Host Managed Live Migration Panel

Sponsored by NVM Express organization, the owner of NVMe® specifications

Host Managed Live Migration Panel Agenda

- Open Ecosystem Alignment (Klaus Jensen - Samsung)
- Real Customer Use Cases:
  - Microsoft (Lee Prewitt)
  - Google (Nicolae Mogoreanu)
  - Nvidia (Chaitanya Kulkarni)
- Questions & Answers

Speakers

Klaus Jensen

SAMSUNG

Lee Prewitt

Microsoft

Nicolae Mogoreanu

Google

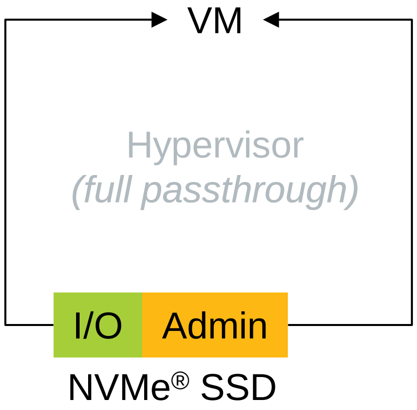
Chaitanya Kulkarni

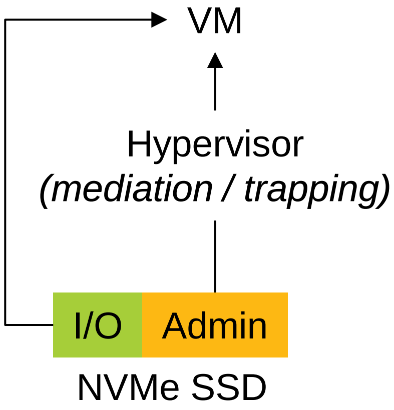
NVIDIA

Samsung Live Migration Use Case: Host Integration, Dirty Tracking and Virtualization

What the **Open Ecosystem** Must Solve

1. Migration Management Host **Integration**
    - Full function **pass-through** or **mediation**





What the **Open Ecosystem** Must Solve

2. Dirty Tracking
   - **Translation Agent** or **Device** assisted

3. NVMe® Controller and PCIe® Function **Virtualization**
   Generational and/or cross-vendor compatibility, MC privilege restriction
   - may be provided by **device**, or
   - if device is **mediated**, can be done in **host software**

Microsoft Live Migration Use Case: VM Support

Why Use Live Migration?

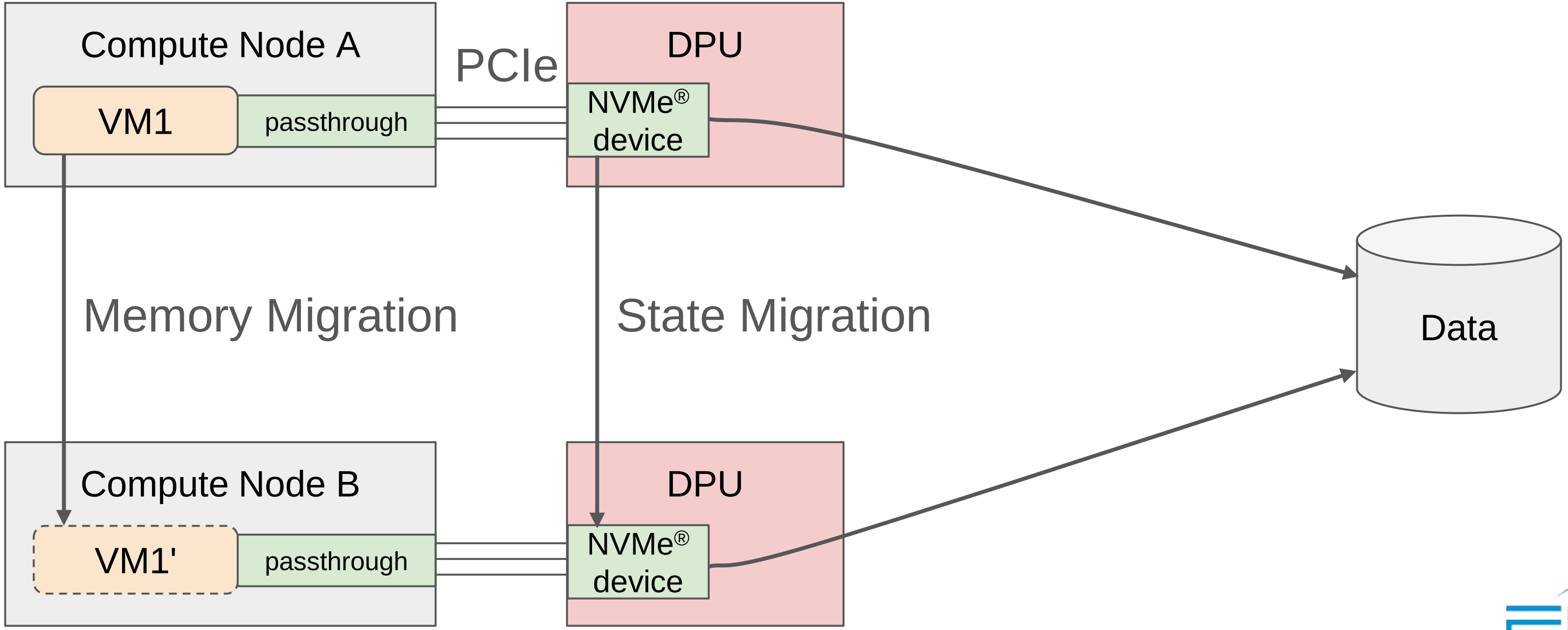
- Customers expect long up times on their VMs with no interruptions
- While very reliable overall, server nodes are complex and have issues:
  - Hardware failures; both immediate and predicted (ML)
  - Firmware updates; security, bugs, features
  - Resource exhaustion; load balancing
- Live migration allows for robust VM support on imperfect hardware

