NVMe Over Fabric and Direct Attached Host with NVMe SSD SR-IOV

Brian Pan

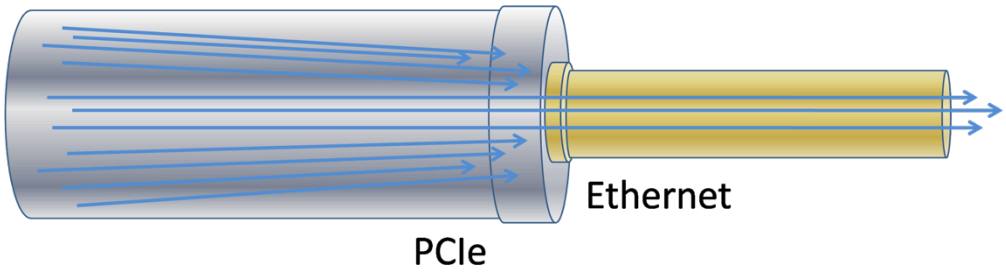
Aug 2018

H3

CONFIDENTIAL

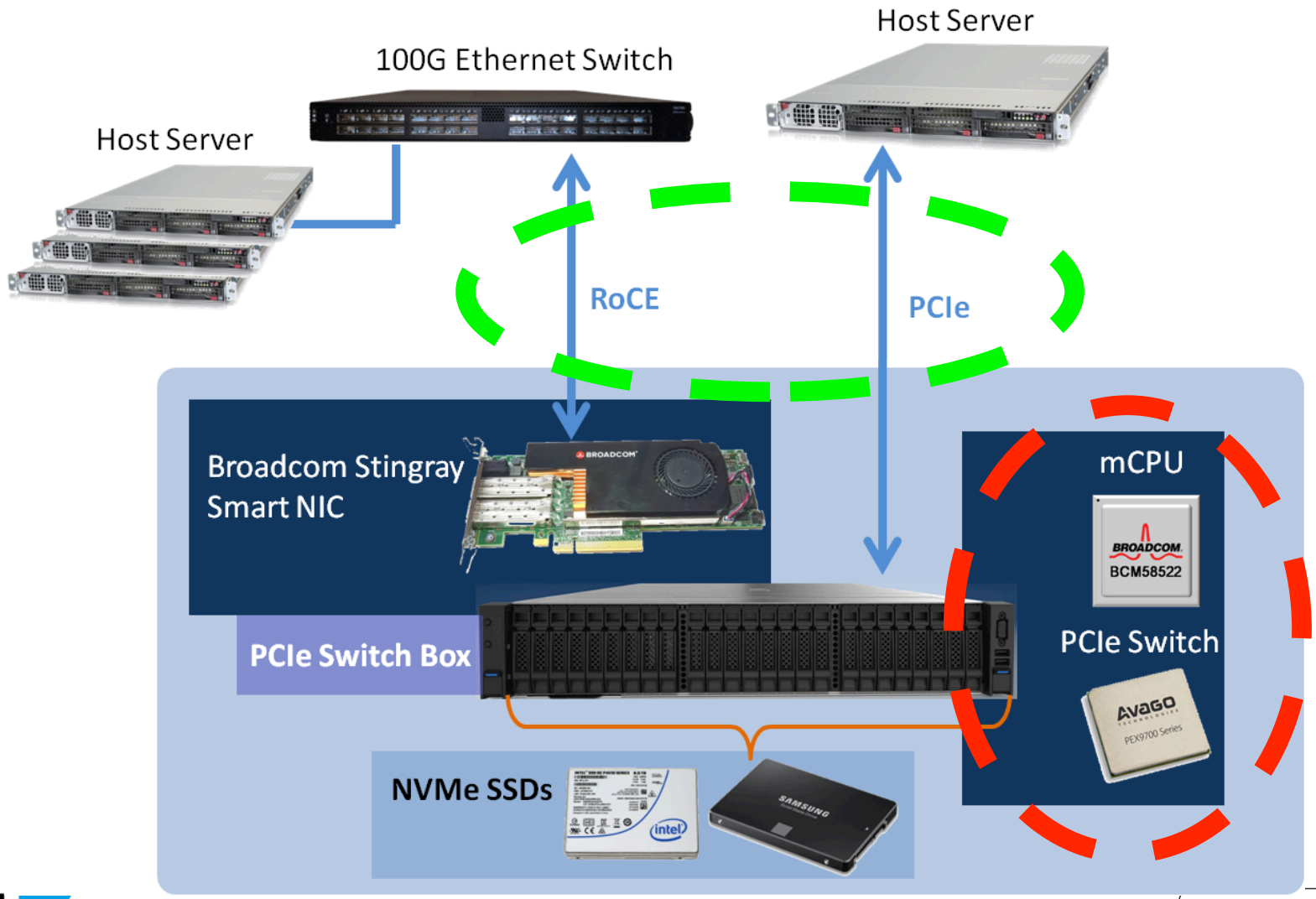
Complexity and Bottleneck

Power sequence, bus number, memory address, hotplug, OS and driver



Performance bottleneck

System Architecture-- Direct PCIe and NVMe-oF



System Specification

- Host connection
  - 2x 100G Smart NIC for NVMe over fabric
  - 2x PCIe Gen3 x16 for host connection
- NVMe SSD
  - 16x U.2 NVMe SSD (PCIe Gen3 x4)
- System specification
