A Member of NARLabs
National Center for
High-performance Computing

The DPU (SmartNIC) Application - Key-Value Store in NCHC

Speaker: Kuo-Teng Ding
RD Members - Yi-Lun (Serena) Pan、Fang-An Kuo
NTU Member - Hung-Shin Chen

2021/7/6

NARLabs

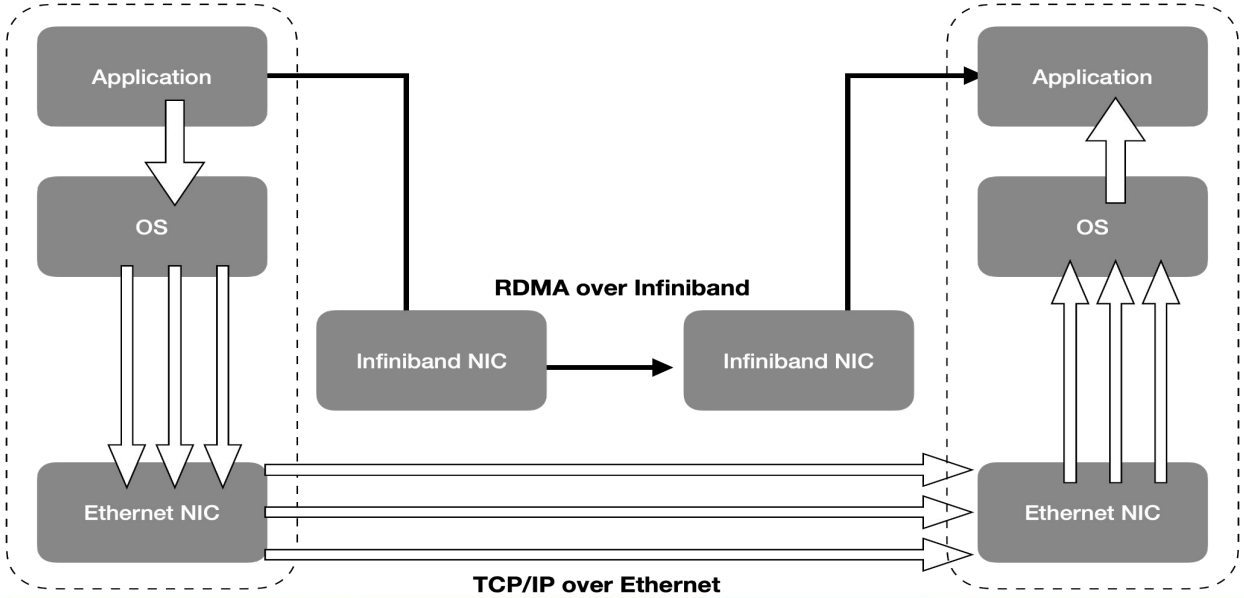
Outline

- Pre-School
- The Current Stage
  - Motivation
  - Baseline Benchmark
- The Next Stage

Pre-School

InfiniBand have RDMA capability, which is less latency without remoteCPU involved compared to TCP/IP

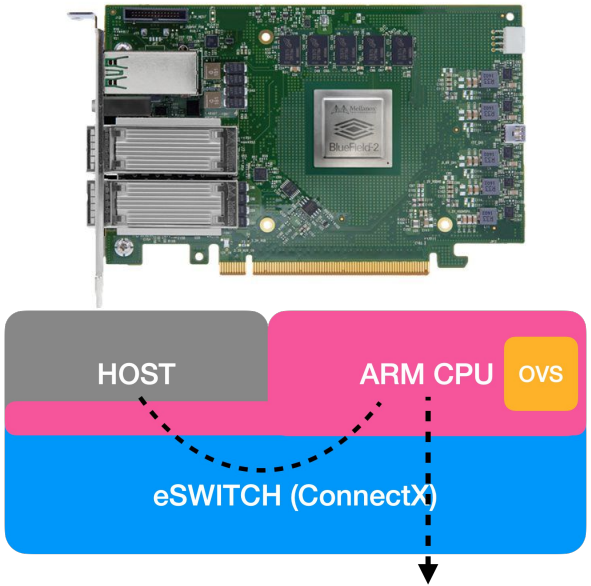
DPUs (Data Processing Units) provide efficient data transmission

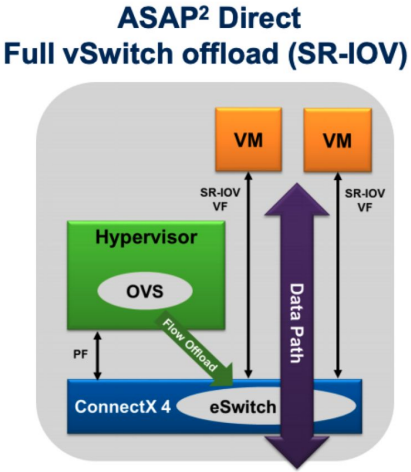


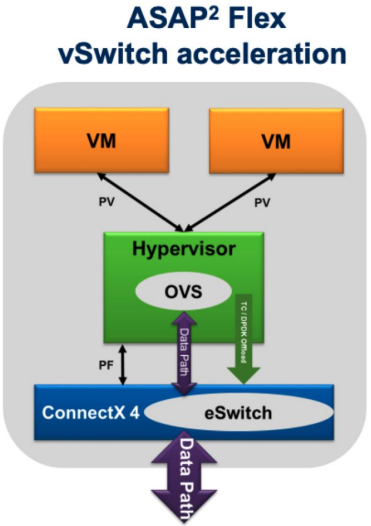
Pre-School

NVidia Mellanox ConnectX is an Infiniband adapter ASIC (NIC).

NVidia Mellanox BlueField is a SoC integrated with ConnectX nics and ARM CPU.







