Elastic Fabric Adapter: A Viable Alternative to RDMA over InfiniBand for DBMS?

TECHNISCHE UNIVERSITÄT DARMSTADT

Master Thesis Presentation

By: **Dwarakanandan B.M**

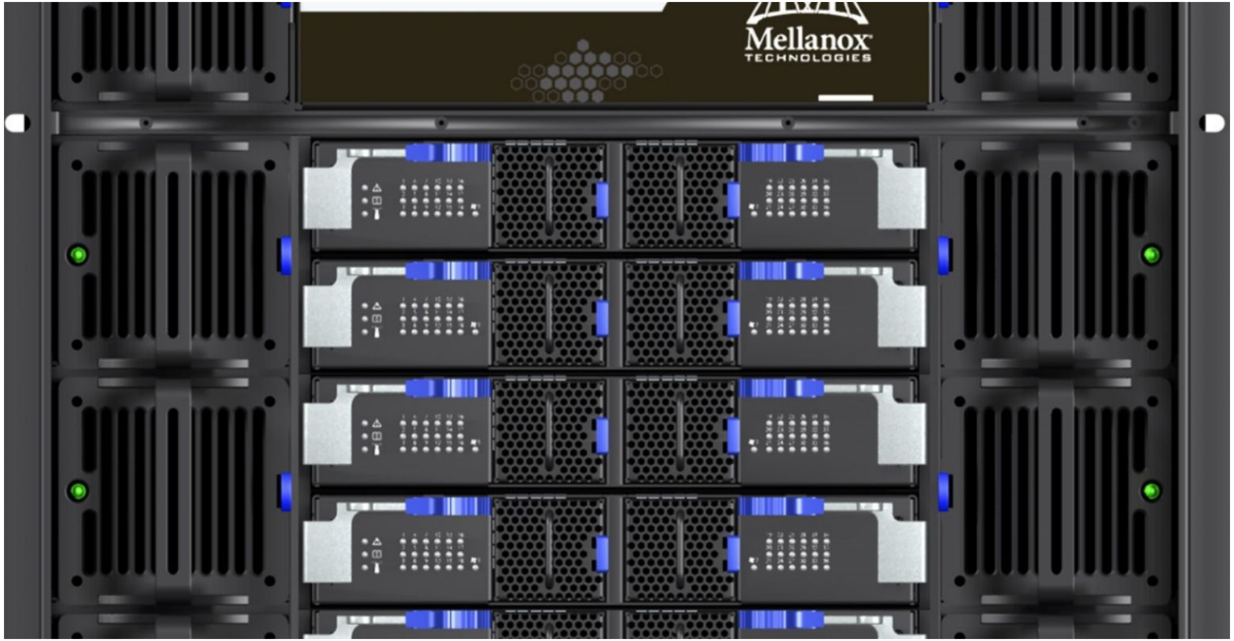
Advisor-1: **Tobias Ziegler**

Advisor-2: **Prof. Dr. Carsten Binnig**

DAMON
Data Management
On New Hardware
18th Workshop
ACM SIGMOD 2022

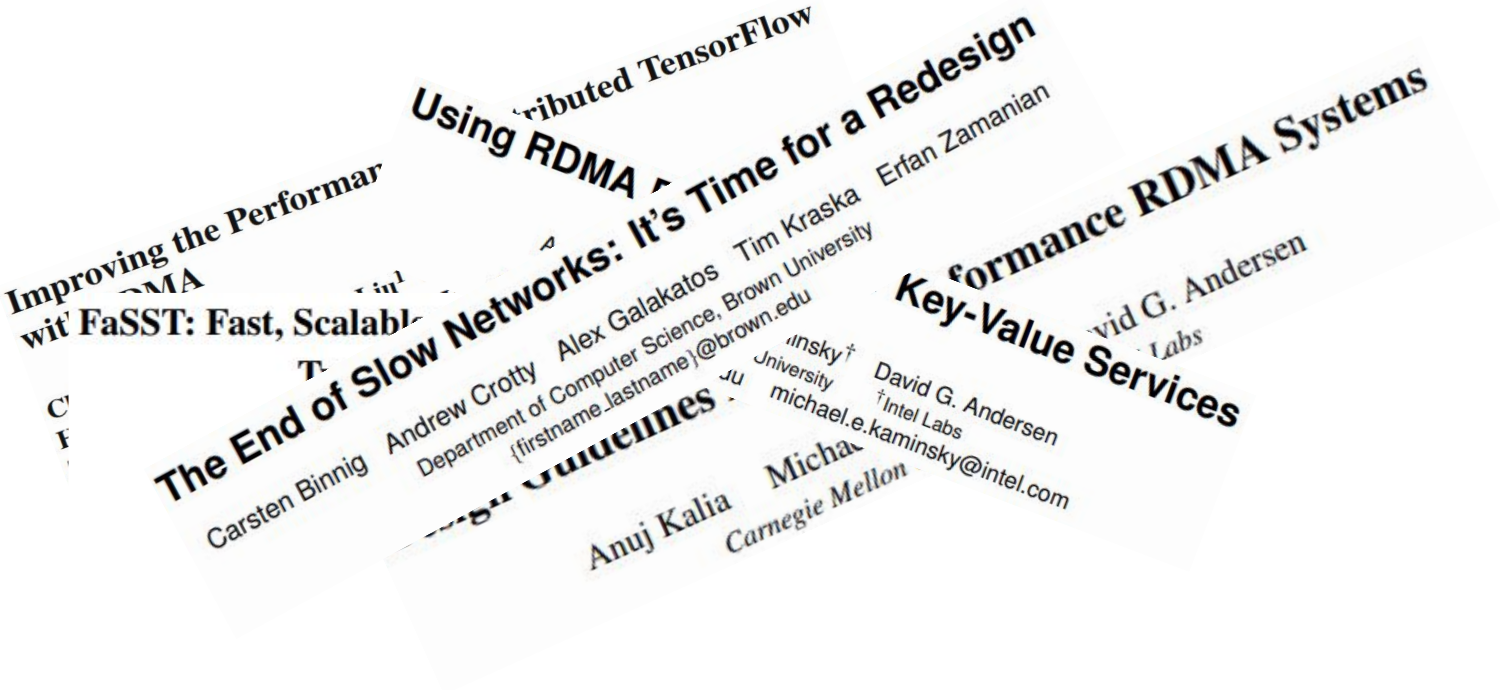
Data Management Lab
Technical University of Darmstadt

Trend 1 : High Performance Networks



Nvidia Mellanox CS7500 100 GB/s InfiniBand Smart Director Switch

Networks are Fast …. So?



