Characterizing Off-path SmartNIC for Accelerating Distributed Systems

Xingda Wei[1,2], Rongxin Cheng[1,2], Yuhan Yang[1], Rong Chen[1,2], and Haibo Chen[1]

[1]Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University
[2]Shanghai AI Laboratory

Abstract

SmartNICs have recently emerged as an appealing device for accelerating distributed systems. However, there has not been a comprehensive characterization of SmartNICs, and existing designs typically only leverage a single communication path for workload offloading. This paper presents the first holistic study of a representative off-path SmartNIC, specifically the Bluefield-2, from a communication-path perspective. Our experimental study systematically explores the key performance characteristics of communication among the client, on-board SoC, and host, and offers insightful findings and advice for designers. Moreover, we propose the concurrent use of multiple communication paths of a SmartNIC and present a pioneering guideline to expose new optimization opportunities for various distributed systems. To demonstrate the effectiveness of our approach, we conducted case studies on a SmartNIC-based distributed file system (LineFS) and an RDMA-based disaggregated key-value store (DrTM-KV). Our experimental results show improvements of up to 30% and 25% for LineFS and DrTM-KV, respectively.

1 Introduction

Remote Direct Memory Access (RDMA) has been widely adopted in modern data centers [23, 71, 20], pushing network bandwidth (towards 400 Gbps [44]) and distributed system performance [17, 76, 75, 64, 80, 82, 77] to the next level. However, the high-speed network requires more CPU resources to saturate a fast RDMA-capable NIC (RNIC) [38], which places a significant CPU burden on distributed systems [32]. One-sided RDMA can alleviate CPU pressures by enabling the RNIC to directly read and write host memory in a CPU-bypass way. However, the limited offloading capabilities may cause network amplifications and thus degrade system performance [61, 28].

The continuous improvements in RDMA [67] and the essential power and memory walls of CPUs have led to the emergence of SmartNICs—the RNICs with programmable capabilities. These NICs offer systems the opportunity to offload more complex computations to the NIC. Currently, there are two main types of SmartNICs. The first one is the *on-path* SmartNIC [42], which directly exposes the processing units (NIC cores) for handling RDMA packets to the systems. Unfortunately, programming low-level NIC cores with firmware [38, 61] and isolating the offloaded program from normal RDMA requests pose significant burdens on developers. To simplify system development, the *off-path* SmartNIC [52, 53, 9, 51] attaches a programmable multicore SoC (with DRAM) next to the RNIC cores, which is off the critical path of RDMA. Thanks to this separation, the SoC is independent of normal RDMA requests and can further deploy a full-fledged OS to make the developments easier [32]. Specifically, developers can treat the SoC as a separate server. In this paper, we focus on off-path SmartNIC[1] due to its generality and programmability.

There have been valuable studies on characterizing off-path SmartNICs [38, 37, 32, 68, 2], with a focus on their ability to offload computation. A key finding is that the computing power of off-path SmartNICs is weaker than that of the host [38, 37, 32]. This means that off-path SmartNICs do not improve the speed of a single network path, such as that between NIC and the host. For example, iPipe [38] found that the path between the host and SoC has a relatively high latency due to the support for more developer-friendly RDMA.

Although prior work has been valuable in utilizing SmartNICs for distributed systems, it has primarily focused on offloading computation to the SmartNIC's SoC. However, it is surprising that the fundamental function of SmartNICs, namely networking, has been overlooked despite its significant impact on overall performance. In fact, networking on the SmartNIC is intricate, because it provides multiple communication paths. For example, SmartNICs support using RDMA to access the memory of the host or SoC, as well as exchanging data between the host and the SoC.

To this end, this paper conducts the first systematic study on characterizing the performance of communication paths of SmartNIC. Unlike previous studies that simply report basic performance numbers [37, 32, 68], we systematically analyze the performance implications of SmartNIC architecture on different paths. Specifically, we investigate why and when one path may be faster than another, identify the bottlenecks for each path, examine how the heterogeneity of the SoC brings performance anomalies in paths related to the SoC, and finally explore how paths interact with each other. The main highlights of our results are:

- *Different paths exhibit diverse performance characteristics.* The RDMA path from the NIC to the SoC is up to 1.48× faster than the path to the host.

[1]This paper will use "SmartNIC" (or "SNIC" for brevity) to specifically refer to off-path SmartNICs.

- *The SoC introduces new performance anomalies to paths related to it*. The low-level hardware details of the SoC, including the memory access path and PCIe MTU, differ from those of the more powerful host CPU. Without considering such factors, RDMA requests involving the SoC suffer from up to 48% bandwidth degradation.
- *The paths between the SoC and the host may underutilize the PCIe.* RDMA from the SoC to the host (and vice versa) crosses the NIC internal PCIe twice. It can only utilize half of the PCIe bandwidth and requires processing up to 6× more PCIe packets than the others. DMA only passes the PCIe once, but it is not always faster than RDMA due to the weaker SoC DMA engine (compared to the one on the RNIC) and also suffers from packet amplifications.

Based on our performance characterization, we found that prior approaches, which mainly optimize a single path for a specific functionality of distributed systems, failed to fully exploit SmartNICs. This is because a single path cannot utilize the computing and networking capability of SmartNICs. Further, only considering a single path may ignore resource interference between different paths (e.g., the PCIe and PCIe switches). As a result, LineFS can only utilize up to 117 Gbps of bandwidth on a 200 Gbps SmartNIC. A similar issue exists in SmartNIC-based disaggregated key-value store: while choosing a path to offload all key-value (KV) store operations to the SmartNIC SoC can eliminate the network amplification in existing RDMA-based key-value stores, the wimpier computing power of SmartNIC SoC limits its overall throughput.

Based on the observations from our study, we further propose an optimization guideline to help designers smartly exploiting multiple paths of SmartNICs. Instead of optimizing distributed systems along a single path, it holistically exploits multiple paths for functionalities with different characteristics and carefully considers cross-path interference. To demonstrate the efficacy of our guideline, we conduct two case studies by optimizing two state-of-the-art systems, namely LineFS [32] and DrTM-KV [11, 76]. Due to the exposed new optimization spaces, following our guideline can improve the performance of LineFS and DrTM-KV by up to 30% and 25% accordingly.

Contributions. We summarize our contributions as follows:

- A comprehensive performance characterization of representative off-path SmartNICs, with a particular focus on various communication paths.
- The first optimization guideline for smartly exploiting the multiple paths of SmartNICs with managed cross-path resource interference.
- Two case studies on SmartNIC-accelerated distributed systems (i.e., file system and key-value store) with notable performance improvements, demonstrating the efficacy of our guideline.

Assumptions and generalizability of our work. We assume an off-path SmartNIC with the following architecture: the SoC is linked with NIC cores via a PCIe switch, and there is heterogeneity between SoC and host CPUs. We believe this is a representative architecture, as many older (e.g., NVIDIA Bluefield-1 [55], Broadcom Stingray [9]), current (e.g., NVIDIA Innova2 [51], Bluefield-2 [52]), and upcoming SmartNICs (e.g., Bluefield-3 [53], Marvell OCTEON 10 DPU [43]) use a similar setup. We conducted experiments on Bluefield-2 [52]—the state-of-the-art SmartNIC with this architecture. Meanwhile, we also confirmed that our results hold for Bluefield-1.

However, we acknowledge that significant architectural changes (e.g., on-path SmartNICs) may affect our findings. Nevertheless, we argue that our methodology—first studying the performance implications of each communication path and then smartly exploiting multiple paths of SmartNICs—can be generalized to other SmartNICs. Our benchmarking code, tools, and systems are available at https://github.com/smartnickit-project.

2 Background and Context

2.1 RDMA-capable NICs (RNICs)

RDMA (Remote Direct Memory Access) is a low-latency (2 μs) and high-bandwidth (200 Gbps) network widely adopted in modern data centers [23]. One intuitive way to utilize RDMA is to accelerate message passing with its *two-sided* primitives (SEND/RECV), such as RDMA-based RPC [27, 17, 12, 47, 30]. Alternatively, the one-sided primitives (READ/WRITE[2]) allow the RNIC to access the host memory bypassing the host CPU. Specifically, the NIC core internally uses the direct memory access (DMA) feature of the PCIe link to access the host memory (see Figure 1(a)).

Though RDMA has boosted the performance of many distributed systems [18, 76, 62, 29], usually by orders of magnitude, it still has the following two problems especially when the RNICs scale up to higher performance.

Issue #1: Host CPU occupation. For two-sided primitives, distributed systems need non-trivial CPUs to saturate a powerful NIC. Our measurements show that a 24-core server can only saturate 87 million packets per second (Mpps) on a 200 Gbps RNIC (ConnectX-6), while NIC cores can process more than 195 Mpps.[3] A recent work further shows that a distributed file system requires 2.27× CPU cores to handle network packets, when the network bandwidth scales from 25 Gbps to 100 Gbps [32]. Although deploying more powerful CPUs can alleviate this issue, RNIC bandwidth is also rapidly growing, currently reaching up to 400 Gbps [44].

Issue #2: Network amplification. Using one-sided RDMA primitives alleviates the host CPU pressure by allowing sys-

[2]We use READ/WRITE to indicate RDMA READ/WRITE in this paper.
[3]Detailed hardware setups can be found in §2.4.

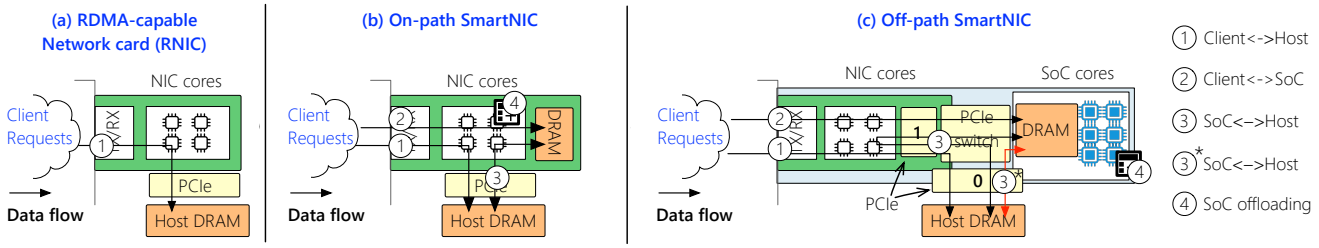

Figure 1: Architecture of different NICs: (a) RDMA-capable NIC (RNIC), (b) on-path SmartNIC, and (c) off-path SmartNIC (our focus).

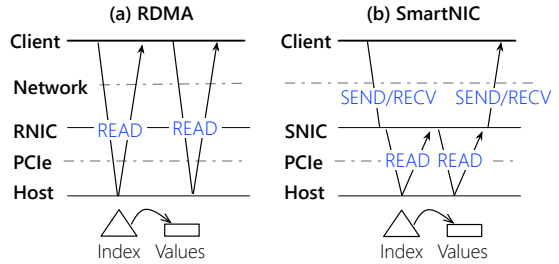

Figure 2: An illustration of a get request in a distributed in-memory key-value store that is accelerated by using either (a) RNICs (w/ network amplification) or (b) SNICs (w/o network amplification).

tems to offload memory accesses to the RNIC. However, the limited offloading capability constraints system performance, as a single request may involve multiple round trips of READs/WRITEs to complete (usually termed *network amplification*). Figure 2(a) exemplifies the execution of a *get* request on a distributed in-memory key-value store with one-sided RDMA READs. The client first uses one (or multiple) READ(s) to query the index for a given key. Based on the index returned by the previous READs, an additional READ is issued to retrieve the value.

2.2 From RNICs to SmartNICs

To address the limitations of RNICs, SmartNIC adds an on-board memory (4–64 GB) together with various computation units (e.g., SoC) to the NIC. By exposing them to the developers, SmartNIC enables offloading customized computations onto it. Specifically, SmartNICs can be categorized as follows.

On-path SmartNIC. As shown in Figure 1(b), the on-path SmartNIC exposes the NIC cores to the systems with low-level programmable interfaces, allowing them to directly manipulate the raw packets. As the name implies, the offloaded code is *on* the critical path of the network processing pipeline. Example NICs include Marvell LiquidIO [42] and Netronome Agilio [48]. The benefit is that the offloaded code is closer to the network packets. Therefore, inline requests that only interact with the NIC, such as writing to the on-board memory (②), are extremely efficient [38, 61].

