Flexible Block Storage Offload for Datacenters with Bluefield-2 and NVIDIA SNAP

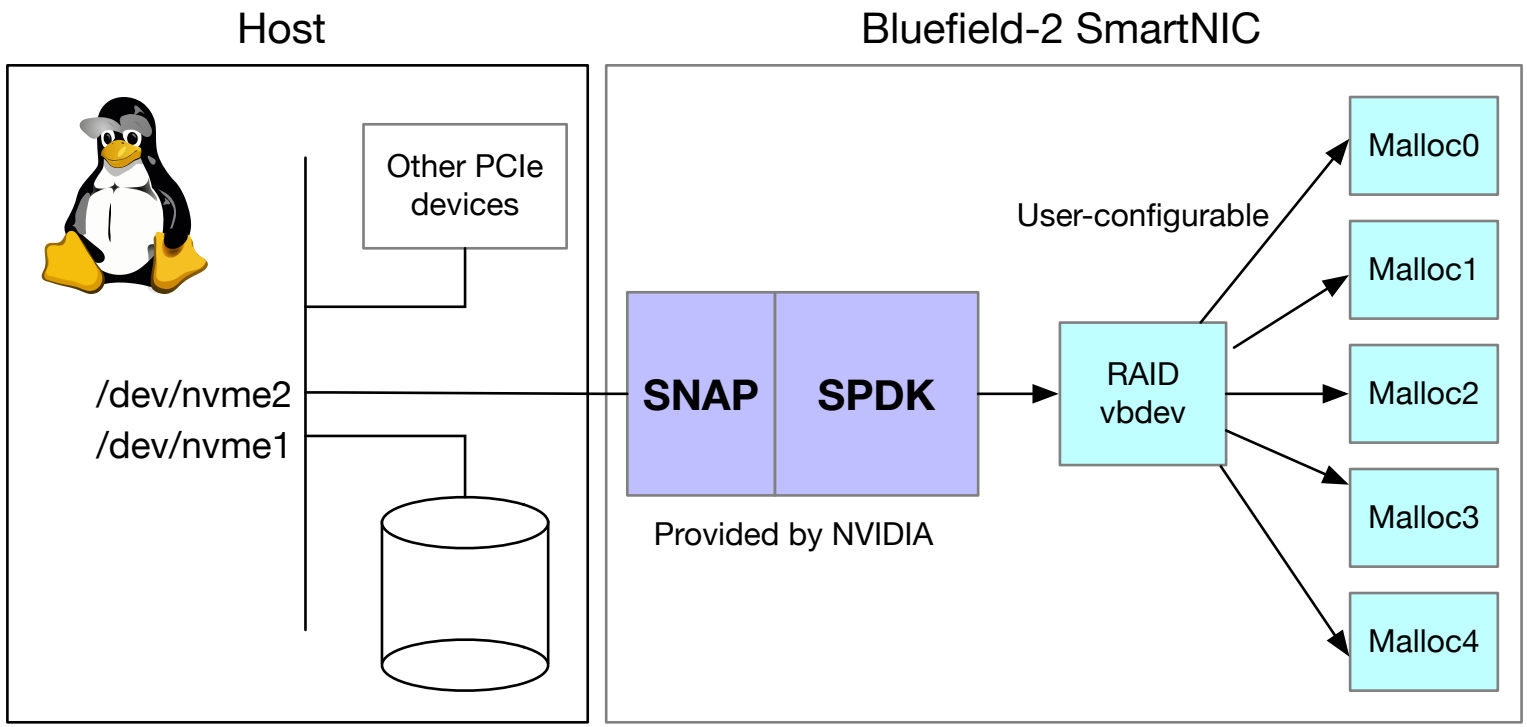
Vojtech Aschenbrenner, Sadman Sakib, John Shawger

