SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Distributed Block Storage using NVMe-over-Fabric

Sujoy Sen, Senior Principal Engineer

Mohan J Kumar, Fellow

Intel

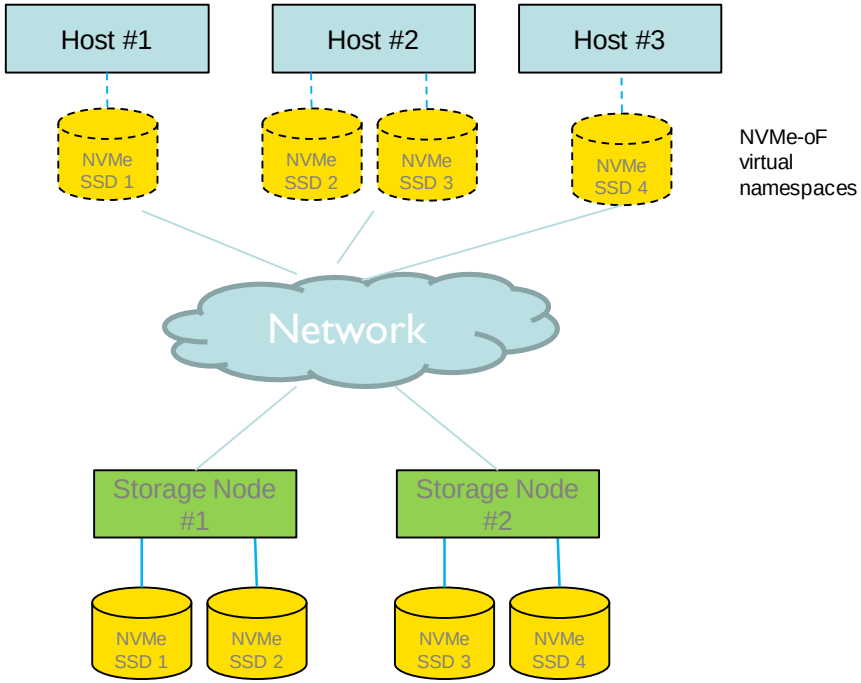
***Acknowledgements*: Scott D Peterson, Reddy Chagam**

Introduction

- NVMe+NVMe-oF+Distributed Scale-Out Storage (DSS)
  > *Element of* {NVMe, NVMe-oF, DSS}
- Overview of NVMe-oF and DSS
- Issues with DSS clients
- Solution Options
- Proposed Concept
- Summary

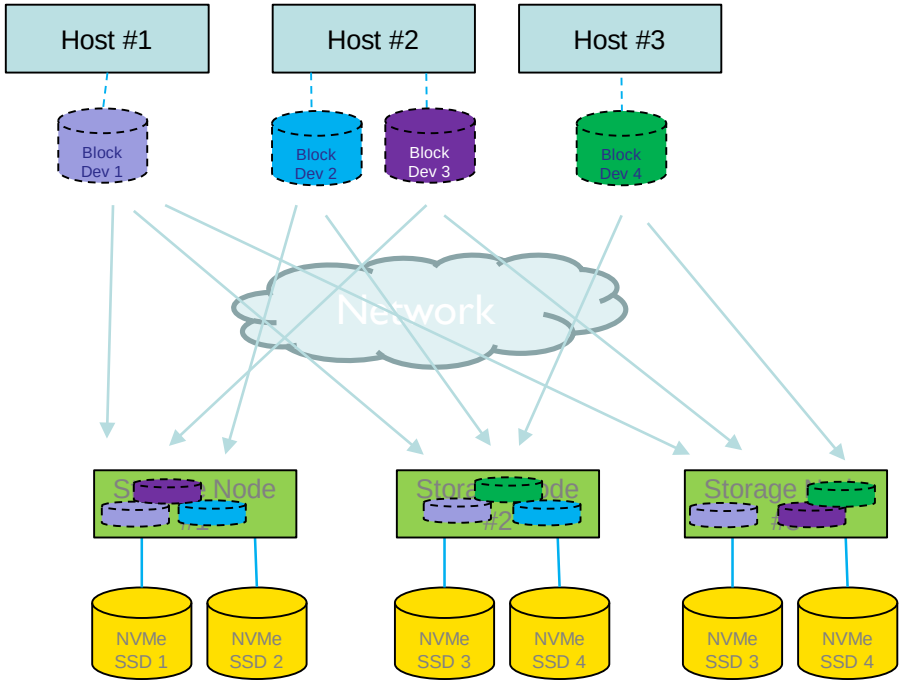
NVMe-over-Fabric

- Surface NVMe Drives to remote systems
  - Appears as NVMe drive/namespace to remote hosts
- Low latency, efficient and high performance
- 1:1 mapping between physical and virtual namespace
- Low footprint initiator (user or kernel mode)
- Industry standard interface (NVMe) and fabric protocol
  - Widespread industry adoption and eco-system support