However, on-path SmartNIC has two limitations. First, the offloaded code (④) competes NIC cores with the network requests sent to the host (①). If offloading too much computation onto it, the normal networking requests sent to the host would suffer a significant degradation [38]. Second, programming on-path NICs is difficult due to its low-level interface.

Table 1: Hardware description of Bluefield-2 [52].

Component	Hardware description
NIC cores	ConnectX-6 (2× 100 Gbps RDMA ports)
SoC cores	ARM Cortex-A72 processor (8 cores, 2.75 GHz)
SoC memory	1× 16 GB of DDR4-1600 DRAM
PCIe1	PCIe 4.0 ×16 (256 Gbps bandwidth)

Off-path SmartNIC. As shown in Figure 1(c), the off-path SmartNIC offers an alternative: it packages additional compute cores and memory in a separate SoC next to the NIC cores. Therefore, the offloaded code is *off* the critical path of the network processing pipeline. From the NIC perspective, the SoC can be viewed as a second full-fledged host with an exclusive network interface. To bridge the NIC cores, SoC and host together, a PCIe switch is integrated inside the SmartNIC to properly dispatch network packets. Example NICs include NVIDIA Bluefield [52, 53] and Broadcom Stingray [9].

Compared to the on-path counterparts, the offloaded code does not affect the network performance of the host as long as it does not involve network communications (②). Thanks to this clear separation, the SoC can run a full-fledged kernel (e.g., Linux) with a full network stack (i.e., RDMA), simplifying system development and allowing for offloading complex tasks [32]. However, accelerating distributed systems with off-path SmartNICs is typically more challenging than using the on-path counterparts. This is because the PCIe switch prolongs all communication paths (i.e., ①, ②, and ③), causing potential performance degradation.

2.3 Target SmartNIC: NVIDIA Bluefield-2

We conduct our study on Bluefield-2, a typical off-path SmartNIC optimized for offloading general-purpose computations. Figure 1(c) illustrates its overall hardware architecture, with detailed hardware configuration shown in Table 1.

Hardware. Bluefield-2 equips a mature RNIC (ConnectX-6) as its NIC cores for high-speed networking. These cores support all RDMA operations. Its programmability comes from an integrated on-borad SoC, which has 16 GB DRAM and an ARM Cortex-A72 (8 cores, 2.75 GHz). A PCIe 4.0 switch bridges the NIC cores, SoC and host together, enabling

Table 2: Machine configurations in our two rack-scale RDMA-capable clusters.

Name	Nodes	RDMA-capable NIC	Host PCIe (PCIe0)	Host CPU	Host Memory
SRV	3	1× ConnectX-6 (200 Gbps) 1× Bluefield-2 (200 Gbps)	PCIe 4.0 ×16 (256 Gbps)	2× Gold 5317 v4 (12 cores, 3.6 GHz)	128 GB DDR4-2933
CLI	20	1× ConnectX-4 (100 Gbps)	PCIe 3.0 ×16 (128 Gbps)	2× E5-2650 v4 (12 cores, 2.2 GHz)	96 GB DDR4-1600

bi-direction data transfer of up to 256 Gbps. Note that the SoC is linked to the PCIe switch via an internal link, rather than through PCIe.[4] Specifically, the hardware counters provided by Bluefield [54] also imply that it has only two PCIe links: one linking RNIC with the switch (PCIe1) and the other linking the switch with the host (PCIe0).

Software. The SoC runs a full-fledged Linux, allowing developers to treat it as a normal ARM server. The kernel also hosts a full RDMA stack, making it convenient for enabling RDMA-based communication. In addition, Bluefield provides DOCA [57] SDK for advanced usage, such as DMA.

Communication primitives: RDMA and DMA. All communication paths related to the SoC are conducted using RDMA to simplify system development. As shown in Figure 1(c), clients can issue one-sided or two-sided RDMA requests to the SoC (②), similar to a twin server on the host. Meanwhile, the SoC can also interact with the host via RDMA, and vice versa (③). However, exchanging data between the SoC and the host must pass through the RNIC (PCIe1 and NIC cores) for RDMA support, which adds a hidden bottleneck to this path. Fortunately, we found that Bluefield further provides DMA support (③*) with DOCA [57], allowing the SoC to use DMA to access the host memory (and vice versa), bypassing the RNIC.

Existing state of exploring Bluefield. Previous studies [38, 37, 32, 68] have mainly focused on the computing power of Bluefield (④ in Figure 1), revealing the relative weakness of the SoC cores in terms of performing offloaded tasks and sending network requests. This is because the frequency and number of cores are inferior to those of the host CPU. Due the power constraints of SmartNICs, it is unlikely that the relative performance comparison between the NIC and host CPU will change. Hence, we take this as a premise during our investigation.

In contrast, few studies have considered various communication patterns in Bluefield (i.e., ①, ②, and ③), which are the main focus of our work. Thostrup *et al*. [68] found that accessing the SoC memory (②) using READ is faster than accessing the host memory (①) in the same way. iPipe [38] shows that using RDMA to communicate between the host and SoC (③) has high latency due to the software overhead of supporting RDMA. This paper systematically explores the performance characteristics of Bluefield and summarizes insightful lessons and advice for future system developers.

[4]This has been confirmed by the NIC vendor.

2.4 Notation and testbed

Notations. This paper follows Bluefield's hardware specification when describing low-level hardware details related to Bluefield-2. As shown in Figure 1(c), "PCIe1" refers to the PCIe link connecting the NIC cores to the PCIe switch, and "PCIe0" refers to the link connecting the switch and the host's PCIe controller. The ARM cores, along with the on-chip memory of Bluefield-2, are collectively referred to as "SoC." The machine hosting Bluefield-2 is referred as the "host." Furthermore, we use the terms "requester" and "responder" to refer to the machine issuing the RDMA requests and the destination hardware component, respectively. For example, in Figure 1(c), the requesters of paths ① and ② are any RDMA-capable machines (also called clients), and the responders are the host and SoC, respectively. For path ③, the requester and responder are the host and SoC, respectively, and vice versa.

Testbed. Table 2 presents the machine configurations in our testbed. To best utilize SmartNIC, we deploy Bluefield-2 on the servers (SRV) with matching PCIe link (PCIe 4.0) by default. These machines can replace Bluefield-2 with 200 Gbps ConnectX-6 (RNIC) for comparisons. Other machines (CLIs) serve as clients that issue RDMA requests to the servers. All machines in SRV and CLIs are connected through a Mellanox SB7890 100 Gbps InfiniBand Switch. Note that the network performance of the evaluated 200 Gbps NIC is not limited since they connect to the switch with two 100 Gbps ports.

Table 3: The findings and advice from our study. Claims supported by sufficient evidence are denoted by **E**, while those supported by hypotheses are denoted by **H**.

SNIC Paths	Findings/Advice	E/H
① (§3.1)	Throughput of RDMA is lower than RNIC	H
	Latency of RDMA is higher than RNIC	E
② (§3.2)	One-sided RDMA performance is better	H
	Avoid memory accesses to close addresses	E
	Avoid large READ requests	H
③/③* (§3.3)	RDMA overuses the PCIe bandwidth	E
	Avoid large READ/WRTIE requests	H
	Enable doorbell batching carefully for RDMA	E
	Use DMA (③*) to improve PCIe utilization	E
①+② (§4.1)	Improve throughput by using paths ① and ② concurrently (esp. in opposite directions)	H
①/②+③ (§4.1)	Selectively offload traffic to ③	E

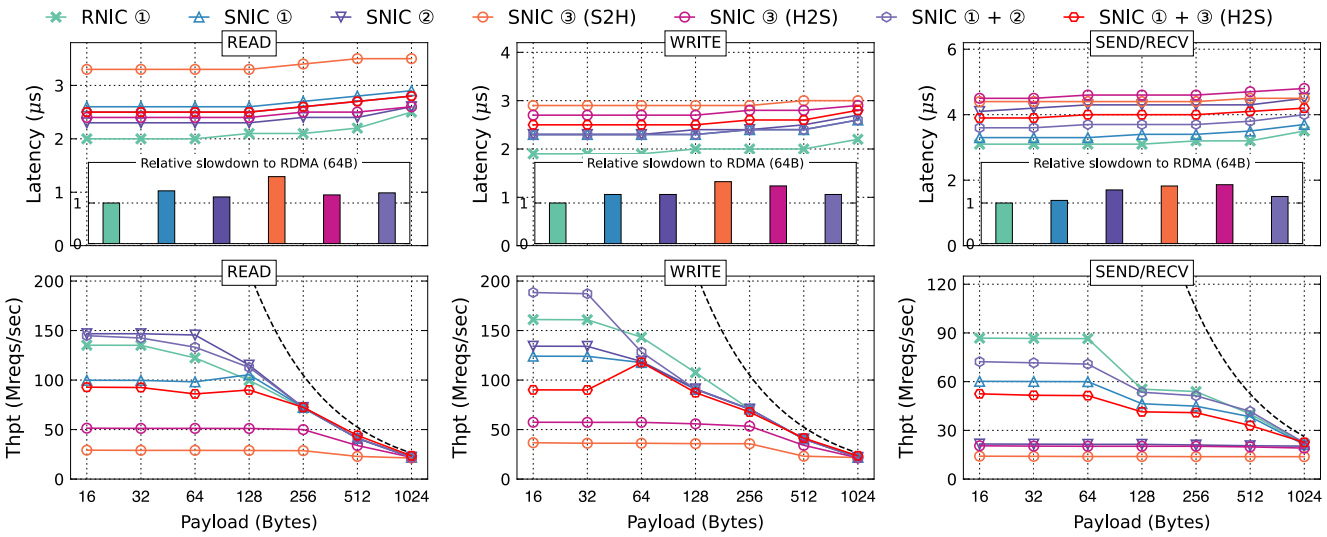

Figure 3: The end-to-end latency (upper) and peak throughput (lower) of random inbound RDMA requests on different NICs. The symbols ①, ②, and ③ in the legend correspond to the communication paths listed in Figure 1.

3 Characterizing SmartNIC Performance

As mentioned in §2.3, it is well-known that the *computing power* of NIC is wimpier than that of the host CPU. Therefore, we focus on analyzing the *communication efficiency* of SmartNIC. Figure 3 shows the end-to-end latency and peak throughput of sending different RDMA requests (e.g., READ, WRTIE, and SEND/RECV) using either RNIC or SmartNIC through different communication paths.

Evaluation setup. We conducted our experiments on the clusters described in Table 2, using a state-of-the-art RDMA communication framework [76]. For one-sided operations (READ and WRITE), the requester communicates with one responder using RDMA's reliable connection (RC) queue pairs (QPs). The responder addresses are randomly chosen from a 10 GB address space by default. For two-sided operations (SEND/RECV), the responder implements an echo server that utilizes all available cores for handling messages, and the requester communicates with it via unreliable datagram (UD) QPs for better performance [29, 76, 30]. For end-to-end latency, we deploy one requester machine to prevent interferences from queuing effects. For peak throughput, we use up to eleven requester machines to saturate the responder. Finally, we enable all well-known optimizations, including address alignment [81], unsignaled requests [27] and huge pages [17] to prevent side effects from misusing RDMA.

3.1 Communication from Client to Host (path ①)

Latency. To compare communication with the host, we conduct an apple-to-apple comparison between Bluefield-2 (SNIC ①) with ConnectX-6 (RNIC ①), as they share the same NIC cores [52]. Their performance gap best illustrates the "performance tax" paid by the SmartNIC architecture. As shown in Figure 3, SNIC ① has 15–30%, 15–21%, and 6–9% higher latency than RNIC ① for READ, WRITE, and SEND/RECV, respectively. The increased latency on SNIC comes mainly from the PCIe switch and PCIe1 between the host and NIC cores. The one-way PCIe latency is approximately 300 ns, which is non-trivial for small RDMA requests (1–2 μs). Note that the result is measured indirectly. Specifically, the end-to-end read latency on SNIC and RNIC is 2.6 μs and 2.0 μs, respectively. Compared to RNIC, READ on SNIC passes through the PCIe switch twice (see Figure 4). Thus, the cost of each pass is around 300 ns, which matches the number reported in recent literature [69]. Furthermore, the increased latency of WRITE on SNIC is lower than that of READ, because it omits one pass through PCIe switch for completion [49]. The latency of SEND/RECV on SNIC also increases, but mainly due to the larger CPU costs at the responder; the latency to post a request (via MMIO) on SNIC is higher than RNIC (399 cycles vs. 279 cycles).