NVMe® Live Migration

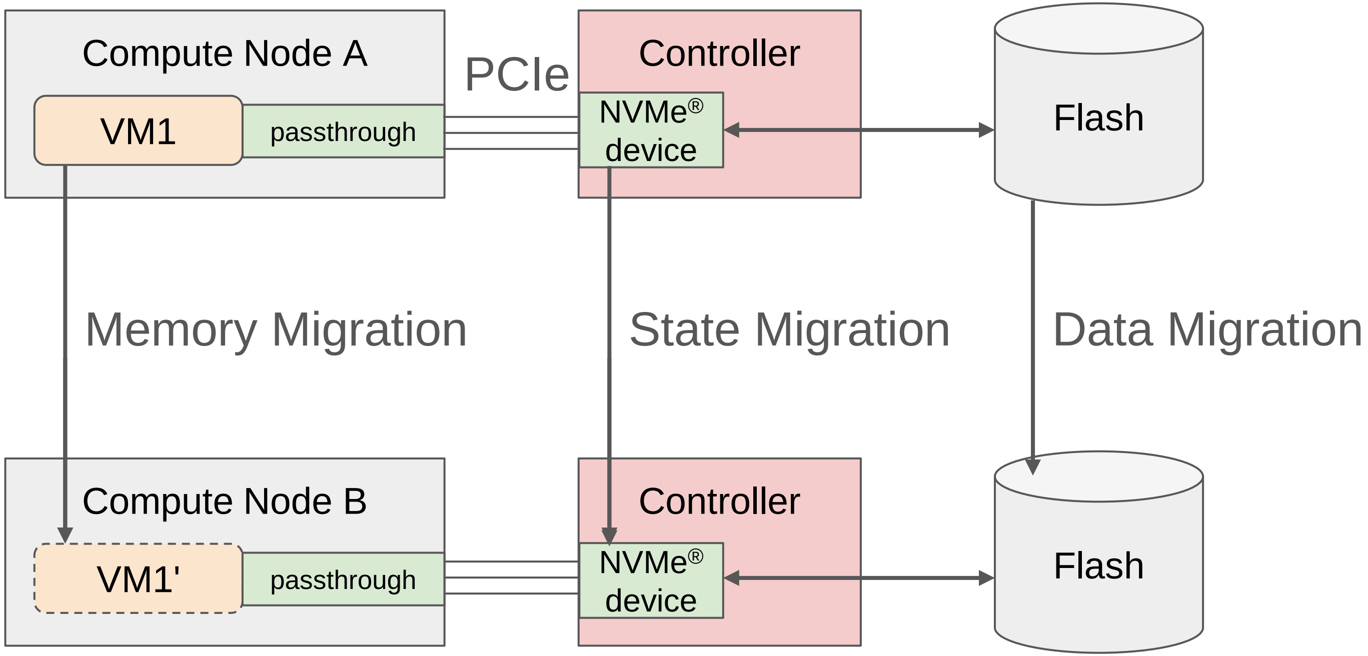
- One standard for use across multiple CSPs
  - Reduces work for vendors (common FW, reduced validation)
- Allows for secure separation of Host controller and Guest VM controllers (MPF, SR-IOV)
- Allows for independent encryption and sanitization
- Allows for Host controller to have access to telemetry for debuggability