Distributed Scale Out Storage

- Block Storage provided by a cluster of co-operating nodes
- Block device distributed across nodes for HA and performance
- Capacity and Performance can be scaled by adding storage nodes
- Solution specific client (host) access module and interface
- Solution specific communication protocol
- Open Source and commercial offerings

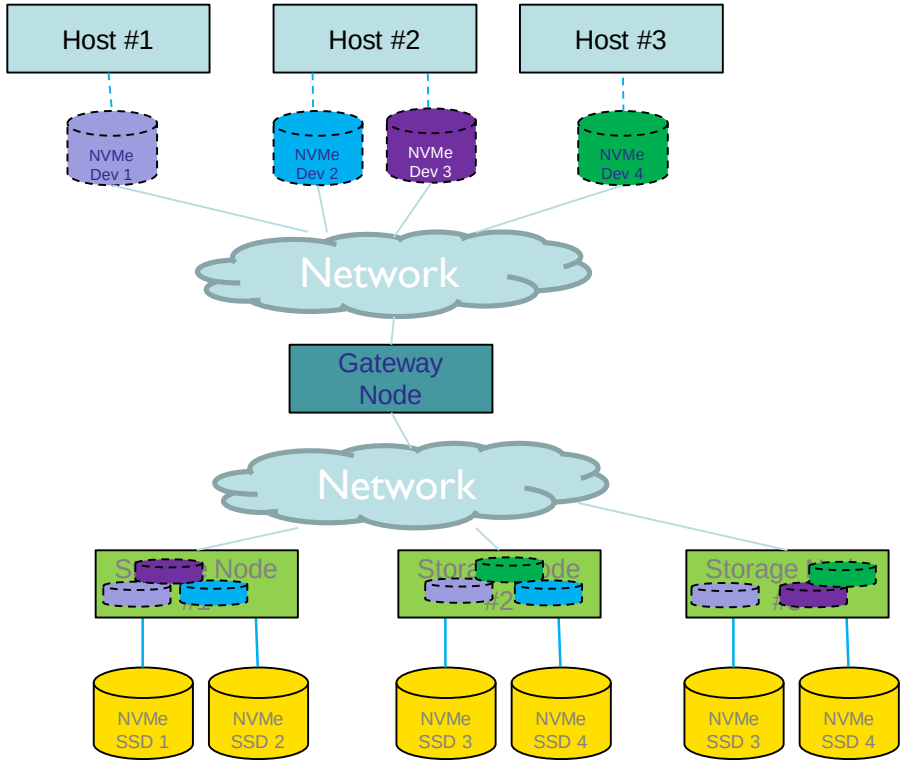


Impact of solution-specific storage client

- Requires drivers for every operating system and hypervisor
- Requires solution-specific management
- Host is strongly coupled with the storage service
  - Adds lifecycle management complexity
  - Storage cluster attack surface extended to host
- May have significant resource footprint at the host
  - Takes away valuable resources from applications
  - Challenging to embed in Infrastructure Accelerators