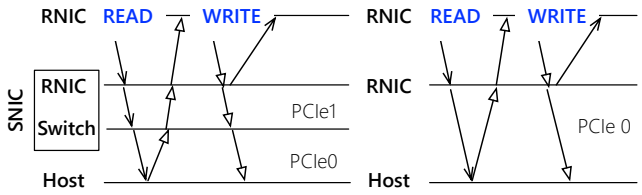

Figure 4: The exec. flow of READ/WRITE on SNIC and RNIC.

Throughput. As shown in Figure 3, for READ, WRITE, and SEND/RECV, SNIC ① has 19–26%, 15–22%, and 3–36% lower throughput than RNIC ① for payloads less than 512 bytes, respectively. We suspect the lower throughput is due to the longer latency in processing RDMA requests caused by PCIe switch. However, for larger requests, the results are similar to using RNIC as both are bottlenecked by the network bandwidth.

Figure 5: (a) An illustration of three examples of different combinations of data flows on different paths, and (b) the peak throughput of different combinations of data flows on different communications paths.

Bottleneck. The lowest bandwidth limit of NIC, PCIe1, and PCIe0 will first become the bottleneck for communication from client to host. On our testbed, the bottleneck is the network: 200 Gbps. On the other hand, we find an interesting phenomenon: the total inbound bandwidth of the requester can approach *twice* the limit—400 Gbps—because the links are *bi-directional* [58]. Specifically, if packets flow in opposite directions, e.g., the READ and WRITE packets in Figure 5(a), they can be multiplexed on the same link. To illustrate this, we dedicate two requesters (each with 12 threads to saturate the one-way bandwidth) to issue 4 KB packets. As shown in Figure 5(b), if two clients send READ and WRITE requests separately, a total of 364 Gbps bandwidth is measured on a 200 Gbps NIC (see READ+WRITE of SNIC ①). In contrast, if both clients send the same type of requests (either READ or WRITE), only about 190 Gbps is measured. Note that though this phenomenon is widely known in traditional networking (i.e., messaging), where the messages are typically two-sided, it is largely ignored by many RDMA-based systems, because RDMA request can be one-sided.

Takeaways. Being "smart" incurs performance degradation for communicating with the host for small requests. For small requests, we demonstrate that extending RNIC (ConnectX-6) to SNIC (Bluefield-2) causes performance degradation by up to 36% and 30% in throughput and latency, respectively. In general, for distributed systems that only use the path ①, it is recommended to use RNIC. Although the overhead may be negligible for large requests or for networking with longer latency, RNIC is cheaper and more energy-efficient than SNIC.

3.2 Communication from Client to SoC (path ②)

Latency. For sending requests from the client to SoC (SNIC ② in Figure 3), the latency of READ decreases by up to 14% compared to the host (SNIC ①). The *reason* is that it skips PCIe0. Yet, it is still 4–15% higher than RNIC, because requests still must go through the PCIe switch at PCIe1. For WRITE, SNIC ② provides similar performance as SNIC ① due to the asynchronous completion of cores (see Figure 4). For SEND/RECV, SNIC ② has 21–30% higher latency than SNIC ① due to the weaker computing power of SoC.

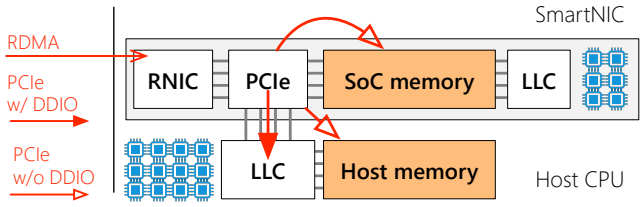

Figure 6: Different paths to access host and SoC memories.

Throughput. SNIC ② has better throughput than SNIC ①, reaching 1.08–1.48× for payloads less than 512 bytes. Interestingly, the READ of SNIC ② is even higher than that of RNIC ① before reaching the peak network bandwidth. For this undocumented results, we suspect that it is due to the closer packaging of SoC memory and the PCIe switch. Specifically, the SoC is linked to the PCIe switch via an internal link, rather than through PCIe. Note that a confident analysis relies on the hardware details of Bluefield, which unfortunately are not available now. For WRITE, SNIC ② is still lower than that of the RNIC ①. Our *hypotheses* are twofolds. First, SoC has fewer DRAM channels compared to the host (1 vs. 4), limiting the concurrency of write accesses. Nevertheless, READ is not affected because read accesses on DRAM are faster than write accesses [25, 73]. Second, SoC can only utilize a portion of NIC cores (see §4.1). Finally, SEND/RECV has a poor performance on ②: it just achieves up to 64% of the host (SNIC ①). This is due to the wimpy computing power of SoC, since the throughput of SEND/RECV is bottlenecked by the responder CPU to send the reply.

Bottleneck. As shown in Figure 1(c), since SNIC ② only flows through NIC and PCIe1, the bottleneck is their lower bandwidth limit, which is still Bluefield-2's 200 Gbps NIC. Therefore, as shown in Figure 5(b), the performance of SNIC ② is the same as that of SNIC ①, namely the total of 400 Gbps and 200 Gbps bandwidth for opposite direction and same direction communication, respectively.

In addition to the basic RDMA performance of the SNIC, we found several factors that could also prevent distributed systems from achieving the aforementioned performance.

Advice #1: Avoid memory accesses to close addresses. The

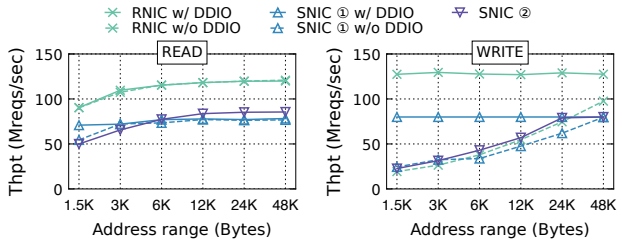

Figure 7: The peak throughput of accessing host memory and SoC memory via SNIC, READ (a) and WRITE (b).

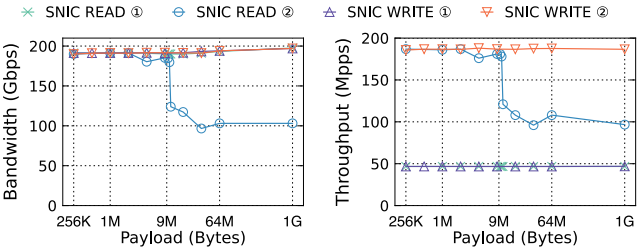

Figure 8: The bandwidth (a) and PCIe packet throughput (b) for accessing (READ and WRITE) the host (SNIC ①) and SoC (SNIC ②) via SmartNIC. For brevity, we omit the result of SEND/RECV since it is the same as WRITE for large payloads [27].

wimpy SoC cores may impact the memory access behavior of one-sided RDMA primitives, because it usually supports fewer features compared to the more powerful host CPU cores. Specifically, Data Direct I/O (DDIO) [26] is widely supported by the host CPUs, which allows the NIC to directly read/write data from/to its last level cache (LLC), as shown in Figure 6. SoC cores may also equip with similar features (e.g., ARM CCI [5]), but whether to do so is vendor-specific. The SoC cores of our hardware (ARM Cortex-A72 in Bluefield-2) do not support DDIO. We find that one-sided RDMA without DDIO suffers performance drop if the requested memory addresses fall into a small range (i.e., they are close together). This is *because* DRAM requires a (not-too-small range) to utilize all memory modules concurrently. LLC is faster than DRAM, so we suspect the impact is smaller.

Figure 7 shows the peak throughput of accessing host memory and SoC memory via SNIC with the increase of address ranges.[5] For WRITE, the throughput of SNIC ② using SoC drops to 22.7 M reqs/s (from 77.9 M reqs/s) when address range decreases to 1.5 KB (from 48 KB). In contrast, the performance of SNIC ① using Host CPU is hardly affected when DDIO is enabled. For READ, the degradation is relatively smaller. The throughput of SNIC ② drops from 85 M reqs/s to 50 M reqs/s when decreasing the range from 48 KB to 1.5 KB. This is because DRAM can serve reads faster than writes [25, 73]. Finally, we also plot the RNIC results as a reference. When requests addresses are close, we can see that ① also suffers a significant performance drop on WRITE when DDIO is disabled.

[5]Note that we attach Bluefield to CLI machines for the evaluation because we are unable to disable DDIO on the SRV machines.

Table 4: PCIe Maximum Transfer Unit (MTU) on our testbed, and the number of PCIe packets required to transfer N bytes via different communication paths of Bluefield-2. Our simplified model omits control-path packets (e.g., two-sided message arrival notification).

	Host CPU cores (H_{MTU})		**SoC cores** (S_{MTU})
PCIe MTU	512 B		128 B
	SNIC ①	**SNIC ②**	**SNIC ③**
PCIe1	$\lceil N/H_{\text{MTU}} \rceil$	$\lceil N/S_{\text{MTU}} \rceil$	$\lceil N/H_{\text{MTU}} \rceil + \lceil N/S_{\text{MTU}} \rceil$
PCIe0	$\lceil N/H_{\text{MTU}} \rceil$	–	$\lceil N/H_{\text{MTU}} \rceil$

Advice #2: Avoid large READ requests. It is common practice to use requests with large payloads to fully exploit network bandwidth. For example, using requests with payloads larger than 16 KB is enough to saturate a 200 Gbps RNIC even using a few threads. Unfortunately, we observed that the READ performance of SNIC ② collapses with request payload larger than 9 MB, as shown in Figure 8(a). We *suspect* that NIC cores suffer from head-of-line blocking when processing large READ requests. For a READ request, the NIC issues a PCIe read transaction to fetch the data, which is further segmented into multiple PCIe packets. The maximum size of a PCIe packet is determined by the PCIe Maximum Transfer Unit (MTU), negotiated by the linked hardware devices during bootstrap [49]. Table 4 lists the PCIe MTU on our testbed. SoC cores (the endpoint of SNIC ②) use a smaller PCIe MTU (128 B) due to its weaker CPU. As a result, NIC core that processes a large DMA read sent to SoC memory (SNIC ②) must wait for more PCIe packets to arrive, resulting in lengthy processing stalls. Since the overall NIC packet processing power is not the bottleneck: as shown in Figure 8(b), the requests with payloads smaller than 9 MB still can achieve a high processing rate while it collapses for the others, so we suspect some blocking happens at the NIC core. Note that WRITE requests are not affected since DMA does not wait for the completion [83, 49].

On the contrary, the host uses a larger PCIe MTU (512 B), so it does not suffer from bandwidth degradation (SNIC ①). As shown in Figure 8(b), the NIC can issue 46.7 million PCIe packets per second to the host (SNIC ①). The aggregated bandwidth reaches 191 Gbps, bottlenecked by the network.

Takeaways. For READ and WRITE, sending requests to SoC is typically faster than that to the host (or even faster than via RNIC) because SoC is "closer" to the NIC (without PCIe0). In contrast, using SEND/RECV to communicate with SoC is slower due to weak SoC cores. Furthermore, designers still need to carefully consider the heterogeneity between host CPU cores and SoC cores to avoid performance anomalies. Specifically, memory accesses to a small address range may suffer performance degradation due to the lack of DDIO support on SoC cores. In addition, sending large READ requests to SoC may underutilize the bandwidth so the request should be proactively segmented into smaller ones.

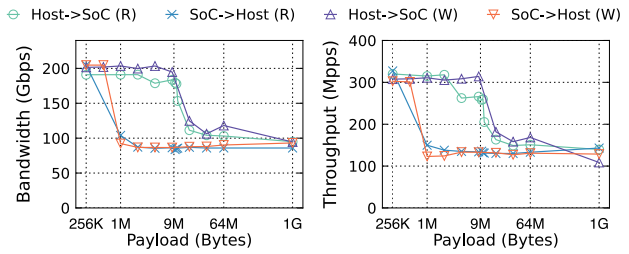

Figure 9: The bandwidth (a) and PCIe packets throughput (b) for sending (R)EAD/(W)RITE requests between the host and SoC.

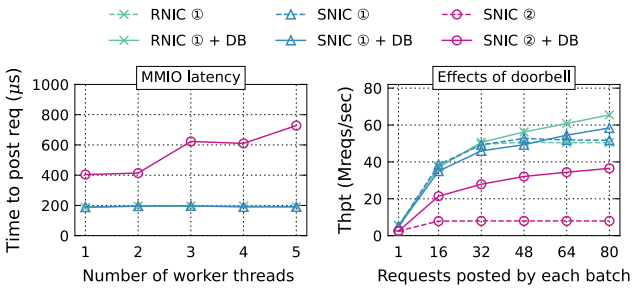

Figure 10: The latency of posting requests to NICs (a) and the impact of doorbell batching (DB) on the requester.

