A DBMS-centric Evaluation of BlueField DPUs on Fast Networks

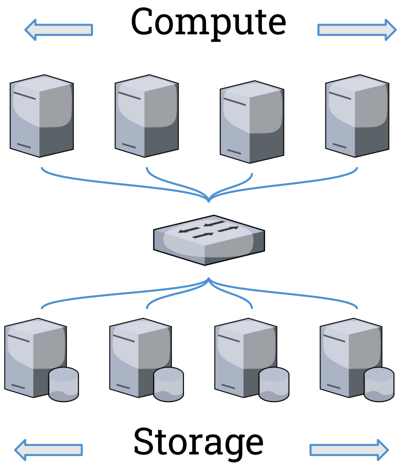
Technical University of Darmstadt

Lasse Thostrup, Daniel Failing, Tobias Ziegler & Carsten Binnig

DATA MANAGEMENT
TU DARMSTADT

State of DBMSs

- Target **disaggregated** setups
- Scale compute & memory independently
- Networks are increasingly on the **hot-path**
- Fast networks & RDMA provides **state-of-the-art** scale-out performance



Programmable Networks

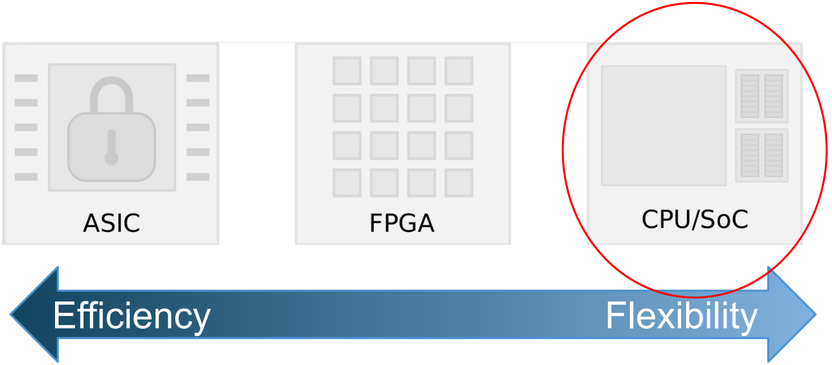
Devices

SmartNICs

Data Processing Units (DPU)

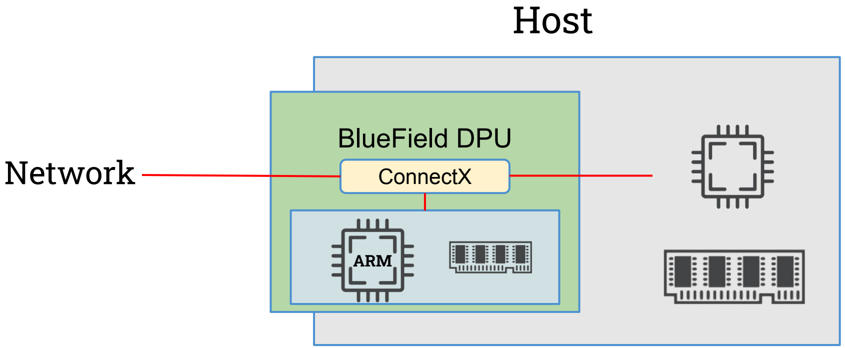
Programmable Switches

Programmability & Processing

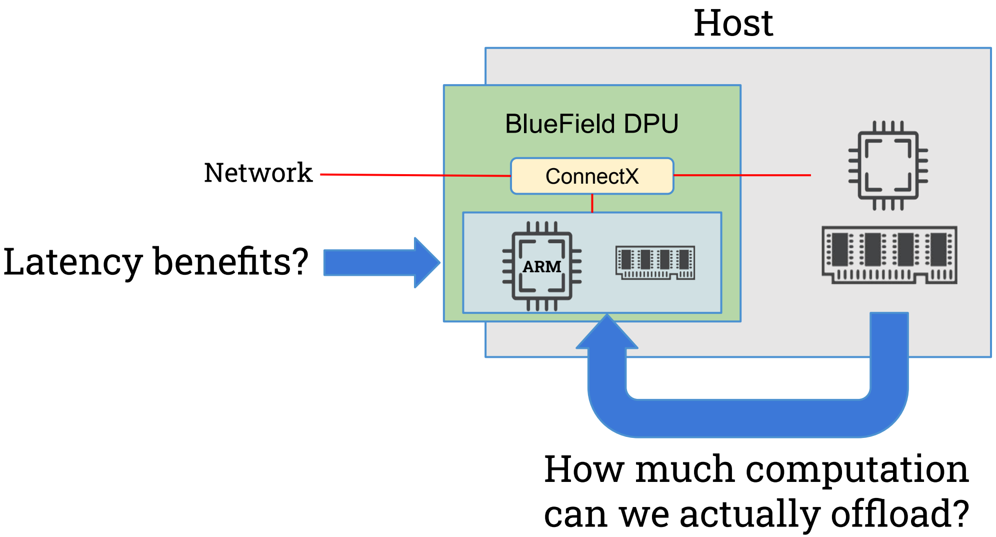


Nvidia BlueField DPU

- **Typical applications**: networking, storage, security
- Equipped with various **hardware accelerators**
- Embedded **ARM** subsystem:
    - 8 cores ARM A-72 CPU
    - 16 GB DRAM



