IOWN GLOBAL FORUM™

RDMA over Open APN PoC Reference

Classification: APPROVED REFERENCE DOCUMENT

Confidentiality: PUBLIC

Version 1.0

July 21, 2022

[RDMA/APN PoC Reference]

Legal

Contents

List of Figures

List of Tables

1. Purpose, Objectives, and Scope

1.1. Purpose

The goal of IOWN GF is to develop new fundamental technologies to improve communication, computing, and energy efficiency. Providing quantum leaps in each of these areas, the IOWN GF technologies are expected to enable new use-cases and new markets. Two areas of use-cases that IOWN GF targets are cyber-physical systems (CPS) and AI-integrated communication (AIC):

CPSs are systems that intelligently monitor or control massive numbers of subjects, such as vehicles, industrial facilities, or network infrastructure. The systems need to transfer and analyze huge amounts of captured data to generate response actions, such as alerting authorities of emergencies. Therefore, the implementation of such interactive CPSs requires communication channels with very high bandwidth, and often also very low latency.

AIC technologies enable users to interact with each other over long distance as if they were together at the same location. Use-cases of such technologies are VR live music concerts, remote and haptic control of maintenance robots, cloud video-gaming, and XR navigation. AIC technologies require communication channels with high bandwidth and low latency to create a convincing feeling of presence for the above use-cases.

However, current communication technologies are insufficient to realize the use-cases that IOWN GF envisions: current protocol stacks incur latency penalties and cause limitations regarding usable bandwidth both inside the network and inside the servers running the actual applications. This is particularly true for long-range communication over tens to thousands of kilometers.

In January 2022, the IOWN GF proposed a new technology called “RDMA-over-Open-APN”. The purpose of the PoC proposed in this document is to confirm and demonstrate that RDMA-over-Open-APN indeed removes the obstacles of long-range communication for IOWN GF use-cases. The results of such demonstrations are expected to provide viable feedback for further IOWN GF technology development and to strengthen the IOWN GF ecosystem.

1.2. Objectives

Currently, algorithms and tuning parameters for RDMA-over-Open-APN have only been derived theoretically. Accordingly, the objectives of the proposed PoC activities are:

- gather real-world experience by implementing the proposed technology
- discover best practices for tuning algorithm parameters to achieve highest performance
- determine the dependence of maximum bandwidth on distance of the RDMA-over-Open-APN technology to provide reference implementation model (RIM) designers with guidelines regarding performance to expect.

1.3. Scope

This PoC aims to evaluate the performance of RDMA over Open APN framework based on some traffic patterns with certain communication distances between sender and receiver. The scope of this PoC is depicted in Figure.1, and the details of spec. are written in Section 3. Key Requirements. RDMA is applicable for some use cases such as main-memory-to-main-memory, XPU-to-XPU. Therefore, we define step-by-step PoC plan in Table.1. Figures 2, 3, and 4 show RDMA transfer pattern in each step.

