`	Yes	Yes	Firmware burner
`saquery`, `sminfo`, `smpdump`, `smpquery`, `smparquery`	Yes (x4), No	Maybe	Issue subnet admininistrator queries
`ibcacheedit`	Yes	Maybe	Edit `ibnetdiscover` output
`ibccconfig`, `ibccquery`	Yes	Yes, No	Congestion control

the LID should be .... [because] LIDs are assigned by the subnet manager and not the node itself" [12]. Several commands can be issued to print host LID(s): `ibaddr`, `ibdiagnet`, `ibnodes`, `ibstat`, `ibnetdiscover`, `ibv_devices` and `ibv_devinfo`. Since these addresses are assigned dynamically, it is likely that they are stored in a file instead of in the firmware (like GUIDs). (The `/sys/class/infiniband/ directory` was ascertained from a Mellanox script elsewhere in the filesystem.) In the case of the personal computer that was tested, these variables were `mlx5_0` and `mlx5_1`, respectively. The files permissions were initially set to read-only for everyone (`-r--r--r--`) and the owner was `root`.

The following commands were executed to grant full read/write access:

```
sudo chmod +w /sys/class/infiniband/mlx5_0/ports/1/lid
sudo chmod 777 /sys/class/infiniband/mlx5_0/ports/1/lid
```

The file permissions were changed, but the contents (`0x2` corresponding to the LID for the device/port) were not changed after the following attempts to edit them:

```
sudo gedit /sys/class/infiniband/mlx5_0/ports/1/lid
sudo echo 0x9 >> /sys/class/infiniband/mlx5_0/ports/1/lid
```

It is possible that the LID file was locked from an application or it existed in firmware.

Warren [29] altered the GUIDs in host channel adapter firmware using the following commands:

```
mstflint =d $PSID -blank_guids i /usr/share/ib_firmware/
mstflint -d $PSID -guids fake_GUID_1 fake_GUID_2 ...
```

The first command blanks the GUIDs in the firmware and the second command spoofs the GUIDs. Note that PSID (parameter-set identification) is a unique identifier for configuring the firmware.

Warren's work, which was published in 2012, involved an old version of OFED. The firmware location has been changed since. For example, the `/usr/share/ib_firmware/` directory does not exist and a system-wide search did not reveal an obvious replacement. Nevertheless, `mstflint` still supports GUID changing options.

Another method to spoof addresses involved verbs programming. InfiniBand connections are established via queue pairs. MacArthur et al. [10] state that queue pairs are "comparable to port numbers in TCP and UDP, and make it possible to multiplex many independent flows to the same destination [host channel adapter]." The `union ibv_gid *gid` output parameter of the `ibv_query_gid()` function or the return value of the `ibv_get_device_guid()` method could be changed later. Likewise, the `struct ibv_port_attr *port_attr` parameter of the `ibv_query_port()` function contains the LID of the port (`lid`) as well as the LID of the subnet manager (`sm_lid`). Unfortunately, the `ibv_create_qp()` function fails to create a queue pair when given incorrect inputs because validation occurs whenever an attempt is made to use resources.

4.4 Network Traffic Sniffing

InfiniBand host channel adapters usually have two physical ports to maximize fabric effectiveness. If a host is connected to more than one networking device, then an attacker who has command line access on one of them would most likely not be able to listen in on all the traffic to or from the host. Ethernet and InfiniBand switches are typically configured via out-of-band connections to the management ports. However, many network administrators prefer to have the ability to make changes remotely. As a result, Ethernet switches commonly allow access via Telnet (port 23, unencrypted), SSH (port 22, encrypted) or even through a web browser over HTTP/HTTPS (port 80, unencrypted; or port 443, encrypted). These services do not run on InfiniBand switches and, even if they were, there would be no way to connect to them.

Nevertheless, the `ibdump` tool enables a user to monitor host channel adapter traffic. The traditional `tcpdump` does not work because packets are crafted in hardware, not in software.