  - 2U form factor
  - 1+ 1 redundant 1400W
  - 4+1 redundant fan

NVMe SSD Pooling by PCIe Switch

Spec	Details
Dimension	**2U**
Front Port	**16 U.2 NVMe SSD**
Back Port	**Stingray, management, power**
Power	**1400W, 1+1 Redundant**

SMART NIC– Stingray
- ARM SoC with 100G smart NIC for ROCEv2

NVMe SSD device allocation
- Dynamically assign VF of NVMe SSD to smart NIC

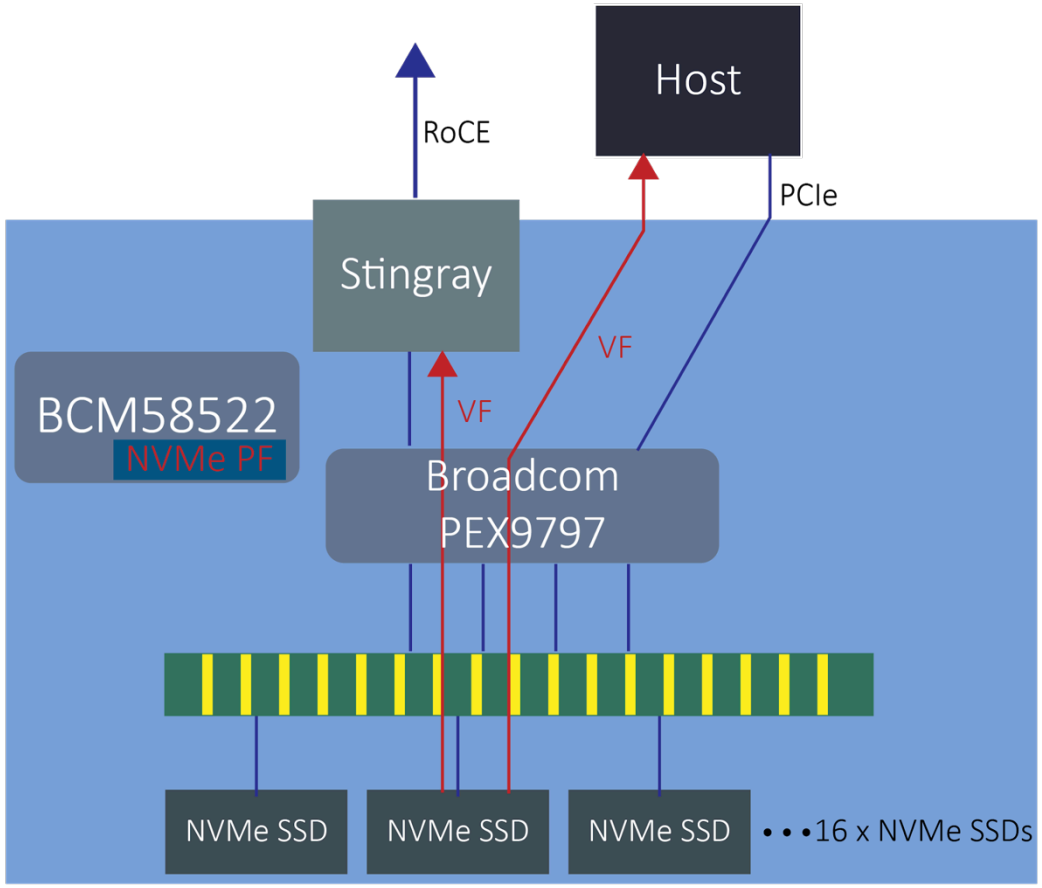
Hot-plug NVMe SSD
- Remove/ add/ re-allocate VF of NVMe SSD from one NIC to another NIC without system shutdown

