Chapter 11

SECURING INFINIBAND TRAFFIC WITH BLUEFIELD-2 DATA PROCESSING UNITS

Noah Diamond, Scott Graham and Gilbert Clark

Abstract InfiniBand is employed in applications outside of high performance computing, including in critical infrastructure assets. This requires efforts at securing InfiniBand networks with encryption and packet inspection. Unfortunately, the performance benefits realized via the use of remote direct memory access by InfiniBand are at odds with many kernel-stack-based IP datagram encryption and network monitoring technologies. As a result, it is necessary to offload these tasks to other hardware. A promising candidate is the NVIDIA Mellanox Bluefield-2 data processing unit, which combines high-performance processors, network interfaces and flexible hardware accelerators, and runs a tailored version of Linux that provides several network management applications.

This chapter characterizes the ability of Bluefield-2 data processing units to encrypt and monitor remote direct memory access traffic. The results demonstrate that the hardware accelerators of Bluefield-2 data processing units can support throughputs of nearly 86 Gbps when encrypting remote direct memory access over Converged Ethernet Version 2 traffic with Internet Protocol security (IPsec) encryption. Offloading IPsec encryption to the hardware accelerators on Bluefield-2 data processing units is a promising method for achieving confidentiality, integrity and authentication in InfiniBand networks with minimal interaction from host processors.

Keywords: InfiniBand, Bluefield-2 data processing unit, encryption

1. Introduction

InfiniBand (IB) is an industry-leading high-bandwidth, low-latency interconnect for hyperscale data centers and high performance computing clusters. Previous research has identified that native InfiniBand

Published by Springer Nature Switzerland AG 2022
J. Staggs and S. Shenoi (Eds.): Critical Infrastructure Protection XVI, IFIP AICT 666, pp. 277–300, 2022.
https://doi.org/10.1007/978-3-031-20137-0_11

plaintext key exchange is vulnerable to man-in-the-middle, denial-of-service and replay attacks [8]. Concerns about these vulnerabilities were minimal during the development of InfiniBand because data centers were assumed to be physically secure. However, interest in securing InfiniBand with encryption is growing because it is increasingly employed in applications outside high performance computing, including in critical infrastructure assets.

Encryption supports data confidentiality, integrity and authentication, but it is computationally expensive and increases the compute load on central processing units (CPUs). Data processing units (DPUs) are new programmable processors designed to assist CPUs in meeting the workloads of large data centers. Data processing units are systems on chips that incorporate a high-performance programmable processor, network interface card (NIC) and flexible hardware accelerators. They are designed to support data-center-specific tasks such as virtualization, networking, storage and security. Data processing units have the potential to join central processing units and graphics processing units as a pillar of networked computing.

Preliminary experiments have demonstrated that Bluefield-2 data processing units can encrypt Ethernet traffic by offloading Internet Protocol security (IPsec) operations to hardware accelerators. This chapter characterizes the ability of the hardware accelerators in Bluefield-2 data processing units to encrypt remote direct memory access (RDMA) traffic without placing a burden on their host CPUs.

2. Background and Related Work

This section provides an overview of InfiniBand networks, convergent technologies and related security issues.

2.1 InfiniBand Network Overview

The switched-fabric InfiniBand architecture (IBA) provides reliability, availability, performance and scalability far beyond bus-oriented input/output architectures for server input/output and inter-server communications [16]. The InfiniBand architecture is maintained by the InfiniBand Trade Association (IBTA), which is led by a steering committee that includes Broadcom, HPE, IBM, Intel, Marvell Technology, Mellanox Technologies and Microsoft [6]. As of 2022, six of the top ten supercomputers in the world use InfiniBand as their core interconnect and InfiniBand is used in thousands of data centers, high performance computing clusters and embedded applications [18]. Figure 1 provides an overview of the InfiniBand fabric.

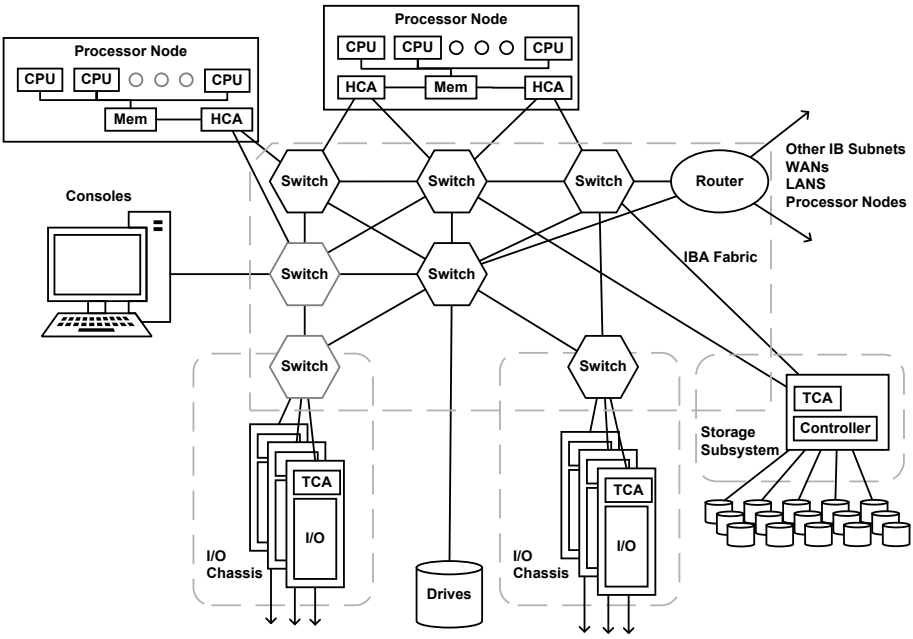

Figure 1. InfiniBand fabric overview (adapted from [5]).

Bus topologies enable multiple devices to connect to a shared physical medium, with all the devices in the same collision domain. This requires the end devices to implement collision detection and collision avoidance protocols. The requirement to support shared access and recover from collisions limits Ethernet bus network throughput to around 1 Gbps [15]. To improve performance, many networks employ star topologies that eliminate collisions via dedicated physical connections to end nodes. Dedicated links also provide flexibility in the choice of protocols implemented at each node [7].

Networks are limited by the speed of processors, input/output interfaces and network protocols. Fortunately, network device performance has steadily improved as manufacturers create chip sets with smaller feature sizes, more efficient computer architectures and faster clock speeds. As these improvements materialize, the governing bodies of network protocols must make careful decisions with respect to future protocols, considering the effects of compatibility with established network protocols. The growing demand for improved network performance and frustrations with the limitations of legacy technologies in the high performance computing domain led to the formation of the InfiniBand Trade Association to promote the use of the InfiniBand architecture.

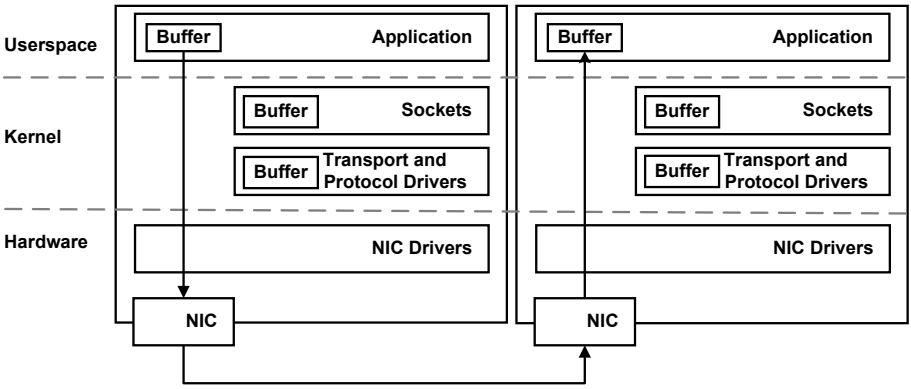

Figure 2. Remote direct memory access traffic flow (adapted from [5]).