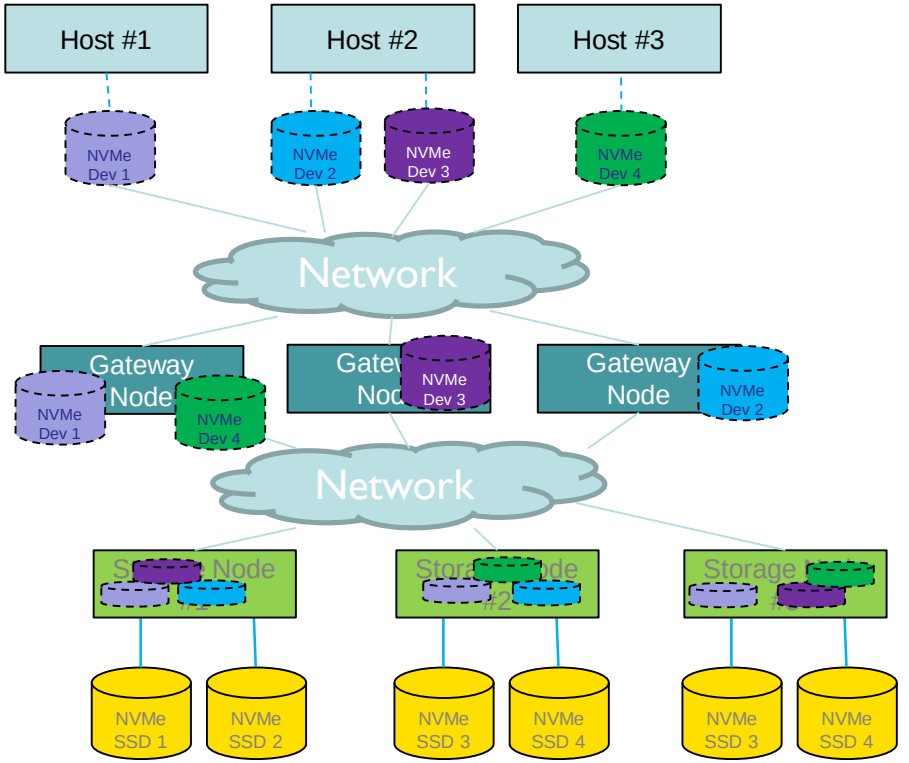
Gateways

- Standards based client replaces custom client
  - iSCSI, NVMe-oF etc.
- Smaller client footprint
- Decoupled from the storage service cluster
- Lower Performance
  - Adds latency due to an extra hop/node
  - Gateway can become a bottleneck
- Management Complexity and increased cost



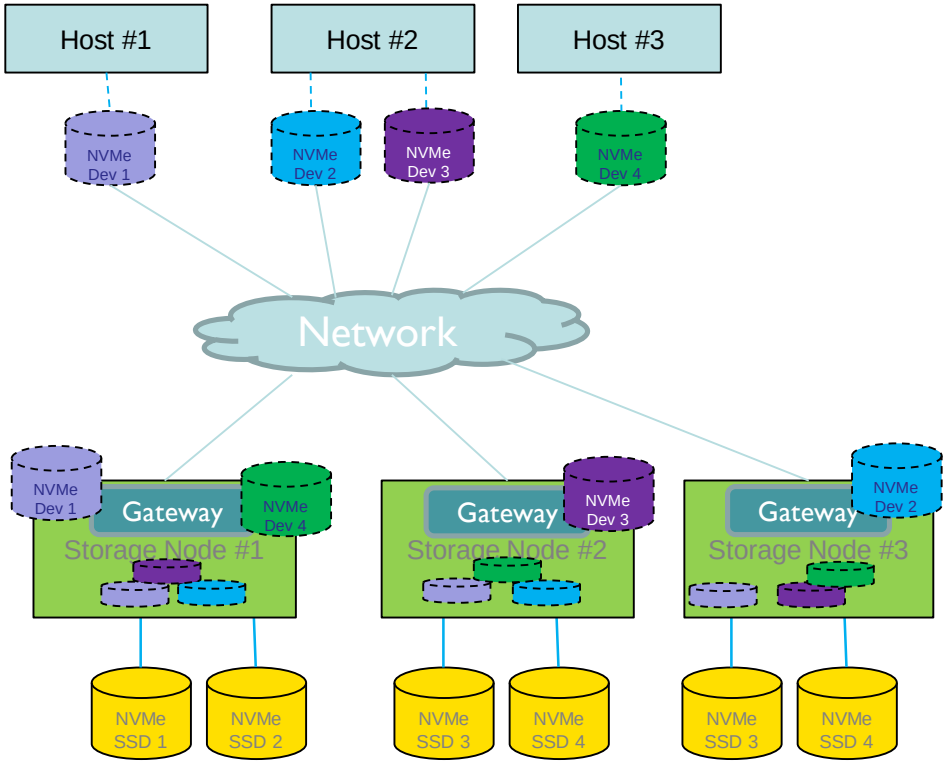
Multiple Gateways

- Add multiple gateways
- Performance
  - Alleviates gateway bottleneck
  - Extra-hop Latency
- Various assignment policies
  - Random
  - Dynamic Load balanced
- More management complexity
- Higher cost