Tuesday, Dec 12 2023

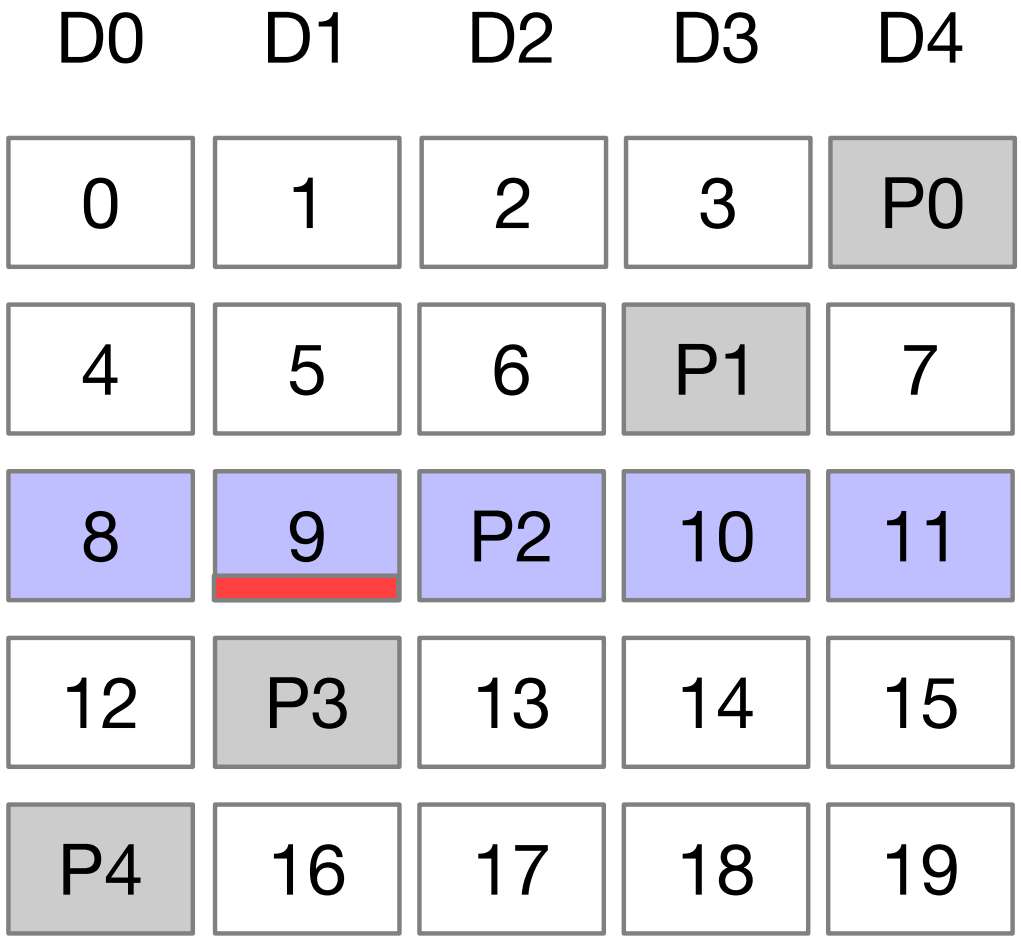
Motivation

- Storage for VMs is traditionally handled by the hypervisor, however there are several drawbacks
  - Consumes host CPU cycles
  - Requires hypervisor reboot to adjust storage stack
  - Context switches/hypercalls affect system performance
- Goal: VMs can use disaggregated storage without host performance overhead
  - SNAP and SR-IOV present an NVMe interface to each VM on host
  - Flexible storage stack requires no VM cooperation
  - No extra context switches/hypercalls on the host
  - Customizations easily applied on Bluefield-2