After running the following command:

```
ibdump -d $device -w  $filename.pcap
```

the packet capture (PCAP) file was loaded into Wireshark, which has an InfiniBand plugin. The plugin was able to decipher all the bits in the layer-2 to layer-4 headers that are not reserved for vendor-specific fields.

4.5 Network Mapping

Reconnaissance is an important step for defenders and attackers. Network administrators need to discover or confirm what is on their networks and attackers need to survey the network landscape to identify potential targets. Passive mapping, in the form of traffic sniffing, is not an easy option in InfiniBand, because, to be effective, it requires putting a network tap on a switch, installing a hardware splitter and altering forwarding tables to mirror traffic.

Several OFED diagnostic tools were designed to perform discovery and their use should not raise any alarms. However, there are two minor limitations with tools such as `ibdiagnet`, `iblinkinfo` and `ibnetdiscover`. First, the outputs are in text form, which is neither user friendly nor intuitive. Second, they list devices one at a time, not necessarily in the order in which they are connected to each other.

InfiniBand does not have a graphical mapping tool like ZenMap. Adapting the open-source ZenMap Python code to InfiniBand was considered. However, it was deemed to be overkill because of the lack of hosted services and virtual ports in InfiniBand and the fact that the OFED tools provide all the needed network information without multiple parameter options. A relatively simple mapping program was found on GitHub [2], which was enhanced to display hostnames, LIDs and port details in addition to the GUIDs. Figure 4 shows the output of the modified program.

4.6 Malicious Subnet Manager

The `ibportstate` tool was chosen to replace the master subnet manager. The `ibstat` command provides the subnet manager LID, so an attacker can then run a tool such as `ibnetdiscover` to determine the switch or switches to which the subnet manager's host is connected. Next, the corresponding switch port(s) are disabled to shut off all communications to the host, causing the subnet manager to fail during polling. For example, the following command instructs LID 3 to disable its first port:

```
ibportstate 3 1 disable
```

Having disabled the master, the malicious version of OpenSM that is waiting in the slave mode takes over as master assuming there are no other backups.

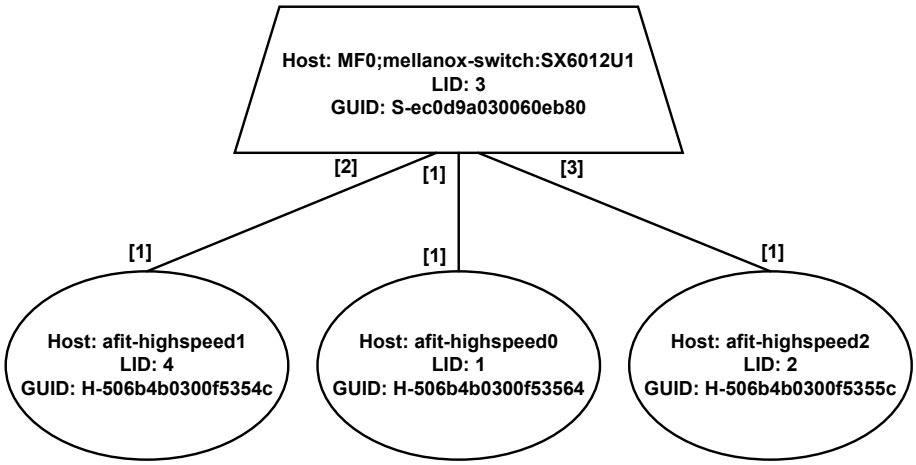

Figure 4. AFIT InfiniBand network.

This process took about ten seconds in the experiments and the switch port(s) were later enabled to restore normal traffic flow. The attack has some limitations, especially if the subnet manager is running on a switch. Disabling all the ports of a switch may be irreversible without physical access.