3.3 Communication between SoC and Host (path ③)

We first describe our measurements and findings of RDMA and then compare RDMA (③) with DMA (③*).

Latency. As shown in Figure 3, the latency of sending requests from SoC to the host (SNIC ③ S2H) is very high, especially for READ, since the requester (SoC) takes longer to issue an RDMA request to the NIC. The latency in the opposite direction (from the host ot SoC, SNIC ③ H2S) is reduced but still 4–17% higher than SNIC ②. Although the intra-machine communication saves one network round-trip, it adds additional PCIe transfers. Specifically, the request on SNIC ② flows the requester-side PCIe (not shown in Figure 1(c)), the network, PCIe1, and the PCIe switch, while the request on SNIC ③ (H2S) flows PCIe0, the PCIe switch, PCIe1 twice (in and out), and the PCI switch (again).

Throughput. For requests with payloads less than 512 bytes, the throughput of SNIC ③ (both S2H and H2S) is dominated by the requester's capability to post networking requests. This is because a single requester machine (either SoC or the host) cannot saturate the NIC with small requests[6], the READ throughput of SNIC ③ only reaches 29 M reqs/s and 51.2 M reqs/s for S2H and H2S, respectively, still far from its limit. For WRITE and SEND/RECV, the results are similar. For larger requests, they are bottlenecked by the PCIe bandwidth, which will be discussed in more detail next.

Bottleneck. As shown in Figure 5(b), for packets flowing in a single direction, communication between host and SoC is bottlenecked by PCIe bandwidth (256 Gbps) rather than the uninvolved NIC (200 Gbps). Therefore, the peak bandwidth of SNIC ③ is slightly higher than SNIC ① and ② (204 Gbps vs. 191 Gbps). Readers might be interested in why the results of SNIC ③ cannot be close to 256 Gbps. We *suspect* that it requires much more PCIe packets than the others. For packets flowing in opposite directions, SNIC ③ can not utilize twice the limit as the other paths (i.e., SNIC ① and ②). This is because RDMA overuses the PCIe: each request passes through PCIe1 twice (in and out), exhausting the bi-directional link.

Advice #3: Avoid large READ/WRTIE requests. Communications between the host and SoC (SNIC ③) also suffers from bandwidth degradation for large READ requests like SNIC ②, possibly due to the head of line blocking as we have discussed before. Moreover, this issue appears with large WRITE requests because the SmartNIC must first read data from the requester and then write it to the responder. As shown in Figure 9(a), the READ/WRITE performance of SNIC ③ collapses to about 100 Gbps for large requests. Table 4 shows the number of PCIe packets required to transfer N bytes via different communication paths. For SNIC ③, the NIC generates more packets due to passing through PCIe1 twice. Further, the performance of S2H collapses earlier than H2S as it will pass through PCIe1 first. Suppose we transfer data at 200 Gbps from SoC to the host. The SoC cores first transfer 195 M PCIe packets per second (pps) to the NIC (PCIe1), then the NIC forwards data back to the PCIe switch via PCIe1 again with 49 Mpps (the host supports 512 B MTU), and finally, the switch forwards 49 Mpps through PCIe0. Therefore, SmartNIC should process at least 293 Mpps for transferring data at 200 Gbps, which is 3× and 1.5× higher than SNIC ① and SNIC ②, respectively. This is further confirmed by our measurements of the hardware counters. As shown in Figure 9(b), for sending 256 KB READ requests from SoC to the host, the bandwidth reaches 204 Gbps, and the NIC transfers about 320 M PCIe packets per second.

Advice #4: Enable doorbell batching carefully. The time of posting each request to the NIC is dominated by Memory-Mapped IO (MMIO) [76, 28]. The SoC suffers a high MMIO latency when communicating with the host (see Figure 10(a)). A known optimization is doorbell batching (DB) [28]: to send a batch of B requests, the requester first chains them together in memory, then use one MMIO to ask the NIC to read these requests with DMA in a CPU-bypass way. DB reduces the number of MMIOs required from B to 1. Thus, for RNIC ① and SNIC ②, DB is always helpful and can bring 2–30% performance improvement (see Figure 10(b)). For the communication between host and SoC (SNIC ③), DB is still helpful at the SoC-side. As shown in Figure 10(b), when sending a batch of READs to the host, DB improves the SoC performance by 2.7–4.6× for batch sizes 16–80. The huge improvement is partly due to the CPU-bypass feature of DMA, and also because the NIC is faster in using DMA to read requests stored on SoC memory (see §3.2). However, DB is not always helpful at the host-side, because it is slower to

[6]We use up to eleven requester machines for SNIC ① and SNIC ②.

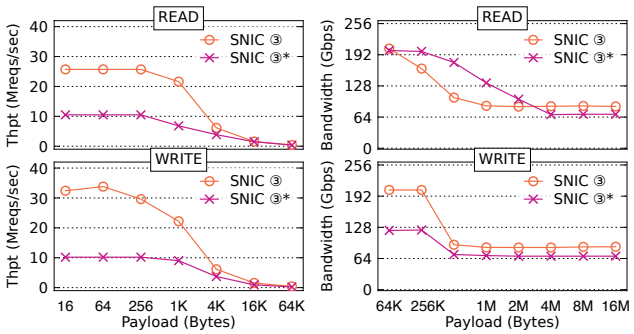

Figure 11: The throughput and bandwidth comparisons of using RDMA (③) and DMA (③*) when communicating from SoC to the host, the only supported primitive of DMA of our SNIC.

read host memory using NIC DMA (see §3.1). For batch sizes of 16, 32, and 48, DB decreases the throughput of host-SoC communication by 9%, 7%, and 6%, respectively.

RDMA (③) vs. DMA (③*). Besides RDMA, SoC can use DMA (③*) to read/write data from the host (and vice versa) via the DMA engine inside the SoC. It has the benefits of reducing two PCIe passes (PCIe1) and bypassing RNIC compared to RDMA (see Figure 1), resulting in a lower latency, e.g., 1.9 μ vs. 2.6 μ for 64 B SoC to host READ. However, we find the SoC DMA engine has a weaker processing power than RNIC (RDMA). For brevity, we only present results on SoC to host. The results of host to SoC is the same as SoC to host since host DMAs are offloaded to SoC for execution [56]. As shown in Figure 11, for WRITE, the peak throughput of DMA is only 47–59% of that of RDMA for requests with payload less than 4 KB. The results of READ is similar. DMA WRITE even fails to saturate the PCIe limit (256 Gbps) for payloads between 16 KB and 1 MB. We suspect it is due to the poor processing capability of the SoC's DMA engine, yet we cannot confirm this without knowing the confidential internal design of the SoC. Another observation from the bandwidth results is that DMA also suffers from the anomalies of RDMA (see Advice #3): For payloads larger than 1 MB, there is a significant performance drop for both READ and WRITE.

For bandwidth, ③* has a higher theoretical upper bound than ③: it is bottlenecked by the bidirectional bandwidth of PCIe, as it bypasses the PCIe1. However, Figure 5 shows that it fails to achieve so (only 178 Gbps for READ + WRITE). This suggests that the slow DMA engine will first become the bottleneck. Nevertheless, bypassing PCIe1 still has the benefits of reducing interferences to other paths. We will discuss them in §4 in detail.

Takeaways. First, enabling doorbell batching is critical for SNIC ③ at the SoC side, because SoC has wimpy computation power. Yet, it is negatively impacted at the host side for small batch sizes. Second, SNIC ③ has a different bottleneck than SNIC ① and SNIC ②. It is always bottlenecked by the uni-directional bandwidth of PCIe, while others are limited by the minimal bi-directional bandwidth of network and PCIe.

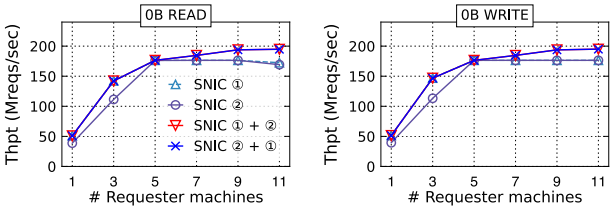

Figure 12: Throughput for (a) READ and (b) WRITE with the increases of requester machines.

If this factor is not adequately considered, distributed systems will underutilize the NIC bandwidth (see §5.1). Third, though DMA utilizes PCIe better than RDMA for SoC to communicate with the host, it has a lower throughput due to the weaker DMA engine at the SoC. Finally, we should avoid transferring large requests between the host and SoC, for both RDMA and DMA and for both READ and WRITE.

4 A Guideline for Smartly Exploiting Multiple Paths of SmartNIC

Previous approaches mainly leverage a single path of SmartNIC to optimize a specific functionality of distributed systems. However, this cannot fully exploit the computing and networking capabilities of SmartNICs. Furthermore, only considering a single path may ignore interference on resources (e.g., PCIe and PCIe switch) between different paths. Therefore, we first holistically study the performance characteristics of concurrently using multiple paths, and then lay out an optimization guideline for designers to smartly use SmartNICs.

4.1 Characterizing concurrent communication paths

Concurrent communication with the host and the SoC (①+②). We focus on the throughput results (see the lower part of Figure 3) since the latency results are roughly the average of the two paths. We evaluate the peak throughput by assigning half of the clients to send requests to the host while the others to send to the SoC. We can see that the total peak throughput of concurrently using ① and ② (SNIC ①+②) is typically faster than each of them. For READ, WRITE, and SEND/RECV, SNIC ①+② outperforms the lower of them by up to 1.45×, 1.50×, and 3.3×, respectively.

For SEND/RECV, a concurrent path utilize both of the host and SoC to process the requests, so the performance improvement is clear. However, the READ/WRITE performance improvement is non-intuitive and undocumented, since two paths should compete for NIC cores. Our *suspicion* is that the SmartNIC internally reserves some NIC cores for each endpoint. Therefore, sending requests to the host and the SoC concurrently can further increase *peak* throughput by enabling more NIC cores. To quantify this, we design a microbenchmark that first increases the requester machines to saturate the NIC and then changes the responder, as shown in Figure 12. All requests use 0 B payload to avoid interference of DMA, i.e., the request will return before passing PCIe1 [6].

For READ, five requester machines are sufficient to saturate NIC cores when using SNIC ① or SNIC ② alone. Therefore, for concurrently using SNIC ① and SNIC ②, we first dedicate five requester machines for one responder, and then add requesters for the other responder. Both cases (SNIC ①+② and SNIC ②+①) offer similar performance, with 4–13% and 5–10% higher throughput than using SNIC ① or SNIC ② alone. For WRTIE, all results are almost the same.

Finally, as expected, the aggregated throughput of the two paths (SNIC ① and SNIC ②) is much higher than concurrently using them (352 Mpps vs. 195 Mpps), indicating that most NIC cores are still shared, i.e., each can communicate with two endpoints, and only a few is dedicated. This also implies that concurrently using multiple resources of SmartNIC is non-trivial.

Concurrent inter- and intra-machine communication (①/②+③). There exist four concurrent combinations of inter- and intra-machine communication. For brevity, we focus on the results of SNIC ①+③/H2S, other combinations are similar. To study the concurrent usage of the two paths, we first deploy sufficient clients (five requester machines) to saturate the network for SNIC ①. Afterward, we start the requester on the host (one machine with 24 threads) sending RDMA requests to the SoC (SNIC ③/H2S). Our measurements reveal that concurrently enabling intra-machine communication degrades the performance of inter-machine communication. As shown in Figure 3, for READ, WRITE, and SEND/RECV, the throughput of small requests (less than 512 bytes) drops 7–15%, 4–27%, and 9–14%, by comparing SNIC ① and SNIC ①+③(H2S). For large requests, the performance is always bottlenecked by the network bandwidth, so the degradation is negligible.

The SNIC ③ affects other communication paths of SmartNIC, because it relies on the NIC (PCIe1 and the PCIe switch) for RDMA support. In comparison, SNIC ③* communication can leverage DMA to reduce such interferences. For example, for READ with payloads 16–64 B, we only observe a 5–6% throughput drop, after adding SNIC ③* to ①.