Host processors are often responsible for virtualization, networking, storage and security applications. The computational power of the processors limits the performance of traditional data centers and high performance computing clusters. Remote direct memory access is a technology that enables data to be transferred with minimal host processor involvement. InfiniBand implements remote direct memory access in hardware to minimize intervention by host processors.

Figure 2 shows how remote direct memory access traffic moves between applications and avoids latencies incurred by buffers in the operating system kernel. Although the host processor may be responsible to authorize the transfer, the hardware-based remote direct memory access implementation bypasses the host kernel for execution.

Figure 3 shows a side-by-side comparison of the Ethernet and InfiniBand network stacks using the five-layer TCP/IP stack as a reference. Between the application and transport layers, InfiniBand uses verbs in place of Ethernet sockets. InfiniBand verbs are the basis for specifying application programming interfaces [2]. Additionally, InfiniBand has a number of transport services. The two primary types are reliable and unreliable connections, which are analogous to TCP and UDP, respectively. At the network layer, native InfiniBand employs local identifiers, global identifiers and globally unique identifiers that are analogous to and used in place of IP and MAC addresses. InfiniBand uses a subnet manager to configure local subnets. At least one subnet manager must be present in a subnet to manage all the switch and router setups, and reconfigure the subnet when a link drops or a new link appears [10].

InfiniBand is an interconnect for end nodes that includes processors, memory subsystems and input/output devices. At a minimum, subnets

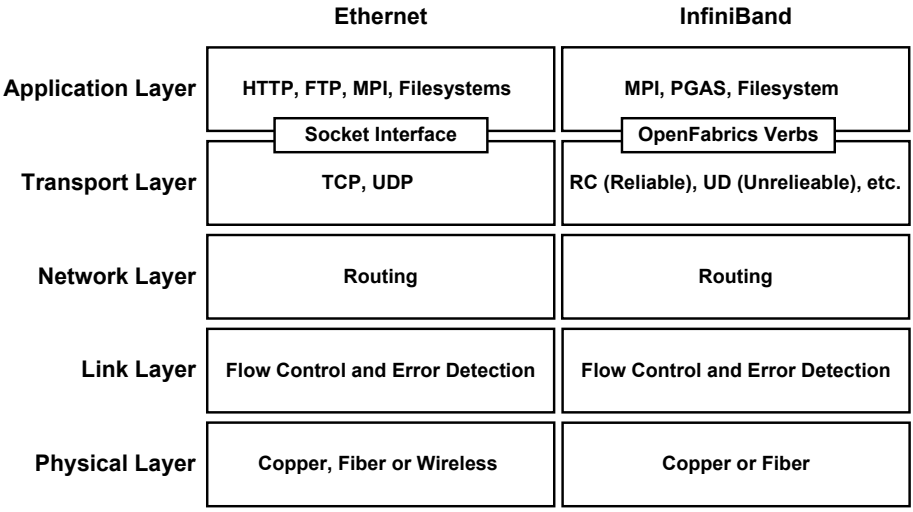

Figure 3. Ethernet and InfiniBand network stacks (adapted from [15]).

comprise two or more end nodes connected via channel adapters and managed by a subnet manager running on at least one device. InfiniBand subnets may also include a switch to take full advantage of the star topology. End nodes can be connected to multiple switches to create a switched fabric network [16].

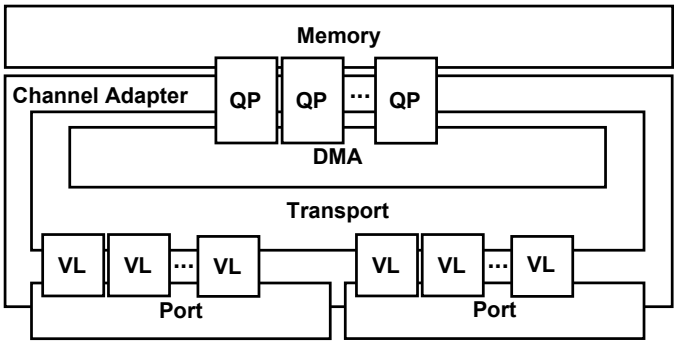

Figure 4. High-level design of a channel adapter (adapted from [15]).

Every end node in an InfiniBand network must have a channel adapter to generate and consume InfiniBand traffic. Figure 4 shows the high-level design of a channel adapter. A channel adapter typically has a few physical links that are multiplexed into independent data streams called virtual lanes (VLs). Each virtual lane is assigned a quality of service

on a packet-boundary basis. Most channel adapters support up to 16 virtual lanes per physical link [10].

InfiniBand offloads traffic control from a software client using execution queues [10]. Figure 5 illustrates the InfiniBand communications stack in which control is offloaded from a software client to a work queue (WQ) managed by InfiniBand. Each communications channel has a queue pair (QP) comprising a send queue and receive queue assigned at the corresponding end node. A client places transactions in the work queue in the form of a work queue entry (WQE) that is processed by the channel adapter. When the transaction is completed, the channel adapter notifies the client by placing an entry in the completion queue (CQ) [10]. Complete hardware implementations of the InfiniBand network stack streamline InfiniBand communications models. They also enable applications to interface with InfiniBand solely using InfiniBand verbs.

Remote direct memory access in InfiniBand requires memory partitions to be protected and registered. Memory registration involves four steps:

- **Registration Request:** The client application sends a virtual address and length to the operating system kernel.
- **Virtual to Physical Mapping:** The operating system kernel handles memory mapping and reserves regions of physical memory for remote direct memory access transactions. This adds a level of security because a process cannot map memory that it does not own.
- **Channel Adapter Cache Mapping:** The channel adapter caches the virtual to physical mapping and issues an alphanumeric handle that includes a local key and remote key.
- **Handle Return:** The handle is returned to the client application.

It is important to note that all InfiniBand keys, namely, partition-level and queue-pair-level keys, are sent in plaintext across a network when remote direct memory access transactions are initiated. This presents an inherent security risk in that an adversary can gain access to the physical memory used by the end nodes in the network [15].