Motivation



We consider two use-cases:

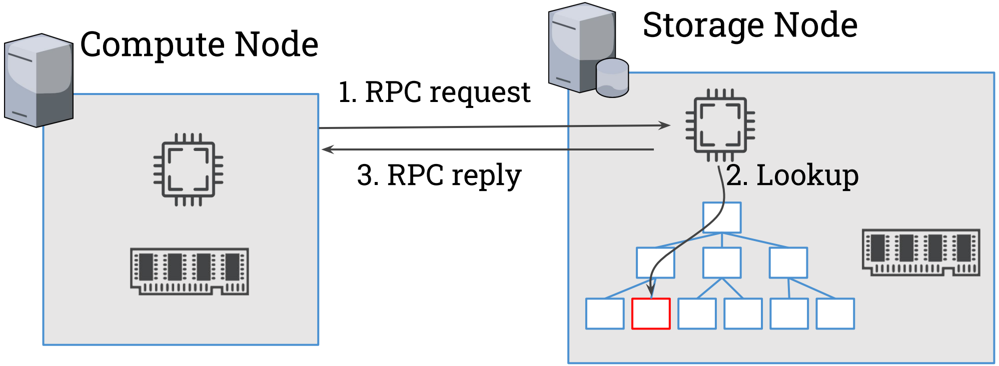
- Remote B-tree
- Remote Sequencer

Expectations:

- Added computational power and memory yields faster processing
- Closer proximity to the network yields lower latency

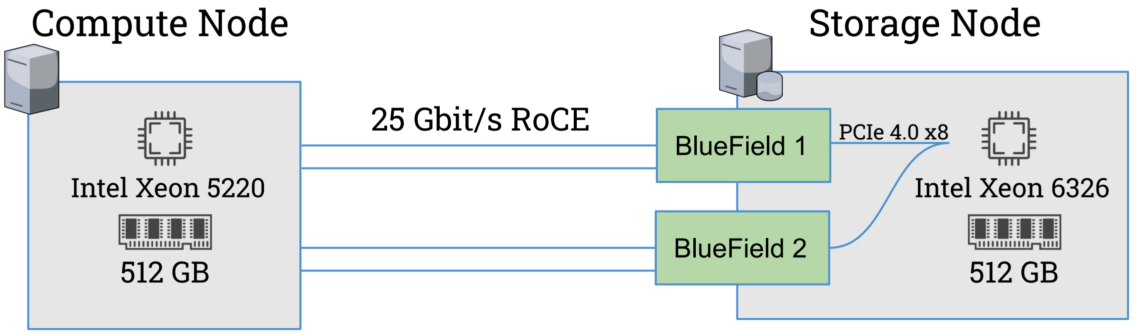
Remote B-tree

- Main data-structure in OLTP databases
- Access B-tree through Remote Procedure Calls (RPC)
- Efficient RPC-framework with state-of-the-art optimizations

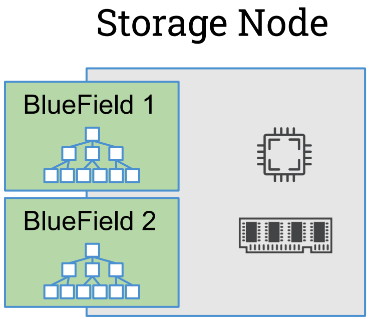


Experimental Setup

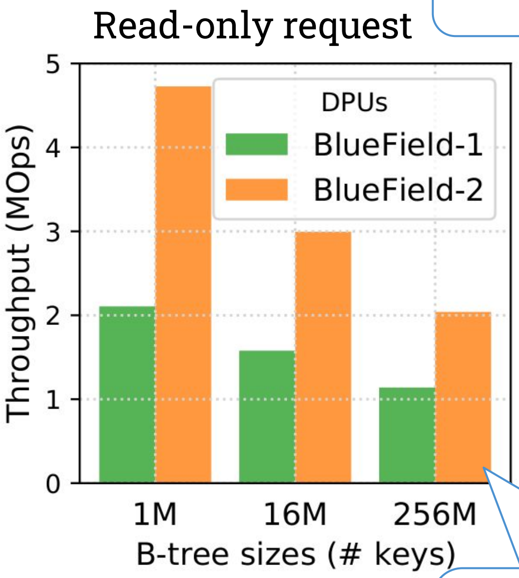
- Mirrors disaggregated memory setup
- 50 Gbit/s between Compute Node & BlueFields
- Use 8 threads on Storage Node & BlueFields

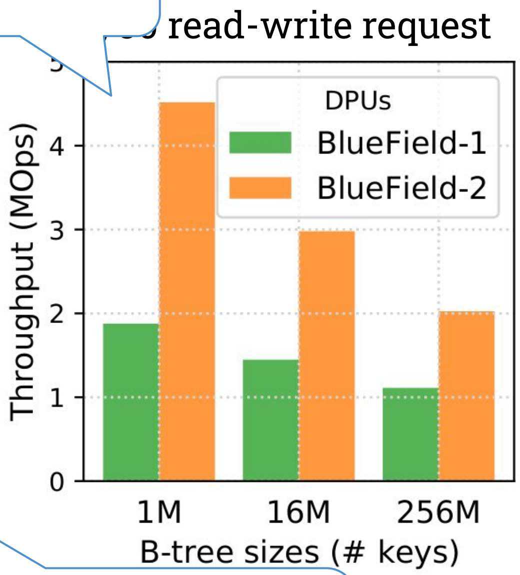


Remote B-tree w/ RPC - Throughput



How does the RPC performance compare between BlueField 1 & 2?