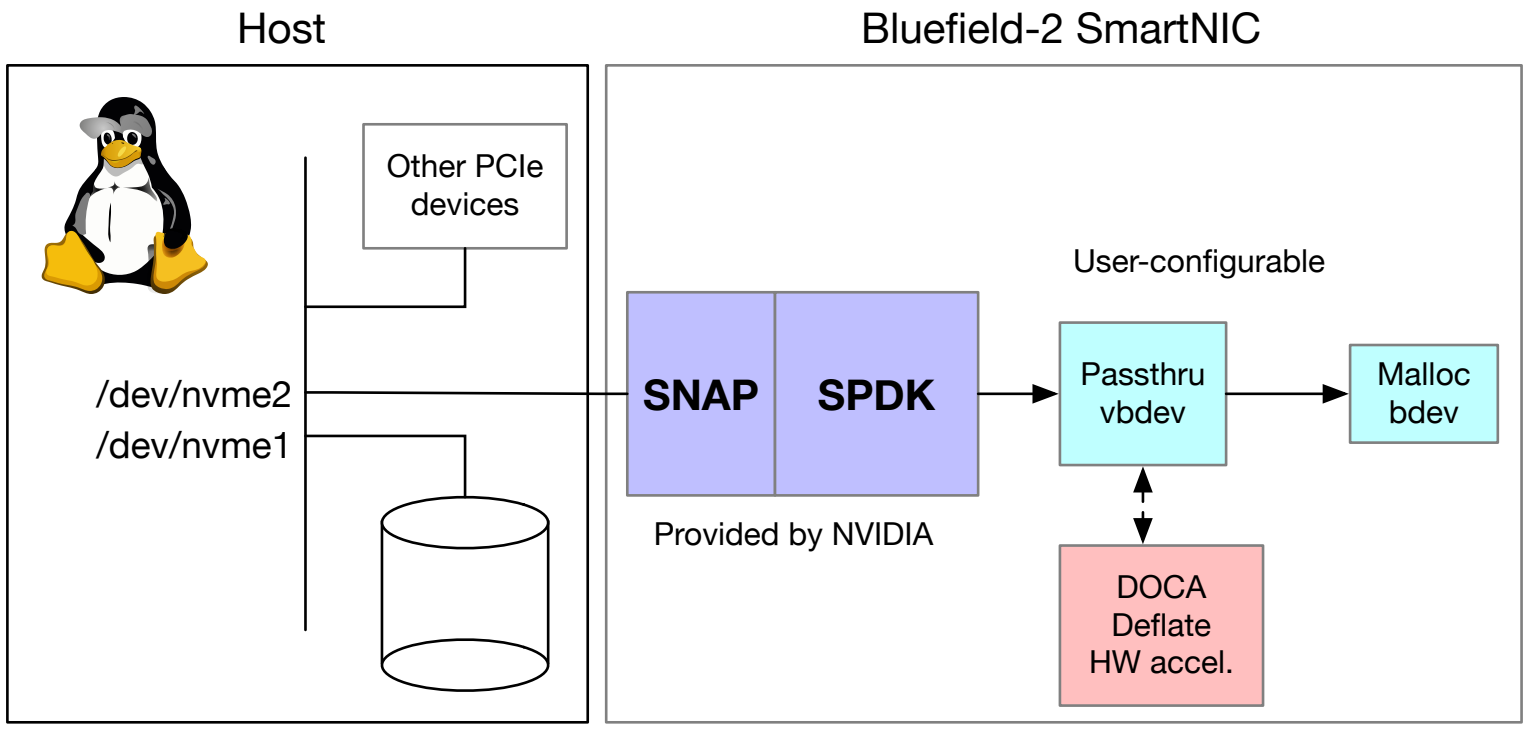
RAID5 bdev


Host
Bluefield-2 SmartNIC
Other PCIe devices
/dev/nvme2
/dev/nvme1
SNAP
SPDK
Provided by NVIDIA
User-configurable
RAID vbdev
Malloc0
Malloc1
Malloc2
Malloc3
Malloc4