Improving the Performan
with RDMA

FaSST: Fast, Scalabl

tributed TensorFlow

Using RDMA

The End of Slow Networks: It's Time for a Redesign

Carsten Binnig Andrew Crotty Alex Galakatos Tim Kraska Erfan Zamanian

Department of Computer Science, Brown University

{firstname_lastname}@brown.edu

formance RDMA Systems

vid G. Andersen

Labs

Key-Value Services

insky†

University

David G. Andersen

†Intel Labs

michael.e.kaminsky@intel.com

Guidelines

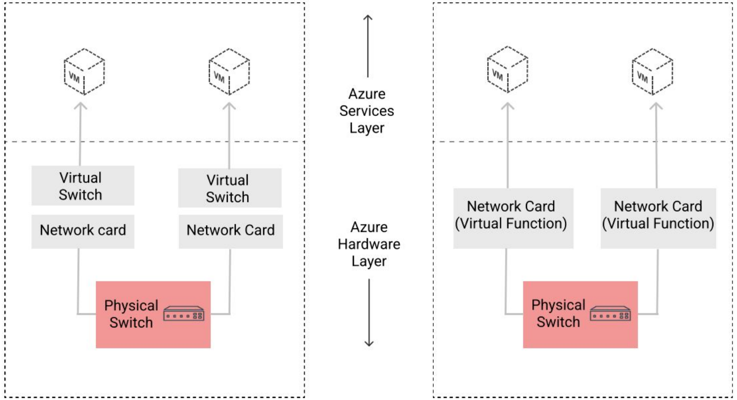
Anuj Kalia Micha

Carnegie Mellon

Trend 2 : The Cloud

- In the cloud, RDMA is offered only by Microsoft Azure
- Supported by very limited number of instance types

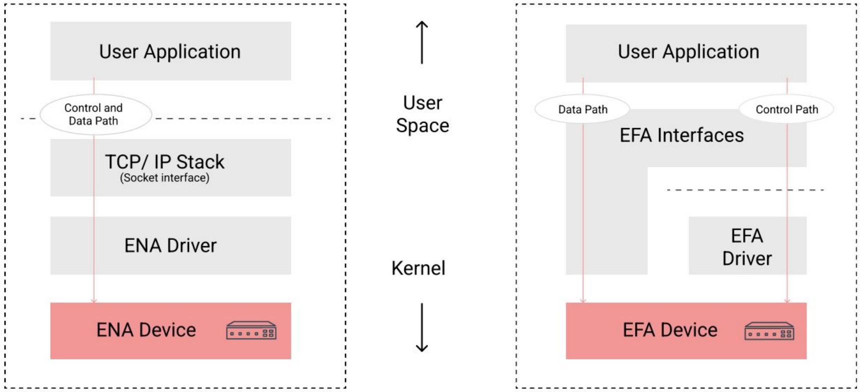
Cloud native High-Performance networks



Without Accelerated Networking

Accelerated Networking (SR-IOV)

Fig:1 Azure Accelerated Networking (SR-IOV)



Elastic Network Adapter (ENA)

Elastic Fabric Adapter (EFA)
OS-Bypass

Fig:2 AWS Enhanced Networking (Kernel-Bypass)

Motivation

- In 2019 AWS announced a new network fabric called Elastic Fabric Adapter (EFA)
- EFA has primarily been marketed towards HPC workloads by AWS.
- Most research also primarily been driven by the HPC community
- No official hardware specifications provided by AWS
- In-depth evaluation of EFA needed to better understand its implications on system design

Research question

Can EFA be a viable alternative to RDMA over InfiniBand for data-intensive systems?

Elastic Fabric Adapter

- Modified ENA Adapter (AWS Nitro)
- Scalable Reliable Datagram protocol
- Low level *ibverbs* library
- Emulated features using high level frameworks - *libfabric*