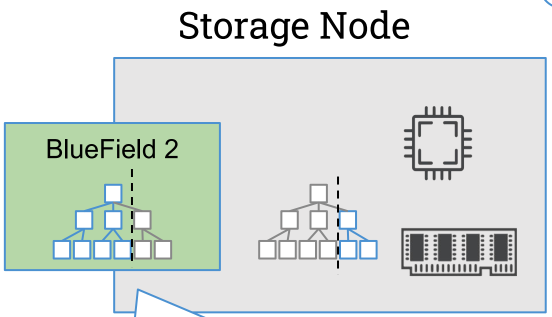


Up to 2.4x faster

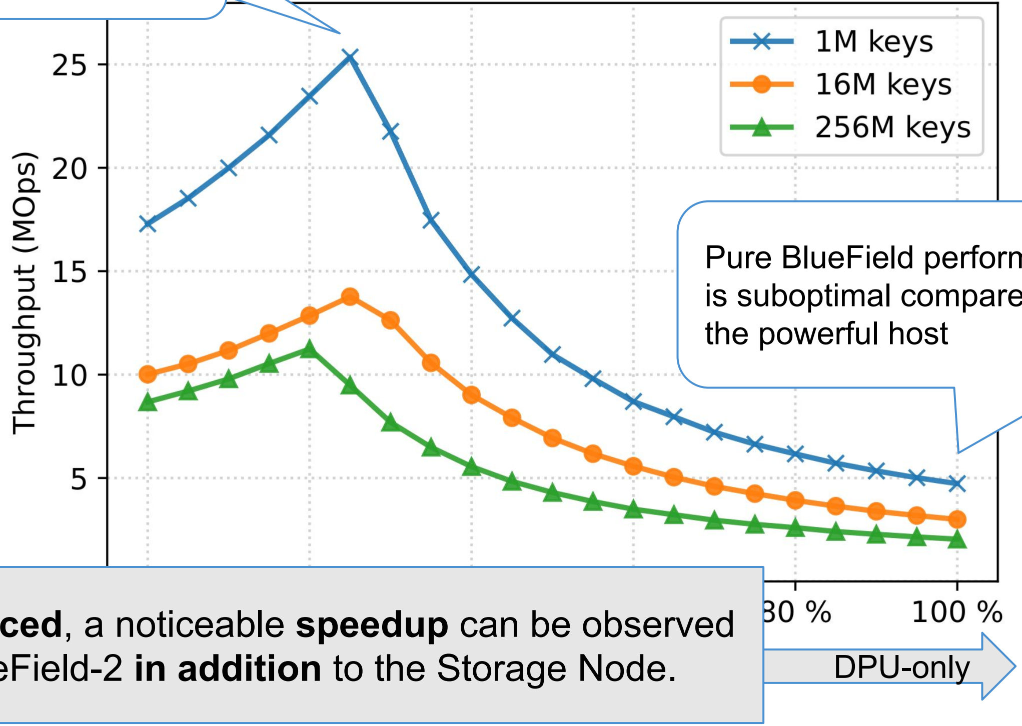
Smaller difference for bigger trees (memory bound)

Remote B-tree w/ RPC - Throughput

Throughput increase by up 47%



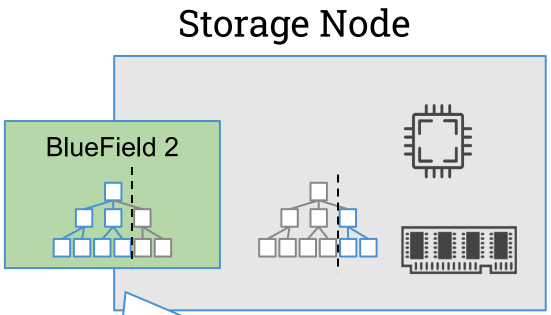
How can the BlueField 2 aid in handling B-tree RPC requests by partitioning?



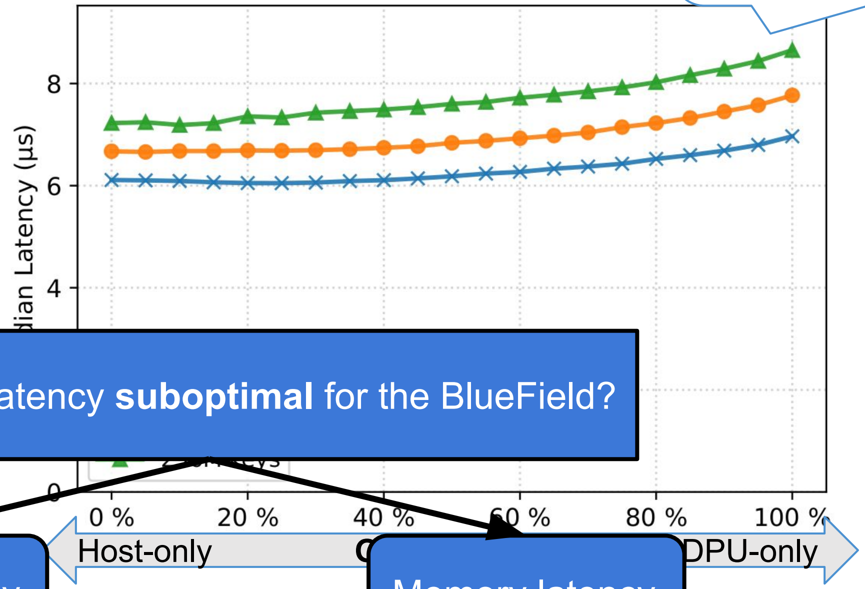
Pure BlueField performance is suboptimal compared to the powerful host