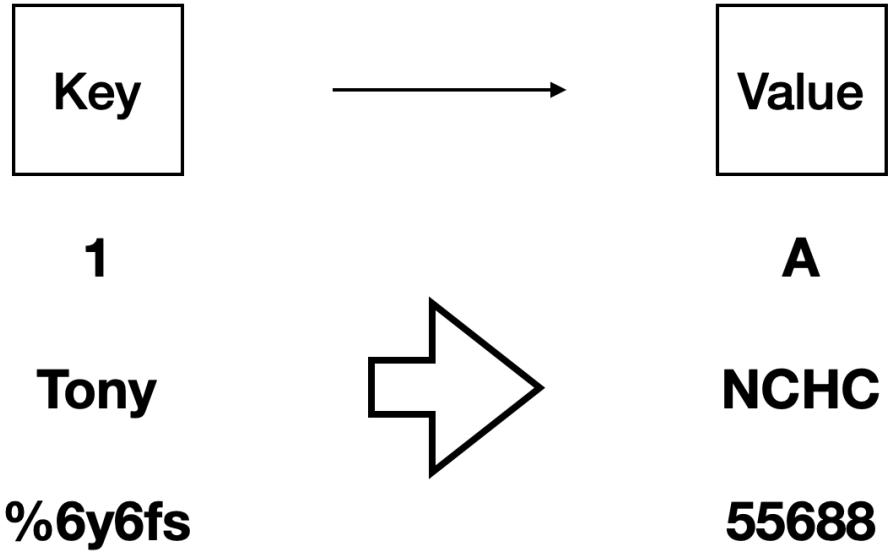
Motivation

Current works focus on designing applications that runs standalone on SmartNIC. The problem of how to efficiently utilize traditional hardwares (x86 CPUs and common InfiniBand NICs) and SmartNICs at the same time remains unclear.

We present a key-value store system purposal which efficiently utilize both hardwares to achieve high-throughput.

Purpose

We want to utilize SmartNIC and workstation hardware to achieve a high-throughtput key-value store system.



Overview

- A KVS system leverage RDMA and DPU capability
- DPU onboard gerernal-purpose cpu can provide further operation
- Enhancing existed KVS system
- Mitigate the CPU performance gap with RDMA

Recent Related Research

- Li et al's KV-direct based on programmable NIC (FPGA)
- Kalia et al's HERD based on operation decoupling
- Cassell et al's Nessie also based on decoupled client-driven operations
- Dragojević et al's FaRM based on shared memory space
- Mitchell et al's Pilaf proposed a self-verifying scheme using checksum.

Hardware Spec.

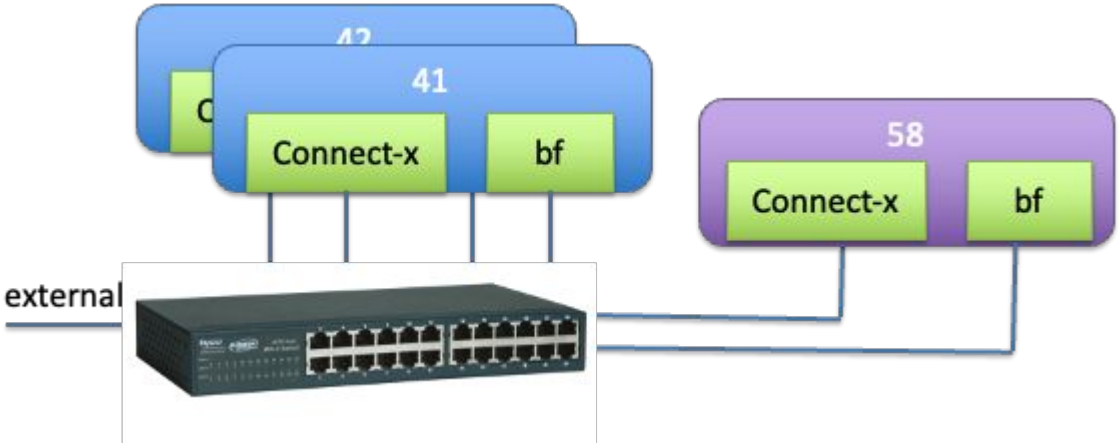
The perf_test and rdma_bench benchmark mentioned in HERD is used in baseline survey.

Hardware setting below:

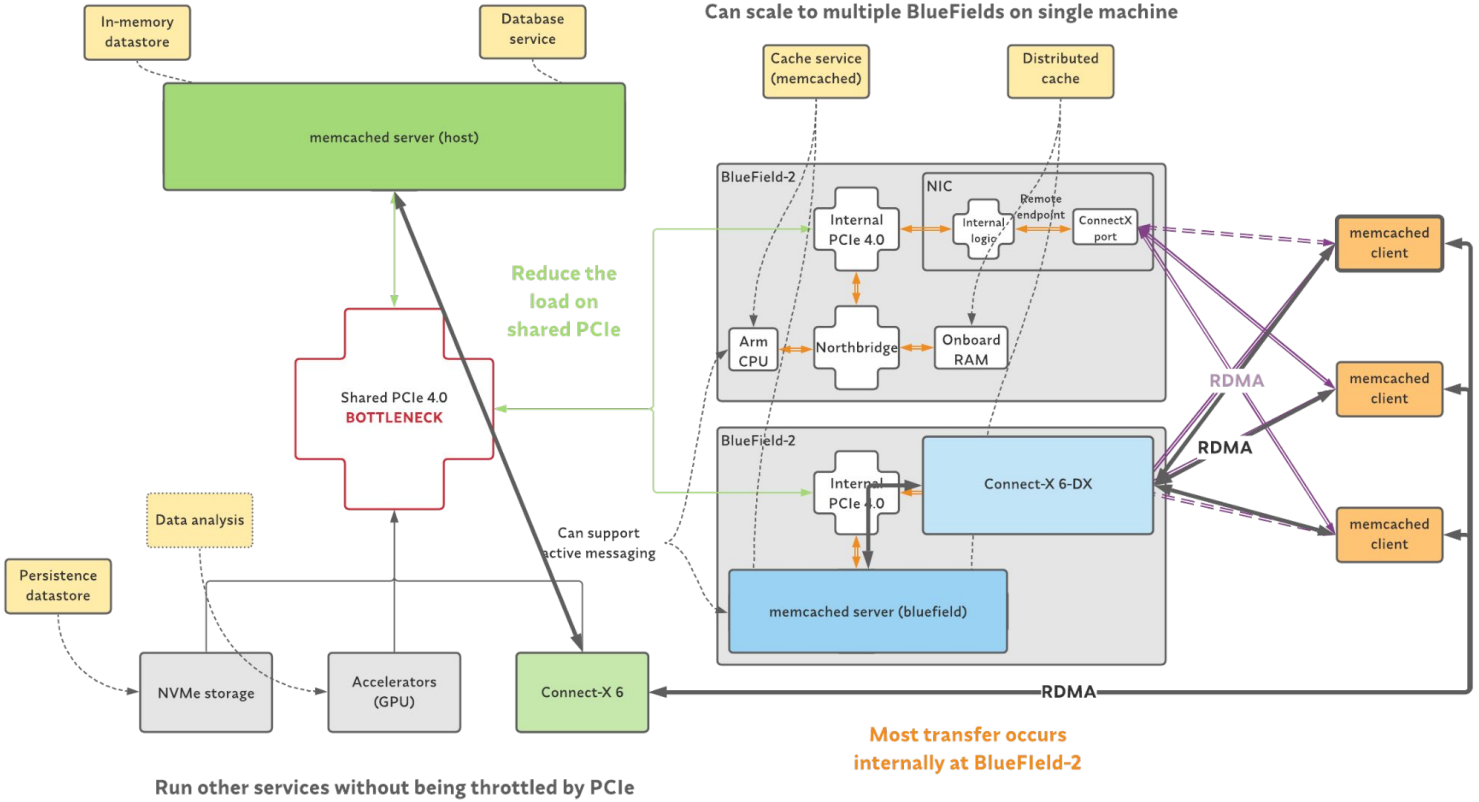
- AMD EPYC 7702P 64-Core: RTX 2080, BlueField-2, ConnectX-6 (HDR), 256GB DDR4 (140.110.18.41/42)
- Mellanox SB7800 100G EDR switch
- Intel Xeon Gold 6148 CPU: DGX A100*2, BlueField-2, ConnectX-6 (HDR), 384GB DDR4 (140.110.18.58)