2.2 Convergent Technologies

The layered abstraction of the Open Systems Interconnection network model enables new network protocols to be integrated with legacy systems. The InfiniBand architecture was developed with this in mind –

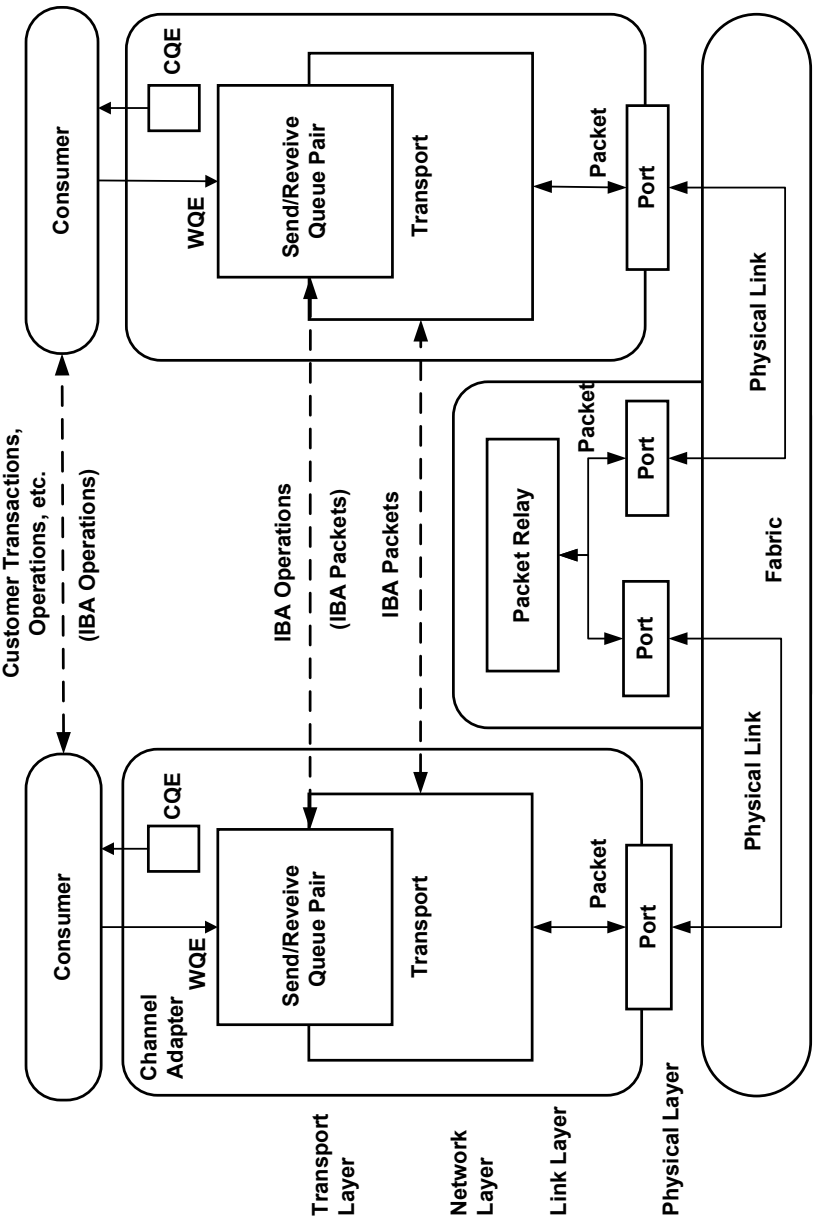

Figure 5. InfiniBand architecture transactions (adapted from [5]).

it is very flexible and backward-compatible with the conventional five-layer network stack. In fact, most channel adapters are compatible with InfiniBand and Ethernet.

Virtual Protocol Interconnect is a distributed messaging technology that supports InfiniBand and Ethernet. It enables the auto-sensing of Layer 2 protocols and may be configured to work with InfiniBand or Ethernet. This enables multi-port channel adapters to use one port for InfiniBand and the other for Ethernet. The integration of Virtual Protocol Interconnect in data centers and clusters allows InfiniBand and Ethernet networks to be hosted on the same hardware [10].

Remote direct memory access over Ethernet (RoE), remote direct memory access over Converged Ethernet (RoCE) and RoCE Version 2 (RoCEv2) are products of the convergence of the InfiniBand network and transport layers with the Ethernet link layer. RoE, which encapsulates InfiniBand packets in Ethernet frames, works natively in Ethernet environments and provides all the benefits of InfiniBand verbs. Congestion control, multicast, prioritization and fixed-bandwidth quality of service are optional in (regular) Ethernet, but are required in the native InfiniBand link layer. RoE, RoCE and RoCEv2 are often used interchangeably, but Converged Ethernet is a lossless link layer. Converged Ethernet uses all the features of the link layer of native InfiniBand [15].

RoCE does not carry an IP header so it cannot be routed across the boundaries of Ethernet Layer 2 subnets using regular IP routers. RoCEv2 is a straightforward extension of the RoCE protocol that replaces InfiniBand global route headers with IP and UDP headers. This enables RoCEv2 packets to traverse IP Level-3 routers [4]. The UDP transport header serves as a stateless encapsulation layer for the RDMA over IP protocol. These convergent communications approaches only affect packet format on the wire because remote direct memory access packets are generated and consumed below the application programming interface. Therefore, applications can operate over any form of remote direct memory access service in a completely transparent manner [4].

2.3 Security Concepts

This section briefly discusses IPsec encryption and security issues related to InfiniBand.

IPsec. Figure 6 shows the format of an IPsec datagram using an encapsulating security payload (ESP) and the tunnel mode. The IPsec datagram meets the requirements of an IPv4 datagram. In the IPsec datagram, the payload comprises an encapsulating security payload header,

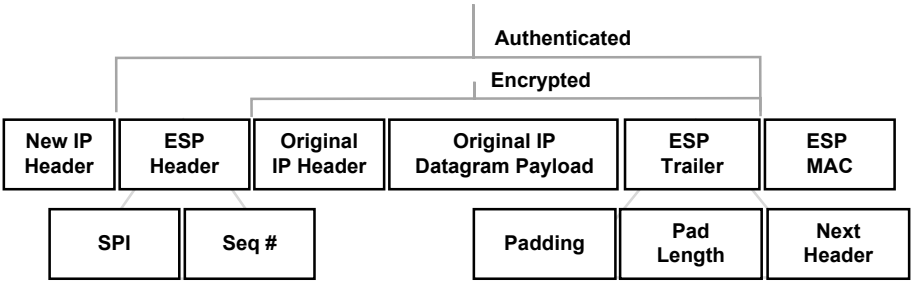

Figure 6. IPsec datagram format (adapted from [7]).

original IP datagram, encapsulating security payload trailer and authentication field.

IPsec headers and trailers create additional overhead that must be considered when configuring the maximum transfer unit (MTU) of a network interface. In total, the protocol suite can add more than 100 bytes of overhead to IP datagrams. Therefore, care must be taken to ensure that a payload combined with the IPsec headers does not exceed the maximum transfer unit of the network link. If the maximum transfer unit is exceeded, packets could be fragmented or dropped.

InfiniBand Security. Demand for high-performance, scalable and reliable networks for diverse applications has attracted considerable interest in InfiniBand networks. During the rapid development of the InfiniBand architecture, developers paid more attention to performance and cost efficiency than to security. For this reason, numerous security loopholes in the InfiniBand architecture have been identified. In fact, the design of secure clusters has recently emerged as a critical issue [17].

The confidentiality, integrity and availability triad covers the key security requirements for secure transmission across networks, and authentication is commonly added to the security triad to provide additional security:

- **Confidentiality:** Only the sender and intended recipient(s) may correctly decode or decrypt message contents.
- **Integrity:** A received message is correct and not altered.
- **Availability:** Services are accessible and available to authorized users.
- **Authentication:** The sender and recipient(s) are able to confirm the identities of each other.

Lee et al. [8, 9] have identified security vulnerabilities in the InfiniBand architecture that stem from its plaintext key management scheme. These vulnerabilities can be exploited with modest effort. Two major vulnerabilities are related to authentication.