When **load is balanced**, a noticeable **speedup** can be observed by using the BlueField-2 **in addition** to the Storage Node.

Remote B-tree w/ RPC - Latency



What is the **latency** whe offloading a B-tree to the BlueField 2?



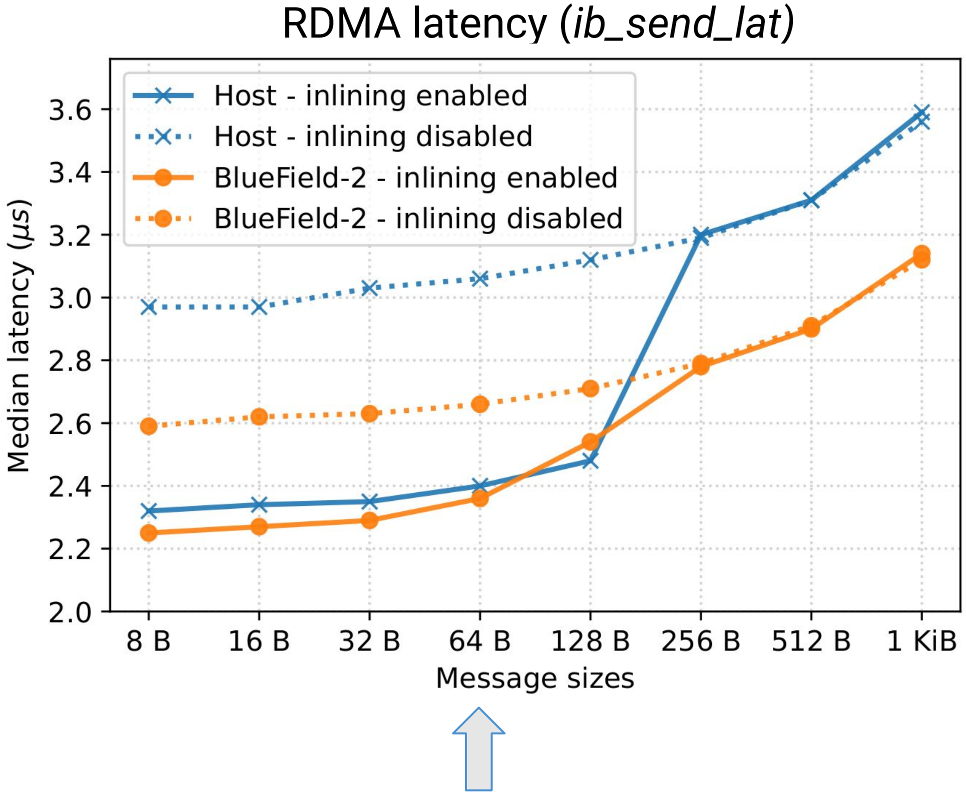
Latency increases between 14% and 20%

Why is the latency **suboptimal** for the BlueField?

Network latency

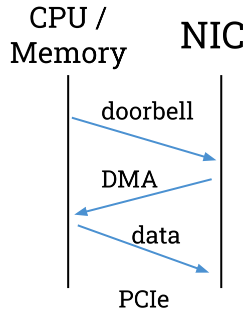
Memory latency

RDMA Latency

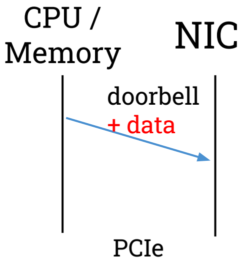


- ~0.4 µs lower latency for the BlueField without inlining
- Inlining reduces the latency gap (enabled by default)
- With 64 B messages, practically no latency difference