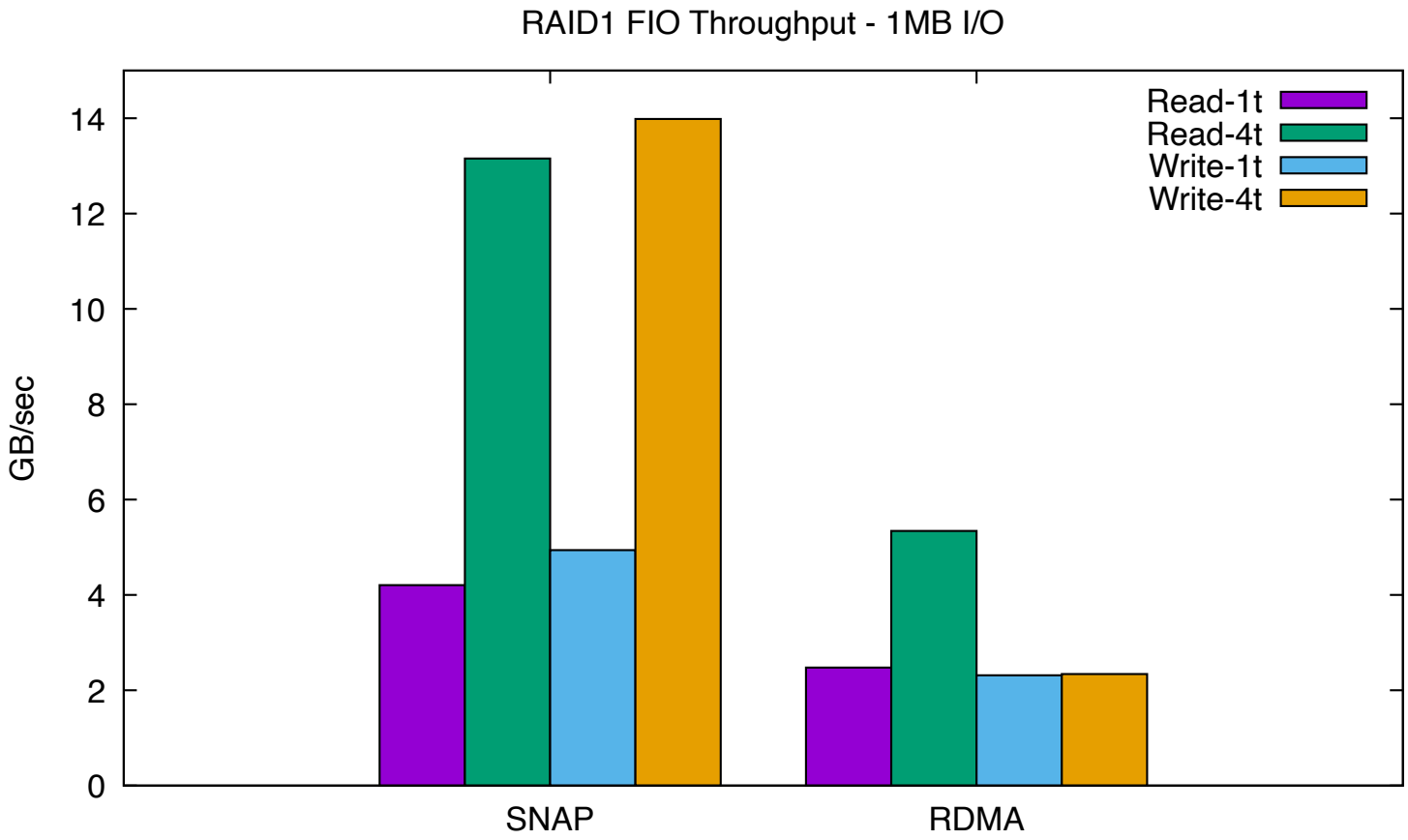
Safe RAID5


D0
D1
D2
D3
D4
0
1
2
3
P0
4
5
6
P1
7
8
9
P2
10
11
12
P3
13
14
15
P4
16
17
18
19