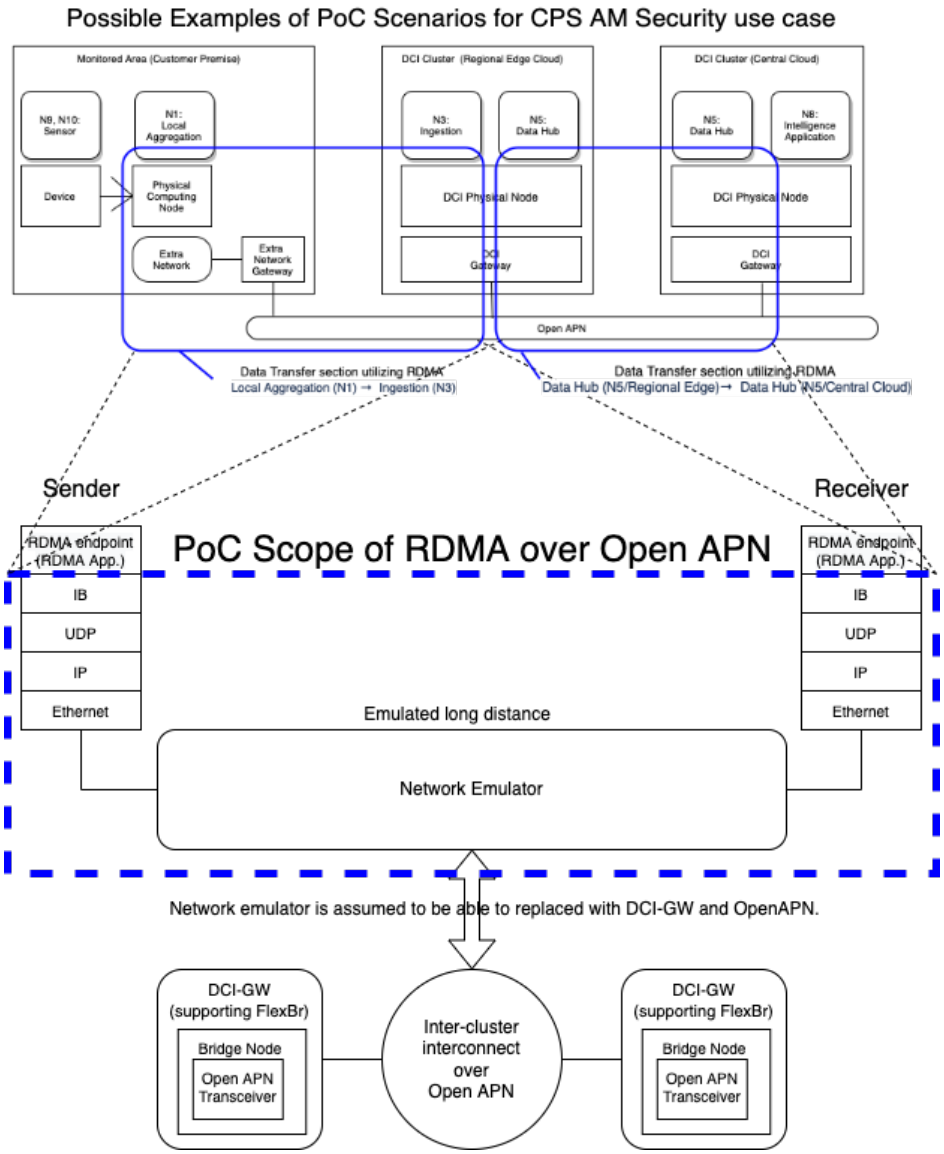


Figure 1: PoC scope of RDMA over Open APN

Table 1: Step by step procedure for RDMA over Open APN

STEP	COMMUNICATION TYPE
Step 1	main memory to main memory
Step 2	main memory to XPU XPU to main memory XPU to XPU
Step 3	main memory to NVMe (NVMe-oF Client to NVMe Device) [1]

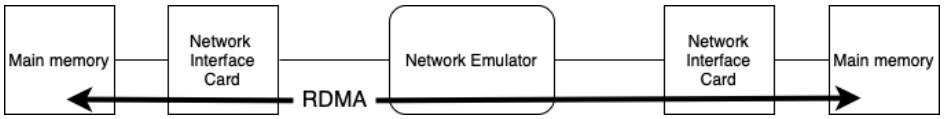


Figure 2: (Step 1) Main memory to main memory

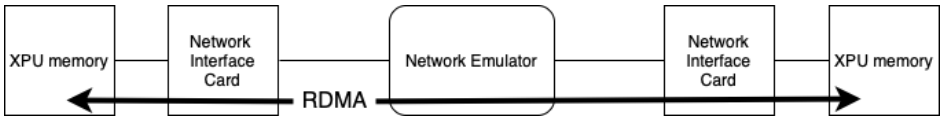
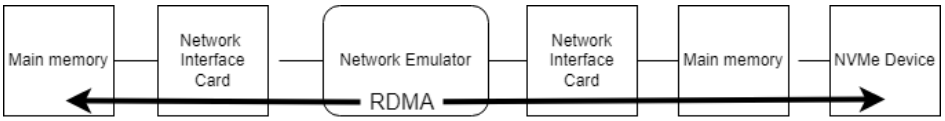


Figure 3: (Step 2) XPU to XPU



Note :
(1) An application layer communication (NVMe-oF Client, NVMe-oF Server) is conducted between main memory on the left and NVMe Device on the right.