Hardware Spec.

We fixed 41/42 machine as a client/server and test performance when 41bf, 41/42 and 58 as a corresponding server/client.



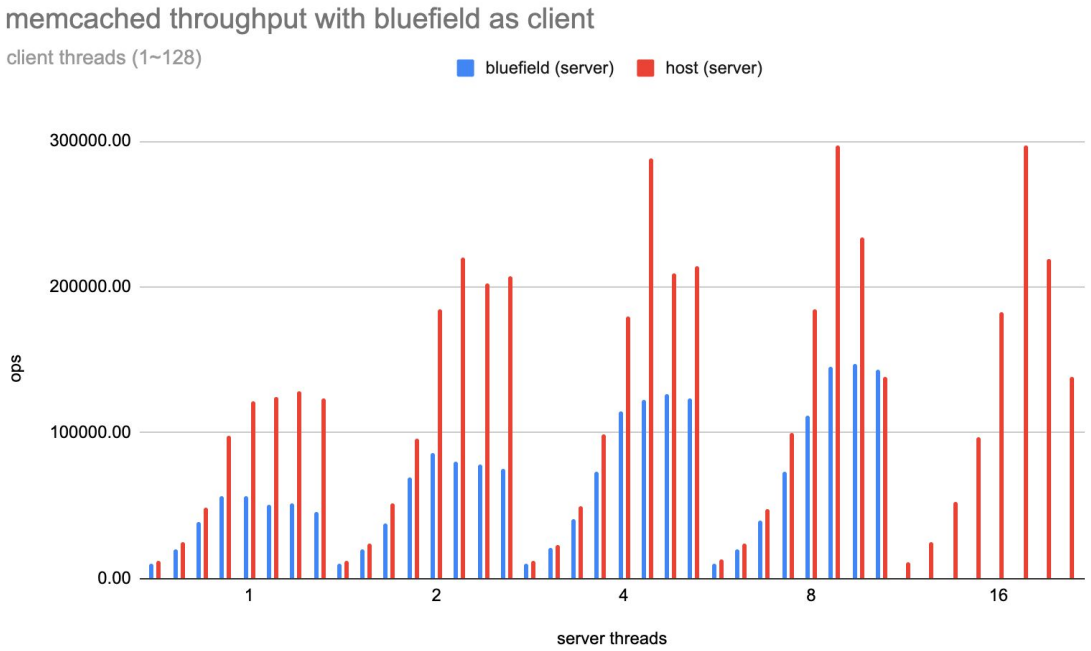
IPoIB memcached experiments



IPoIB memcached experiments

Value size: 1000 bytes

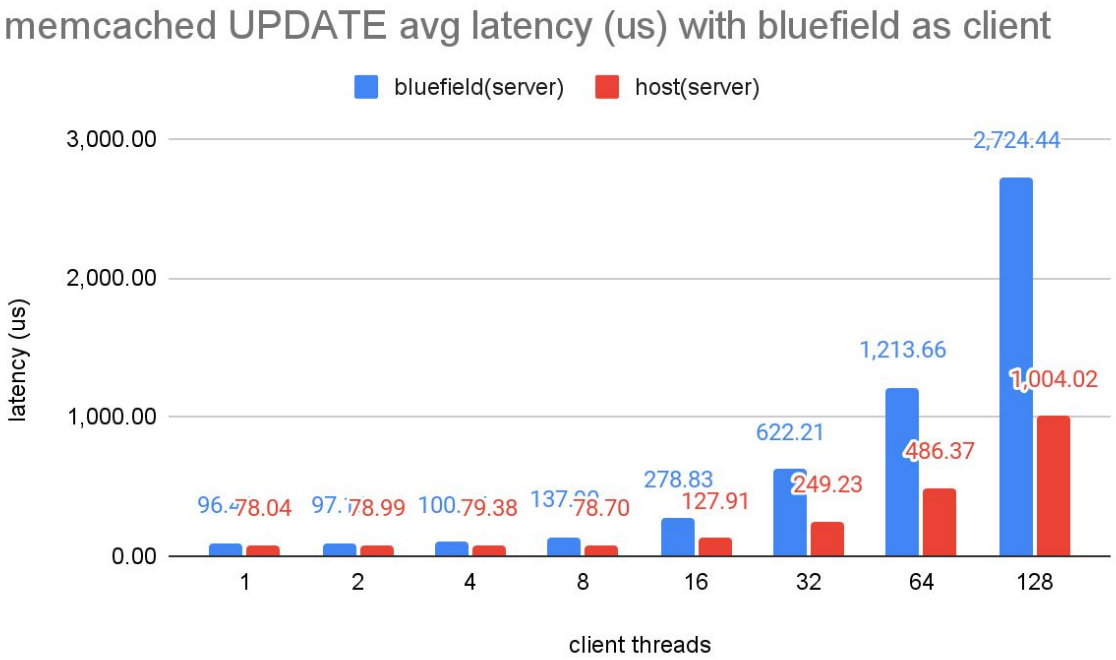
Memcached is bottlenecked by CPU core synchronization.