Bottleneck. Assuming each path has only one type of request, e.g., either READ or WRITE. For SNIC ①+②, each part has the same bottleneck (the NIC), so the bandwidth limit is 400 Gbps (bi-directional). For SNIC ①+③, it is bottlenecked by SNIC ③, which is limited on the uni-direction of PCIe (256 Gbps) since it occupies both directions of PCIe1 (see Figure 5(b)). Nevertheless, if SNIC ① is used in opposite directions (i.e., READ and WRITE), SNIC ①+③ can reach a higher limit. For example, the aggregated bandwidth can achieve 456 Gbps (in theory) if we restrict the bandwidth of data transfer on SNIC ③ to 56 Gbps. This suggests that selectively offloading small portion of data to SoC may be optimal. Finally, if possible, it is usually better to combine SNIC ① or ② with DMA (①/②+③*) despite DMA being slower than RDMA (see §3.3). This is because DMA has better PCIe utilization (without passing PCIe) and RNIC utilization (without using RNIC).

Takeaways. Sending requests from clients to the host and the SoC concurrently (SNIC ①+②) can better utilize NIC cores to handle small RDMA requests, especially when used in opposition directions (e.g., one for READ and one for WRITE). On the contrary, uncontrolled use of intra-machine (host-SoC) communications (SNIC ③) may harm inter-machine communications, which is the intrinsic purpose of using SmartNIC. Specifically, if the uni-directional bandwidth of PCIe is smaller than the bi-directional bandwidth of the NIC, using SNIC ③ can introduce a hidden bottleneck. Therefore, we should always consider using SNIC ③ only when spare resources are made available. Specifically, if the inter-machine communication saturates the NIC, the bandwidth used by SNIC ③ should no larger than $P - N$, where P and N are the limit of the PCIe and the network, respectively. For example, it should be 56 Gbps on our testbed. Using SNIC ③* can reduce the interference between paths, but SNIC ③* also has limitations: it is slower than SNIC ③.

Finally, in real-world distributed systems, it is common that a single communication path cannot fully saturate all resources of SmartNIC. For example, SNIC ② is the fastest but limited by small memory and wimpy cores on the SoC. On the other hand, only using SNIC ① as RNIC would waste all resources on the SoC. Therefore, we should concurrently use multiple paths provided by the SmartNIC, but carefully avoid interference between them.

4.2 An optimization guideline

This section presents our optimization guideline for smartly utilizing multiple communication paths of SmartNIC to improve the performance of distributed systems. Specifically, given the functionality (e.g., file replication in a distributed file system) of a target distributed system that needs to be accelerated by SmartNIC, we recommend designers consider the following steps:

1. Devise potential alternatives for SmartNIC to support the given functionality, and optimize them based on performance characteristics uncovered by our study.
2. Evaluate and rank alternatives based on system-specific criteria.
3. Select and combine alternatives in turn until the resource of SmartNIC is saturated.

System-specific criteria. The criteria can be the desirable properties that the system designer aims to achieve, or the restrictions of the systems. For replication in a distributed file system, the properties include low host CPU overhead and high network bandwidth utilization [32]. For a disaggregated key-value store, the properties include less network amplification, low latency and high throughput. The restriction the host has little or no CPU that we can use [86].

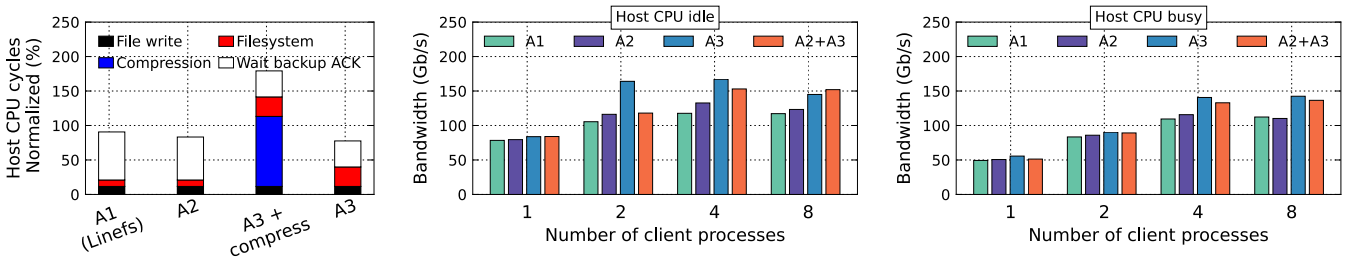

Figure 13: The host CPU usage breakdown of different alternatives when replicate 8 MB data (a). Write bandwidth when the host is idle (b) and busy (c) for A1, A2 and A2 + A3.

Discussion. We currently only consider the combination of alternatives in a greedy way, which is sufficient for most networked functions in real-world distributed systems. Further, SmartNIC usually offers a limited number of available options. Note that efficiently combining alternatives is challenging. For different systems, different alternatives may consume different resources on the SmartNIC, while a combination of them may involve different levels of resource contentions. Our previous analysis—including the bottleneck of different communication paths and concurrently utilizing multiple paths on the SmartNIC—will guide designers to avoid most performance contention. Nevertheless, how to systematically choose and combine different paths is our future work.

5 Case Studies

To demonstrate the efficacy of our study and the optimization guideline, this section presents two detailed case studies.

5.1 Distributed file system

Overview. File replication is a key pillar in distributed file systems for fault tolerance. With the emergence of RDMA and non-volatile memory (NVM), an appealing trend is to use RDMA to directly replicate file updates on remote NVM for better performance [32, 3, 40, 4], i.e., RDMA primitives can directly write NVM just like DRAM, with network and NVM bandwidth fully utilized [81].

Devise alternatives. The desirable properties of file replication are high performance, high network utilization and low host CPU overhead. There are three alternatives to implement file replications on our SmartNIC, as illustrated in Figure 14.

1. **Alternative (A1).** It comes from the state-of-the-art distributed file system on SmartNIC, LineFS [32], which completely offloads the file replication to SoC. The SoC will compress and replicate the file to reduce data transferred through the network with low host CPU usage. After receiving a replication request, the primary SoC reads the file from host (③), compresses it (④), and writes the file to remote backups with chain replication [72] (②). Specifically, if there are multiple backups, the second backup will further re-replicate the log to the next backup on the chain and so on.
2. **Alternative (A2).** Guided by our study, we can replace the ③ in A1 with ③* to reduce interference on the PCIe bandwidth, specifically, PCIe1 on the SmartNIC.
3. **Alternative (A3).** The host can directly write the file from the host to the remote backup with WRITE (①) [40]. Note that this approach typically skips file compression to prevent non-trivial host CPU overhead (see Figure 13 (a)).

Baseline. LineFS [32] is a state-of-the-art distributed file system based on NVM and SmartNIC. It adopts A1 to replicate the files. We further implement A2 and A3 on its open-source codebase[7], and rewrite its backend with more efficient RDMA implementation to scale to 200 Gbps networking, e.g., with asynchronous and batched RDMA operations.

Optimization on each alternative. By default, LineFS adopts a chunk size of 16 MB in its open-source codebase for A1. Based on our Advice #3 described in §3.3, we shrink it to 256 KB for a better performance over ③. This optimization further applies to A2 and A3.

Analyse alternatives. A1 is the most straightforward way to offload file replication, reducing the data transferred through the network (d vs. $d \times ratio$). Thus, the ideal peak bandwidth is $N/ratio$, where N is the bandwidth limit of SmartNIC. However, A1 does not consider the costly PCIe occupation of ③ (§3.3), which even fails to saturate the network bandwidth for file transfer. Denote the primary's PCIe limit (uni) as P. A1's file transfer bandwidth d is limited by $\frac{P}{1+ratio}$, because each data packet must pass the *PCIe1 out link* twice. As shown in Figure 14, one is from SoC to RNIC (d bytes) and another from SoC to the remote ($d \times ratio$ bytes). On our platform ($p = 256\,Gbps$), so A1 is only better than file is not compressed (whose performance is bottlenecked by the network $N = 200\,Gbps$) when the compression ratio is lower than 28%. Worse even, A1 cannot saturate the network bandwidth of SmartNIC when encountering a bad compression ratio ($\geq$ 28%). For example, without compression ($ratio = 1$), the peak of A1 is only 128 Gbps.

Figure 13 (b) presents the results of A1 on the file write benchmark of LineFS. This benchmark does not compress the file. We can see that A1 only achieves 117 Gbps with 8 clients when the host is idle.

A2 addresses the poor PCIe utilization of A1 by replacing ③ with ③*. As shown in Figure 13 (b), A2 is 1.01–1.13 ×

[7] https://github.com/casys-kaist/LineFS

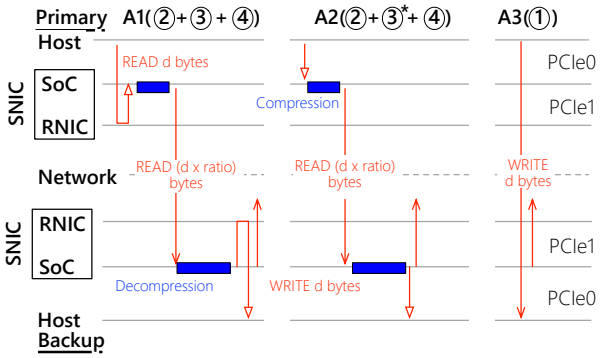

Figure 14: Overview of alternatives for file replication with SmartNIC. *ratio* is defined as `compressed size` / `uncompressed size`. We omit the control path messages as they are trivial.

faster than A1 under different number of clients. However, A2 fails to achieve a close to 200 Gbps result (peak at 133 Gbps) due to the following two reasons. First, the WRITE of ③* cannot fully utilize the full PCIe bandwidth on our platform (see Figure 11). Second, the poor computation power of SoC may also become the performance bottleneck of file replication.

A3 bypasses the PCIe occupation problem of A1, and the slow DMA WRITE and weak SoC issues of A2. Meanwhile, its data path is shorter (see Figure 14). As shown in Figure 13(a), it takes 40% shorter time to wait for the log acknowledgment compared to A2. As a result, A3's replication bandwidth is 5–41% faster than A2 under different client setups. The drawback is that A3 takes more CPU cycles even without considering compression (Filesystem), see Figure 13(a). This is because A1 and A2 can digest the file log on the SoC. Thus, the overall process time reduction of A3 is 8% (decreased from 40%) compared to A2.

Select and combine alternatives. Since A2 is always better than A1, we will only consider combining A2 with A3. As we have analyzed before, A3 is faster than A2. Therefore, increasing the ratio of A3 in a combined path (A2 + A3) always improves the performance, as shown in Figure 15. However, if file compression for high network utilization is enabled, it has high host CPU utilization, as shown in Figure 13 (a). Disabling compression for A3 will lower the network utilization, also illustrated in Figure 15. Specifically, when increasing the percentage of path A3 in clients, the network utilization is reduced from 50% to 0% considering a fixed 50% compression ratio.

Considering A2 has better network utilization, we follow a greedy approach that first saturate the SoC with A2 for better network utilization. Afterward, clients use A3 to do the file replication. This approach can achieve the best of both worlds: the combined path is faster than A2 with network better utilized than A3.

Evaluation results. Figure 13 (b) and (c) further present the file replication benchmark results of A2 + A3 when the host CPU is idle and busy, respectively. We follow the same setup as LineFS [32]'s benchmark and add a CPU-intensive

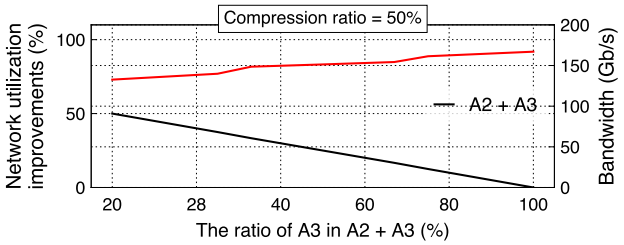

Figure 15: Analysis of the network utilization and performance when combining A2 and A3.

workload (streamcluster [7]) to the host CPU to emulate a busy experimental setup. A2 + A3 is 7–30% and 4–21% faster than original LineFS when CPU is idle and busy, respectively, thanks to the more efficient usage of SmartNIC and a smart utilization of multiple execution paths.

5.2 Disaggregated key-value store

Overview. RDMA-based disaggregated key-value stores (R-KVS) are prevalent in modern data centers [75, 70, 17, 86]. In R-KVS, one or more memory servers store both indexes (usually hash table) and values. Clients on other machines use READs to traverse the index and retrieve the corresponding value to handle requests (i.e., get), see A1 in Figure 16.