Google Live Migration Use Case: Remote and Local Storage

Remote Storage Use Case



Local Storage Past and Present



Google Industry Alignment Focus Areas

Google Compute Engine (GCE)

- Controller presentation on the admin queue
- Antagonist and untrusted workload isolation
- Controller insight debuggability / telemetry

Internal

- Root of trust and encryption
- Left shift, reduce time to market.
    - Reduce iterations, expose requirements and validation

NVIDIA Live Migration Use Case: Live Migration Flow

Why Use NVM Express® with Virtual Function I/O (VFIO Mode)?

- Virtual machines often make use of direct device access when configured for the highest possible I/O performance

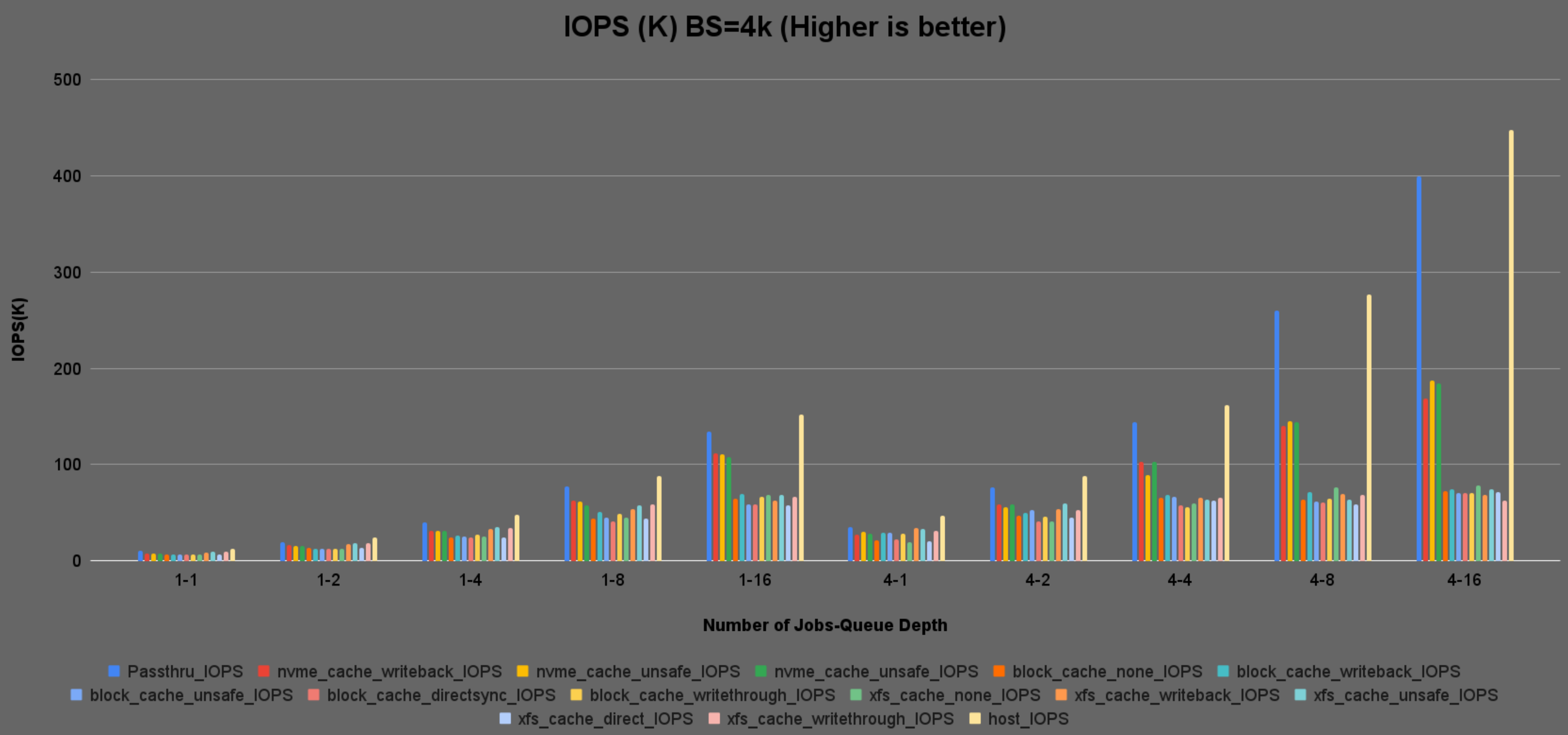
- From a device and host perspective, this simply turns the VM into a userspace driver, with the benefits of significantly reduced latency, higher bandwidth