As mentioned above, InfiniBand partitioning keys are sent in plaintext. An adversary who compromises partitioning keys would be able to transmit unauthorized InfiniBand remote direct memory access traffic. To address the problem, Lee et al. [8, 9] proposed a partition-level and queue-pair-level symmetric key management/distribution scheme. The partition-level key management scheme ensures that all communications in a partition use the same shared secret key. The queue-pair-level key management scheme guarantees confidentiality and integrity in a partition using temporary session keys between queue pairs. Simulation results have verified that the secret key management schemes harden the InfiniBand architecture with only marginal performance degradation induced by the encryption and authentication algorithms.

Several communications models such as RoCEv2 combine features of InfiniBand and high-speed Ethernet. As a result, most channel adapters and data processing units support InfiniBand as well as high-speed Ethernet. RoCEv2, unlike native InfiniBand, uses IP addresses at the network layer and is compatible with IPsec encryption. Mireles et al. [12] characterized the abilities of the NVIDIA Mellanox Innova Flex SmartNIC and Innova IPsec Ethernet Adapter to offload and encrypt RoCEv2 traffic with IPsec-enabled hardware. Mireles and colleagues found that the Innova Flex SmartNIC and Innova IPsec Ethernet Adapter were unable to offload RoCEv2 traffic to the IPsec-enabled hardware.

Hintze et al. [3] investigated the offloading and encryption of RoCEv2 traffic using the NVIDIA Mellanox Bluefield-1 data processing unit suite of IPsec-enabled hardware accelerators. They found that the Bluefield-1 data processing unit was also unable to encrypt RoCEv2 traffic in hardware.

3. Testbed Design

This research has sought to characterize the ability of Bluefield-2 data processing units to encrypt RoCEv2 traffic in hardware and to demonstrate methods for monitoring remote direct memory access traffic. To achieve these goals, a number of throughput tests were performed using NVIDIA Mellanox InfiniBand fabric utilities and network topologies.

Figure 7 shows the network topology used to characterize the performance of Bluefield-2 data processing unit hardware accelerators. The Bluefield-2 devices in the two workstations ride on sixteen lanes of pe-

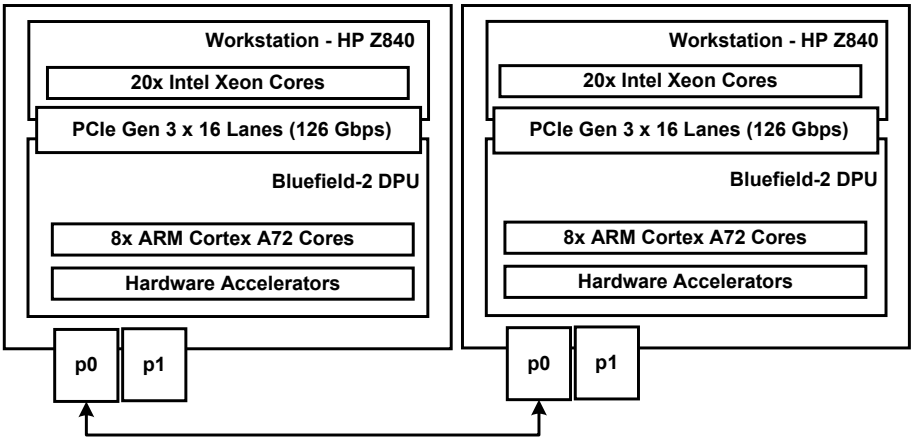

Figure 7. Performance measurement topology.

ripheral component interconnect express (PCIe) Gen 3. Each host has one data processing unit. The two Bluefield-2 data processing units are connected in tandem by a 100 Gbps fiber optic cable.

Figure 8 shows the network topology used to monitor remote direct memory access traffic and verify encryption. The network incorporates an intermediate workstation with a Bluefield-1 data processing unit. The Bluefield-1 device rides on sixteen lanes of PCIe Gen 3 and the data processing unit ports are connected by 100 Gbps fiber optic cables.

An NVIDIA Mellanox Bluefield-1 data processing unit combines a ConnectX-5 DX network adapter with an array of advanced reduced instruction set computer machines (ARM) cores and hardware accelerators. A Bluefield-1 device operates as an independent system that communicates with its host over 16 lanes of third or fourth generation PCIe, offering theoretical transfer rates of 128 Gbps or 256 Gbps, respectively. The card incorporates two multi-function 100 Gbps ports, 16 GB local DDR4 RAM, 16 Cortex A72 ARM cores and local persistent storage. Each core has a 48 KB I-cache and 32 KB D-cache. The ARM CPU also features a 1 MB L2 cache per two cores and two banks of 6 MB L3 cache with sophisticated eviction policies. The card has a tailored version of Ubuntu 18.04 provided by NVIDIA Mellanox that supports the development of new applications and the deployment of existing applications directly on the card. The applications can process and modify traffic before it is seen on the host. Thus, Bluefield-1 devices can host a variety of applications and services for networking, storage and security [11].

An NVIDIA Mellanox Bluefield-2 data processing unit employs a ConnectX-6 DX network adapter. The Bluefield-2 device communicates

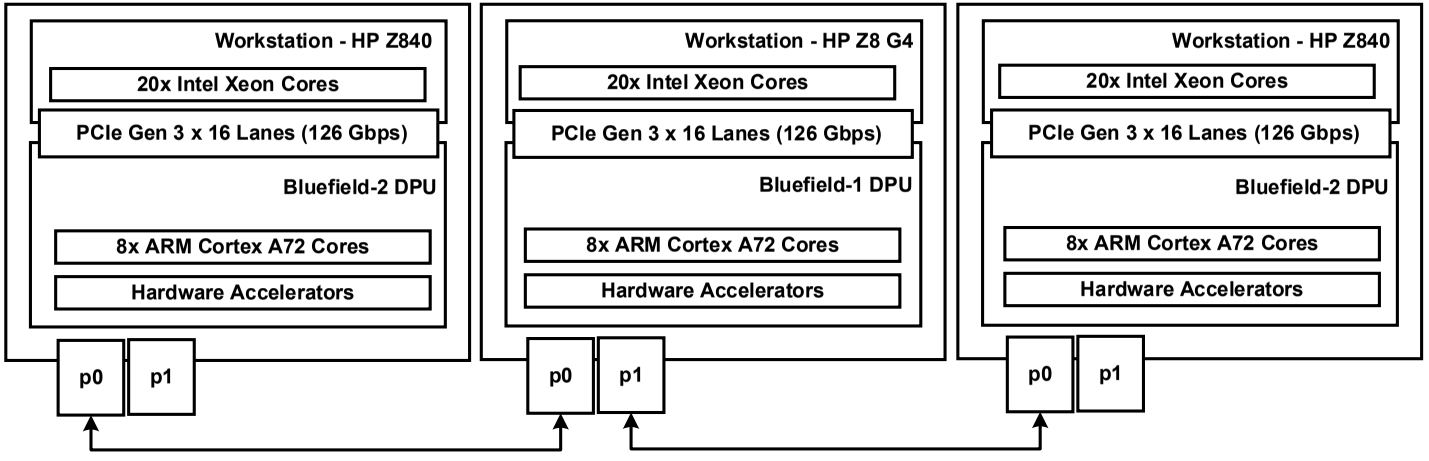

Figure 8. Network monitoring topology.

with its host over 16 lanes of third/fourth generation PCIe. The card incorporates two multi-function 100 Gbps ports, 16 GB local DDR4 RAM, eight ARM Cortex A72 pipeline processors and local persistent storage. Each core has a 48 KB I-cache and 32 KB D-cache. The ARM CPU also features a 1 MB L2 cache per two cores and a 6 MB L3 cache with multiple eviction policies. The transfer rate of the Bluefield-2 DDR4 RAM is 3,200 Tbps. The card has a tailored version of Ubuntu 20.04 provided by NVIDIA Mellanox [14].