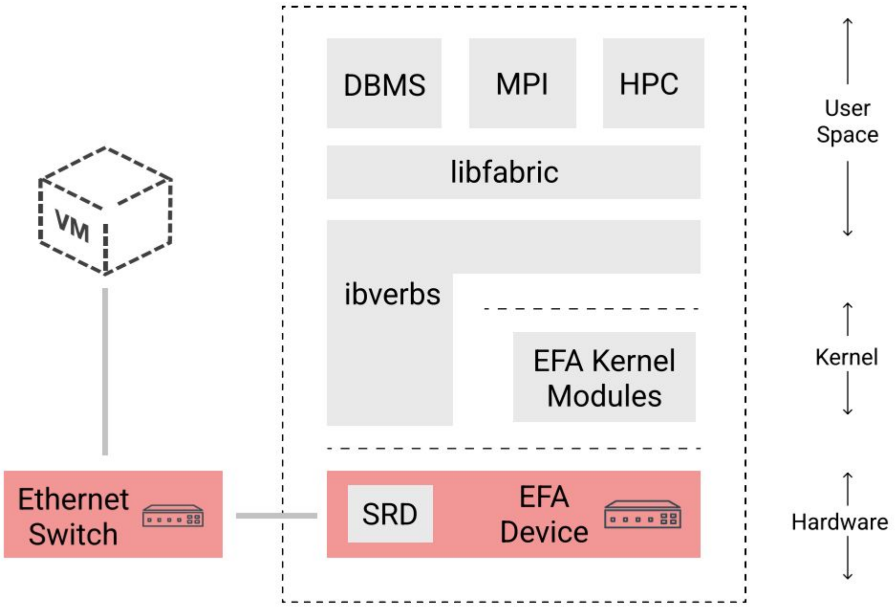
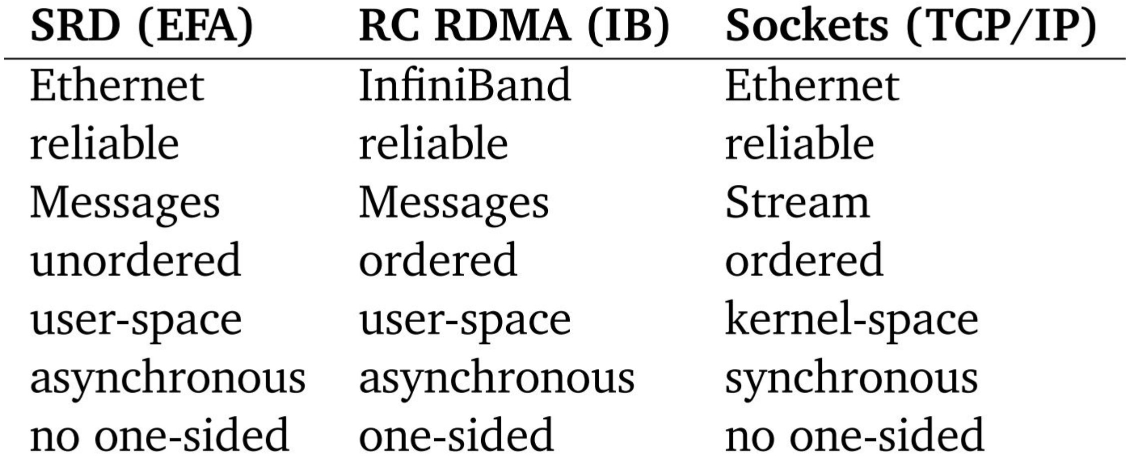


Fig:3 EFA Hardware and Software Stack

SRD (EFA) in a Nutshell



SRD (EFA)	RC RDMA (IB)	Sockets (TCP/IP)
Ethernet	InfiniBand	Ethernet
reliable	reliable	reliable
Messages	Messages	Stream
unordered	ordered	ordered
user-space	user-space	kernel-space
asynchronous	asynchronous	synchronous
no one-sided	one-sided	no one-sided

Fig:4 SRD compared to reliable connected RDMA and traditional TCP/IP sockets

Evaluation Methodology

- Isolate the fundamental properties of EFA and RDMA:

  InfiniBand Verbs Performance Tests (*perftest)*

- Multi-threaded evaluations and other EFA specific features:

  libefa and ***efa-bench***

- Evaluate EFA's programming interfaces:

  OSU Micro-Benchmarks (OMB)

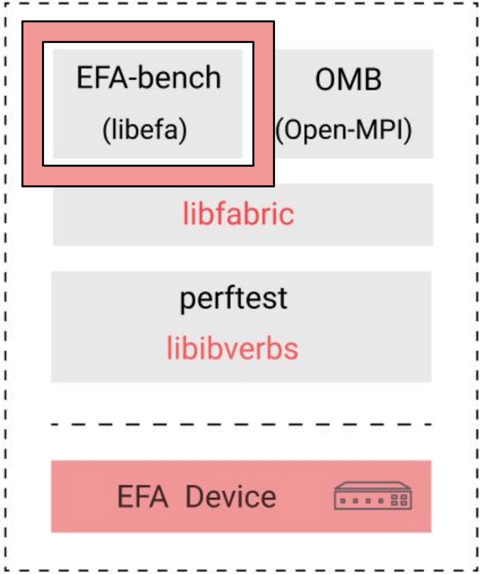


Fig:5 Evaluation software stack

Libfabric and *efa-bench*

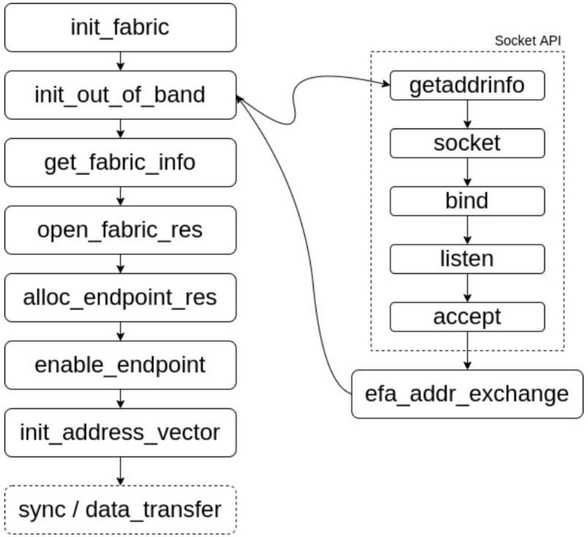


Fig:6 libfabric - Fabric setup flow

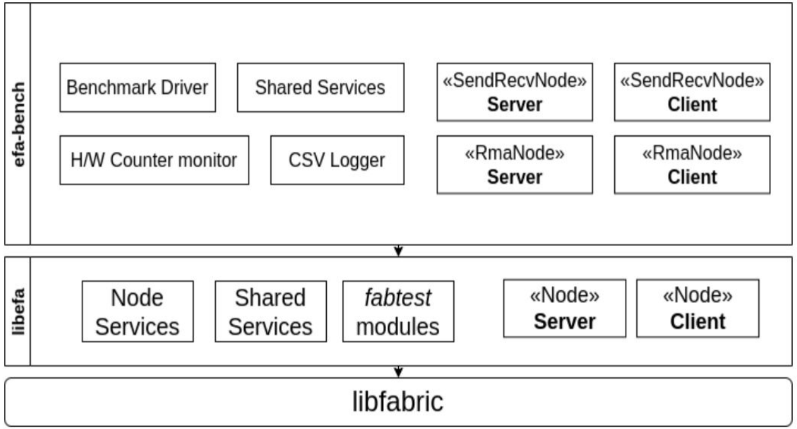


Fig:7 Architecture of efa-bench

Setup

RDMA bare-metal

56 CPUs

1 TB memory

Mellanox ConnectX-5
(MT27800 100G)