Why Use NVM Express® with VFIO Mode ?

- Applications, particularly in the high-performance computing field, also benefit from low-overhead, direct device access from user space
- Examples include network adapters (often non-TCP/IP based) and compute accelerators
- NVMe® Protocol is particularly designed for the high performance where users can get maximum performance out of storage

Performance Matrix

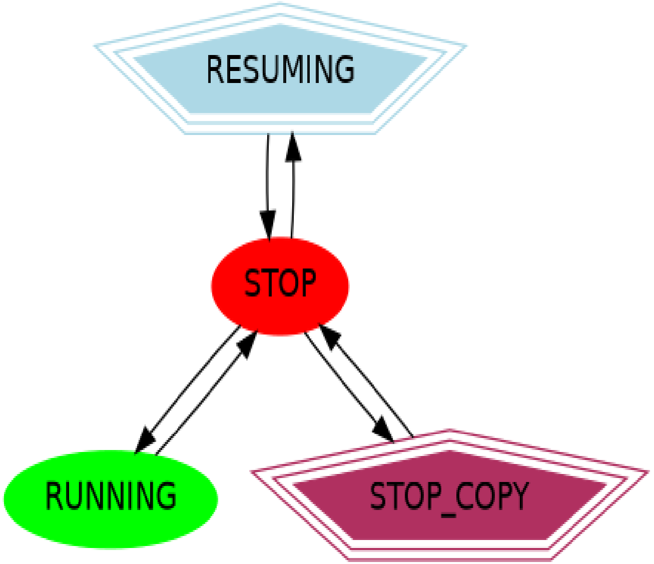
- IOPS (K)/Bandwidth (MB/s)/Latency
- CPU Guest User/ System
- CPU Host User/System
- IOPS Per Core/Bandwidth Per Core
- Block Size 4k, jobs 1 and 4
- Queue Depth 1/2/4/8/16
- Backend Categories:-
    - Pass-through (VFIO)
    - QEMU Userspace NVMe driver NVMe controller (3 Modes)
    - QEMU virio-blk on NVMe controller (5 Modes)
    - File created on XFS formatted on NVMe controller (5 Modes)


IOPS (K) BS=4k (Higher is better)
IOPS(K)
500
400
300
200
100
0
1-1
1-2
1-4
1-8
1-16
4-1
4-2
4-4
4-8
4-16
Number of Jobs-Queue Depth
Passthru_IOPS
nvme_cache_writeback_IOPS
nvme_cache_unsafe_IOPS
nvme_cache_unsafe_IOPS
block_cache_none_IOPS
block_cache_writeback_IOPS
block_cache_unsafe_IOPS
block_cache_directsync_IOPS
block_cache_writethrough_IOPS
xfs_cache_none_IOPS
xfs_cache_writeback_IOPS
xfs_cache_unsafe_IOPS
xfs_cache_direct_IOPS
xfs_cache_writethrough_IOPS
host_IOPS