Without inlining



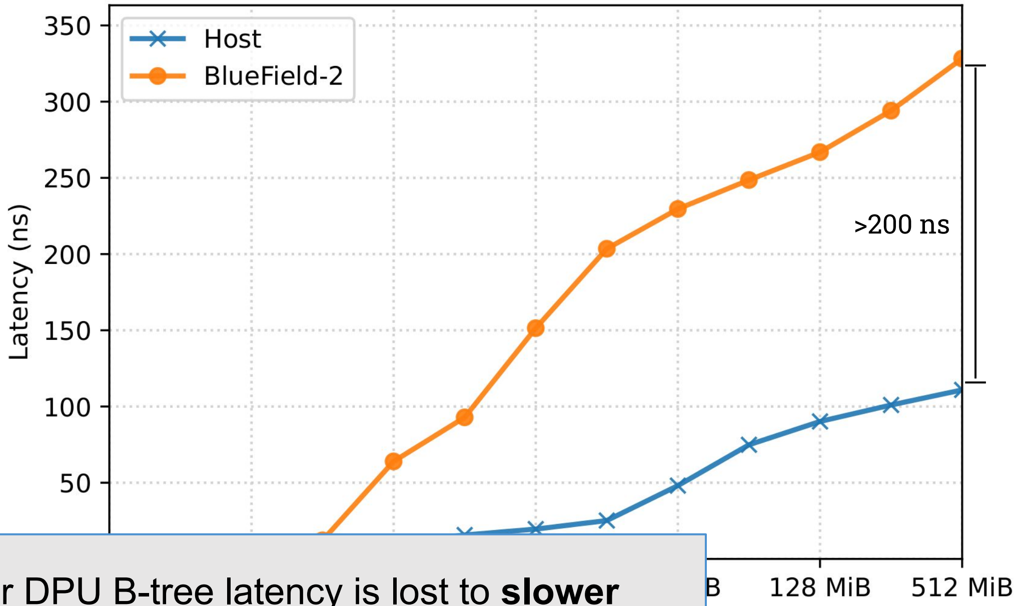
With inlining



Memory Latency

- Random reads in increasing memory block sizes
- Different cache sizes:
  - BlueField LLC: **6 MB**
  - Host LLC: **24 MB**
- BlueField-2 inhibits much higher latency when spilling out of the cache

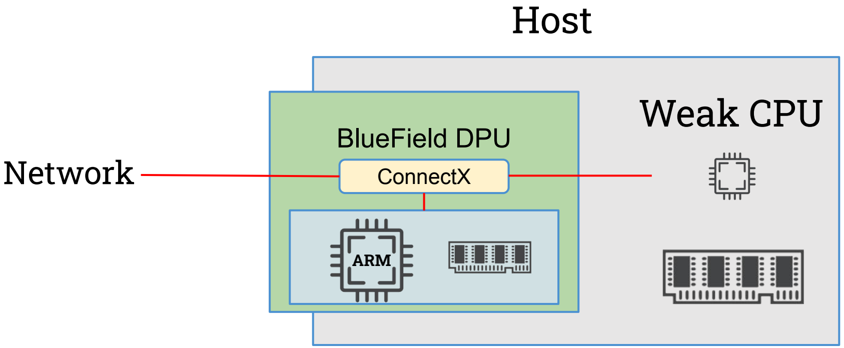
Memory latency (*tinymembench*)



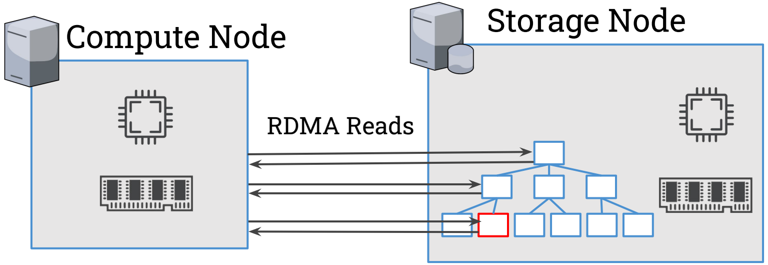
The **expected** lower DPU B-tree latency is lost to **slower memory access** in B-tree lookups and a similar network latency.

Near-zero Computation

- Storage servers typically employed with very **little compute power**
- RPC heavily involves remote CPU → Instead: **One-sided RDMA**
- One-sided RDMA accesses have **no remote CPU load**