IPoIB Memcached Latancy

Value size: 1000 bytes

The latency is bottlenecked by Ethernet transmission.



Summary: IPoIB Memcached Latancy

- Memcached is bottlenecked by CPU core synchronization.
- The latency is bottlenecked by Ethernet transmission.
- Our experiments is similar with Mitchell et al's work, they conclude that memcached RTT of IPoIB/Ethernet is over 60 µs

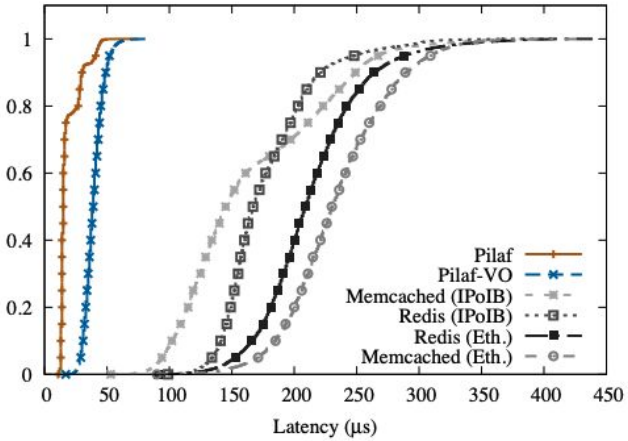
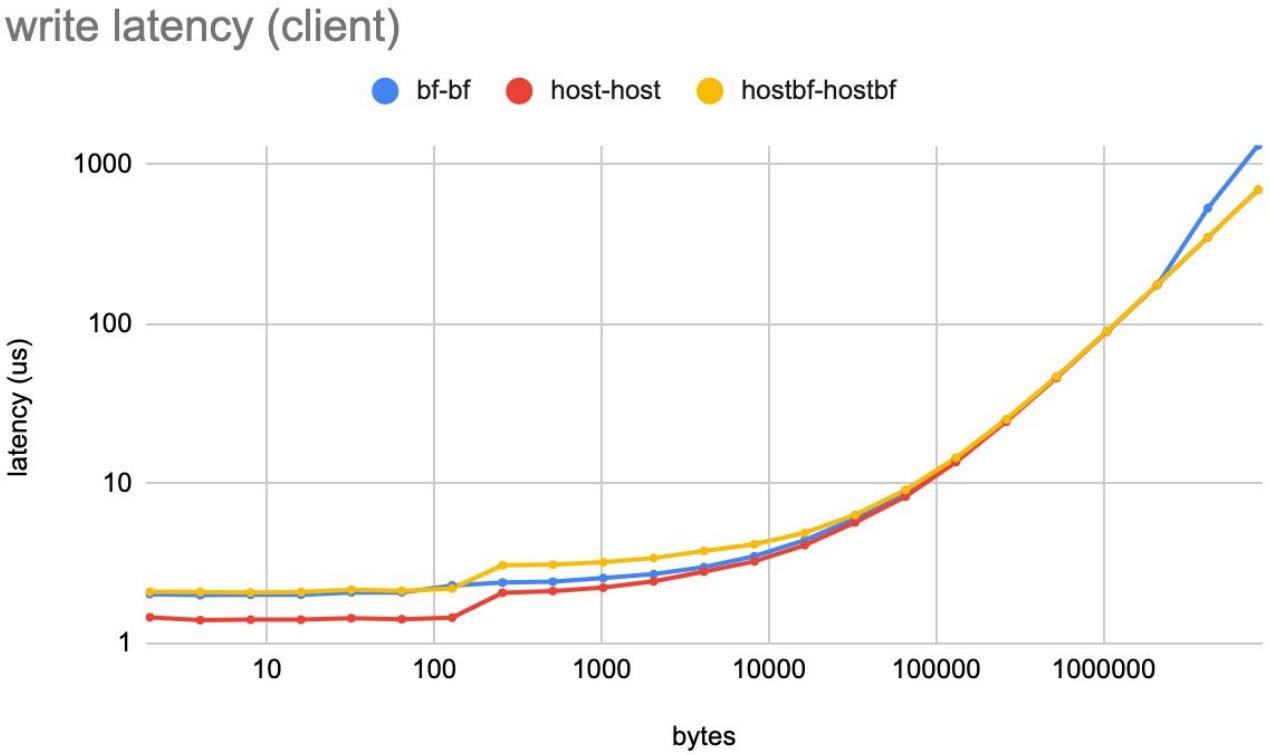