Devise alternatives. The desired properties are high throughput, low latency and minimal network amplification. The restriction is that we can barely use the host CPU (i.e., disable SEND/RECV for path ①). SmartNIC enables five alternatives for R-KVS, as illustrated in Figure 16.

1. **Alternative (A1).** The client treats SmartNIC as a normal RNIC and uses READs to handle the get request (①). This approach suffers from network amplification.
2. **Alternative (A2).** One intuitive approach for offloading is to send the get request to the SoC using SEND/RECV (②). The SoC can then traverse the index and read the value on the host via RDMA or DMA READ. This approach effectively eliminates network amplification.
3. **Alternative (A3).** One drawback of A2 is that reading data from SoC to the host is slower reading from the host's local memory. An optimization is to offload the indexes to the SoC memory (④). This approach is similar to index caching at the clients [11, 62, 75], but caching the indexes at the SmartNIC is more effective. Each client has a small memory that can only cache hundreds of entries in a disaggregated setting [86], while SmartNIC has a relatively large SoC memory (e.g., 16 GB on Bluefield-2) that can cache all the indexes.
4. **Alternative (A4).** Accessing the index on the SoC using SEND/RECV (②) cannot fully utilize the NIC cores of SmartNIC, because the peak throughput of SEND/RECV is only 21.6 M reqs/s. Therefore, we can use READs to traverse the index on the SoC (②), and another READ to retrieve the value on the host (①).

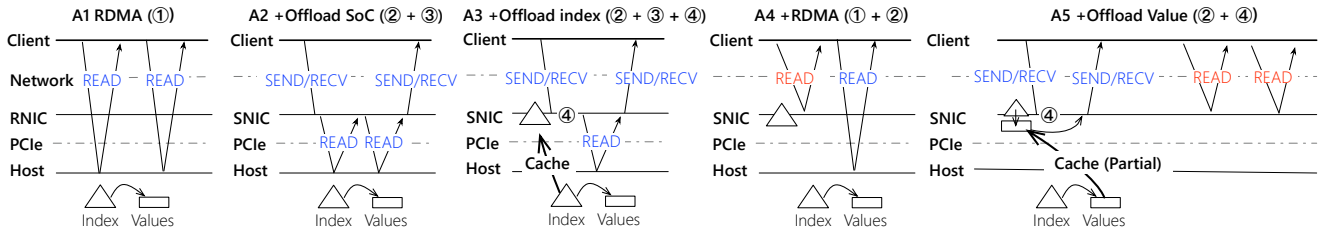

Figure 16: Alternatives (A1–A5) for offloading a get request of RDMA-based disaggregated key-value store to off-path SmartNIC.

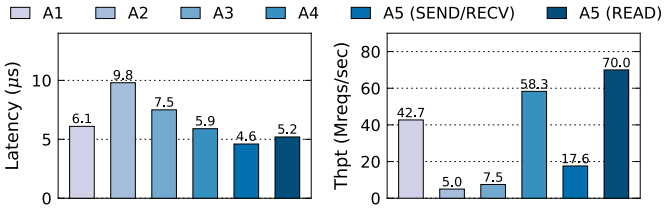

Figure 17: (a) Latency and (b) throughput comparisons between different alternatives on YCSB C. For A5, we restricted the selection of clients keys to always hit the cached values on SmartNIC.

This approach still has network amplification, but can utilize the fast path (②) to improve performance (see §3.2).

5. **Alternative (A5).** Similar to index caching, SoC memory can further cache a portion of values (e.g., the values of hot keys). This approach avoids using the costly communication path (③) of the previous alternatives.

Baseline. DrTM-KV [11] is a state-of-the-art KV store optimized for RDMA: it adopts cluster-chaining hash index such that the client typically finds the value position of a given key in one READ. Specifically, for a get request, the client first READs a 64 B bucket (based on the hash of the key), finds the remote address of the corresponding value in it, and then fetches the value with another READ. DrTM-KV supports index caching at the client to skip the first READ [80], but it may not be always feasible in a disaggregated environment due to memory constraints [86], so we disable it.

Optimization on each alternative. We implement A1–A5 on DrTM-KV guided by our study (§3). Specifically, we carefully enabled doorbell batching for alternatives related to the SoC CPU (A2, A3 and A5). Besides, we apply Advice #1 for A4 and A5, which replicate a few hot keys to multiple replications to avoid sending requests to a small range of memory. We use DMA (③*) instead of RDMA (③) to implement A2 and A3 as it is always faster due to lower latency. For example, A2 throughput is improved by up to 79% with ③*. A2 and A3 does not suffer from the low DMA throughput discovered in §3.3 because the SoC will first become the bottleneck.

Analyse alternatives. We use YCSB C [16] (100% get) with default Zipfian request distribution ($\theta = 0.99$) for all the experiments. The payload sizes of keys and values are 8 B and 64 B, respectively, similar to prior work [41, 45, 30, 66, 75]. Following the microbenchmark setup, we use one client machine to measure the latency and deploy up to eleven client machines to measure the peak throughput.

Figure 17 demonstrates that none of the path can achieve both high throughput and low latency. A5 (SEND/RECV) achieves the lowest latency (4.6 μs) because it completely eliminates the network amplifications problem and costly host-SoC communications (③). However, its peak throughput (17.6 M reqs/s) is significantly lower than some other alternatives. Specifically, the peak throughput of A5 (READ) and A4 reach 70 M reqs/s and 58.3 M reqs/s, respectively. They have a higher throughput because the RDMA path to SoC (①) is faster (§3.2). Note that A5 is not always achievable, which requires caching all the key-values at the SoC memory. Therefore, A4 is a suitable design if the SoC cores become the bottleneck (④). A1 has a higher latency and lower throughput than A4, since RDMA to the host (①) is relatively slow. A2 and A3 are bottlenecked by the slow host-SoC communication (③, see §3.3), which is not suitable for offloading KV store requests.

Select and combine alternatives. Our analysis suggests that the optimal combination is A4 and A5. Initially, the first few clients use A5, whereas the later clients use A4. The exact switch point can be estimated by using queuing theory [24] to model the capacity of SoC and the capability of RNIC, as in prior work [46].

In addition, using A5 presents a challenge as clients are unaware of which values are cached at SoC. Although A3 can be used as a fallback path for cache misses, it will result in significant performance degradation (see Figure 17). To tackle this issue, we provide a simple solution: when a cache miss occurs, the SoC returns the address of the value to the client, which then issues a READ to retrieve the value accordingly, similar to A4. In real-world skewed workloads (e.g., YCSB [16]), cache misses are rare.

Evaluation results. Figure 18 shows the latency and throughput results on YCSB C. We plotted the graph by increasing the number of client machines. The combination of A4 + A5 achieves a peak throughput of 68 M reqs/s, which is 25%, 36%, and 12% higher than RNIC, A1, and A4, respectively. Note that we omit A2 and A3 as they are bottlenecked by SoC cores and have extremely low peak throughput. The benefits

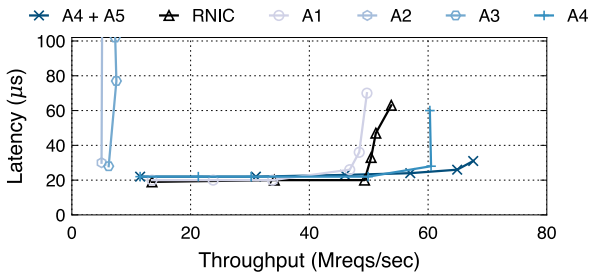

Figure 18: Performance of YCSB C using different alternatives. Note that A5 cannot run a full workload alone, since SoC memory is not large enough to cache all values. For A4 + A5, one client uses A5, and the rest use A4.

of A4 + A5 mainly come from utilizing faster SoC RDMA and SoC cores for reducing network amplifications.

6 Discussion

Generalizability. Although our study primarily focuses on one particular SmartNIC, Bluefield-2 [52], we believe that our findings and advice can be applied to other off-path SmartNICs that share a similar hardware architecture. These SmartNICs extend RDMA-capable NICs, such as Stingray PS225 [9] (which extends NetXtreme 100 Gbps RNIC [8]), by attaching a heterogeneous SoC and bridging SoC and RNIC together with a PCIe switch. We have confirmed that all our results hold on Bluefield-1 [55]. Moreover, the next generation of Bluefield (Bluefield-3) still follows the same architecture, except for using faster RNIC (400 Gbps ConnectX-7), PCIe (5.0), and SoC (ARMv8.2+ A78). Even though other SmartNICs may have different parameters than Bluefield-2, our methodology, analysis tools (open-sourced), and performance models (e.g., Table 4) also apply to them.

Furthermore, DPDK [1] is another popular communication primitive over SmartNIC. From a NIC's perspective, DPDK is similar to SEND/RECV over UD. Therefore, we believe that most of our findings are still applicable to DPDK as well. Unfortunately, we do not have an Ethernet-based testbed to confirm this further.

Suggestions for hardware vendors. Our study has uncovered several anomalies that can be mitigated through hardware improvements, which we suggest vendors consider. For example, current host to SoC DMA must offload to SoC for execution [56], while supporting CXL [15] can utilize the more powerful host CPU DMA engine for it. However, doing so in a programmer-friendly way [21] will require strong cooperation between the SoC OS and host OS. To the best of our knowledge, no SmartNIC supports CXL yet. Moreover, supporting CCI [5] can mitigate the performance degradation problem described in Advice #1. Furthermore, aligning the SoC PCIe MTU with the host is likely to improve PCIe performance when transferring large payloads. Finally, we encourage vendors to disclose more hardware details of SmartNICs to help explain and confirm the findings of our study.

7 Other Related Work

SmartNIC offloading. Offloading computation to SmartNICs has attracted significant attention in academia and industry. The offloaded tasks include network functions [59, 34, 19], microservices [14, 39], and others [33, 36, 22, 35, 61, 74, 65]. We share the same vision—improving the performance of distributed systems by offloading computation and communication to SmartNICs, but further exploit the multiple communication paths of SmartNICs. In addition, most prior work has focused on leveraging a single path of on-path SmartNICs, so our work can inspire future research on multi-path offloading for on-path SmartNICs.

RDMA offloading. Before the emergence of SmartNICs, many distributed systems offloaded remote memory accesses to one-sided RDMA primitives [75, 64, 17, 63, 46, 50, 13, 76, 82, 79, 84, 40, 85, 86]. However, prior work has observed the poor semantics of one-sided RDMA and has therefore leveraged advanced RDMA features (e.g., WAIT [60, 31], DCT [77, 78]) or introduced new RDMA primitives [65, 10]. These efforts are orthogonal to our work and could also benefit from our findings when using SmartNICs in the future.

8 Conclusion

Designing high-performance distributed systems with SmartNICs requires an in-depth understanding of low-level hardware details. This paper presents a comprehensive study of off-path SmartNIC. Unlike prior work, we explore how the SmartNIC architecture and the heterogeneity of its computation units can impact communication performance related to its components. We further propose the first optimization guideline for designers to smartly exploit multiple communication paths of SmartNICs for distributed systems, and demonstrate our guideline by improving two distributed systems. In general, our study can help system designers develop a better understanding of SmartNICs before applying them in high-performance distributed systems.

Acknowledgment

We sincerely thank our shepherd, Rachit Agarwal, and the anonymous reviewers for their comments and suggestions to improve the paper. We also thank Zhuobin Huang for discussing DPDK on SmartNIC, Jun Lu and Dong Du for providing the Bluefield-1 hardware, Dingji Li, Erhu Feng, Jinyu Gu, and Fangming Lu for their valuable feedback on earlier versions of the paper. This work was supported in part by the National Key Research & Development Program of China (No. 2022YFB4500700), the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China (No. 62202291, 62272291, 61925206), as well as research grants from Huawei Technologies and Shanghai AI Laboratory. Corresponding author: Rong Chen (rongchen@sjtu.edu.cn).

References

[1] Data plane development kit. https://www.dpdk.org/, 2023.

[2] AMARO, E., LUO, Z., OUSTERHOUT, A., KRISHNAMURTHY, A., PANDA, A., RATNASAMY, S., AND SHENKER, S. Remote memory calls. In *HotNets '20: The 19th ACM Workshop on Hot Topics in Networks, Virtual Event, USA, November 4-6, 2020* (2020), B. Y. Zhao, H. Zheng, H. V. Madhyastha, and V. N. Padmanabhan, Eds., ACM, pp. 38–44.