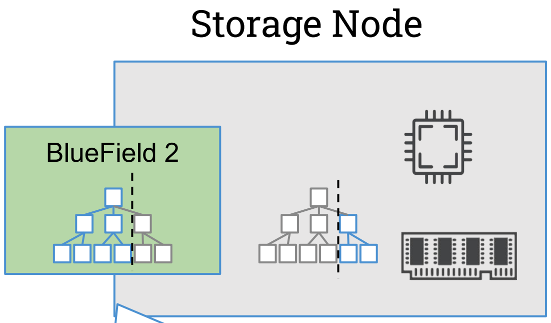


Remote B-tree w/ One-sided RDMA



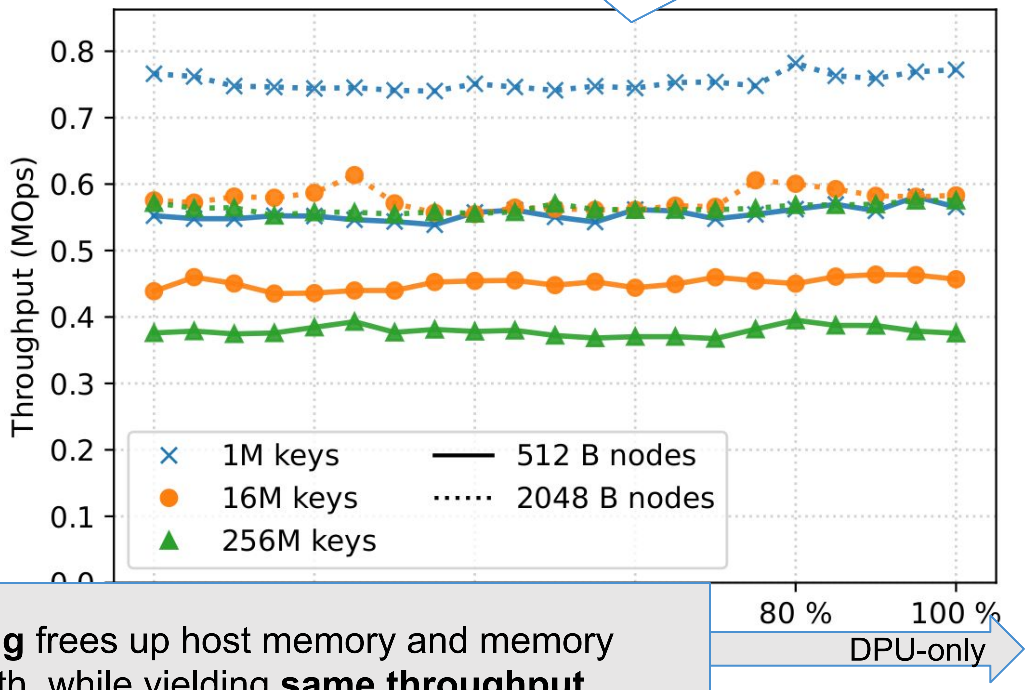
- Performs the binary search for child on Compute Node
- A network round-trip for each level
- **No CPU load** on Storage Node or BlueField

Use case 1: Remote B-tree w/ One-sided RDMA



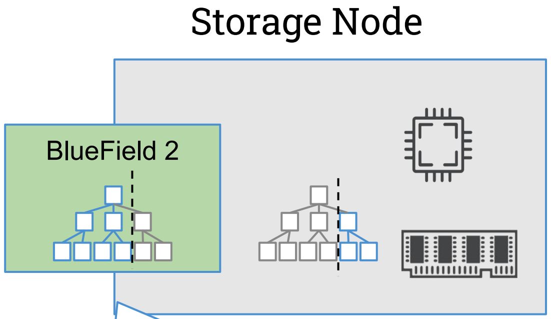
What is the **throughput** when offloading a B-tree to the BlueField 2?

Stable throughput



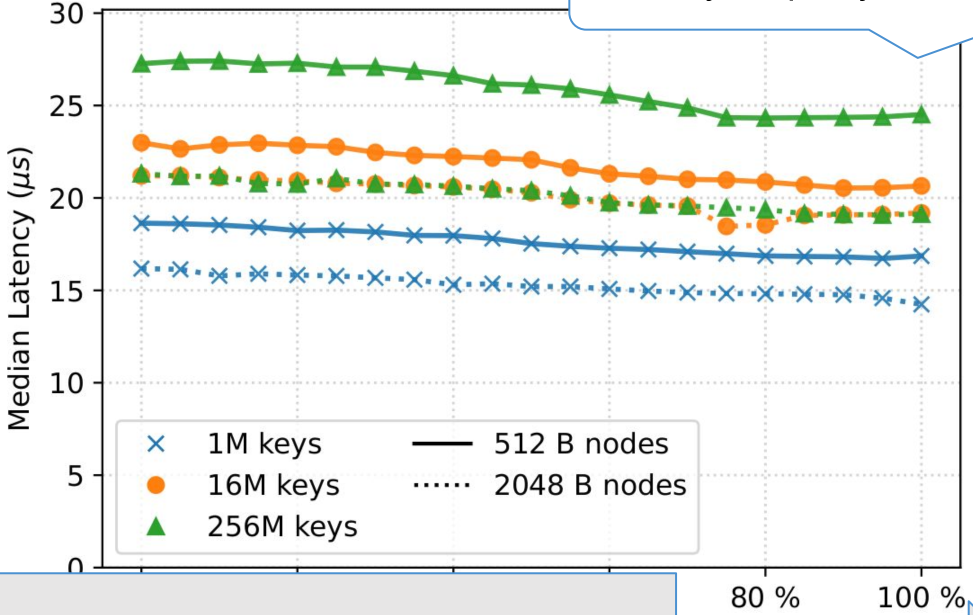
Offloading frees up host memory and memory bandwidth, while yielding **same throughput.**

Use case 1: Remote B-tree w/ One-sided RDMA



What is the **latency** when offloading a B-tree to the BlueField 2?