VFIO NVM Express® Live Migration FSM

- Supporting Live Migration includes creating vfio-nvme implementation that will support VFIO live migration Finite State Machine (FSM). See next slide
- This also includes support from the NVM Express® protocol that will allow us to execute the subsequent command that are sent from the VFIO FSM

Simplistic View of VFIO Live Migration FSM



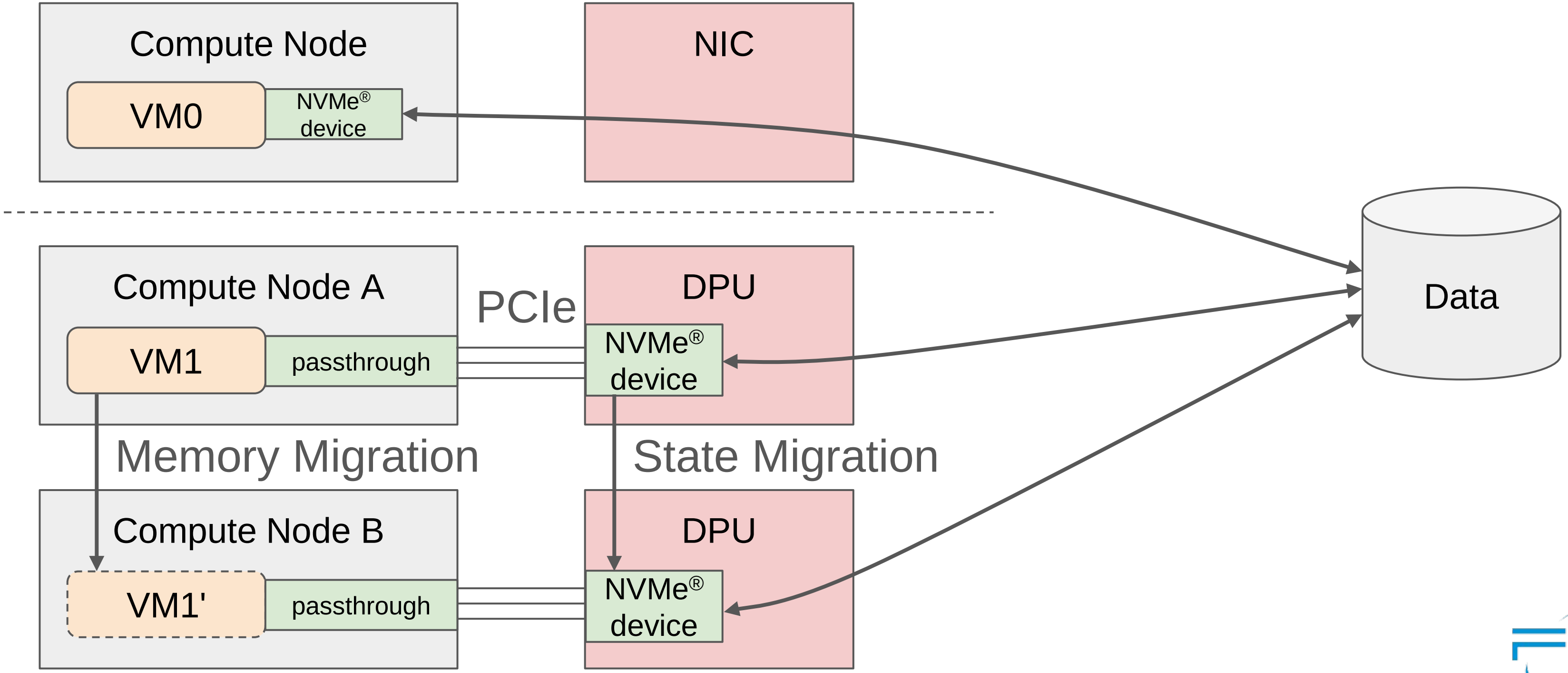
Questions?

