SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

NVMe as a Cloud Interface

Presented by Jake Oshins, Microsoft

Interface Choice

- Ethernet line rates are climbing quickly
  - 200Gb/s is common, today
  - Paravirtual interfaces aren't scaling
- NVMeoF works
  - Complicated to deploy
  - Relies on various NIC drivers
- NVMe has a driver in every OS
  - Intended for local storage
  - Works well, though, even for remote storage
  - Can integrate storage encryption

NVMe for Remote Storage: Semantic Mismatch

- Assumptions about namespaces
  - All come from a uniform pool of media
  - All have similar latency
  - All have the same caching properties
  - Total number of namespaces derived from size of pool divided by smallest useful volume
- Assumptions about queues
  - Count tied to number of cores in the machine
  - Used symmetrically across namespaces

Scaling Challenge

- Open Compute Project's Cloud SSD
  - 64-255 functions
  - 2-4 PCIe lanes
- SR-IOV
  - 64-2000 virtual functions
  - 2-16 PCIe lanes

- Smart NICs front for petabytes of storage
  - A controller per storage object isn't practical
- 2-socket, 1U host can have hundreds of cores
  - Hosting thousands of containers, each with multiple processes

Nested Virtualization is Very Difficult

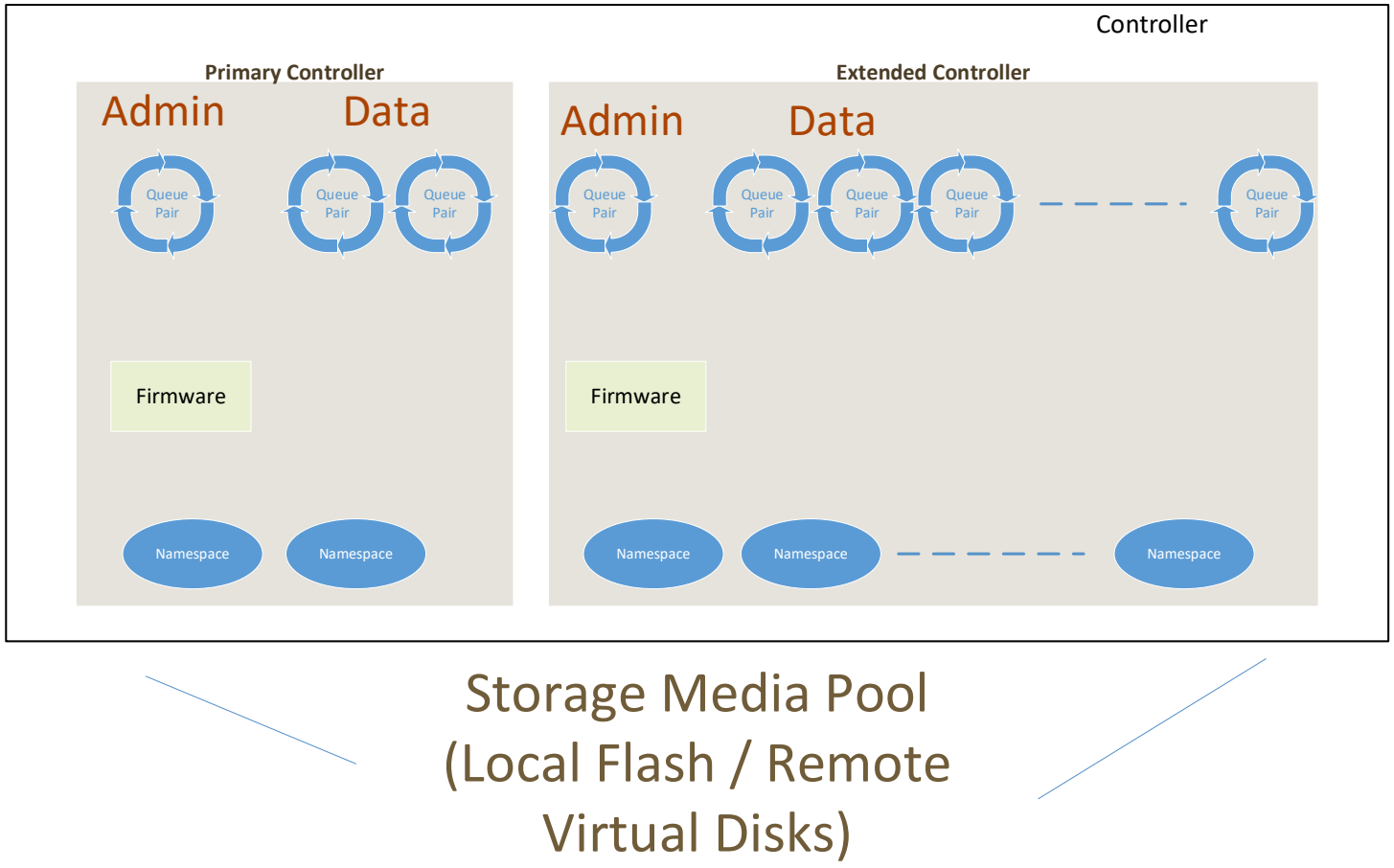
- Containers sometimes exist within VMs
  - VMs sometimes exist within VMs
- Processor's view of an NVMe controller consists of:
  - 1 Control/Status page
  - Doorbells
  - MSI-X table
- Queue IDs leak into Completion Queue Entries
- Admin Queue is difficult to intercept in a hypervisor
- SPDK, user-mode clients, are another level of nesting