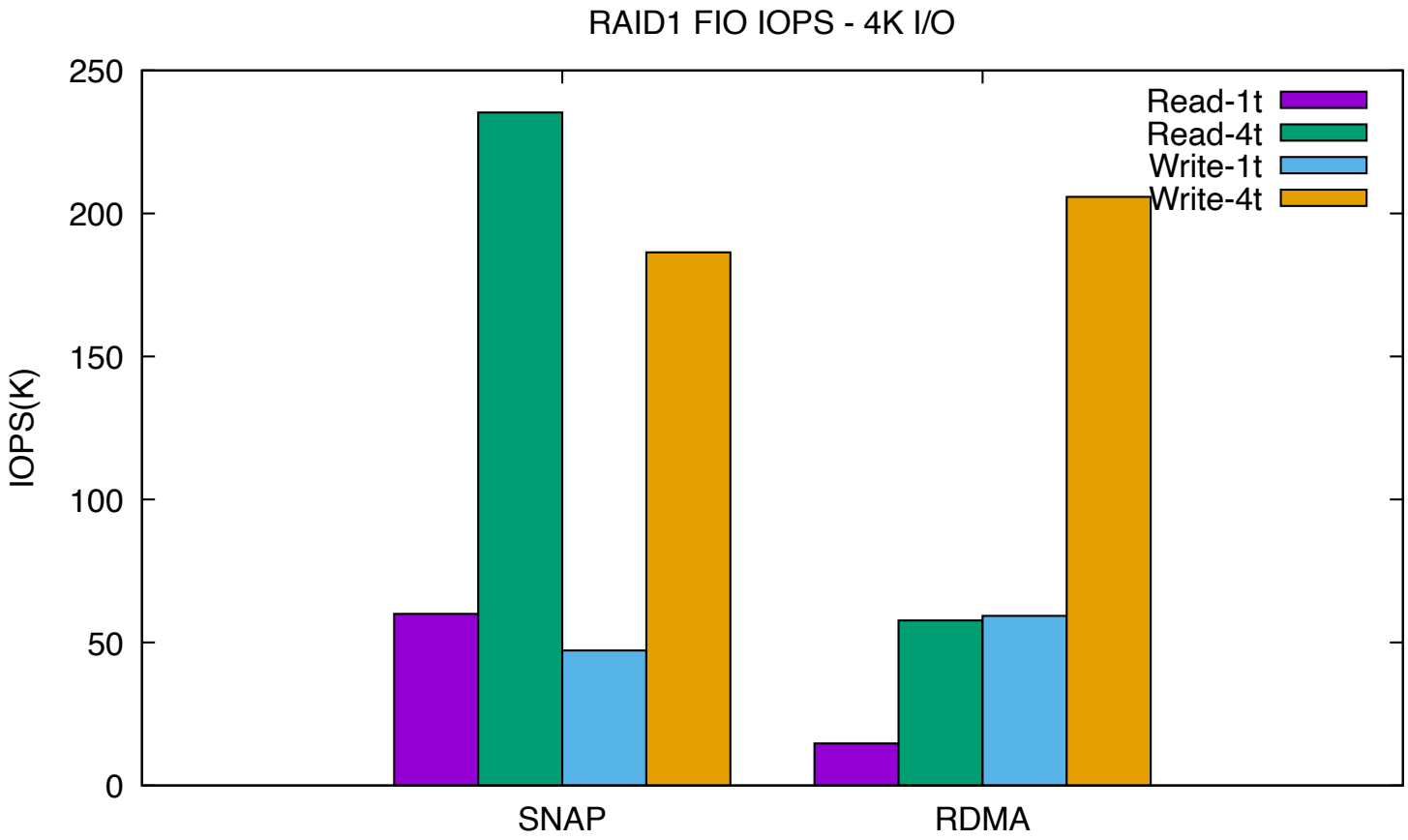
Compression bdev



SNAP vs. RDMA Throughput



SNAP vs. RDMA IOPS



- SNAP 4K latency (read): 16us
- RDMA 4K latency (read): 63.7us