(2) NVMe stack is over RDMA transport layer [1].

(3) On the receiver side, NVMe queue is deployed in main memory in the same side. Therefore, application layer communication is via main memory on the right side of this figure [2].

Figure 4: (Step 3) main-memory to NVMe Device

2. Reference Case: CPS Area Management Security

We will develop reference cases, making reference to the RIM for the CPS Area Management (AM) Security use case [3] as the first reference scenario. In CPS AM Security use case, it is assumed that there are two RDMA communication flows between Monitored Area and Regional Edge Cloud, and between Regional Edge Cloud and Central Cloud for high volume data traffic including video streaming from many sites (e.g., 1000 cameras per Monitored Area) written in Figure.1. The RIM for CPS AM Security use case also shows a deployment example for Japan and estimates the traffic volumes as follows:

1. Traffic volume between Monitored Area and Regional Edge Cloud
    a. Up to 67 Gbps (= 60 + 7) [3]

2. Traffic volume between Regional Edge Cloud and Central Cloud
    a. Up to about 349 Gbps (=about 3.6Gbps/Monitored Area x 96 Monitored Area) [3]

3. Key requirements

When we realize the technical feature PoC of RDMA over Open APN, key requirements of this PoC are defined in the following:

1. Traffic
    a. Pattern
        i. Burst traffic. Each application has specific traffic patterns and traffic sizes. It is difficult to emulate all of network traffic patterns and volumes. Burst traffic is one of the hardest traffic patterns to handle. Therefore, it is sufficient for measurement of transfer capacity.

    b. (RDMA) Message size (transfer size that user application requests)
        i. Message size is set to a value from 4,096 (4K) bytes to 524,288 (512K) bytes.
        ii. Note
            1. Message size is a length of buffer set for each WR (work request) in RDMA WQE (Work Queue Element). RDMA stack copies memory area as a unit of this message size of sender to the memory region of receiver.
            2. RDMA message is not equal to packet/frame size. If RDMA message size is bigger than MTU size, RDMA message is divided into multiple packets/frames.

    c. (Ethernet) MTU size
        i. 4,196 byte or more
            1. The Ethernet standard normally sets a maximum frame size as 1500 bytes excluding Ethernet header and FCS. However, if we want to get higher throughputs, it should be supported bigger frame size than normal maximum Ethernet frame size.

2. Link speed of wired network (LineSpeed)
    a. (step1,2) 100Gbps.
        i. Commercial network interface cards generally support this line speed.

    b. (Step3) 50Gbps
        i. Maximum transport on NVMe is 64Gbps.
            1. NVMe shows upper limit.

    c. (Option) 200Gbps

    d. PCI Express (PCIe) generation(gen) 4 or higher is required.
        i. The uni-directional speed of a lane in PCIe gen3 is 8 Gbps. The speed of 16-lanes in PCIe gen3 is 128 Gbps ideally, and 126.031 Gbps including encoding overhead (128b/130b).
        ii. The uni-directional speed of a lane in PCIe gen4 is 16 Gbps. The speed of 16-lanes in PCIe gen4 is 256 Gbps ideally, and 252.062 Gbps including encoding overhead (128b/130b).

3. Emulation of communication distance of APN
    a. In this PoC, it is assumed that a network emulator is inserted instead of real network infrastructure of APN so that we can focus on investigating the RDMA performance independent from APN infrastructure. A network emulator inserts delay of 0.5 msec. / 100km.

    b. (Mandatory) evaluation of 100km.

    c. (Optional) evaluation from 20km to 1534km.
        i. This covers requirement of CPS AM Security in RIM TF.

    d. Note
        i. Analyze the relation of transmission distance and maximum achievable throughput.