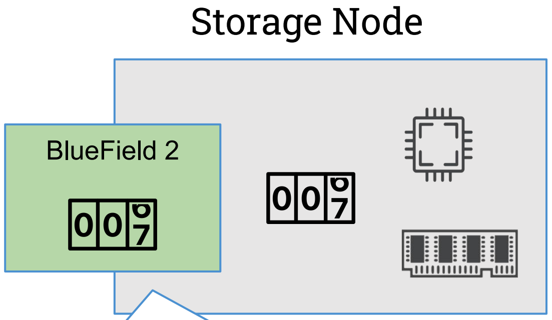
Latency drops by 11-13%



DPU-only

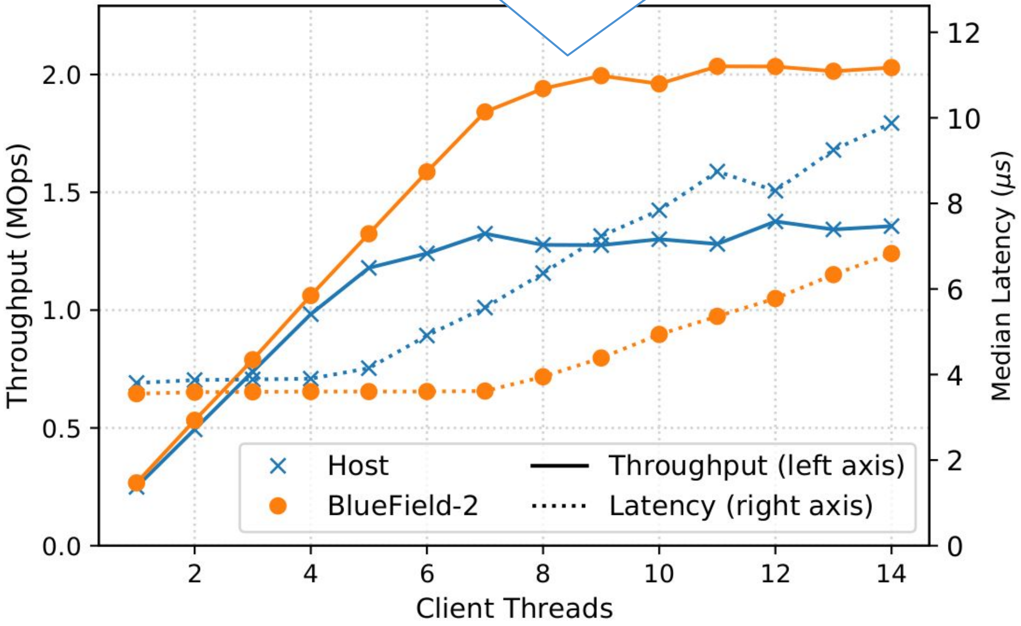
The BlueField-2 provides **competitive** offloading characteristics for **one-sided access.**

Use case 2: Remote Sequencer w/ One-sided RDMA



What is the **throughput** and **latency** when incrementing a remote counter with RDMA fetch & add?

Almost 50% throughput increase



Conclusion

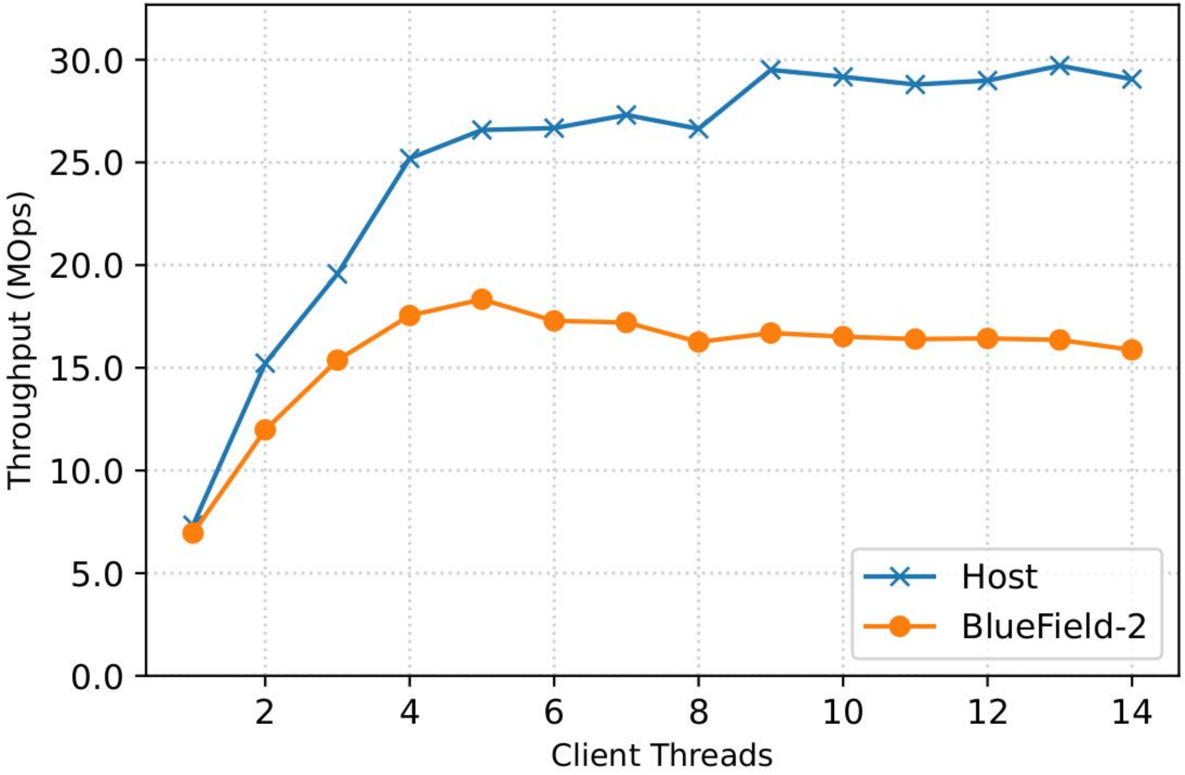
The BlueField DPU can **accelerate** typical DBMS operations

- Main benefits shown for one-sided accesses
- Two-sided operations can provide speed-ups when used in union with host