Host-device port configuration
- Assign the PCIe slot as PCIe host connection or device ports dynamically (restart required)

Management API
- Follow the redfish standard and integrate with Intel RSD management software

H3

NVMe-oF System Architecture

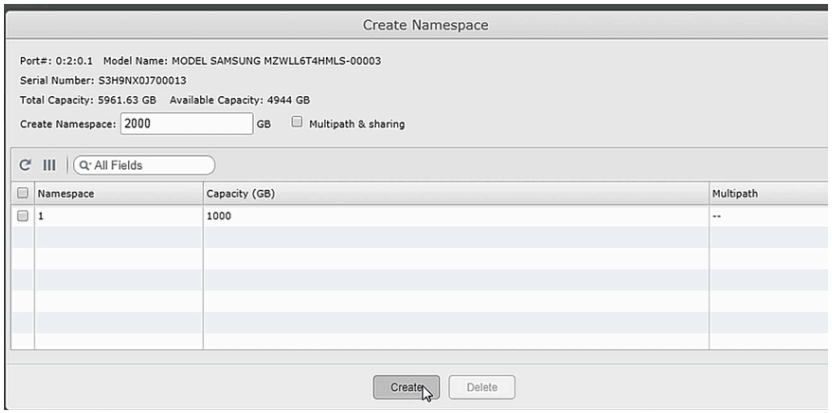


System

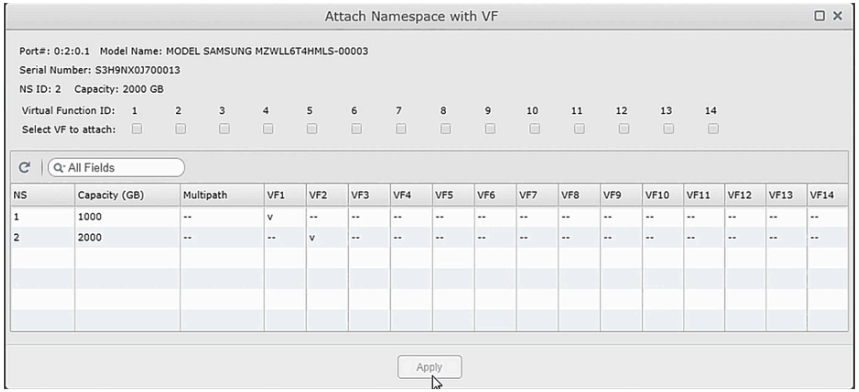
- 1x 25G Smart NIC
- 16x NVMe SSD with SR-IOV capability
- 1x 96 Lane PCIe switch
- PF of all NVMe SSD are installed in PCIe mCPU
- Each PF is with 5 VF. 5 VF are assigned to Smart NIC and host.