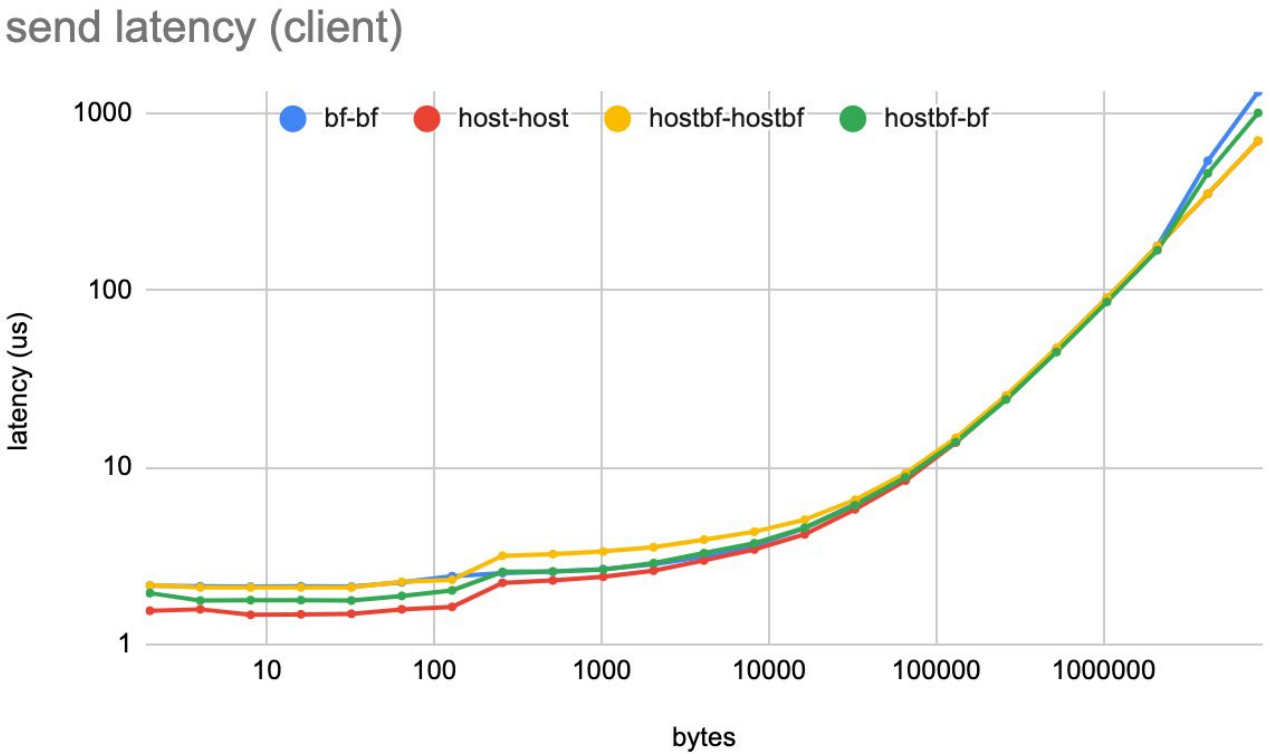
Figure 11: CDF of Pilaf latency compared with Memcached, Redis and Pilaf-VO in a workload consisting of 90% `gets` and 10% `puts`. The average value size is 1024 bytes. The experiments involved 10 clients.

NARLabs

RDMA Operation Latency

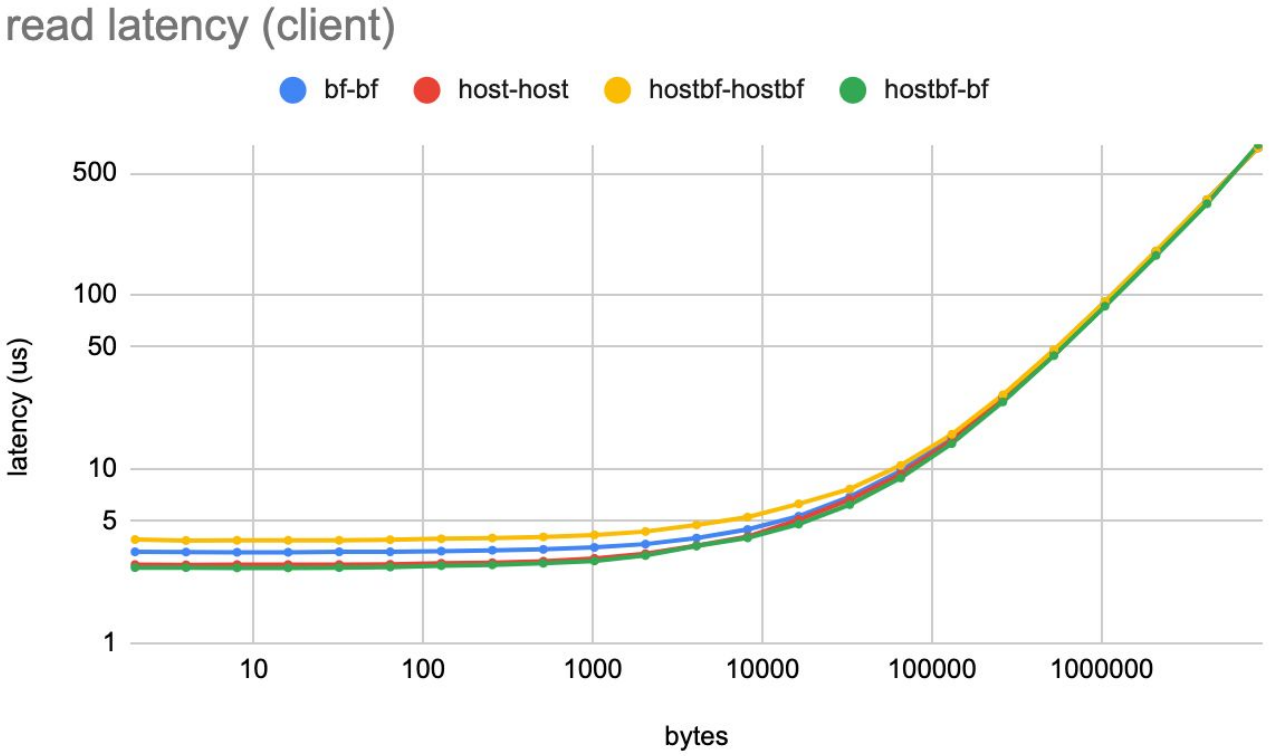


RDMA Operation Latency



NARLabs

RDMA Operation Latency



Summary: RDMA Latency performance

- There is no obvious performance difference in operation latency
- Compared to Ethernet RDMA, Infiniband RDMA achieve very low latency
- The latency of SEND/WRITE operation dramatically increase when transaction data is larger than 1000 bytes and for READ opeartion is 10000 bytes.