4.7 Denial-of-Service Attacks

There are several ways to execute denial-of-service attacks. The experiments investigated the use of the `ibping` command with the flood option. The following command sends echo request packets to and receives echo reply packets from LID 4 (victim computer) back-to-back without any delay:

```
ibping -f 4
```

During the experiments, a single instance sent 265 K packets in five seconds, or 53 K per second. Running three terminals simultaneously produced 1.6 M packets in 8.5 seconds, or 217 K per second, when the ends of the capture were removed to account for starting and stopping the commands manually. Four terminals generated 290 K packets per second. Finally, five terminals each were run on two hosts, corresponding to a total of ten simultaneous `ibping` instances. In this case, 4.5 M packets were sent in 9.8 seconds or 582 K per second. The volume seems to scale linearly in the limited sampling. No major packet loss or other harmful effects were observed, but hundreds of instances could result in distributed denial of service.

5. Conclusions

Although the InfiniBand Trade Association did not make cyber security its top priority when the InfiniBand architecture was designed, it is evident that

many InfiniBand features are inherently resistant to tampering and attack. Hardware packet crafting, predetermined routing and redundant pathways in the switched fabric topology contribute to the very low network latency and high availability desired by the high-performance computing community. InfiniBand relies somewhat on external defense mechanisms such as firewalls and other forms of network segregation. However, high performance computing clusters were intended to be located in data warehouses behind demilitarized zones and not to provide services as Internet-facing servers. This was part of the justification for not incorporating packet encryption in InfiniBand or providing support for protocols running over TCP/IP.

Nevertheless, some minor security upgrades could make InfiniBand networks more difficult to exploit without significantly degrading network performance. The subnet manager is a critical component and should have some protections in place should the master fail. A possible solution is file verification, where a node would not become the master if its OpenSM file hash values do not match those of the other standby nodes. Additionally, denial-of-service attacks, such as disabling switch ports from any host or invoking ping floods, should not be allowed even if the user is operating with elevated system privileges.

The views expressed in this chapter are those of the authors, and do not reflect the official policy or position of the U.S. Air Force, U.S. Department of Defense or U.S. Government. This document has been approved for public release, distribution unlimited (Case #88ABW-2018-6395).

References

[1] R. Boyce, Vulnerability Assessments: The Proactive Steps to Secure Your Organization, Information Security Reading Room, SANS Institute, North Bethesda, Maryland, 2001.

[2] cyberang3l, InfiniBand-Graphviz-ualization, GitHub (`github.com/cyberang3l/InfiniBand-Graphviz-ualization`), 2016.

[3] D. Deming, InfiniBand software architecture and RDMA, presented at the *Storage Developer Conference*, 2013.

[4] J. Grand, Hardware reverse engineering: Access, analyze and defeat, presented at the *Black Hat DC Workshop*, 2011.

[5] P. Grun, Introduction to InfiniBand for End Users: Industry-Standard Value and Performance for High-Performance Computing and the Enterprise, InfiniBand Trade Association, Beaverton, Oregon, 2010.

[6] E. Hutchins, M. Cloppert and R. Amin, Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains, *Proceedings of the Sixth International Conference on Information Warfare and Security*, 2011.

[7] InfiniBand Trade Association, About InfiniBand, Beaverton, Oregon (`www.InfiniBandta.org/about-InfiniBand`), 2019.

[8] Information Systems Audit and Control Association, Security Vulnerability Assessment, Rolling Meadows, Illinois (cybersecurity.isaca.org/info/cyber-aware/images/ISACA_WP_Vulnerability_Assessment_1117.pdf), 2017.

[9] M. Lee, E. Kim and M. Yousif, Security enhancement in the InfiniBand architecture, *Proceedings of the Nineteenth IEEE International Parallel and Distributed Processing Symposium*, 2005.