Although desirable, the high data throughput supported by the card can quickly overwhelm its processors and memory if all the traffic is directed through the Linux kernel. To address this problem, the card offers several hardware offload and acceleration features that operate directly on network traffic without routine involvement by the ARM CPU. This enables the ARM multi-core CPU to orchestrate the hardware to perform operations on traffic at high rates instead of processing all the traffic directly.

The Bluefield-1 device was installed on an HP Z8 G4 workstation and the Bluefield-2 devices were installed on two identical HP Z840 workstations. The HP Z8 G4 and HP Z840 workstations have up to PCIe Gen 3, which is capable of supporting 128 Gbps using 16 lanes. The PCIe Gen 3 provided sufficient throughput for the research although the Bluefield-1 and Bluefield-2 devices are compatible with PCIe Gen 4. Additionally, the HP Z8 G4 and HP Z840s each have 20 Intel Xeon Cores, 256 GB RAM and a 1 TB hard drive.

The ARM subsystem of a Bluefield data processing unit must have control of the ConnectX network adapter in order to interact with the hardware accelerators. This mode of operation is called SMARTNIC (embedded) mode. In this mode, a virtual bridge is required to forward packets through the Bluefield devices correctly. The Open Virtual Switch (OVS) and Data Plane Development Kit (DPDK) `testpmd` virtual bridges were evaluated in this research.

The Ubuntu Bluefield images provided by NVIDIA Mellanox are preloaded with OVS and DPDK. OVS interfaces with the operating system kernel and DPDK sits directly above the hardware. OVS tends to be straightforward because a user can rely on the kernel to manage application resources. On the other hand, DPDK requires a user to specify and reserve resources for DPDK applications manually. DPDK provides a platform for developing lightweight, custom userspace applications that interface closely with the hardware. This research compares the abilities of OVS and `testpmd` to provide virtual bridges for monitoring remote direct memory access in man-in-the-middle positions.

Table 1. Data gathering and analysis tools.

Tool	Description
`top`	Linux command-line tool that presents a real-time system summary, including CPU utilization and Linux processes
`numactl`	Linux command-line tool that runs processes with specific non-uniform memory access scheduling and memory placement policies
InfiniBand Fabric Utilities	NVIDIA Mellanox application bundle that includes several diagnostic and performance utilities
`tcpdump`	Free command-line packet analyzer
Wireshark	Free packet analyzer that provides useful decryption features

Table 1 describes the tools used to verify network configurations and conduct throughput tests.

4. Experimental Scenarios

The research has sought to evaluate whether Bluefield-2 data processing units may be harnessed to add layers of security to the InfiniBand architecture. The experiments specifically investigated the ability of Bluefield-2 data processing units to encrypt RoCEv2 traffic and compared the performance of OVS and DPDK in forwarding RoCEv2 traffic.

NVIDIA Mellanox provides a tuning tool for Bluefield data processing units that optimizes network performance for a variety of use cases. However, this research did not employ the tuning tool because reverting the system to its original state would have required the reinstallation of the operating systems and software on the workstations and Bluefield data processing units. Although performance improvements were likely given incremental configuration changes, the results can be assumed to be representative of the impacts that encryption and hardware acceleration could have on system performance.

Throughput was employed as the response variable throughout the research. Message size was the primary independent variable for creating performance curves that characterize the performance of Bluefield-2 hardware accelerators and virtual switches.

Two experiments were conducted. The first experiment characterized the ability of Bluefield-2 data processing units to accelerate IPsec encryption of RoCEv2 traffic in hardware and the second experiment

characterized the performance of OVS and DPDK when used as virtual bridges for TCP and RoCEv2 traffic:

- **Hardware Accelerator Characterization:** Previous work has determined that the hardware accelerators on a Bluefield-2 data processing unit significantly mitigate the performance degradation incurred when Ethernet/TCP frame packets are encrypted with IPsec. In fact, the performance for encrypted traffic was identical to the performance for plaintext traffic [1]. This experiment sought to build on the results of the previous research by investigating the ability of Bluefield-2 hardware accelerators to encrypt RoCEv2 traffic.

  The experiment involved two steps. Plaintext RoCEv2 throughput tests were conducted to provide a baseline for Bluefield-2 data processing unit performance. Next, the IPsec transport mode was configured and verified on the Bluefield-2 data processing units before an additional set of RoCEv2 throughput tests was performed.

- **Virtual Bridge Performance:** The Bluefield data processing unit in the middle of the monitoring network design in Figure 8 caused significant performance degradation to RoCEv2 traffic when OVS served as the virtual bridge on the card. This experiment compared the performance of OVS and DPDK when serving as virtual bridges in the monitoring network design. Additionally, the experiment sought to demonstrate the ability of each platform to sniff traffic by running `tcpdump` with OVS and `testpmd` with DPDK. Unlike `tcpdump`, `testpmd` is a lightweight virtual bridge DPDK application, not a packet analyzer. However, it was reasonable to expect that a DPDK traffic analyzer would have similar performance to `testpmd` when sniffing traffic. The experiment was only conducted with plaintext traffic because the virtual bridges merely forwarded packets.

  The experiment involved two steps. A series of throughput tests were conducted using iPerf3 to compare network performance when OVS and DPDK `testpmd` were used with Ethernet and TCP at the link and transport layers, respectively. Next, an additional set of throughput tests was conducted to compare the network performance of RoCEv2 traffic. The drop rates of the virtual bridges were recorded during all the throughput tests.

Performance Evaluation. The Kruskal-Wallis test is a non-parametric alternative for analyzing variance in situations where the normality

assumption does not hold [13]. It employs an F-test analysis of variance that does not require normal residuals. Preliminary throughput tests revealed that network performance using the two designs in Figures 7 and 8 was non-normal. Therefore, the Kruskal-Wallis test is appropriate for analyzing the statistical significance of the data collected in the experiments.

Verification. As noted in previous research [3], Ethernet traffic analyzers cannot sniff remote direct memory access traffic because kernel bypass packets never traverse the TCP/IP stack. The man-in-the-middle network topology shown in Figure 8 addresses this issue. Bridging the network with OVS on a Bluefield-1 forced traffic through the TCP/IP kernel stack on the card. Forwarding remote direct memory access traffic with this method significantly degraded network performance, but it enabled Ethernet traffic analyzers to sniff network traffic. This helped verify that the Bluefield-2 data processing units were configured properly and actually encrypted the packets.

Verifying IPsec encryption involved the following steps:

- Configuration of the Bluefield-2 data processing units with Ethernet at the link layer and configuration of the desired encryption settings.
- Configuration of the network in the monitoring topology (Figure 8).
- Sniffing of traffic sent across the network by running `tcpdump` and saving the sniffed traffic to a PCAP file.
- Verification of IPsec encryption by uploading the PCAP file to Wireshark to decrypt the captured packets using the known encryption key.
- Configuration of the network in the performance measurement topology by connecting the Bluefield-2 data processing units in tandem (Figure 7).
- Execution of the throughput tests with the verified network configuration.