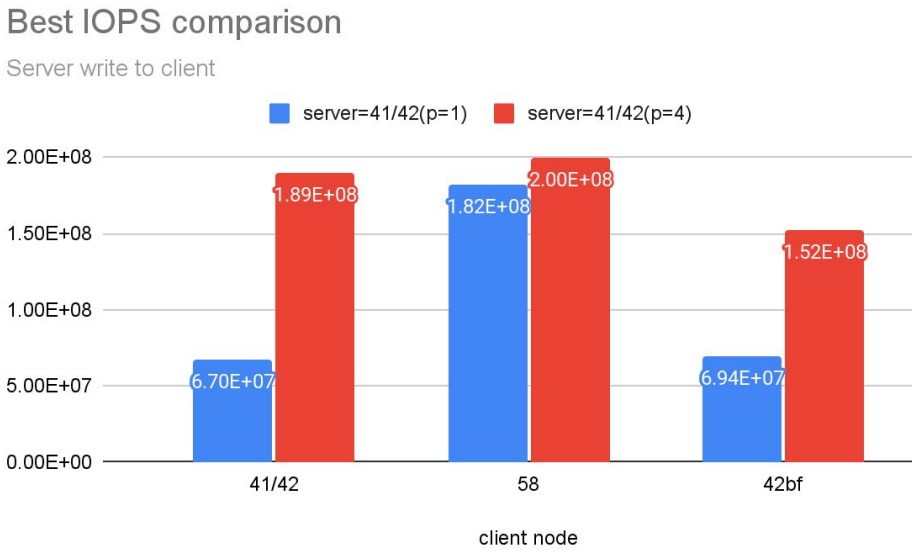
NARLabs

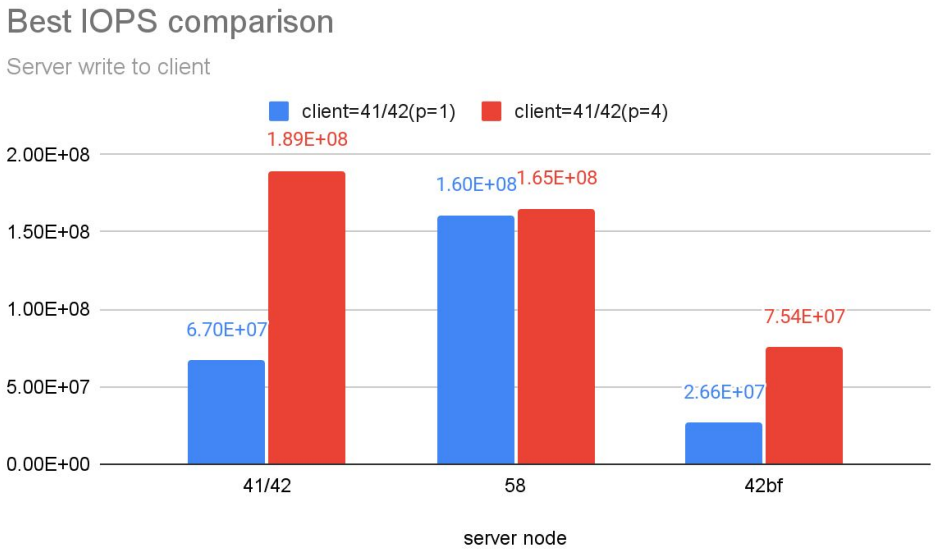
Baseline RDMA Performance: WRITE

Server WRITE to client and use same thread number.

Observation: BF is slightly worse than x86 hosts when used as client, however, the performance is not bad when written as a server.

*p: The number of RDMA work requests are included in one ibv_post_send(). Similar to the concept of batch size.

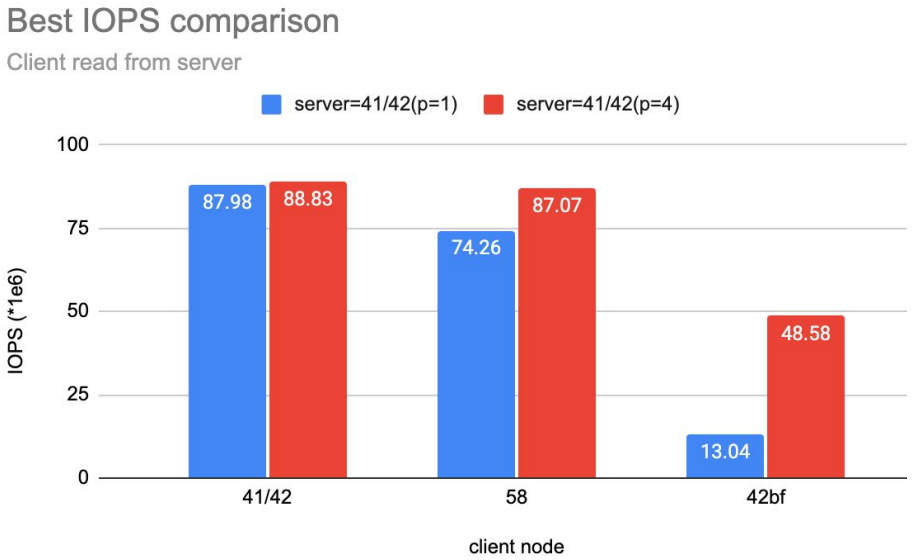


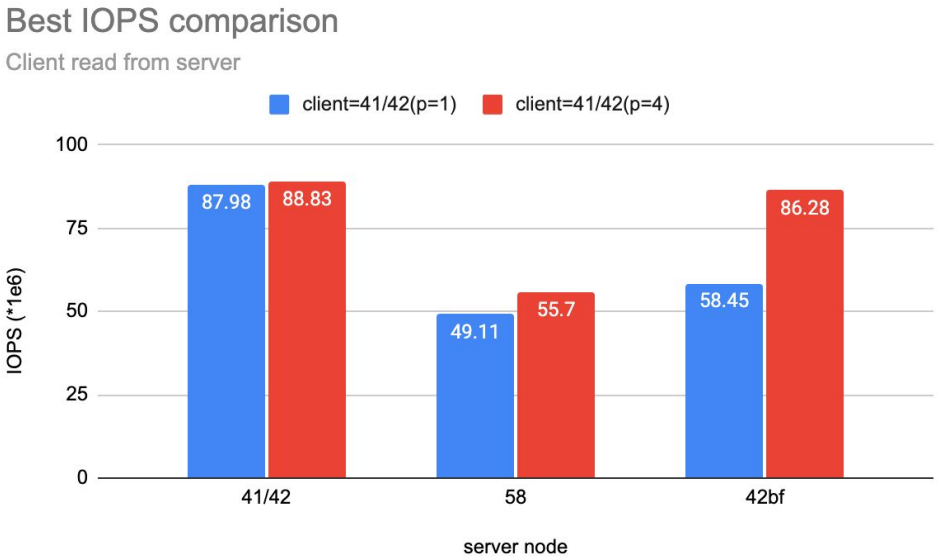


Baseline RDMA Performance: READ

Client READ from server.