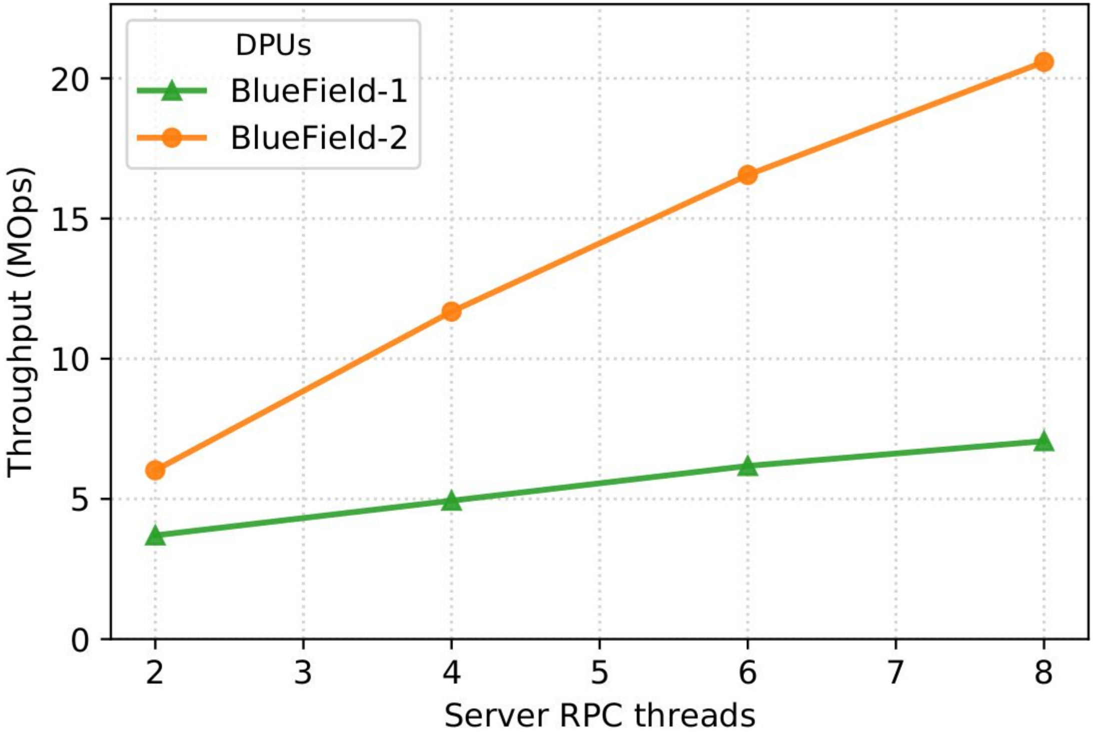
Future work

- Explore other BlueField hardware accelerators (compression, encryption & NVMe)
- BlueField 3 → faster CPU & memory

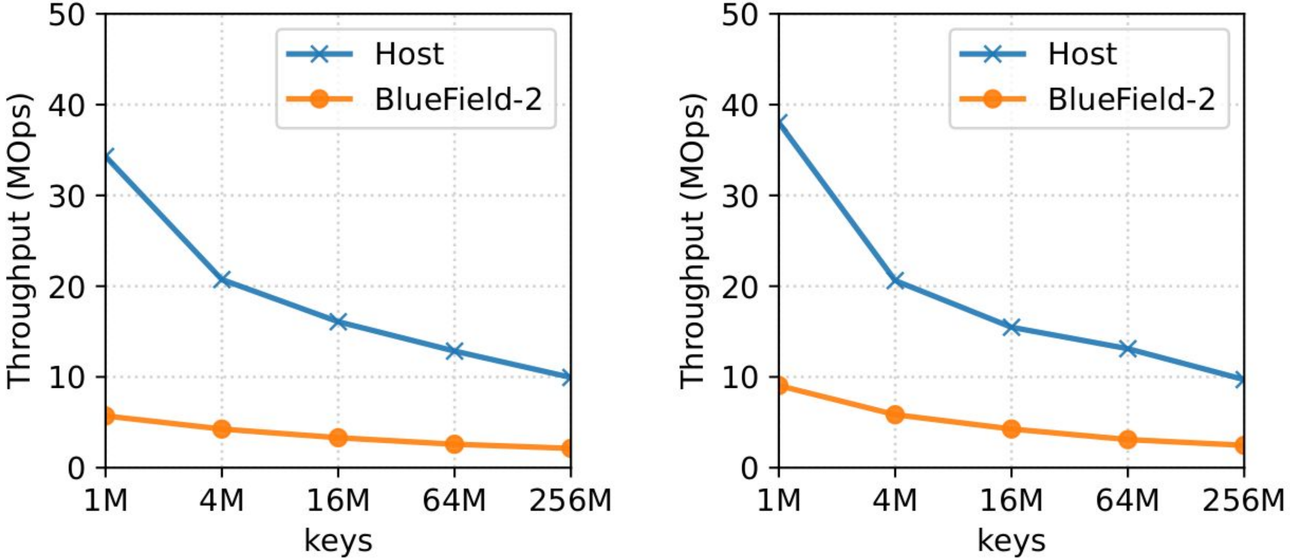
Backup: Throughput of RPC sequencer



Backup: No-Op RPC BlueField-1 vs BlueField-2



Backup: Throughput of local B-tree



(a) Updates

(b) Lookups