[3] ANDERSON, T. E., CANINI, M., KIM, J., KOSTIC, D., KWON, Y., PETER, S., REDA, W., SCHUH, H. N., AND WITCHEL, E. Assise: Performance and availability via client-local NVM in a distributed file system. In *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020* (2020), USENIX Association, pp. 1011–1027.

[4] ANDERSON, T. E., CANINI, M., KIM, J., KOSTIC, D., KWON, Y., PETER, S., REDA, W., SCHUH, H. N., AND WITCHEL, E. Assise: Performance and availability via client-local NVM in a distributed file system. In *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020* (2020), USENIX Association, pp. 1011–1027.

[5] ARM. Corelink CCI-550. https://developer.arm.com/Processors/CoreLink%20CCI-550, 2022.

[6] ASSOCIATION., I. T. Infiniband architecture specification. https://cw.infinibandta.org/document/dl/7859, 2022.

[7] BIENIA, C., KUMAR, S., SINGH, J. P., AND LI, K. The PARSEC benchmark suite: characterization and architectural implications. In *17th International Conference on Parallel Architectures and Compilation Techniques, PACT 2008, Toronto, Ontario, Canada, October 25-29, 2008* (2008), A. Moshovos, D. Tarditi, and K. Olukotun, Eds., ACM, pp. 72–81.

[8] BROADCOM. Bcm57504 - 100gbe. https://en.broadcom.com/products/ethernet-connectivity/network-adapters/bcm57504-100g-ic, 2022.

[9] BROADCOM. Product Brief: Stingray PS225. https://docs.broadcom.com/doc/PS225-PB, 2022.

[10] BURKE, M., DHARANIPRAGADA, S., JOYNER, S., SZEKERES, A., NELSON, J., ZHANG, I., AND PORTS, D. R. K. PRISM: rethinking the RDMA interface for distributed systems. In *SOSP '21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 26-29, 2021* (2021), R. van Renesse and N. Zeldovich, Eds., ACM, pp. 228–242.

[11] CHEN, H., CHEN, R., WEI, X., SHI, J., CHEN, Y., WANG, Z., ZANG, B., AND GUAN, H. Fast in-memory transaction processing using RDMA and HTM. *ACM Trans. Comput. Syst. 35*, 1 (2017), 3:1–3:37.

[12] CHEN, Y., LU, Y., AND SHU, J. Scalable RDMA RPC on reliable connection with efficient resource sharing. In *Proceedings of the Fourteenth EuroSys Conference 2019, Dresden, Germany, March 25-28, 2019* (2019), G. Candea, R. van Renesse, and C. Fetzer, Eds., ACM, pp. 19:1–19:14.

[13] CHEN, Y., WEI, X., SHI, J., CHEN, R., AND CHEN, H. Fast and general distributed transactions using RDMA and HTM. In *Proceedings of the Eleventh European Conference on Computer Systems, EuroSys 2016, London, United Kingdom, April 18-21, 2016* (2016), C. Cadar, P. R. Pietzuch, K. Keeton, and R. Rodrigues, Eds., ACM, pp. 26:1–26:17.

[14] CHOI, S., SHAHBAZ, M., PRABHAKAR, B., AND ROSENBLUM, M. λ-nic: Interactive serverless compute on programmable smartnics. In *40th IEEE International Conference on Distributed Computing Systems, ICDCS 2020, Singapore, November 29 - December 1, 2020* (2020), IEEE, pp. 67–77.

[15] CONSORTIUM, C. Cxl specification. https://www.computeexpresslink.org/download-the-specification, 2022.

[16] COOPER, B. F. YCSB Core Workloads. https://github.com/brianfrankcooper/YCSB/wiki/Core-Workloads, 2021.

[17] DRAGOJEVIC, A., NARAYANAN, D., CASTRO, M., AND HODSON, O. FaRM: Fast remote memory. In *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2014, Seattle, WA, USA, April 2-4, 2014* (2014), R. Mahajan and I. Stoica, Eds., USENIX Association, pp. 401–414.

[18] DRAGOJEVIC, A., NARAYANAN, D., NIGHTINGALE, E. B., RENZELMANN, M., SHAMIS, A., BADAM, A., AND CASTRO, M. No compromises: distributed transactions with consistency, availability, and performance. In *Proceedings of the 25th Symposium on Operating Systems Principles, SOSP 2015, Monterey, CA, USA, October 4-7, 2015* (2015), E. L. Miller and S. Hand, Eds., ACM, pp. 54–70.

[19] FIRESTONE, D., PUTNAM, A., MUNDKUR, S., CHIOU, D., DABAGH, A., ANDREWARTHA, M., ANGEPAT, H., BHANU, V., CAULFIELD, A. M., CHUNG, E. S., CHANDRAPPA, H. K., CHATURMOHTA, S., HUMPHREY, M., LAVIER, J., LAM, N., LIU, F., OVTCHAROV, K., PADHYE, J., POPURI, G., RAINDEL, S., SAPRE, T., SHAW, M., SILVA, G., SIVAKUMAR, M., SRIVASTAVA, N., VERMA, A., ZUHAIR, Q., BANSAL, D., BURGER, D., VAID, K., MALTZ, D. A., AND GREENBERG, A. G. Azure accelerated networking: Smartnics in the public cloud. In *15th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2018, Renton, WA, USA, April 9-11, 2018* (2018), S. Banerjee and S. Seshan, Eds., USENIX Association, pp. 51–66.

[20] GAO, Y., LI, Q., TANG, L., XI, Y., ZHANG, P., PENG, W., LI, B., WU, Y., LIU, S., YAN, L., FENG, F., ZHUANG, Y., LIU, F., LIU, P., LIU, X., WU, Z., WU, J., CAO, Z., TIAN, C., WU, J., ZHU, J., WANG, H., CAI, D., AND WU, J. When cloud storage meets RDMA. In *18th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2021, April 12-14, 2021* (2021), J. Mickens and R. Teixeira, Eds., USENIX Association, pp. 519–533.

[21] GOUK, D., LEE, S., KWON, M., AND JUNG, M. Direct access, High-Performance memory disaggregation with DirectCXL. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)* (Carlsbad, CA, July 2022), USENIX Association, pp. 287–294.

[22] GRANT, S., YELAM, A., BLAND, M., AND SNOEREN, A. C. Smartnic performance isolation with fairnic. In *SIGCOMM '20: Proceedings of the 2020 Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, Virtual Event, USA, August 10-14, 2020* (2020), H. Schulzrinne and V. Misra, Eds., ACM, pp. 681–693.

[23] GUO, C., WU, H., DENG, Z., SONI, G., YE, J., PADHYE, J., AND LIPSHTEYN, M. RDMA over commodity ethernet at scale. In *Proceedings of the ACM SIGCOMM 2016 Conference, Florianopolis, Brazil, August 22-26, 2016* (2016), M. P. Barcellos, J. Crowcroft, A. Vahdat, and S. Katti, Eds., ACM, pp. 202–215.

[24] HARCHOL-BALTER, M. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press, 2013.

[25] HASSAN, H., VIJAYKUMAR, N., KHAN, S. M., GHOSE, S., CHANG, K. K., PEKHIMENKO, G., LEE, D., ERGIN, O., AND MUTLU, O. Softmc: A flexible and practical open-source infrastructure for enabling experimental DRAM studies. In *2017 IEEE International Symposium on High Performance Computer Architecture, HPCA 2017, Austin, TX, USA, February 4-8, 2017* (2017), IEEE Computer Society, pp. 241–252.

[26] INTEL. Intel® data direct i/o technology. https://www.intel.com/content/www/us/en/io/data-direct-i-o-technology.html, 2022.

[27] KALIA, A., KAMINSKY, M., AND ANDERSEN, D. G. Using RDMA efficiently for key-value services. In *ACM SIGCOMM 2014 Conference, SIGCOMM'14, Chicago, IL, USA, August 17-22, 2014* (2014), F. E. Bustamante, Y. C. Hu, A. Krishnamurthy, and S. Ratnasamy, Eds., ACM, pp. 295–306.

[28] KALIA, A., KAMINSKY, M., AND ANDERSEN, D. G. Design guidelines for high performance RDMA systems. In *2016 USENIX Annual Technical Conference, USENIX ATC 2016, Denver, CO, USA, June 22-24, 2016* (2016), A. Gulati and H. Weatherspoon, Eds., USENIX Association, pp. 437–450.

[29] KALIA, A., KAMINSKY, M., AND ANDERSEN, D. G. Fasst: Fast, scalable and simple distributed transactions with two-sided (RDMA) datagram rpcs. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016* (2016), K. Keeton and T. Roscoe, Eds., USENIX Association, pp. 185–201.

[30] KALIA, A., KAMINSKY, M., AND ANDERSEN, D. G. Datacenter rpcs can be general and fast. In *16th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2019, Boston, MA, February 26-28, 2019* (2019), J. R. Lorch and M. Yu, Eds., USENIX Association, pp. 1–16.

[31] KIM, D., MEMARIPOUR, A. S., BADAM, A., ZHU, Y., LIU, H. H., PADHYE, J., RAINDEL, S., SWANSON, S., SEKAR, V., AND SESHAN, S. Hyperloop: group-based nic-offloading to accelerate replicated transactions in multi-tenant storage systems. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, SIGCOMM 2018, Budapest, Hungary, August 20-25, 2018* (2018), S. Gorinsky and J. Tapolcai, Eds., ACM, pp. 297–312.

[32] KIM, J., JANG, I., REDA, W., IM, J., CANINI, M., KOSTIC, D., KWON, Y., PETER, S., AND WITCHEL, E. Linefs: Efficient smartnic offload of a distributed file system with pipeline parallelism. In *SOSP '21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 26-29, 2021* (2021), R. van Renesse and N. Zeldovich, Eds., ACM, pp. 756–771.

[33] LI, B., RUAN, Z., XIAO, W., LU, Y., XIONG, Y., PUTNAM, A., CHEN, E., AND ZHANG, L. Kv-direct: High-performance in-memory key-value store with programmable NIC. In *Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017* (2017), ACM, pp. 137–152.

[34] LI, B., TAN, K., LUO, L. L., PENG, Y., LUO, R., XU, N., XIONG, Y., AND CHENG, P. Clicknp: Highly flexible and high-performance network processing with reconfigurable hardware. In *Proceedings of the ACM SIGCOMM 2016 Conference, Florianopolis, Brazil, August 22-26, 2016* (2016), M. P. Barcellos, J. Crowcroft, A. Vahdat, and S. Katti, Eds., ACM, pp. 1–14.

[35] LI, J., LU, Y., WANG, Q., LIN, J., YANG, Z., AND SHU, J. Alnico: Smartnic-accelerated contention-aware request scheduling for transaction processing. In *2022 USENIX Annual Technical Conference, USENIX ATC 2022, Carlsbad, CA, USA, July 11-13, 2022* (2022), J. Schindler and N. Zilberman, Eds., USENIX Association, pp. 951–966.

[36] LIN, J., PATEL, K., STEPHENS, B. E., SIVARAMAN, A., AND AKELLA, A. PANIC: A high-performance programmable NIC for multi-tenant networks. In *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020* (2020), USENIX Association, pp. 243–259.

[37] LIU, J., MALTZAHN, C., ULMER, C. D., AND CURRY, M. L. Performance characteristics of the bluefield-2 smartnic. *CoRR abs/2105.06619* (2021).

[38] LIU, M., CUI, T., SCHUH, H., KRISHNAMURTHY, A., PETER, S., AND GUPTA, K. Offloading distributed applications onto smartnics using ipipe. In *Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM 2019, Beijing, China, August 19-23, 2019* (2019), J. Wu and W. Hall, Eds., ACM, pp. 318–333.

[39] LIU, M., PETER, S., KRISHNAMURTHY, A., AND PHOTHILIMTHANA, P. M. E3: energy-efficient microservices on smartnic-accelerated servers. In *2019 USENIX Annual Technical Conference, USENIX ATC 2019, Renton, WA, USA, July 10-12, 2019* (2019), D. Malkhi and D. Tsafrir, Eds., USENIX Association, pp. 363–378.

[40] LU, Y., SHU, J., CHEN, Y., AND LI, T. Octopus: an rdma-enabled distributed persistent memory file system. In *2017 USENIX Annual Technical Conference, USENIX ATC 2017, Santa Clara, CA, USA, July 12-14, 2017* (2017), D. D. Silva and B. Ford, Eds., USENIX Association, pp. 773–785.

