FMS
the Future of Memory and Storage

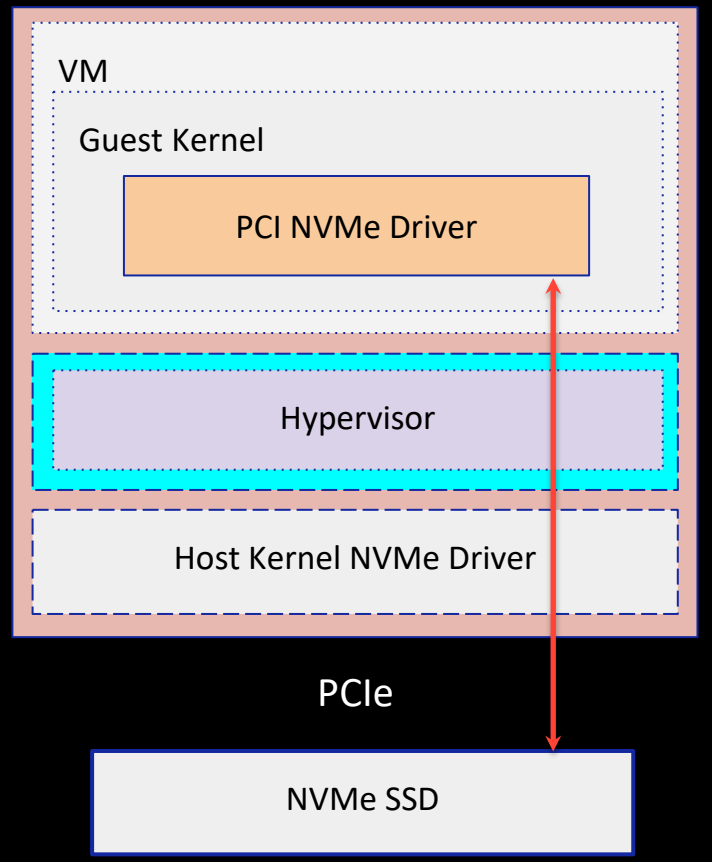
*Lee Prewitt is a Director of Cloud Hardware Storage at Microsoft with 30+ years of storage industry experience ranging from Magneto-Optical to spinning rust to Flash. His former work at Microsoft has included working in the Windows and Devices Group where he was responsible for many of the components in the storage stack including File Systems, Spaces, Storport and Microsoft's inbox miniport drivers (SD, UFS, NVMe, etc.). He currently works in the Azure Hardware group where he is responsible for future Data Center storage initiatives, specifications (OCP, NVMe, EDSFF), and evangelization.*


Lee Prewitt

Director Cloud Hardware Storage
Microsoft

# Hyperscaler Perspective

- Challenge: Consistent, Predictable, & Performant VM Guest Experience
  - Live migration should not be observable by the guest VM
  - Live migration should be efficient and performance isolated
  - Hardware iteration should not be observable by the guest VM
  - Includes SSD upgrade, platform change, etc
- Historical Approach to Achieve these Goals
  - Hyperscaler-specific logic to make changes invisible to the guest VM
  - For lower latency, vendor specific changes to SSDs were introduced
- Standards Solve this Challenge in a Durable Manner
  - Samsung, Google, Microsoft, and others have led standards activities in NVMe and OCP since 2023 to solve this challenge
  - Features are delivered piece by piece to advance the ecosystem
  - Hardware & software are tested at scale in a robust manner




**Robust standards needed to broadly deliver the consistent, predictable, performant VM storage experience**

# Why is Live Migration Needed?

- Customers hate to be interrupted when they are working
  - Need very low interruption rates on imperfect hardware
  - Measured as the number of customer VM events per year
    - Annual Interruption Rate (AIR)
- Live Migration is required to meet AIR goals
- Use cases for Live Migration
  - Actual hardware failure on the node (may not affect all VMs)
  - Predicted hardware failure (allows timely migration before an issue occurs)
  - Scheduled node maintenance (Hardware or firmware upgrade)
  - Load balancing (allows for a mixture of large and small VMs to coexist)

# Why is Para-virtualization Not Enough?

- SSDs are fast
- Para-virtualization is slow
  - Leaves IOPs on the table
- Para-virtualization is expensive
  - Reduces the number of sellable CPU cores on the node

# First Thing is to Get the Hyper-visor out of the IO Path

- TP4165 Tracking LBA Allocation
  - Host migration software only needs to copy the LBAs that are in use
- TP4159 PCIe Infrastructure for Live Migration
  - Namespace Migration
    - Track changed LBAs between copy passes to minimize work
  - Controller Migration
    - Suspend source controller and migrate its state to the destination controller
- Quality of Service Control
  - Allows for VM resource isolation
  - But more importantly it allows for VM rate limiting during Live Migration

# But This is Not Enough

- Need to get out of trapping the Admin Queue as well
  - Not that CPU intensive, but it does count
  - More importantly, need to minimize the hypervisor surface in the face of confidential computing
- But we still need to make sure that the VM does not see a change in hardware after it's been migrated
- PCIe® Exported NVM Subsystem – How to a consistent lie
  - Abstracts Controller and Namespace Inquiry Data
  - Abstracts Get Log Page (page support as well as returned data)
  - Abstracts Command support

# Are We There Yet?

- Still a couple of pieces needed to complete the picture
- Resource allocation
  - Standardized way attach resources (Namespaces, Queue Pairs, etc.) to portions of a device
  - These sub portions may be PCIe Physical Functions, SR-IOV Functions or SIOV Assignable Interfaces
- TDISP support for NVMe
  - Standardize the DEVICE_INTERFACE_REPORT
  - Allows for standardized support for Confidential Compute