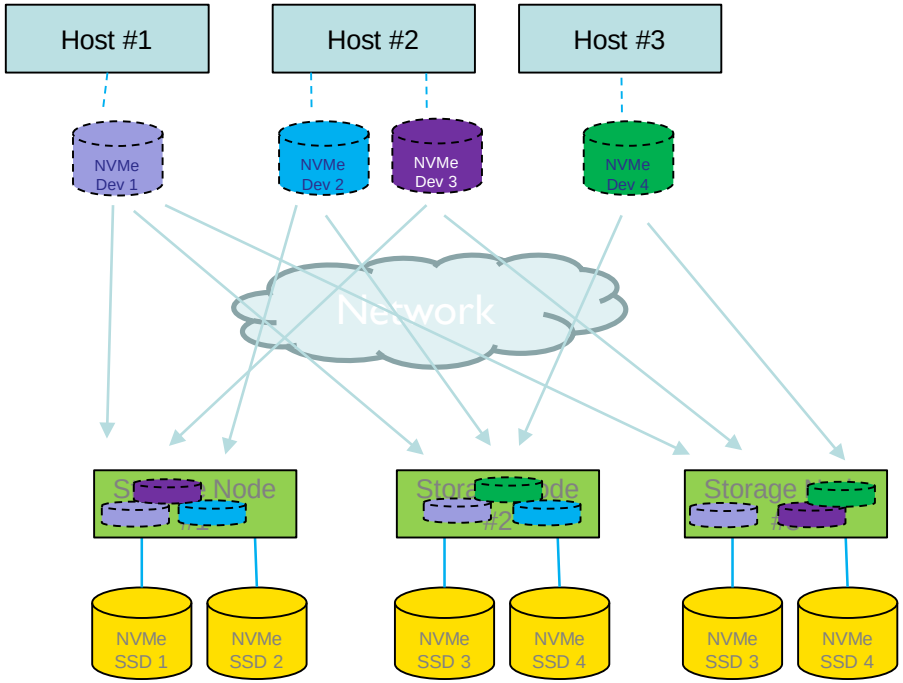
Integrated Gateways

- Integrate gateways into the storage node
- Reduces cost, some reduction in management complexity
- Performance
  - Extra-hop latency due to I/O forwarding
  - Gateway bottleneck adds CPU pressure to storage node



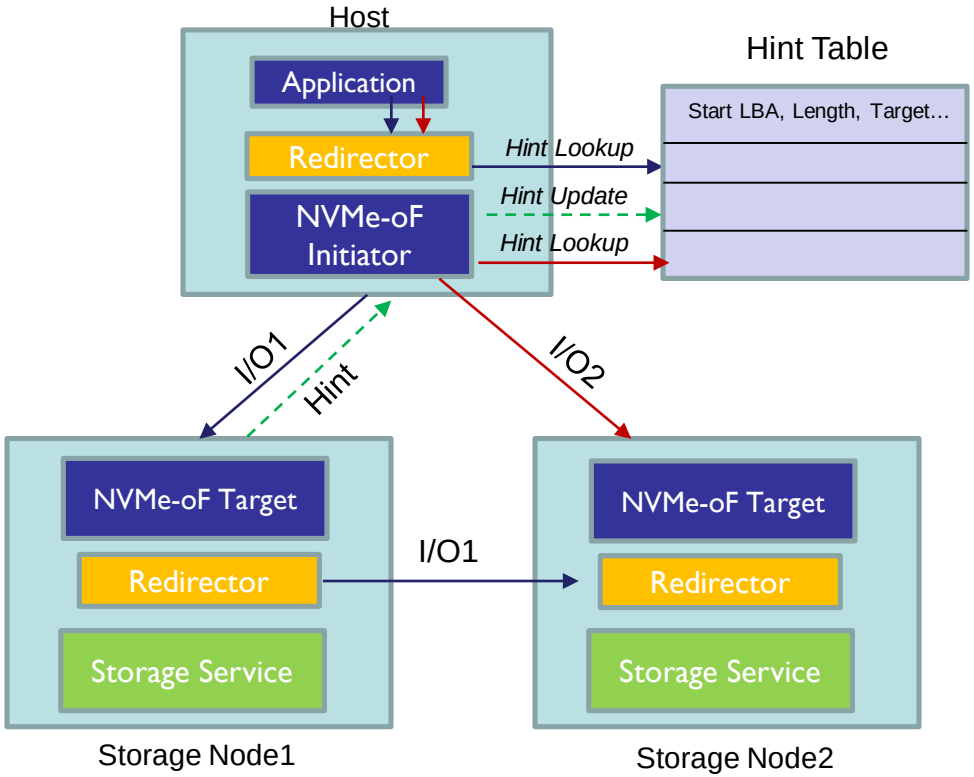
Distributed Integrated Gateways

- Natively distribute NVMe volume to the *correct* storage service node
- Requires NVMe-oF volume to be mapped to Target Tuples
- Requires NVMe-oF volumes to support extensible placement mechanisms



