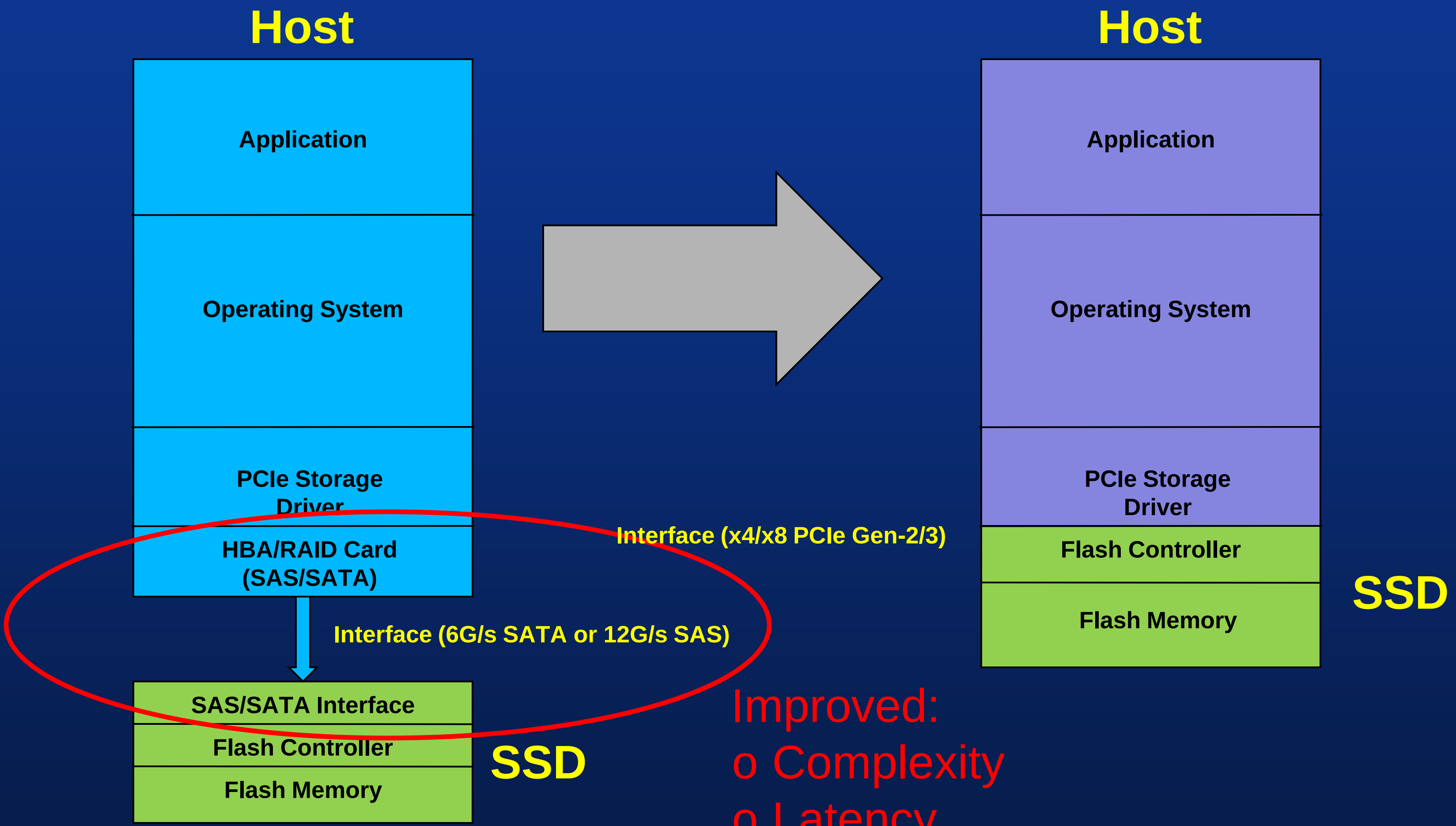
Power Management
PCIe/NVMe Flash Cards

Presenter: Bob Weisickle

CEO/Founder: OakGate Technololgy, Inc.