EFA c5n.18xlarge

72 vCPUs

192 GB memory

EFA 100G Network adapter

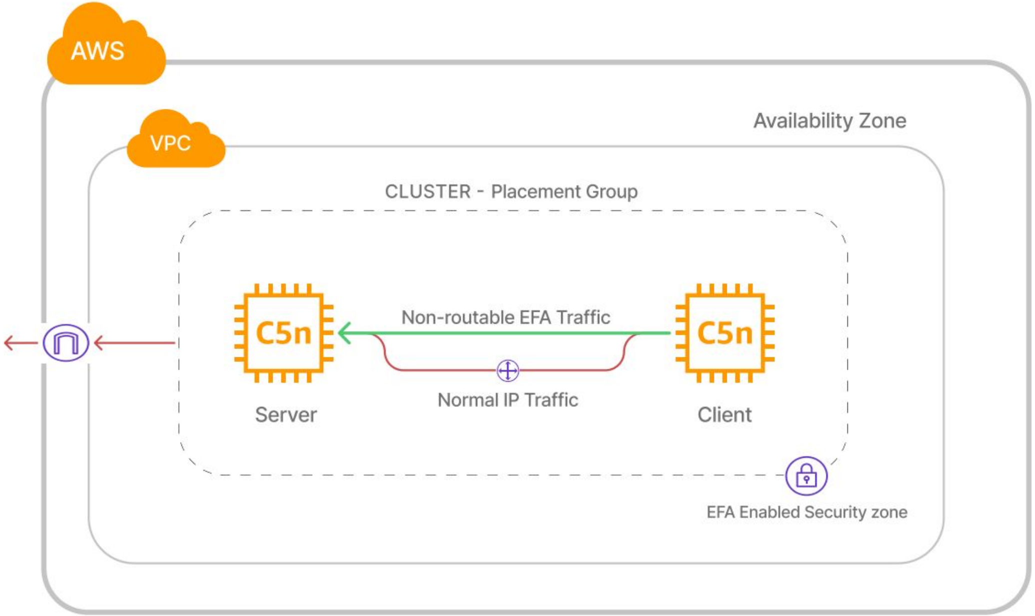
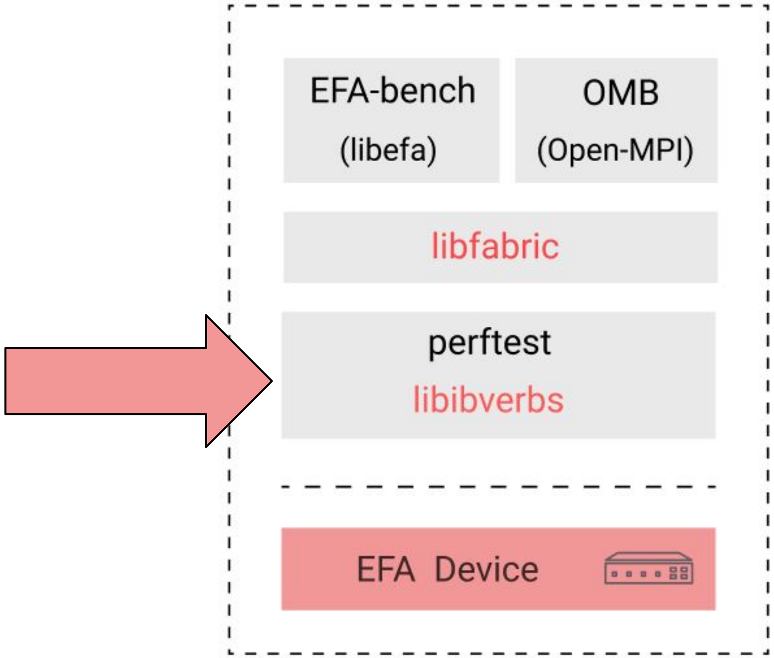


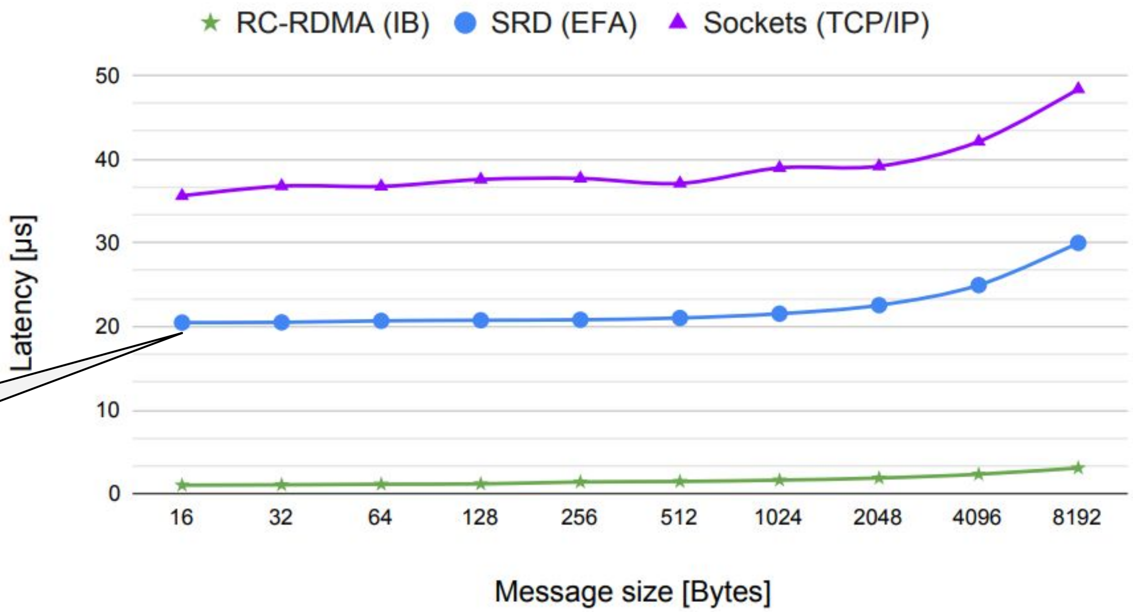
Fig:8 EFA Evaluation setup on the AWS Cloud