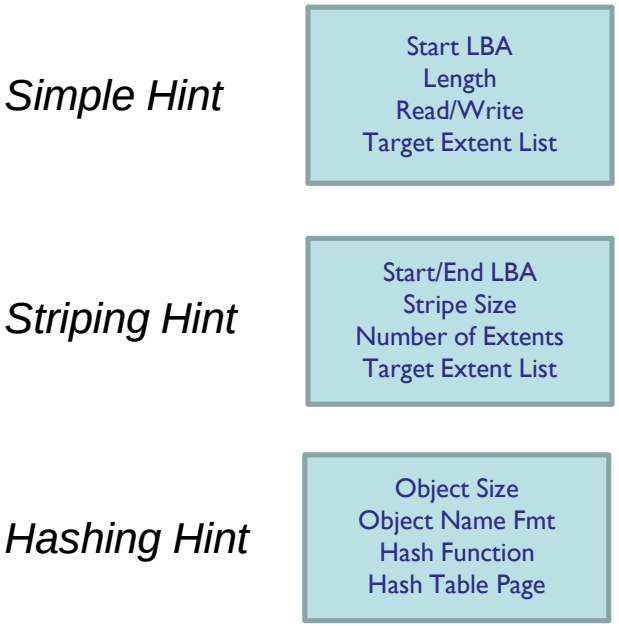
Proposal: Distributed Self-learning NVMe-oF

- Introduce the concepts of
  - Redirector
  - Hints
- Initiator redirector directs I/O to the correct NVMe-oF target based on the hint table
- Target Redirector forwards I/O to the correct storage node and propagates hints backward and forward
  - Allows fully decoupled and legacy clients
- Initiator redirector updates local hint table based on received hints about a volume
  - Eventually learns about the volume's distribution across nodes



Hints

- The hints allows flexible mapping of LBA to nodes
- Algorithmic or specific maps can be supported
  - Simple Hints
  - Striping Hints
  - Hashing Hints
- Allows NVMe-oF to support storage services ranging from pair-wise HA to massive scale out

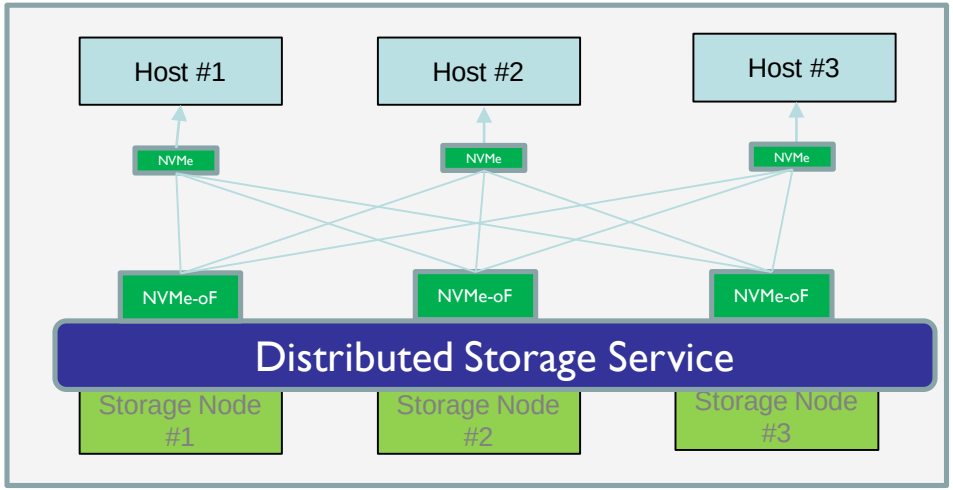


NVMe-oF Changes

- Re-use existing NVMe-oF elements
- Log Pages for Hints
  - Namespace specific log pages
  - Log Page must be consistent during read operation
- AER for Hint notification
- Redirector Capability Discovery
  - Allow Initiators to discover Redirector capable Targets
  - Use “Supported Capabilities” in GetFeatures command

Summary

- Distributed Scale-out Storage can benefit from standards-based host interfaces
- NVMe is the ideal host interface for scale-out storage
- NVMe-oF can be naturally extended to support scale-out storage



	Standard Host Interface	Gateway Availability	Gateway Performance Bottleneck	Low Latency
DSS				
DSS+ NVMe-oF GW				
DSS + Multiple NVMe-oF GW				
DSS+ Distributed SL NVMe-oF				