Validation. Confounding variables and uncontrolled factors can introduce noise in the experiment results. The Kruskal-Wallis analysis of variance test was applied to a full factorial design to identify the factors that have significant effects on throughput, the response variable.

Table 2. Full factorial design.

Number	Maximum Transfer Unit	RDMA Operation	Transport	Iterations
1	256	Read	DC	1,000
2	256	Read	DC	100,000
3	256	Read	RC	1,000
⋮	⋮	⋮	⋮	⋮
16	4,096	Write	RC	100,000

Table 2 shows how the factor levels were set during each trial of the full factorial design. Sixteen treatments were tested in a full factorial test of four, two level factors (2^4). Three replications of each treatment were performed to further reduce noise. In total, 48 RoCEv2 throughput tests (16 treatments $\times$ 3 replications) were performed in the screening test using the InfiniBand fabric utilities.

Only two factor levels were required for the screening test. Screening tests often work best when the factor levels have large differences. The maximum transfer units were set to 256 and 4,096 bytes because they corresponded to the minimum and maximum values supported by a Bluefield-2 data processing unit when handling RoCEv2 traffic. The maximum transfer units were evaluated because network performance is often dependent on packet size. Remote direct memory access read and write are foundational operations. Reliable connection and dynamically-connected transports, which operate similarly to TCP and UDP, respectively, were tested for the connection types. Additionally, 1,000 and 100,000 iterations were tested. Increased throughput test duration may improve the experimental results because longer tests can dilute the noise caused by systems throttling CPU clocks; many end nodes dynamically throttle clock rates to reduce power consumption. Each of these factors was configured using the command-line arguments of the InfiniBand fabric utility applications.

Applying the Kruskal-Wallis test to the results of the full factorial design determined that the maximum transfer unit, remote direct memory access operation type and iterations significantly affected the average RoCEv2 throughput. The effect of the maximum transfer unit was deemed to be significant at a 99.9% ($p = 2.2 \times 10^{-16}$) confidence interval, remote direct memory access operation type at a 99% ($p = 0.0077$) confidence interval and iterations at a 95% ($p = 0.0275$) confidence interval.

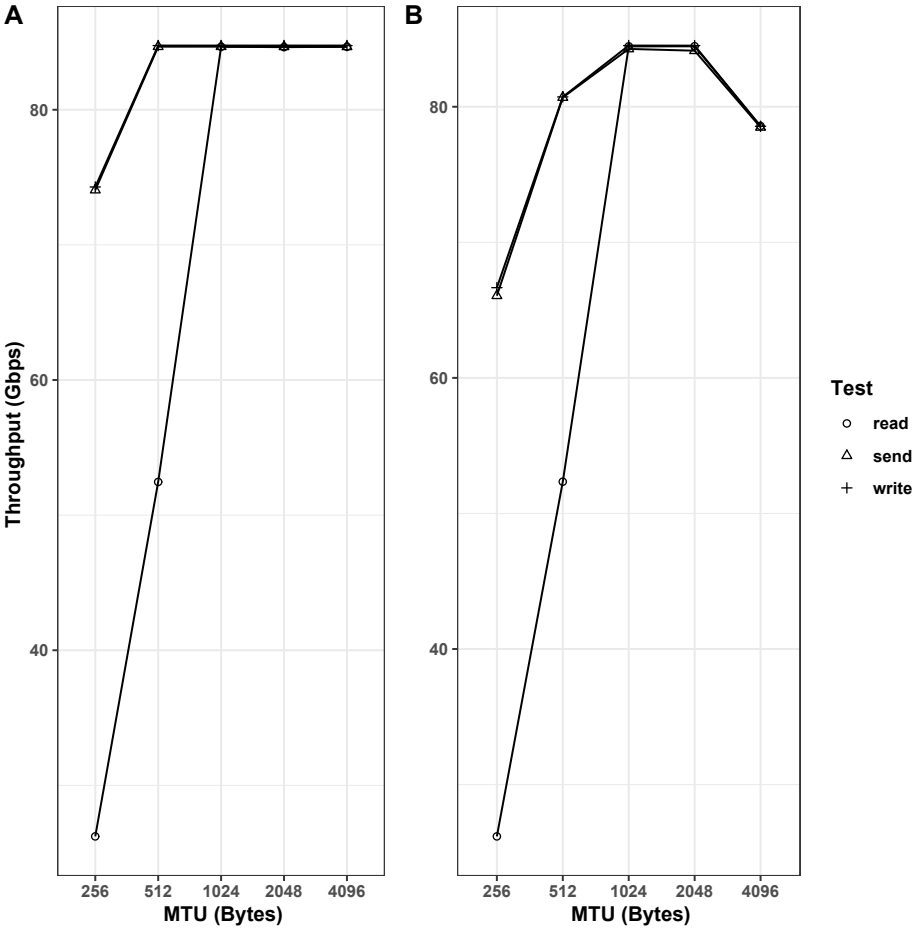

Figure 9. Hardware accelerator performance (A: plaintext, B: IPsec).

The Kruskal-Wallis test assumes that the data is independent of the run order. The experiments ensured independence from run order by randomizing the factor levels.

5. Experimental Results

This section presents the hardware accelerator characterization and virtual bridge performance results.

Hardware Accelerator Characterization. Figure 9 shows the performance curves of the Bluefield-2 hardware accelerators for plaintext and IPsec-encrypted traffic. A total of 45 throughput tests were con-

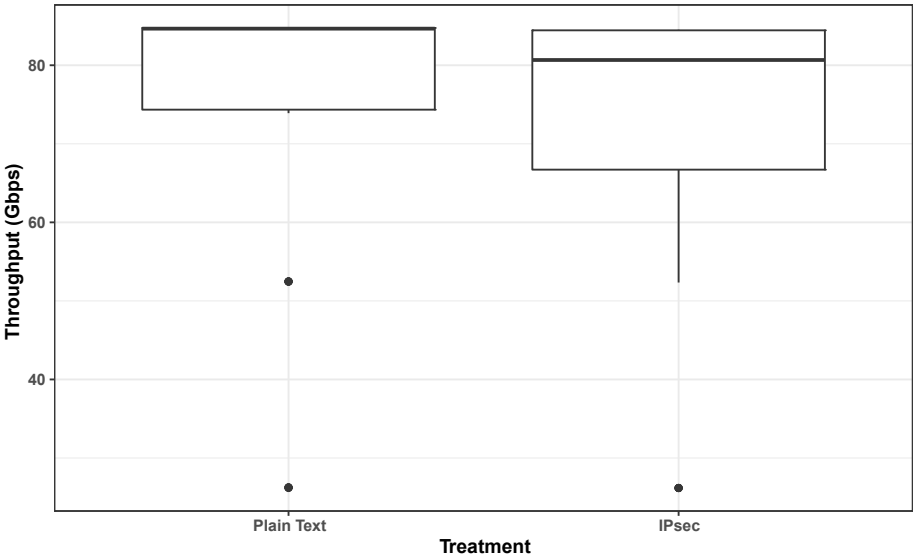

Figure 10. Comparison of plaintext and IPsec RoCEv2 performance.

ducted for each configuration. The Bluefield-2 hardware accelerators encrypted RoCEv2 traffic at a rate of nearly 86 Gbps.