Demo of NVMe SR-IOV

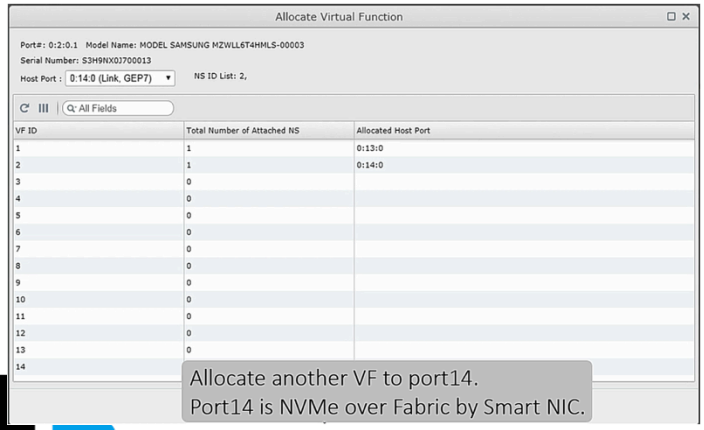
1. Create Namespace in NVMe



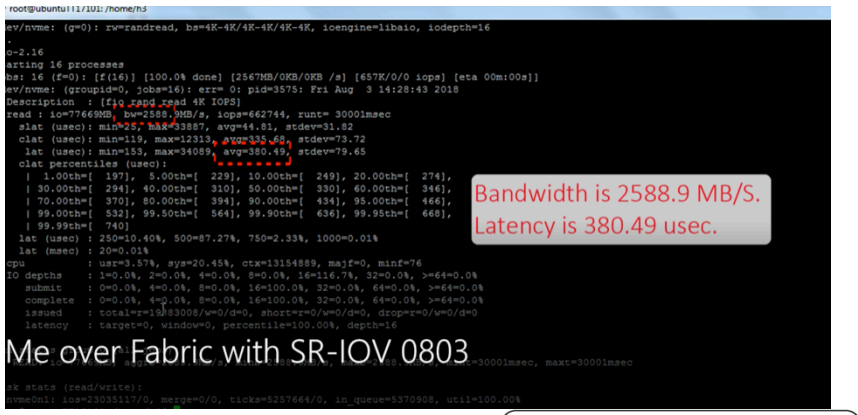
2. Map NS with VF



3. Assign VF to Smart NIC/ host



4. Run FIO test

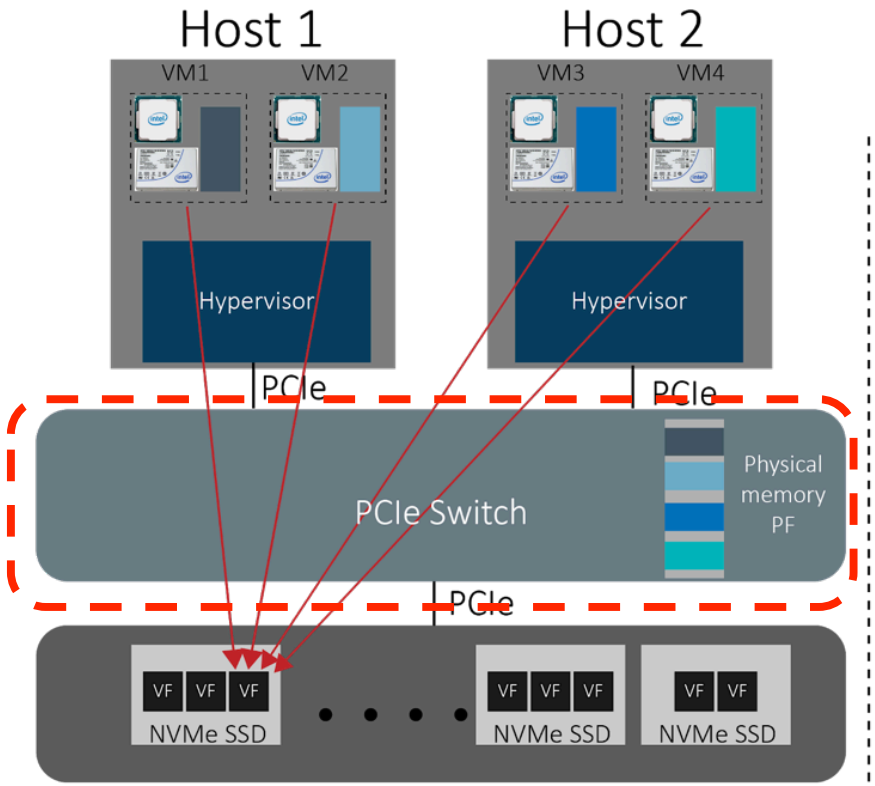


CONFIDENTIAL

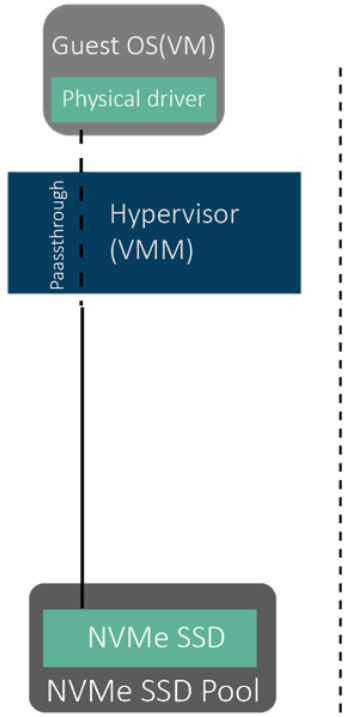
FIO Test Result of Direct Attach vs Smart NIC