[10] P. MacArthur, Q. Liu, R. Russell, F. Mizero, M. Veeraraghavan and J. Dennis, An integrated tutorial on InfiniBand, verbs and MPI, *IEEE Communications Surveys and Tutorials*, vol. 19(4), pp. 2894–2926, 2017.

[11] Mellanox Technologies, Introduction to InfiniBand, White Paper, Document No. 2003WP, Santa Clara, California (www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf), 2003.

[12] Mellanox Technologies, Security in Mellanox Technologies InfiniBand Fabrics, Technical Overview, White Paper, Document No. 3861WP Rev. 1.0, Sunnyvale, California (www.mellanox.com/related-docs/whitepapers/WP_Secuirty_In_InfiniBand_Fabrics_Final.pdf), 2012.

[13] Mellanox Technologies, RDMA Aware Networks Programming User Manual, Rev. 1.7, Sunnyvale, California (www.mellanox.com/related-docs/prod_software/RDMA_Aware_Programming_user_manual.pdf), 2015.

[14] Mellanox Technologies, RoCE vs. iWARP Competitive Analysis, White Paper, Document No. 15-4514WP Rev. 2.0, Sunnyvale, California (www.mellanox.com/related-docs/whitepapers/WP_RoCE_vs_iWARP.pdf), 2017.

[15] Mellanox Technologies, Mellanox OFED for Linux User Manual, Revision 4.4, Software Version 4.4-1.0.0.0, Sunnyvale, California (www.mellanox.com/related-docs/prod_software/Mellanox_OFED_Linux_User_Manual_v4_4.pdf), 2018.

[16] Mellanox Technologies, SX6012 Switch, Product Brief, Sunnyvale, California (www.mellanox.com/related-docs/prod_ib_switch_systems/PB_SX6012.pdf), 2018.

[17] Mellanox Technologies, ConnectX-5 Single/Dual-Port Adapter Supporting 100Gb/s with VPI, Sunnyvale, California (www.mellanox.com/page/products_dyn?product_family=258&mtag=connectx_5_vpi_card), 2019.

[18] Mellanox Technologies, Mellanox OpenFabrics Enterprise Distribution for Linux (MLNX_OFED), Sunnyvale, California (www.mellanox.com/page/products_dyn?product_family=26), 2019.

[19] MITRE Corporation, ATT&CK Matrix for Enterprise, Bedford, Massachusetts (attack.mitre.org), 2019.

[20] National Institute of Standards and Technology, Cyber Supply Chain Risk Management, Gaithersburg, Maryland (csrc.nist.gov/Projects/cyber-supply-chain-risk-management), 2019.

[21] OpenFabrics Alliance, Index of /downloads/management (www.openfabrics.org/downloads/management), 2017.

[22] Oracle, Delivering Application Performance with Oracle's InfiniBand Technology: A Standards-Based Interconnect for Application Scalability and Network Consolidation, Version 2.0, Technical White Paper, Redwood Shores, California, 2012.

[23] G. Pfister, An introduction to the InfiniBand architecture, in *High Performance Mass Storage and Parallel I/O: Technologies and Applications*, R. Buyya and T. Cortes (Eds.), John Wiley and Sons, New York, pp. 617–632, 2001.

[24] QLogic, Fabric Manager User Guide, Firmware Version 6.0, D000007-007 C, Aliso Viejo, California, 2010.

[25] S. Rubenoff, HDR 200G InfiniBand: Empowering Next Generation Data Centers, *insideHPC*, February 25, 2018.

[26] A. Shostack, *Threat Modeling: Designing for Security*, John Wiley and Sons, Indianapolis, Indiana, 2014.

[27] Symantec, Internet Security Threat Report, Volume 23, Mountain View, California, 2018.

[28] TOP500, List Statistics, Sinsheim, Germany (`www.top500.org/statistics/list`), November 2018.

[29] A. Warren, InfiniBand Fabric and Userland Attacks, Information Security Reading Room, SANS Institute, North Bethesda, Maryland, 2012.