Figure 10 shows that the Bluefield-2 hardware accelerators perform slightly better on average with plaintext traffic at a 99.9% ($p = 2.3 \times 10^{-9}$) confidence interval.

Virtual Bridge Performance. Virtual bridge performance was evaluated for Ethernet and RoCEv2 traffic; the monitoring capability was also evaluated:

- **Ethernet:** Figure 11 shows the performance curves of the OVS and DPDK `testpmd` virtual bridges with the network configured for Ethernet and TCP. The performance of OVS and DPDK `testpmd` reaches a maximum under 10 Gbps, but OVS performs slightly better than DPDK `testpmd` on average at a 99.9% ($p = 2.053 \times 10^{-6}$) confidence interval (Figure 12).

- **RoCEv2:** Figure 13 shows the performance curves of the OVS and DPDK `testpmd` virtual bridges with the network configured for RoCEv2. DPDK `testpmd` performs better than OVS at a 99.9% ($p = 2.2 \times 10^{-16}$) confidence interval (Figure 14). The performance of DPDK peaks around 70 Gbps. Interestingly, the performance of the remote direct memory access read operations across the DPDK

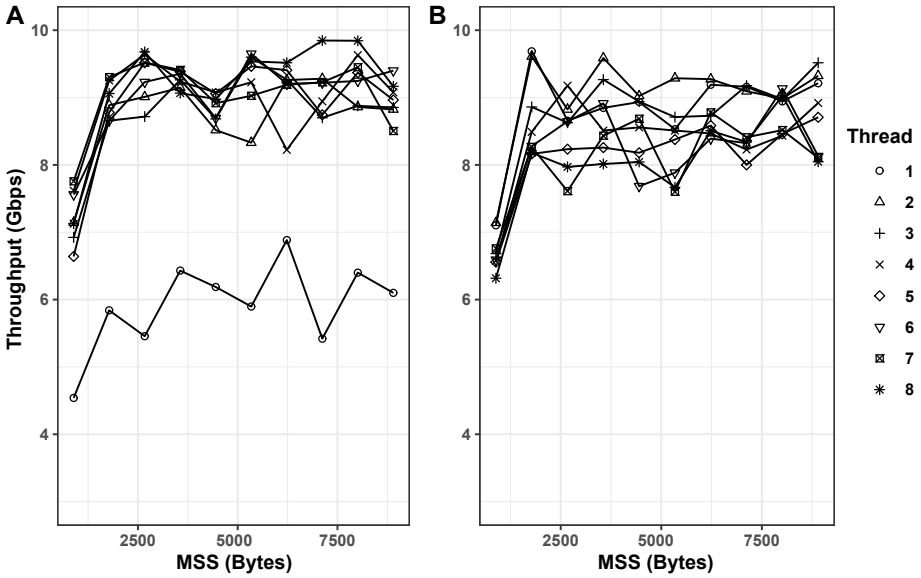

Figure 11. Qualitative Ethernet comparison (A: OVS, B: DPDK).

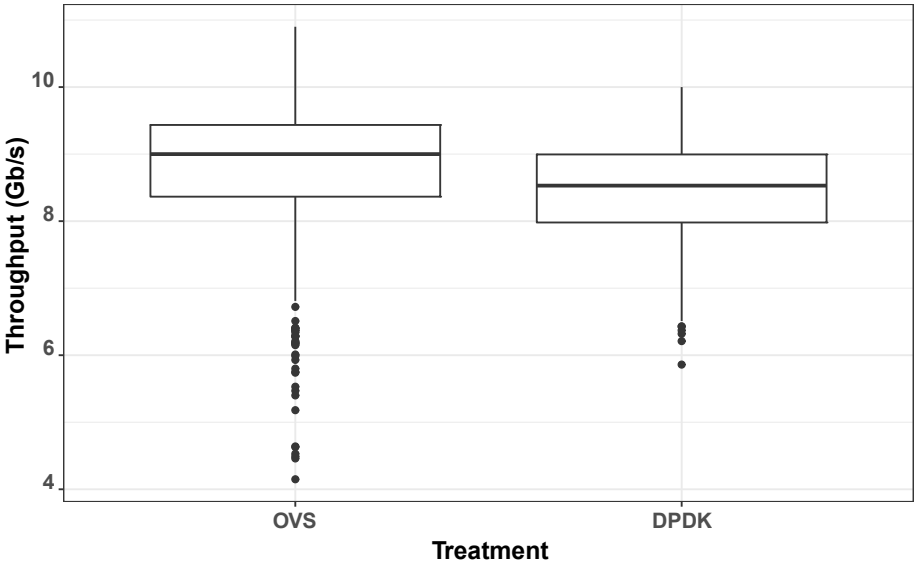

Figure 12. Statistical Ethernet comparison.

`testpmd` virtual bridge are significantly slower than the remote direct memory access write and send operations.

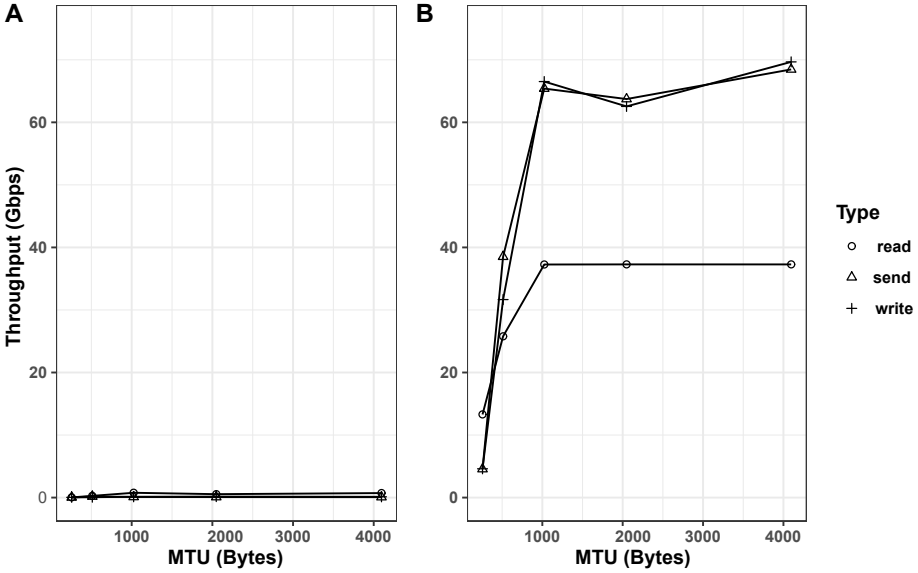

Figure 13. Qualitative RoCEv2 comparison (A: OVS, B: DPDK).