4. RDMA profile [4]
    a. RDMA transport protocol stack
        i. RoCE v2(UDP/IP/Ethernet)

    b. RDMA library for applications
        i. Application directly ordered memory to memory transmission by using RDMA library. Or RDMA is bridged by DMA (QDMA, P2PDMA and so on).

    c. RDMA service type
        i. RC (Reliable Connection)

    d. RDMA operation type
        i. SEND, RDMA WRITE or RDMA READ

    e. Retransmission algorithm
        i. RC Service guarantee data integrity by go back N retransmission.

    f. Queue depth configuration
        i. In the IOWN GF deliverable of Jan. of 2022 [5], the value for queue depth configuration is derived from the following equation:
            1. (RTT * LineSpeed) / MessageSize = Required QueueDepth
                a. See point 1.b.i above for the meaning of MessageSize.
                b. See point 2 above for the meaning of LineSpeed.
            2. Note
                a. This queue depth value is ideal. If you build this PoC evaluation setup and evaluate performance of data transfer, you may face performance issues because there are some overheads in data transfer. In that case, you can configure bigger value than the value as you expected by the equation.

4. Key Benchmarks

In this section, key benchmarks are defined as we expect implementors to follow the RDMA over Open APN PoC.

4.1. Benchmark 1: throughputs for data transferring

Table 2: Benchmark 1: throughputs for data transferring

ITEM	DESCRIPTION
scope	Speed of data traveling from the sender to the receiver. Note - The goal of this PoC is 80% or more of link speed.
metrics	Gbps
measuring method	Measurement starts after establishing RDMA connection

4.2. Benchmark 2: latency between RDMA endpoints

Table 3: Benchmark 2: latency between RDMA endpoints

ITEM	DESCRIPTION
Scope	The time that it takes to transfer data between the RDMA endpoints excluding delay of transmission distance. Note - The goal is 1 millisecond (msec), or less. In this PoC, it is important to evaluate behavior of RDMA over Open APN framework with long distance. And delay of transmission is provided by network emulator. This value is known. Therefore, delay of transmission distance is excluded.
Metrics	msec
measuring method	• (Option1) You can measure the difference between the transmission time in a RDMA endpoint in sender-side and the reception time in a RDMA endpoint in receiver-side. • (Option2) You can calculate half of RTT by sent back data.

4.3. Benchmark 3: power consumption

Table 4: Benchmark 3: power consumption

ITEM	DESCRIPTION
scope	The effect of reducing power consumption.
metrics	percentage

measuring method	You can evaluate the effect of reducing power consumption by measuring the CPU usage rate due to hardware offloading for RDMA transmission. Note - One of big impact factor of power consumption is derived by protocol processing in CPU. In this PoC, PoC implementors should evaluate efficiency of hardware offloading of protocol processing. However, it is not easy to evaluate efficiency of hardware offloading from power consumption point of view because there are many complicated factors. Therefore, measurement of CPU usage as an alternative method is defined to evaluate power consumption.

5. References

[1] NVM Express, “NVMe over Fabrics - Discussion on Transports“, Flash Memory Summit 2018 (https://nvmexpress.org/wp-content/uploads/NVMe-102-1-Part-2-NVMe-oF-Transports_Final.pdf).

[2] Kevin Deierling and Idan Burstein, “Ethernet Storage Fabrics Using RDMA with Fast NVMe-oF Storage to Reduce Latency and Improve Efficiency“, Storage Developer Conference 2017 (https://www.snia.org/sites/default/files/SDC/2017/presentations/Solid_State_Stor_NVM_PM_NVDIMM/Idan_Burstein_EthernetStorageFabricsUsingRDMAwithFastNVMe-oFStorage.pdf).

[3] IOWN Global Forum, “Reference Implementation Model (RIM) for the Area Management Security Use Case“, January 2022 (https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-RIM-for-AM-Use-Case-1.0.pdf).

[4] InfiniBand Architecture Specification Volume 1, Release1.4, InfiniBand Trade Association, April 7, 2020.

[5] IOWN Global Forum, “Data-Centric Infrastructure Functional Architecture“, January 2022 (https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-DCI-Functional-Architecture-1.0-1.pdf).

History

Revision	Release Date	Summary of Changes
1.0	July 21, 2022	Initial Release