nvm EXPRESS®

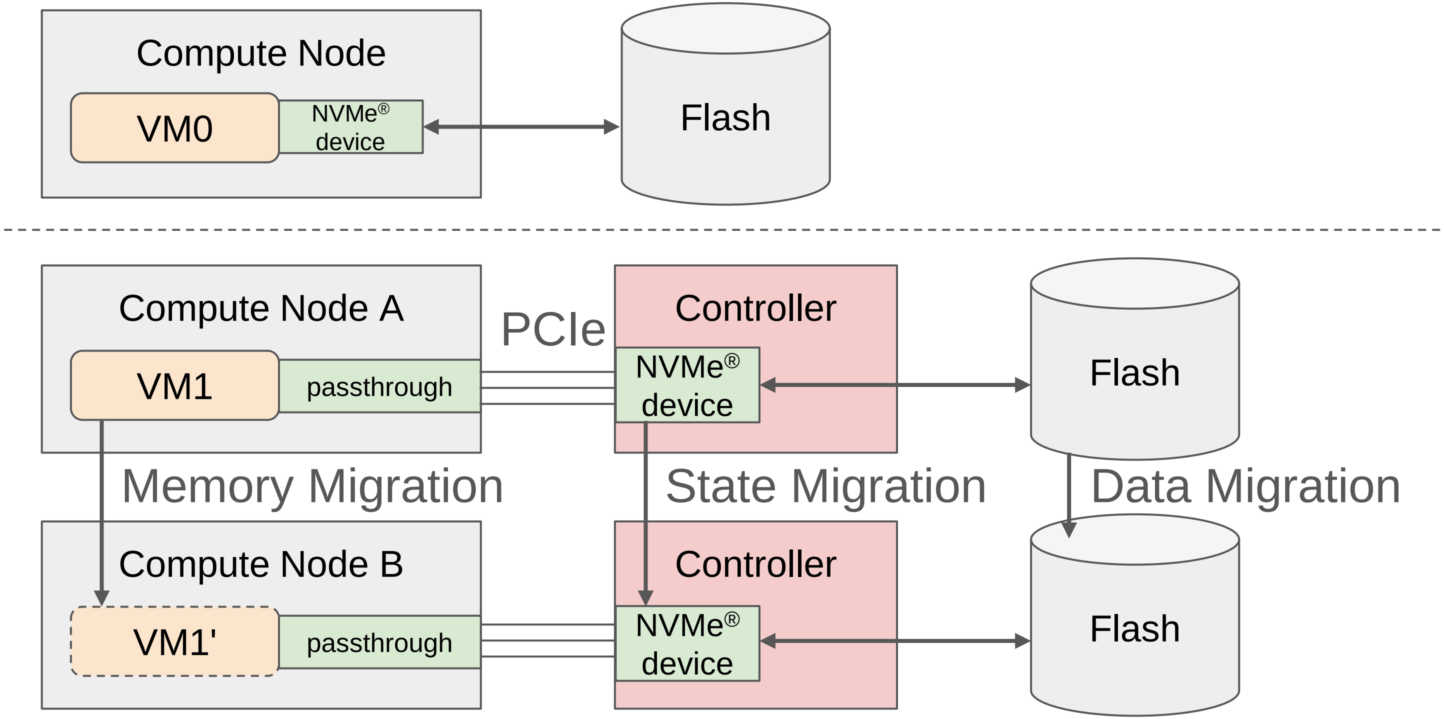
Architected for Performance

Backup Slides

Remote Storage Past and Present



Local Storage Past and Present



Google Industry Alignment Focus Areas

Internal

1. **Security**
    a. **Root of trust**. I am who I say I am and I run a proven firmware. Caliptra?
    b. **Key Management / Encryption**. Keys secure, Encrypt at rest. LOCK?
2. **Isolation**. Read vs Write, head-of-line blocking and inadvertently antagonistic workloads.
3. **Telemetry**
4. **Debuggability**
5. **Left Shift / Time to Market.** Reduce iterations; Speed up cycle times
6. **WAF Reduction**

Cloud

1. **Baremetal Presentation**
2. **VM Presentation**
3. **Live Migration**
4. **Antagonist Isolation**
    a. Rate limiting read/write/trim
    b. Hotspot isolation
5. **Malicious Activity Containment**
    a. Controller takeover. Impact on other VFs and PF + Host
6. **Debuggability**
    a. Windows communicated obscure messages 3 days ago at 7:34 AM - fix it.
    b. My filesystem says it's corrupt