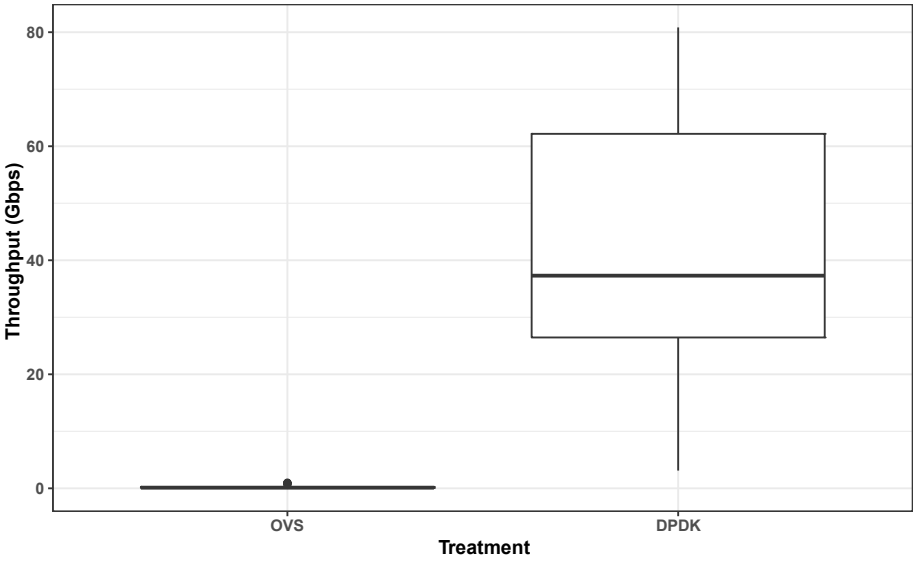

Figure 14. Statistical RoCEv2 comparison.

- **Monitoring Capability:** Table 3 shows the `tcpdump` and `testpmd` capture rates. Note that `testpmd` hardly dropped any TCP

Table 3. Bridge capture capability.

Experimental Treatment	Total Packets Transmitted	Total Packets Received	Capture Rate
`tcpdump` Ethernet	526,628,133	152,672,846	28.99%
`tcpdump` RoCEv2	24,479,800	8,775.338	35.85%
`testpmd` Ethernet	2,055,767,590	2,055,672,563	99.99%
`testpmd` RoCEv2	923,504,628	920,113,923	99.81%

and RoCEv2 packets whereas `tcpdump` dropped a significant majority of the packets sent across the network. Also, `tcpdump` performed the same when forwarding TCP and RoCEv2 traffic.

6. Conclusions

Convergent InfiniBand and Ethernet communications models such as RoCEv2 leverage the superior performance of remote direct memory access and existing TCP/IP network infrastructure. Direct memory access is a kernel bypass technology that prevents many conventional security applications from being able to sniff network traffic. However, it is imperative that this issue is addressed because these hybrid communications models are being deployed in critical infrastructure assets. Encryption and monitoring techniques such as deep packet inspection are mature and commonly-adopted practices in TCP/IP networks. The Bluefield-2 data processing unit provides a configurable platform capable of supporting a variety of security and network management applications. The Bluefield-2 data processing unit stands out from among other InfiniBand channel adapters because of its high performance, programmable ARM CPU and suite of crypto-enabled hardware accelerators. This research has investigated practical methods for securing the InfiniBand architecture by combining the computational capabilities of Bluefield-2 data processing units with conventional encryption and monitoring technologies.

The first experiment demonstrates that Bluefield-2 data processing units can support confidentiality, integrity and authentication in InfiniBand networks with minimal interaction from host CPUs. The performance of Bluefield-2 devices when encrypting RoCEv2 traffic is nearly identical to when it sends plaintext traffic, peaking at nearly 86 Gbps. This is an impressive level of performance given the computational demands of the AES-GCM cypher used by IPsec.

The second experiment demonstrates the performance benefits gained by using DPDK applications. The DPDK `testpmd` application bridges RoCEv2 traffic at nearly 70 Gbps. This is a significant improvement over the performance of OVS, which struggles to forward RoCEv2 traffic, only achieving a few Mbps of throughput.

Clearly, DPDK is able to support the high data rates created by RoCEv2. Future work will further investigate the abilities of Bluefield-2 data processing units and DPDK to encrypt and monitor network traffic. IPsec is incompatible with native InfiniBand because it does not use IP addresses. Encrypting native InfiniBand at the link layer has the potential to provide secure end-to-end communications. MACsec traditionally provides link layer encryption in Ethernet. Perhaps a similar protocol could be developed in a DPDK bare-metal application and implemented on Bluefield-2 data processing units.

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-2021-1014).

References

[1] N. Diamond, S. Graham and G. Clark, Securing InfiniBand networks with the Bluefield-2 data processing unit, *Proceedings of the Seventeenth International Conference on Cyber Warfare and Security*, pp. 459–468, 2022.

[2] 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 (`network.nvidia.com/pdf/whitepapers/Intro_to_IB_for_End_Users.pdf`), 2010.

[3] K. Hintze, S. Graham, S. Dunlap and P. Sweeney, InfiniBand network monitoring: Challenges and possibilities, in *Critical Infrastructure Protection XV*, J. Staggs and S. Shenoi (Eds.), Springer, Cham, Switzerland, pp. 187–208, 2022.

[4] InfiniBand Trade Association, Supplement to InfiniBand Architecture Specification, Volume 1, Release 1.2.1, Annex A17: RoCEv2, Beaverton, Oregon (`cw.infinibandta.org/document/dl/7781`), 2014.

[5] InfiniBand Trade Association, InfiniBand Architecture Specification, Volume 1, Release 1.6, Beaverton, Oregon (`www.infinibandta.org/ibta-specification`), 2022.

[6] InfiniBand Trade Association, InfiniBand Trade Association, Beaverton, Oregon (www.infinibandta.org, 2022.

[7] J. Kurose and K. Ross, *Computer Networking – A Top-Down Approach*, Pearson, Hoboken, New Jersey, 2017.

[8] M. Lee and E. Kim, A comprehensive framework for enhancing security in the InfiniBand architecture, *IEEE Transactions on Parallel and Distributed Systems*, vol. 18(10), pp. 1393–1406, 2007.

[9] M. Lee, E. Kim, K. Yum and M. Yousif, Instant attack stopper in the InfiniBand architecture, *Proceedings of the IEEE International Symposium on Cluster Computing and the Grid*, pp. 105–110, 2005.

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

[11] Mellanox Technologies, NVIDIA Mellanox BlueField Data Processing Unit (DPU), Data Processor Product Brief, Sunnyvale, California, 2020.

[12] L. Mireles, Implications and Limitations of Securing an InfiniBand Network, M.S. Thesis, Department of Electrical and Computer Engineering, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio, 2020.

[13] D. Montgomery, *Design and Analysis of Experiments*, John Wiley and Sons, Hoboken, New Jersey, 2019.

[14] NVIDIA Corporation, NVIDIA Bluefield-2 DPU Data Center Infrastructure on a Chip, Datasheet, Santa Clara, California (www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-2-dpu.pdf), 2022.

[15] D. Panda and S. Sur, Designing Cloud and Grid Computing Systems with InfiniBand and High-Speed Ethernet: A Tutorial, presented at the *Cluster, Cloud and Grid Workshops* (www.ics.uci.edu/~ccgrid11/files/ccgrid11-ib-hse_last.pdf), 2011.

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

[17] B. Rothenberger, K. Taranov, A. Perrig and T. Hoefler, ReDMArk: Bypassing RDMA security mechanisms, *Proceedings of the Thirtieth USENIX Security Symposium*, pp. 4277–4292, 2021.

[18] E. Strohmaier, J. Dongarra, H. Simon and M. Meuer, TOP 500 The List, Prometeus, Sinsheim, Germany (www.top500.org/statistics/list), 2022.