NVMe Premise/Challenge



Host

Application

Operating System

PCIe Storage Driver

HBA/RAID Card (SAS/SATA)

Interface (6G/s SATA or 12G/s SAS)

SAS/SATA Interface

Flash Controller

Flash Memory

SSD

Host

Application

Operating System

PCIe Storage Driver

Interface (x4/x8 PCIe Gen-2/3)

Flash Controller

Flash Memory

SSD

Improved:

o Complexity

o Latency

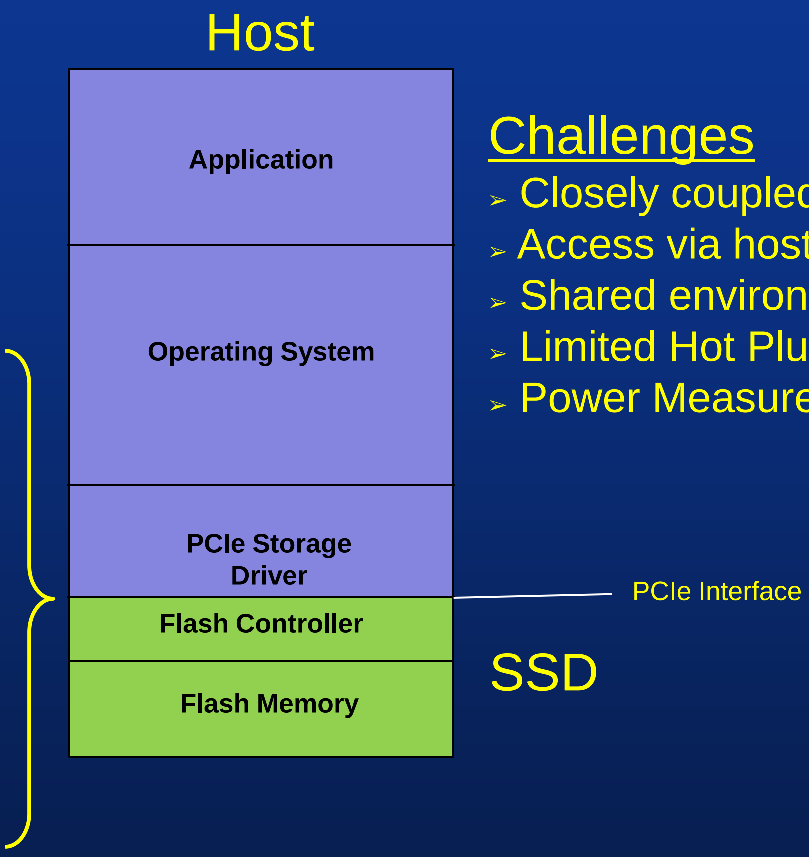