EFA-bench
(libefa)
OMB
(Open-MPI)
libfabric
perftest
libibverbs
EFA Device

Latency evaluation

Impact of message size on Average Latency

RC-RDMA vs SRD vs TCP/IP



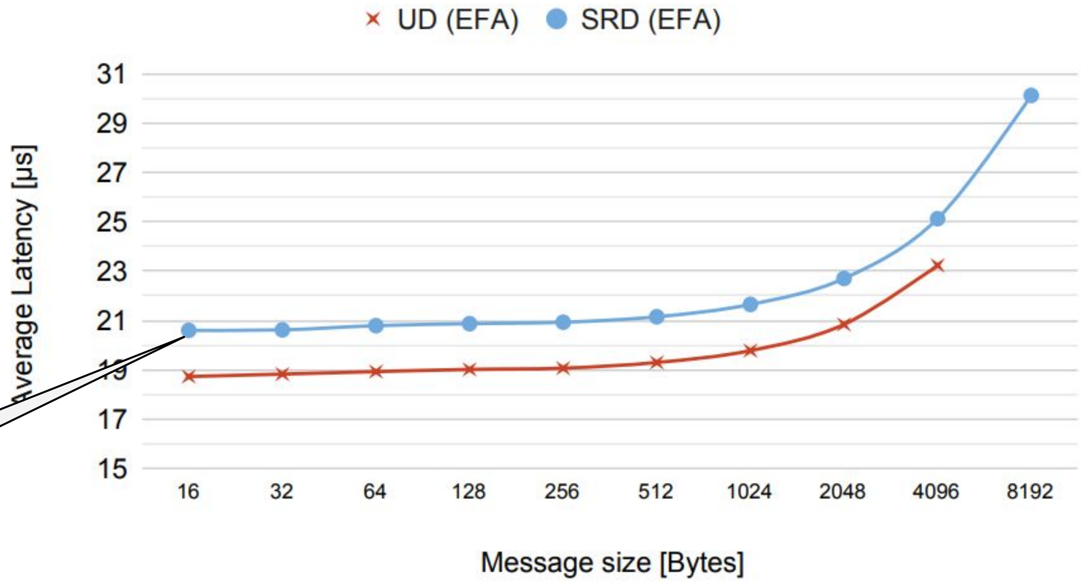
EFA latencies are still an order of magnitude higher than those of RDMA

Fig:9

EFA’s SRD vs UD protocol

Impact of message size on Average Latency

EFA-UD vs EFA-SRD



SRD only has a slightly higher latency of around 2 µs over UD

Fig:10

Synchronous bandwidth

Impact of Message size on Synchronous bandwidth

Similar to the latency gap, RDMA achieves 10x more bandwidth than EFA-SRD.