Performance	Throughput	Latency
Directly attach to host	2853.5 MB/s	345.03 usec
NVMe over Fabric	2588.9 MB/s	380.49 usec

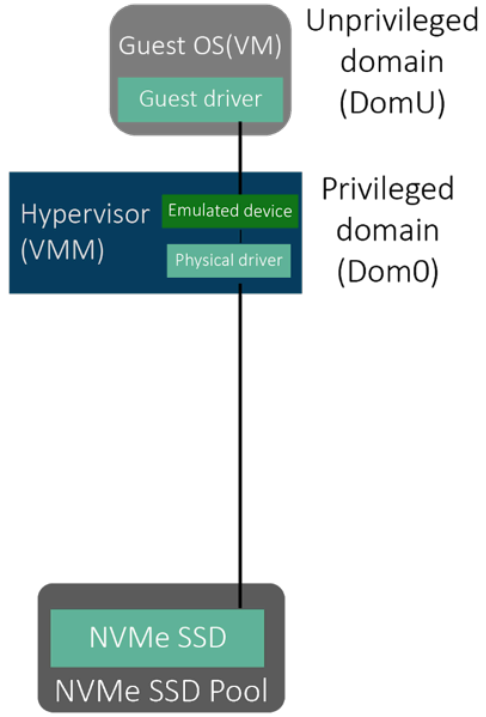
NVMe with or without SR-IOV Capability



NVMe with SR-IOV. VM talk to VF directly. PF is sit on PCIe switch.



NVMe without SR-IOV. Passthrough model



NVMe without SR-IOV. Hypervisor manage VMs to NVMe SSD

Performance Results of NVMe with SR-IOV

NVMe Virtual Functions on Linux KVM (Passthrough)

Read Performance

Tested Functions	4K Read (Random)		
	MB/s	IOPS	Latency (us)
VM_1 access to VF_1	280	68.5k	202
VM_2 access to VF_2	280	68.4k	202
VM_3 access to VF_3	277	67.5k	206
VM_4 access to VF_4	283	69.0kk	200
VM_5 access to VF_5	281	68.5k	202
VM_1 access to VF_1	310	75.6k	180

The performance measured using Fio in CentOS 7.5, with queue depth 16 by 1 worker.

VM_1 access to PF	90	22k	647

CONFIDENTIAL

Performance Results of NVMe without SR-IOV

NVMe Physical Function on Linux KVM (Shareable)

Read Performance

Tested Functions	4K Read (Random)		
	MB/s	IOPS	Latency (us)
VM_1 access to PF	90	22k	647
VM_2 access to PF	97	23.8k	596
VM_3 access to PF	92	22.5k	643
VM_4 access to PF	95	23k	616
VM_5 access to PF	90	22k	649
VM_1 access to PF	119	29k	488

The performance measured using Fio in CentOS 7.5, with queue depth 16 by 1 worker.

Key Benefits of NVMe SR-IOV

- Performance
  - The VF latency is only ⅓ of PF latency in multi-VMs environment
  - The performance is 3 times of PF performance
  - PCIe Gen3 x4 performance can be shared by multi-VMs
- Cost saving
  - Tens of VFs associated with a single PF, extending the capacity of a device and lowering hardware cost
  - With better latency and performance, the utilization rate will be higher to further reduce the hardware cost
  - Reduce NVMe SSD amount by sharing NVMe via PCIe

Key Benefits of NVMe SR-IOV

- Multi-path IO via PCIe
  - The name space on NVMe can be accessed by different hosts through PCIe connection
- Flexibility configuration
  - Dynamic control by the PF through registers designed to turn on the SR-IOV capability, eliminating the need via direct access to hardware from the virtual machine environment.
- Inter-operatability
  - A standard way of sharing the capacity of any given I/O device thus allowing for the most efficient use of that resource in a virtual system

Brian Pan | H3

GM

huaiyangpan

www.h3platform.com

brian.pan@h3platform.com

+886 2 2698 3800#110

H3