Observation: When used as server, BF is comparable to x86 hosts when p is large enough.

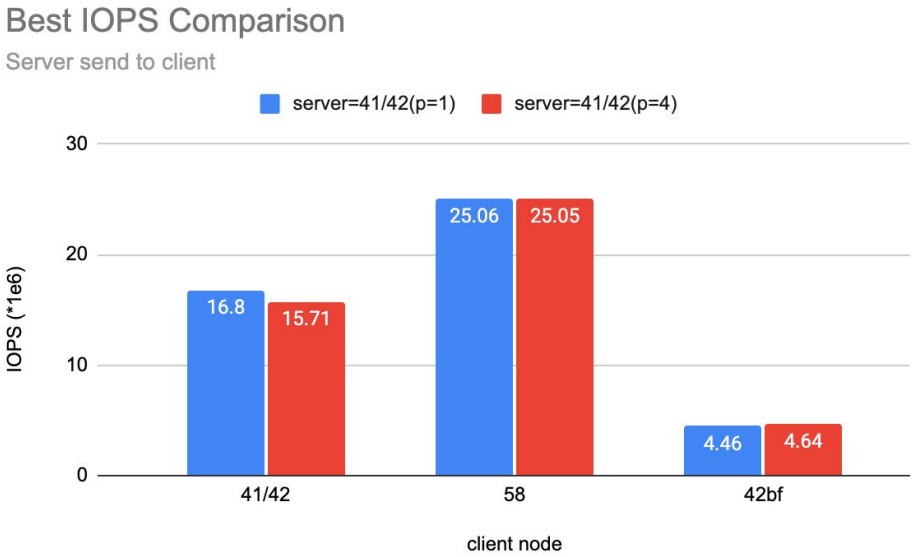


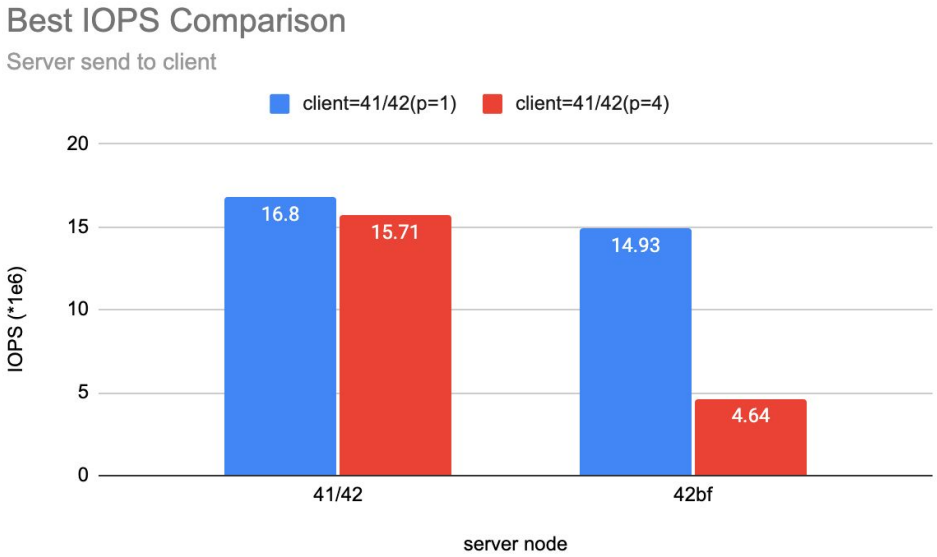


Baseline RDMA Performance: SEND

Server send to client.

- Performance is bounded by client-side performance.
- Moreover, increasing p does not improve the max performance.