- Create “logical controllers” out of large pools of queues and namespaces
  - Similar to NVMe over Fabrics’ “MI – TP 6011 Scalable Resource Management”
  - Have exactly one full PCIe based NVMe controller with a single function on the bus
  - Have “extended controllers” that are just groupings of queues, borrowed from that primary controller
  - The control-plane software associates these queues with NVMe namespaces (virtual disks)
  - A tagging mechanism on the PCIe fabric is used to do address translation

- Only one set of MSI-X table entries mapped into memory space
- Only one set of queue doorbells mapped into memory space
  - All doorbells are in the PF BAR
- Nested virtualization does not require the hypervisor to intercept and process the admin queues
- One PCI Configuration space
- One PCIe-NVMe Controller first page of control/status registers
- Virtualization of queue numbers, MSI-X table indices, PASIDs, etc. can be done entirely in firmware, without adding additional hardware

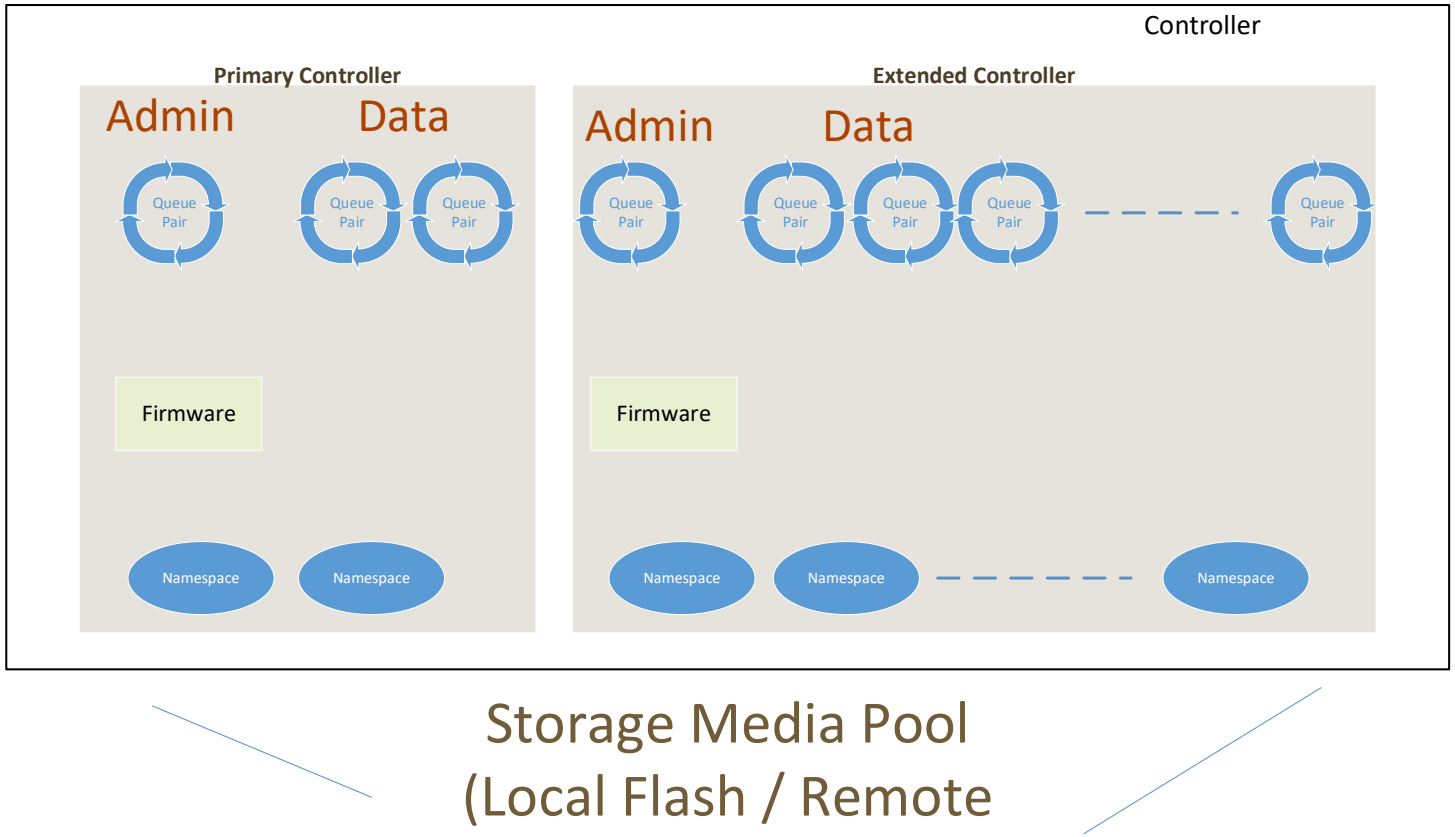
Mechanics

- When creating an extended controller, the driver sends a command to the primary Admin queue, requesting the creation of that extended controller and associating a set of bus tags with it
- Next, the primary driver sets Quality of Service parameters and limits on the number of queues that the extended controller can use
- More commands associate underlying namespaces with the extended controller and assign namespace numbers to them that are relative to that controller
- Finally, the primary Admin queue creates another admin queue for the extended controller



Mechanics

- At this point, the driver for the extended controller sends data queue creation commands to the new admin queue
- Because the creation of the new admin queue and the data queues will be associated with a bus tag, the addresses used for the ring buffers can be relative to the host or relative to a VM or container
- As the device fetches commands from submission queues, these addresses will be translated into guest-relative terms
- The host-relative view of the controller is flat. Any virtualization happens within a hypervisor. No virtual function base address registers are necessary.



Next Steps

- Scalable IOV Working Group in Open Compute Project
  - Existing (v1) Specification isn't quite sufficient.
    - Workstream currently defining missing pieces
- Result will feed into PCIe Specification, CXL Specification
- Then, on to NVM Express

- We're looking for people who are interested in the discussion
- Engage SNIA to discuss collaboration points across working groups

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

Architectures, Solutions, and Community
VIRTUAL EVENT, APRIL 11-12, 2023

Please take a moment to rate this session.

Your feedback is important to us.