o Performance

o Reliability

Performance/Validation Challenge

Requirements

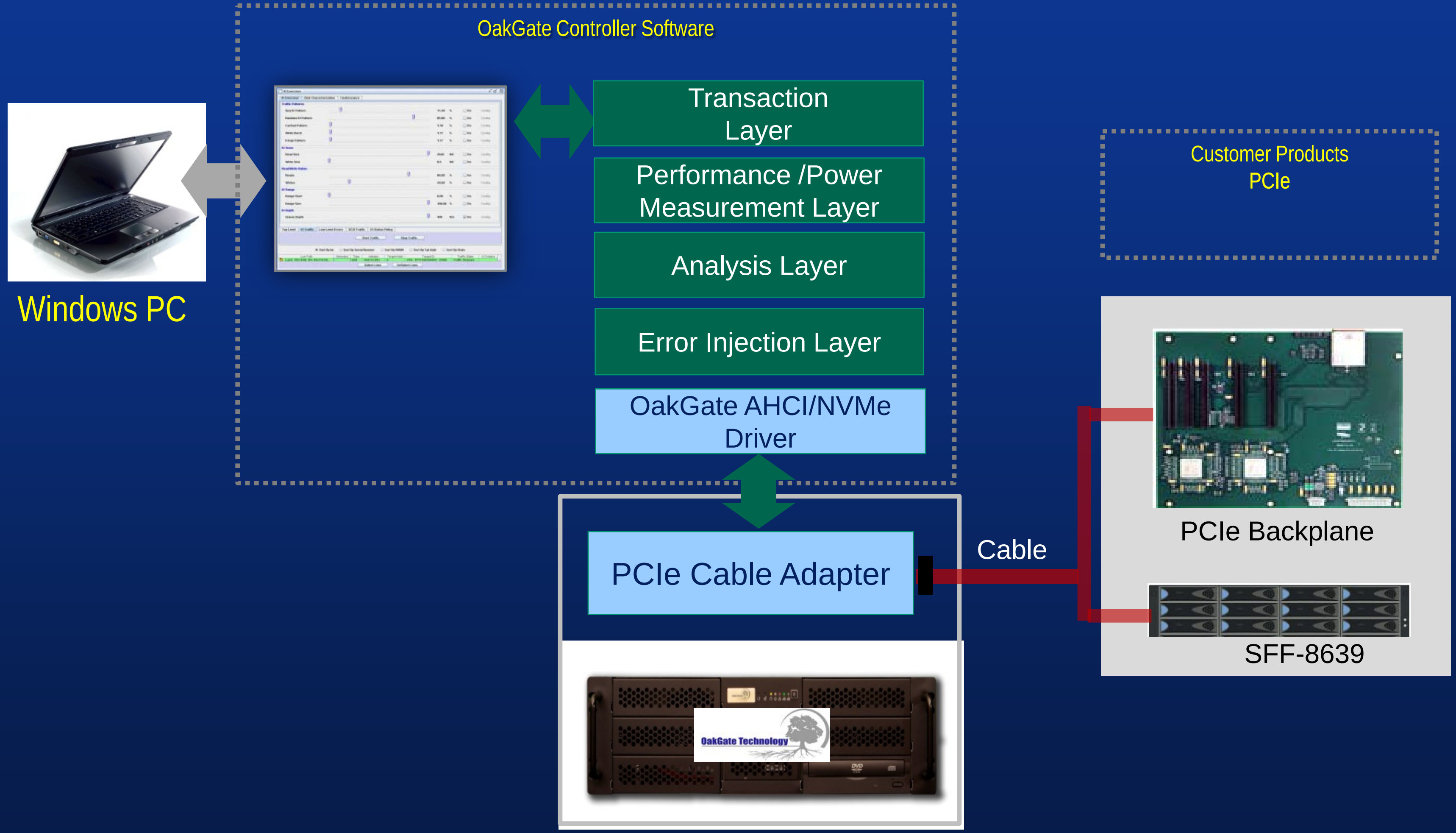
- Access to PCIe Registers
- Queue Configuration
- Hot Discovery
- Error/Fault Injection
- Power Loss/Recovery
- PCIe Resets
- NVMe Performance/Power



Challenges

- Closely coupled to host
- Access via host driver
- Shared environment
- Limited Hot Plug/Discovery
- Power Measurement