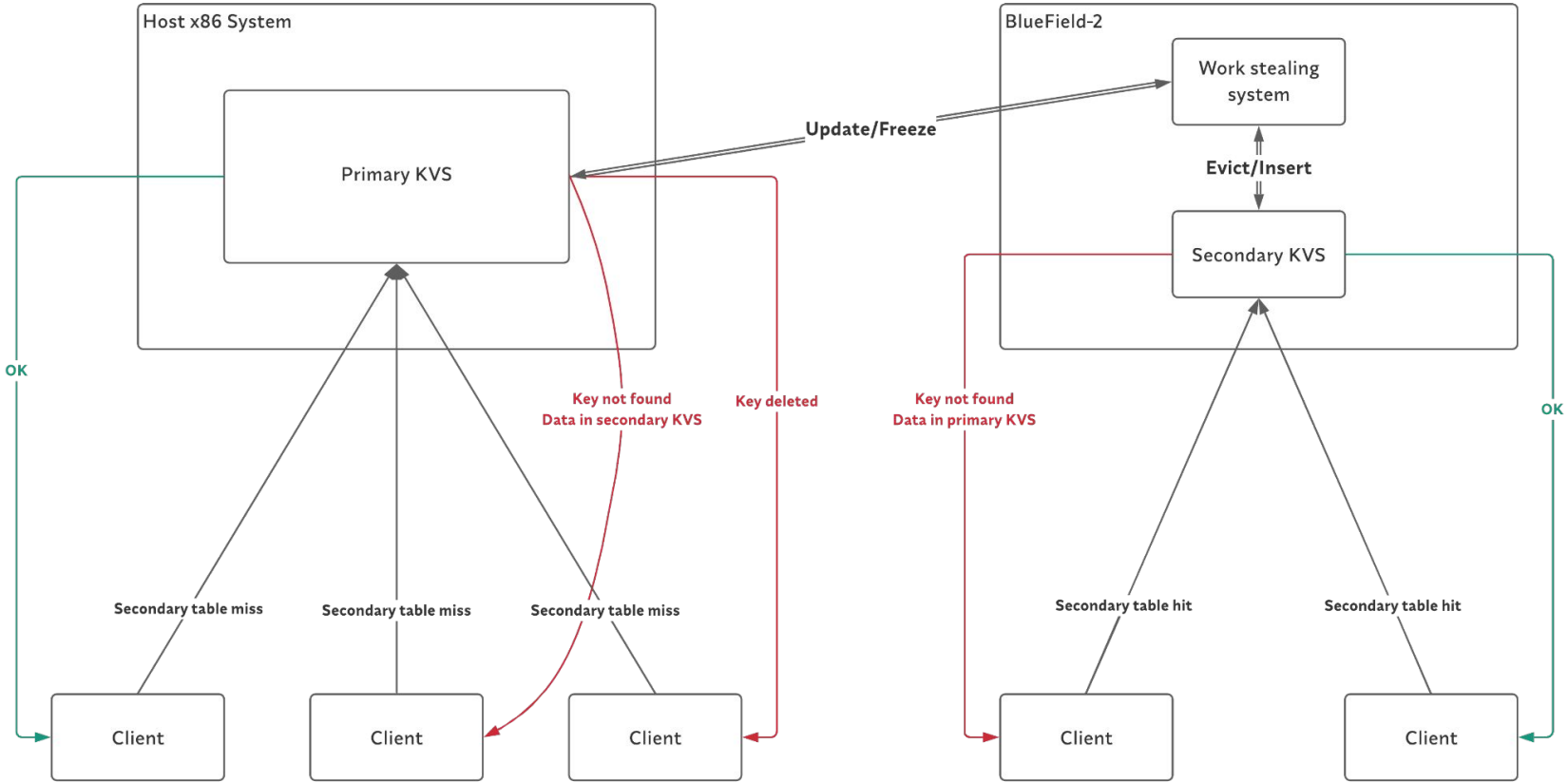
Summary: RDMA Performance

- READ and SEND operation for Bluefield as a server node is comparable to the host workstation.
- WRITE operation for Bluefield as a client is not bad.

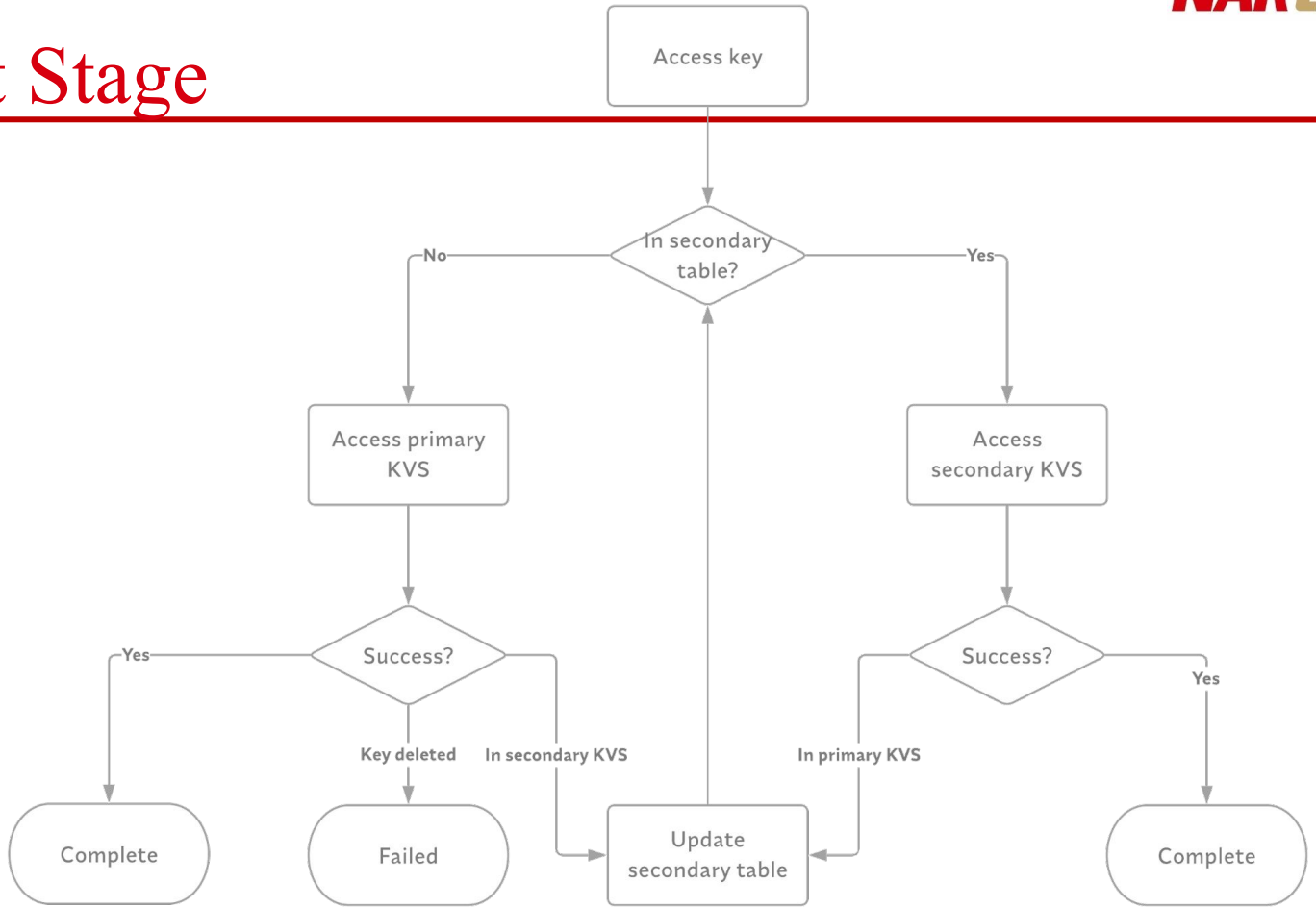
Next Stage

- A hierarchical structure consists of a main workstation and (a) slave smartNIC(s).
- Client KVS system serves a popular subset of KV pairs.
- Primary and secondary KVSs holds disjoint sets of keys.
- Special data migrate operations

Next Stage


Host x86 System
Primary KVS
BlueField-2
Work stealing system
Update/Freeze
Evict/Insert
Secondary KVS
OK
Key not found
Data in secondary KVS
Key deleted
Key not found
Data in primary KVS
OK
Secondary table miss
Secondary table miss
Secondary table miss
Secondary table hit
Secondary table hit
Client
Client
Client
Client
Client

Next Stage


Access key
In secondary table?
No
Yes
Access primary KVS
Access secondary KVS
Success?
Success?
Yes
Yes
Key deleted
In secondary KVS
In primary KVS
Complete
Failed
Update secondary table
Complete

Thank You