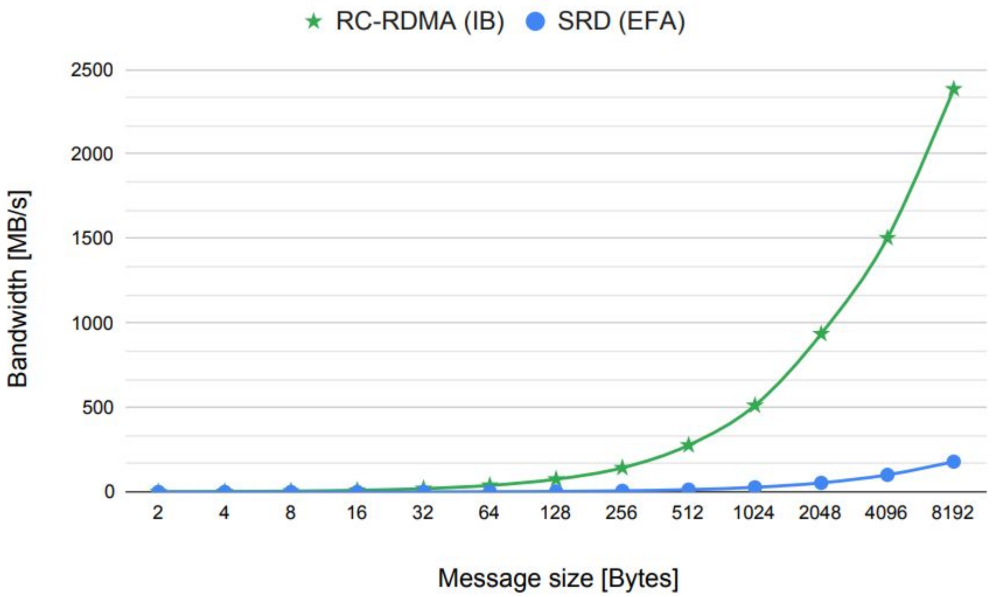


Fig:11

Asynchronous bandwidth - Message rate

Impact of transmission depth (outstanding operations) on Async Bandwidth

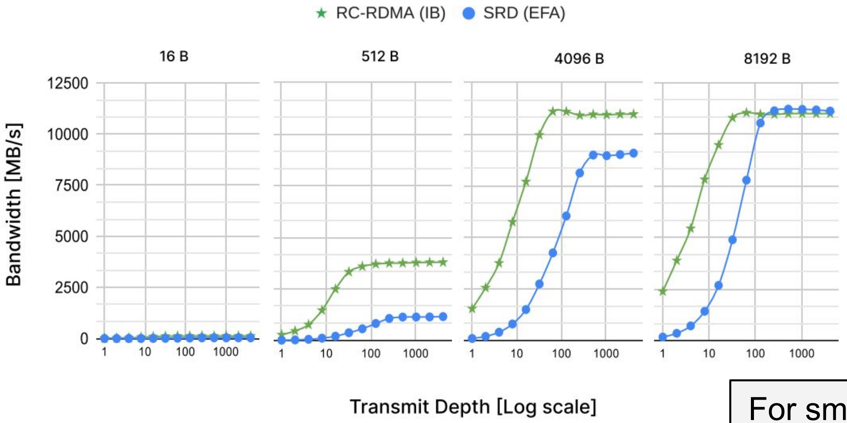
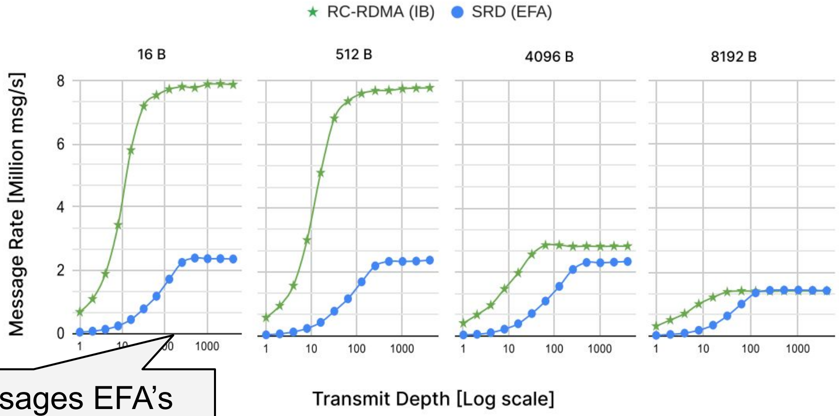


Fig:12

Impact of transmission depth (outstanding operations) on Message Rate



For small messages EFA's msg rate is considerably smaller than RDMA

Fig:13

Network Interface parallelism

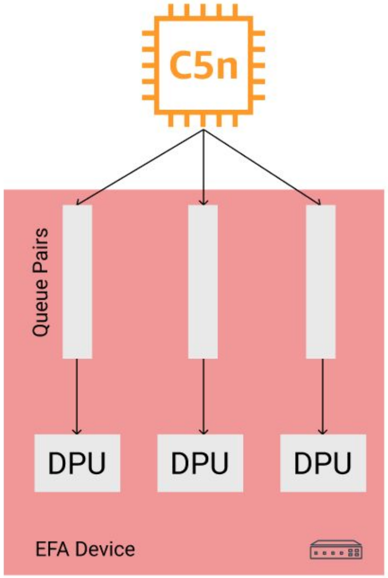
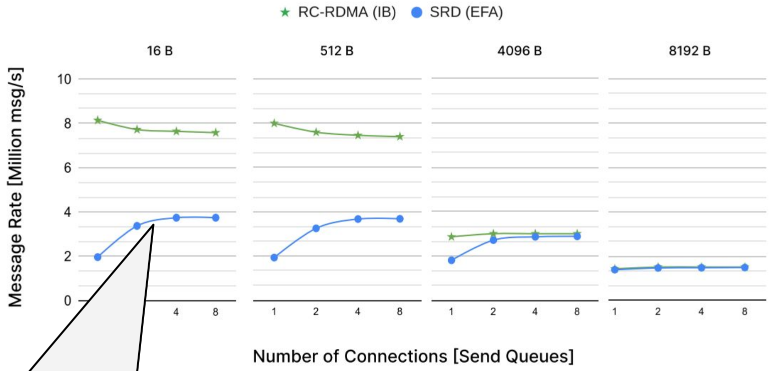


Fig:14 NIC Architecture showcasing connection queues

Impact of connection pairs (Send queues) on Message Rate



We achieve a maximum of 4 M messages per second by exploiting multiple Nitro Card DPUs

ig:15

Multi-threaded evaluation

Impact of thread count on Bandwidth (TX depth - 128)

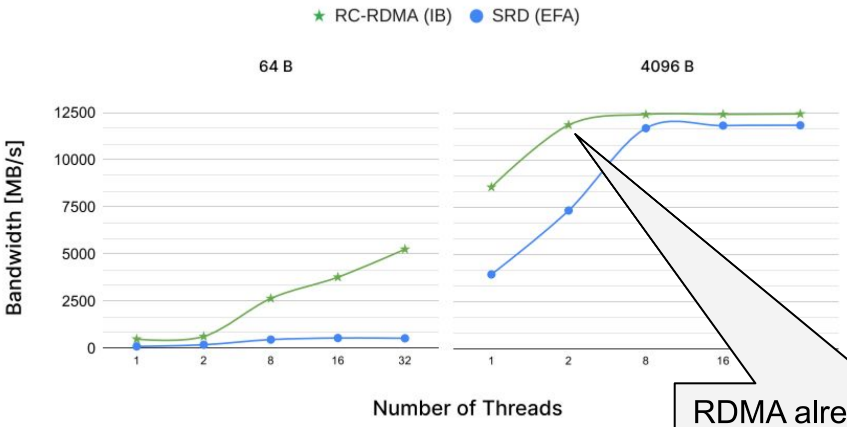


Fig:16

Impact of thread count on Message Rate (TX depth - 128)

