UNIVERSITY OF CALIFORNIA SANTA CRUZ

Evaluating 100Gbps Flash Disaggregation on ARM SoC

Minghao Xie, Heiner Litz, Chen Qian
Flash Memory Summit, August 6 – 8, 2019

Mail: mhxie@ucsc.edu

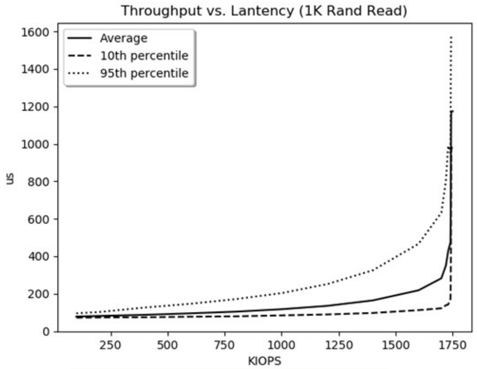
Motivation

Can we achieve a competitive performance on ARM64 to Intel x86?

What are the challenges of porting an application from x86 to aarch64?

What is the TCO optimal platform for Flash storage disaggregation?

Latency Distribution



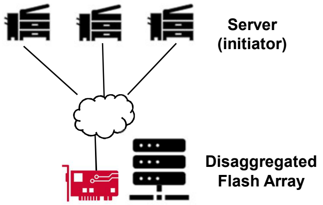
Contribution

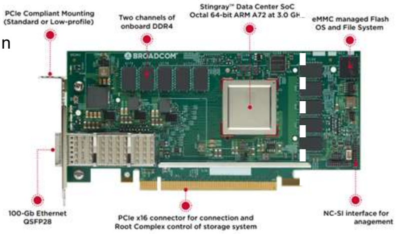
1. Provide a portable flash disaggregation software (across DPDK/SPDK versions & different architectures)
2. Explore the differences between ARM and x86
3. Achieve a state-of-the-art throughput with better TCO

Overview

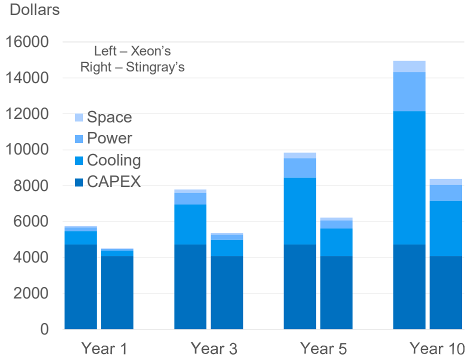
Flash Disaggregation enables sharing flash storage across the data center, improving resource utilization and reducing the total cost of ownership(TCO). The previous work ReFlex[2] tightly integrates the networking layer and storage layer by employing IX data plane architecture with the high-performance storage access framework SPDK.

With an energy-efficient and powerful smart NIC[3], we can think more… like offloading the processing tasks to free the host entirely. We want to build an integrated storage, computation and network platform over ReFlex and find a new tradeoff among these three. This NIC has 8 ARM A72 cores and it is so small and low-powered that using it as a PCI-e root complex with a fanout of several SSDs would largely lower down the previous OPEX in data centers. Without degrading much of the performance, we can come up with a flash disaggregation solution with a lower TCO.

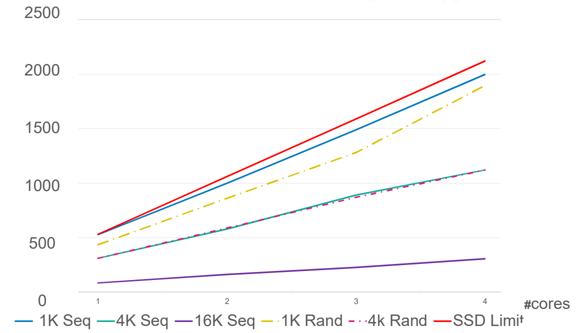



PCIe Compliant Mounting (Standard or Low-profile)
Two channels of onboard DDR4
Stingray™ Data Center SoC Octal 64-bit ARM A72 at 3.0 GH
eMMC managed Flash OS and File System
BROADCOM
100-Gb Ethernet QSFP28
PCIe x16 connector for connection and Root Complex control of storage system
NC-SI interface for anagement

Evaluation



TCO Analysis



Throughput Scaling

Portability is one of the development goal throughout this project. Working on different architectures, it's hard to reuse all the codes because we need assembly code in some critical areas. And available CPU features may vary. We've been modified the legacy codes and upgrade the APIs to the newest and make it portable across hardware/DPDK versions .

Performance is the thing we concerned most. We've achieved a decent performance with offloading features enabled. While good IOPS and throughput does not mean a good latency. In our 4k random read test, the system can enforce a 95th tail latency to less than 400 µs in a 400K IOPS test, which is still better than the conventional SSD read. And it can be scaled up to 2100k IOPS with 4 cores.

TCO is an important factor to build a datacenter . With current price, using this new smart NIC is not able to bring the benefits of CAPEX. But it does deliver better OPEX with lower energy consumptions and smaller space occupation. We will see better TCO if we can further optimize our system. With a more scalable design upcoming, the CAPEX would be very promising soon.

Performance

4,000,000+
Capable IOPS

5-10ms
RTT SLA

44.0%
TCO saving in a 10-year span

95th
Tail Read Latency SLOs

2.3x
Power Saving by Offloading tasks to the smart NIC

22μs
Latency Loss

90+ Gbps
Max Throughput

Future Work

Telemetry data from NVMe SSDs

Dynamic IOPS++ by prediction

Better performance isolation

Scalable to changing workloads

Reference:

[1] Klimovic, A., Kozyrakis, C., Thereska, E., John, B., & Kumar, S. (2016, April). Flash storage disaggregation. In Proceedings of the Eleventh European Conference on Computer Systems (p. 29). ACM.

[2] Klimovic, A., Litz, H., & Kozyrakis, C. (2017). Reflex: remote flash≈ local flash. ACM SIGOPS Operating Systems Review, 51(2), 345-359.

[3] https://www.broadcom.com/products/storage/ethernet-storage-adapters-ics/ps1100r

[4] Patterson, M. K., Costello, D., Grimm, P., & Loeffler, M. (2007). Data center TCO; a comparison of high-density and ow-density spaces. Thermal Challenges in Next Generation Electronic Systems (THERMES 2007), 42-49.

CENTER FOR RESEARCH IN STORAGE SYSTEMS

BROADCOM