[41] MAO, Y., KOHLER, E., AND MORRIS, R. T. Cache craftiness for fast multicore key-value storage. In *European Conference on Computer Systems, Proceedings of the Seventh EuroSys Conference 2012, EuroSys '12, Bern, Switzerland, April 10-13, 2012* (2012), P. Felber, F. Bellosa, and H. Bos, Eds., ACM, pp. 183–196.

[42] MARVELL. Marvell liquidio iii. https://www.marvell.com/content/dam/marvell/en/public-collateral/embedded-processors/marvell-liquidio-III-solutions-brief.pdf, 2022.

[43] MARVELL. Marvell® octeon 10 dpu platform. https://www.marvell.com/content/dam/marvell/en/public-collateral/embedded-processors/marvell-octeon-10-dpu-platform-product-brief.pdf, 2023.

[44] MELLANOX. ConnectX-7 product brief. https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf, 2022.

[45] MITCHELL, C., MONTGOMERY, K., NELSON, L., SEN, S., AND LI, J. Balancing CPU and network in the cell distributed b-tree store. In *2016 USENIX Annual Technical Conference, USENIX ATC 2016, Denver, CO, USA, June 22-24, 2016* (2016), A. Gulati and H. Weatherspoon, Eds., USENIX Association, pp. 451–464.

[46] MITCHELL, C., MONTGOMERY, K., NELSON, L., SEN, S., AND LI, J. Balancing CPU and network in the cell distributed b-tree store. In *2016 USENIX Annual Technical Conference, USENIX ATC 2016, Denver, CO, USA, June 22-24, 2016* (2016), A. Gulati and H. Weatherspoon, Eds., USENIX Association, pp. 451–464.

[47] MONGA, S. K., KASHYAP, S., AND MIN, C. Birds of a feather flock together: Scaling RDMA rpcs with flock. In *SOSP '21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 26-29, 2021* (2021), R. van Renesse and N. Zeldovich, Eds., ACM, pp. 212–227.

[48] NETRONOME. Netronome agilio. NetronomeAgilioSmartNICs.https://www.netronome.com/products/smartnic/overview/, 2022.

[49] NEUGEBAUER, R., ANTICHI, G., ZAZO, J. F., AUDZEVICH, Y., LÓPEZ-BUEDO, S., AND MOORE, A. W. Understanding pcie performance for end host networking. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, SIGCOMM 2018, Budapest, Hungary, August 20-25, 2018* (2018), S. Gorinsky and J. Tapolcai, Eds., ACM, pp. 327–341.

[50] NOVAKOVIC, S., SHAN, Y., KOLLI, A., CUI, M., ZHANG, Y., ERAN, H., PISMENNY, B., LISS, L., WEI, M., TSAFRIR, D., AND AGUILERA, M. K. Storm: a fast transactional dataplane for remote data structures. In *Proceedings of the 12th ACM International Conference on Systems and Storage, SYSTOR 2019, Haifa, Israel, June 3-5, 2019* (2019), M. Hershcovitch, A. Goel, and A. Morrison, Eds., ACM, pp. 97–108.

[51] NVIDIA. Innova-2 flex. https://www.nvidia.com/en-us/networking/ethernet/innova-2-flex/, 2022.

[52] NVIDIA. Nvidia bluefield dpu-2. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-2-dpu.pdf, 2022.

[53] NVIDIA. Nvidia bluefield dpu-3. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-3-dpu.pdf, 2022.

[54] NVIDIA. Performance Monitoring Counters, BlueField SW Manual v2.4.0.11082. https://docs.nvidia.com/networking/display/BlueFieldSWv24011082/Performance+Monitoring+Counters, 2022.

[55] NVIDIA. Nvidia bluefield dpu-1. https://docs.nvidia.com/networking/display/BFVPIDPU/Specifications, 2023.

[56] NVIDIA. Nvidia doca dma programming guide. https://docs.nvidia.com/doca/sdk/dma-programming-guide/index.html, 2023.

[57] NVIDIA. Nvidia doca software framework. https://developer.nvidia.com/networking/doca, 2023.

[58] OLUMIDE OLUSANYA AND MUNIRA HUSSAIN. Need for Speed: Comparing FDR and EDR InfiniBand (Part 1). https://dl.dell.com/manuals/all-products/esuprt_software/esuprt_it_ops_datcentr_mgmt/high-computing-solution-resources_white-papers77_en-us.pdf, 2022.

[59] PHOTHILIMTHANA, P. M., LIU, M., KAUFMANN, A., PETER, S., BODÍK, R., AND ANDERSON, T. E. Floem: A programming system for nic-accelerated network applications. In *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018* (2018), A. C. Arpaci-Dusseau and G. Voelker, Eds., USENIX Association, pp. 663–679.

[60] REDA, W., CANINI, M., KOSTIC, D., AND PETER, S. RDMA is turing complete, we just did not know it yet! *CoRR abs/2103.13351* (2021).

[61] SCHUH, H. N., LIANG, W., LIU, M., NELSON, J., AND KRISHNAMURTHY, A. Xenic: Smartnic-accelerated distributed transactions. In *SOSP '21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 26-29, 2021* (2021), R. van Renesse and N. Zeldovich, Eds., ACM, pp. 740–755.

[62] SHAMIS, A., RENZELMANN, M., NOVAKOVIC, S., CHATZOPOULOS, G., DRAGOJEVIC, A., NARAYANAN, D., AND CASTRO, M. Fast general distributed transactions with opacity. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019* (2019), P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, Eds., ACM, pp. 433–448.

[63] SHAMIS, A., RENZELMANN, M., NOVAKOVIC, S., CHATZOPOULOS, G., DRAGOJEVIC, A., NARAYANAN, D., AND CASTRO, M. Fast general distributed transactions with opacity. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019* (2019), P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, Eds., ACM, pp. 433–448.

[64] SHI, J., YAO, Y., CHEN, R., CHEN, H., AND LI, F. Fast and concurrent RDF queries with rdma-based distributed graph exploration. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016* (2016), K. Keeton and T. Roscoe, Eds., USENIX Association, pp. 317–332.

[65] SIDLER, D., WANG, Z., CHIOSA, M., KULKARNI, A., AND ALONSO, G. Strom: smart remote memory. In *EuroSys '20: Fifteenth EuroSys Conference 2020, Heraklion, Greece, April 27-30, 2020* (2020), A. Bilas, K. Magoutis, E. P. Markatos, D. Kostic, and M. I. Seltzer, Eds., ACM, pp. 29:1–29:16.

[66] TANG, C., WANG, Y., DONG, Z., HU, G., WANG, Z., WANG, M., AND CHEN, H. Xindex: A scalable learned index for multicore data storage. In *Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (New York, NY, USA, 2020), PPoPP '20, Association for Computing Machinery, p. 308–320.

[67] THOMAS, S., VOELKER, G. M., AND PORTER, G. Cachecloud: Towards speed-of-light datacenter communication. In *10th USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2018, Boston, MA, USA, July 9, 2018* (2018), G. Ananthanarayanan and I. Gupta, Eds., USENIX Association.

[68] THOSTRUP, L., FAILING, D., ZIEGLER, T., AND BINNIG, C. A dbms-centric evaluation of bluefield dpus on fast networks. In *13th International Workshop on Accelerating Analytics and Data Management Systems Using Modern Processor and Storage Architectures* (2022).

[69] TIMOTHY PRICKETT MORGAN. Pushing PCI-express fabrics up to the next level. https://www.nextplatform.com/2020/03/27/pushing-pci-express-fabrics-up-to-the-next-level/, 2022.

[70] TSAI, S., SHAN, Y., AND ZHANG, Y. Disaggregating persistent memory and controlling them remotely: An exploration of passive disaggregated key-value stores. In *2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020* (2020), A. Gavrilovska and E. Zadok, Eds., USENIX Association, pp. 33–48.

[71] TSAI, S.-Y., AND ZHANG, Y. Lite kernel rdma support for datacenter applications. In *Proceedings of the 26th Symposium on Operating Systems Principles* (New York, NY, USA, 2017), SOSP '17, ACM, pp. 306–324.

[72] VAN RENESSE, R., AND SCHNEIDER, F. B. Chain replication for supporting high throughput and availability. In *6th Symposium on Operating System Design and Implementation (OSDI 2004), San Francisco, California, USA, December 6-8, 2004* (2004), E. A. Brewer and P. Chen, Eds., USENIX Association, pp. 91–104.

[73] WANG, X., KOTRA, J. B., AND JIAN, X. Eager memory cryptography in caches. In *55th IEEE/ACM International Symposium on Microarchitecture, MICRO 2022, Chicago, IL, USA, October 1-5, 2022* (2022), IEEE, pp. 693–709.

[74] WANG, Z., HUANG, H., ZHANG, J., WU, F., AND ALONSO, G. Fpganic: An fpga-based versatile 100gb smartnic for gpus. In *2022 USENIX Annual Technical Conference, USENIX ATC 2022, Carlsbad, CA, USA, July 11-13, 2022* (2022), J. Schindler and N. Zilberman, Eds., USENIX Association, pp. 967–986.

[75] WEI, X., CHEN, R., AND CHEN, H. Fast rdma-based ordered key-value store using remote learned cache. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)* (Nov. 2020), USENIX Association, pp. 117–135.

[76] WEI, X., DONG, Z., CHEN, R., AND CHEN, H. Deconstructing RDMA-enabled distributed transactions: Hybrid is better! In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)* (Carlsbad, CA, Oct. 2018), USENIX Association, pp. 233–251.

[77] WEI, X., LU, F., CHEN, R., AND CHEN, H. KRCORE: A microsecond-scale RDMA control plane for elastic computing. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)* (Carlsbad, CA, July 2022), USENIX Association, pp. 121–136.

[78] WEI, X., LU, F., WANG, T., GU, J., YANG, Y., CHEN, R., AND CHEN, H. No provisioned concurrency: Fast rdma-codesigned remote fork for serverless computing. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)* (Boston, MA, July 2023), USENIX Association.

[79] WEI, X., SHEN, S., CHEN, R., AND CHEN, H. Replication-driven live reconfiguration for fast distributed transaction processing. In *Proceedings of the 2017 USENIX Annual Technical Conference* (Santa Clara, CA, 2017), USENIX ATC'17, USENIX Association, pp. 335–347.

[80] WEI, X., SHI, J., CHEN, Y., CHEN, R., AND CHEN, H. Fast in-memory transaction processing using RDMA and HTM. In *Proceedings of the 25th Symposium on Operating Systems Principles, SOSP 2015, Monterey, CA, USA, October 4-7, 2015* (2015), E. L. Miller and S. Hand, Eds., ACM, pp. 87–104.

[81] WEI, X., XIE, X., CHEN, R., CHEN, H., AND ZANG, B. Characterizing and optimizing remote persistent memory with RDMA and NVM. In *2021 USENIX Annual Technical Conference, USENIX ATC 2021, July 14-16, 2021* (2021), I. Calciu and G. Kuenning, Eds., USENIX Association, pp. 523–536.

[82] XIE, X., WEI, X., CHEN, R., AND CHEN, H. Pragh: Locality-preserving Graph Traversal with Split Live Migration. In *2019 USENIX Annual Technical Conference, USENIX ATC 2019, Renton, WA, USA, July 10-12, 2019* (2019), pp. 723–738.

[83] XILLYBUS. Down to the tlp: How pci express devices talk. http://xillybus.com/tutorials/pci-express-tlp-pcie-primer-tutorial-guide-1, 2022.

[84] ZAMANIAN, E., BINNIG, C., KRASKA, T., AND HARRIS, T. The end of a myth: Distributed transaction can scale. *Proc. VLDB Endow. 10*, 6 (2017), 685–696.

[85] ZHANG, Y., CHEN, R., AND CHEN, H. Sub-millisecond stateful stream querying over fast-evolving linked data. In *Proceedings of the 26th Symposium on Operating Systems Principles* (New York, NY, USA, 2017), SOSP'17, ACM, pp. 614–630.

[86] ZUO, P., SUN, J., YANG, L., ZHANG, S., AND HUA, Y. One-sided rdma-conscious extendible hashing for disaggregated memory. In *2021 USENIX Annual Technical Conference, USENIX ATC 2021, July 14-16, 2021* (2021), I. Calciu and G. Kuenning, Eds., USENIX Association, pp. 15–29.