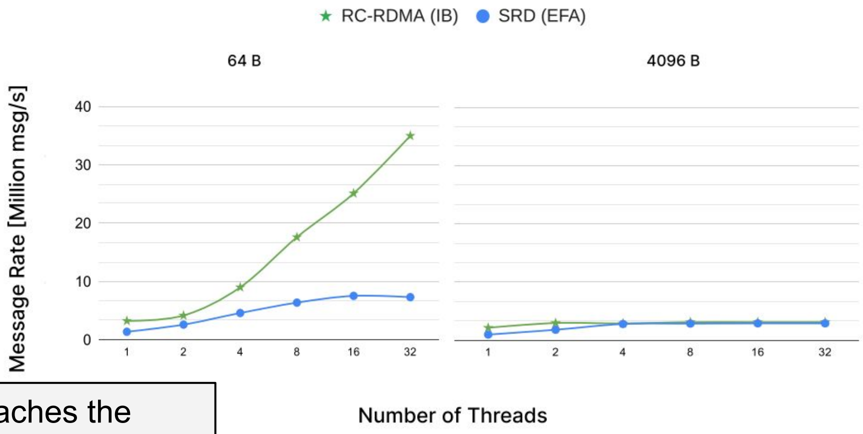
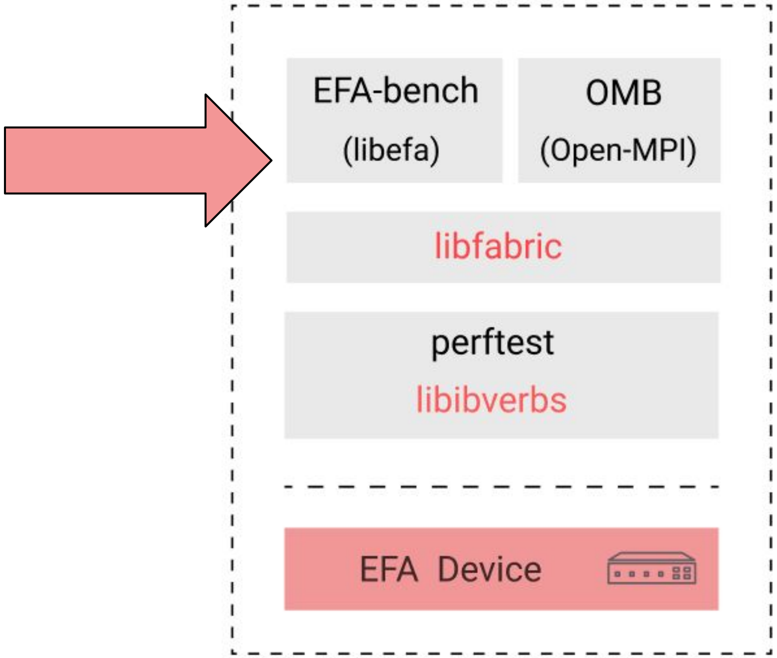


Fig:17

RDMA already reaches the bandwidth limit with 2 threads, EFA requires at-least 4 threads


EFA-bench
(libefa)
OMB
(Open-MPI)
libfabric
perftest
libibverbs
EFA Device

Interface evaluation

Performance implication of EFA interfaces (TX Depth - 256)

If the application does not intend to use any of libfabric's feature set, ibverbs might be the optimal choice

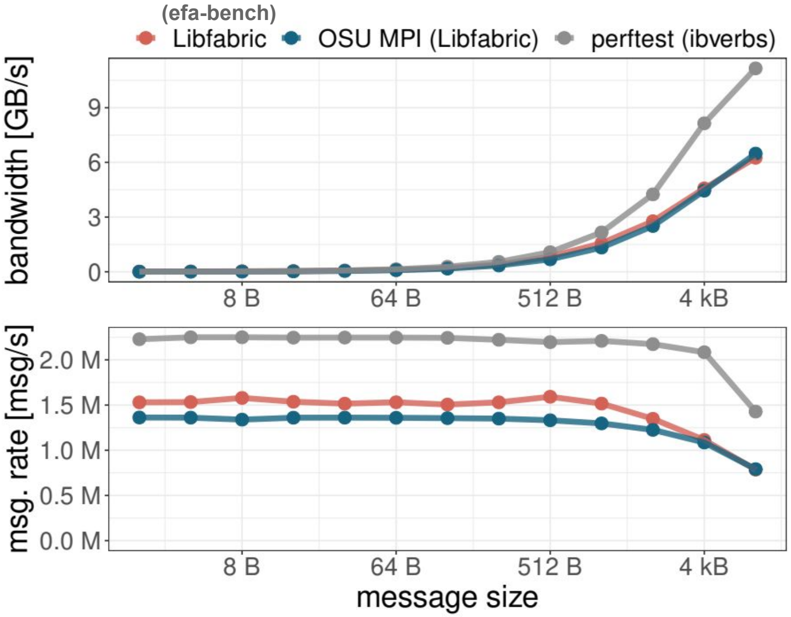


Fig:18

One sided operations - RMA

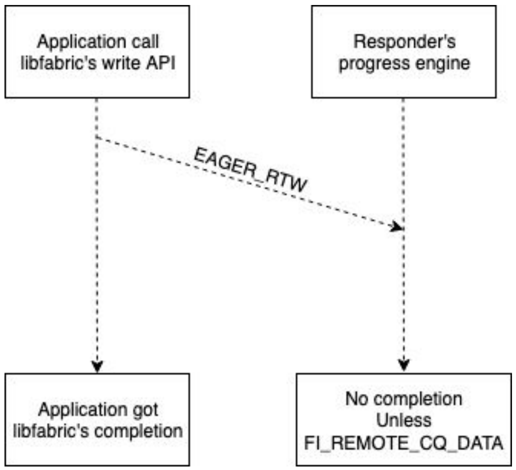
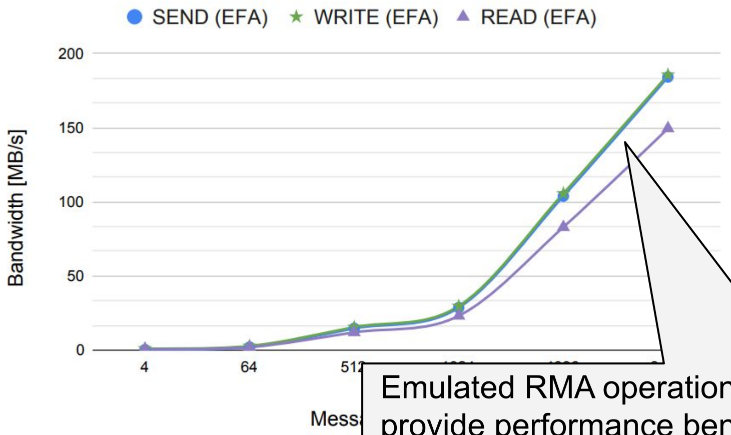

Application call libfabric's write API
Responder's progress engine
EAGER_RTW
Application got libfabric's completion
No completion Unless FI_REMOTE_CQ_DATA

Fig:20 libfabric- EFA RDM Communication Protocol V4

Performance of emulated RMA operations compared to send/recv



Fig

Emulated RMA operations do not provide performance benefits like true one-sided operations, such as in RDMA

Message Segmentation Overhead

Average Link latency at Segmentation Boundary of about **8760** bytes

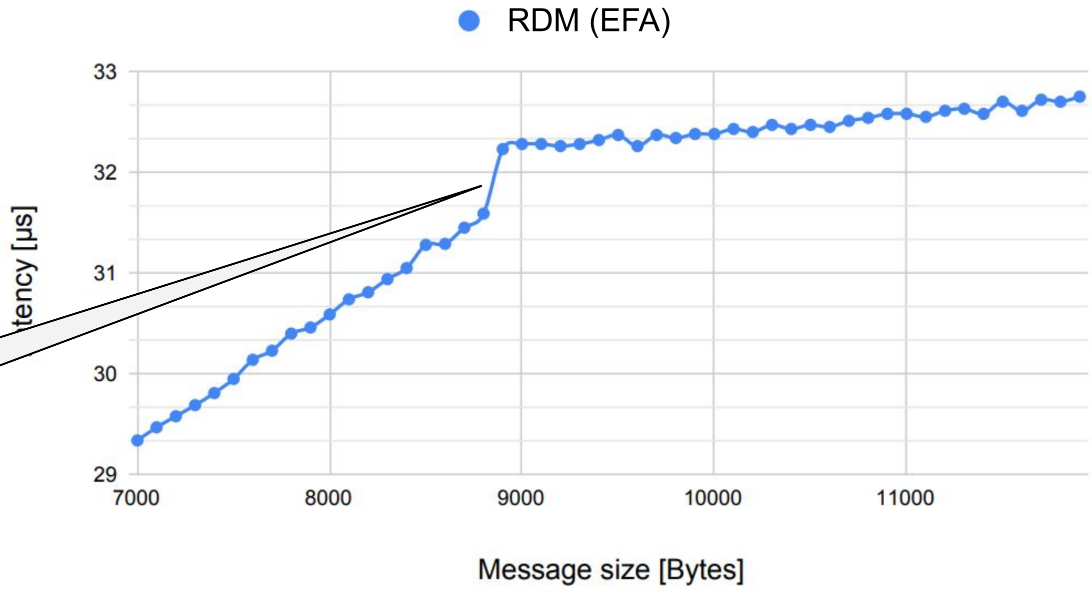


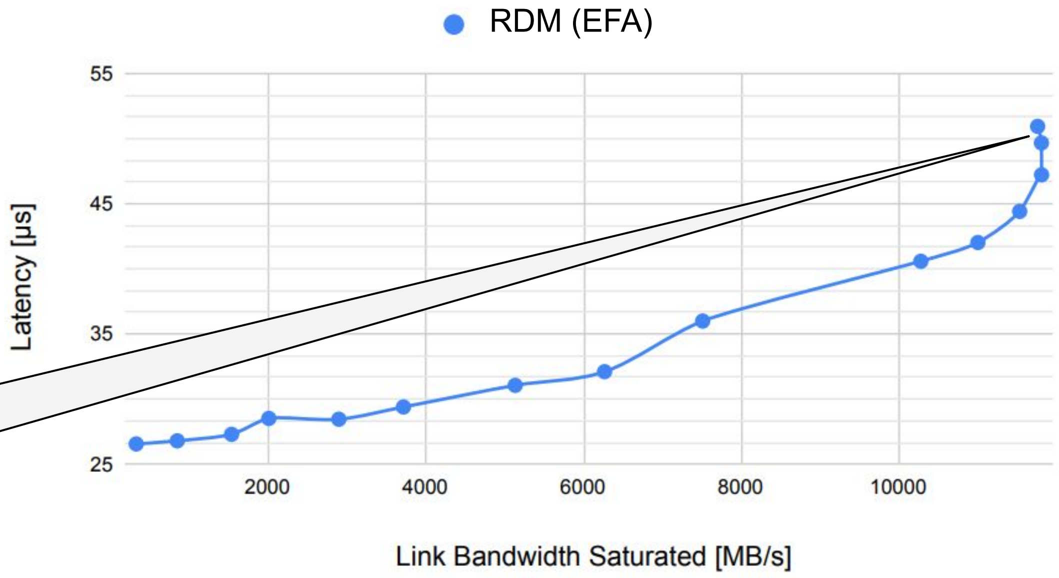
Fig:19

Overhead is only around 1 µs. Given the relatively high baseline of EFA’s latency, this is insignificant

Link Latency at Saturation

Average Link latency at Saturation

(Message size - 4096 Bytes)



At peak saturation, we see latencies of about 50 µs for a 4096 Byte packet, which is almost twice the normal unsaturated latency

EFA In Summary

- **Latency -** Better than TCP, but there is a substantial gap compared to RDMA
- **Bandwidth -** Strongly dependent on the transmission depth and the message size
- **Message Rate -** Multiple flows are needed to fully exploit the NIC
- **No one-sided operations -** Emulated in software with a performance penalty
- **Cloud Native & Proprietary** - Migration to AWS cloud a prerequisite (Vendor Lock-in)

So… is EFA a viable alternative?

An Ongoing Saga: For now EFA still has considerable limitations compared to RDMA over InfiniBand

However in comparison with the only other cloud alternative TCP/IP, EFA paired with its SRD protocol has potential for data-intensive systems