Performance/Validation Environment



OakGate Controller Software

Transaction Layer

Performance /Power Measurement Layer

Analysis Layer

Error Injection Layer

OakGate AHCI/NVMe Driver

Windows PC

Customer Products PCIe

PCIe Cable Adapter

Cable

PCIe Backplane

SFF-8639

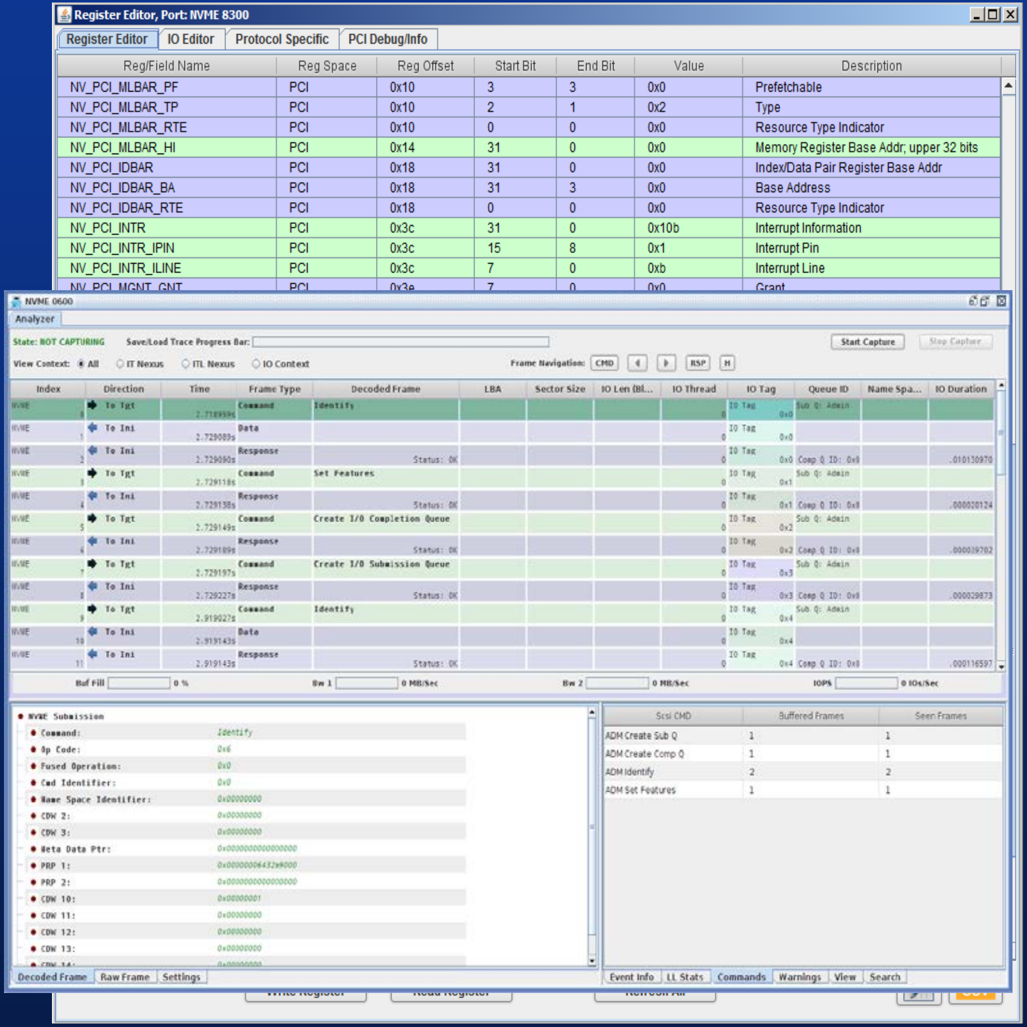
Test System

Flash Memory SUMMIT

OakGate Technology

Base Level PCIe Validation

- Register Access
  - Bit Level verification
  - Error Injection (Bit/Register Level)
  - Base Address Visibility
  - Get/Set Features
- Command Analyzer
  - Posted commands w/ responses
  - Visibility to errors



Register Editor, Port: NVME 8300

Reg/Field Name	Reg Space	Reg Offset	Start Bit	End Bit	Value	Description
NV_PCI_MLBAR_PF	PCI	0x10	3	3	0x0	Prefetchable
NV_PCI_MLBAR_TP	PCI	0x10	2	1	0x2	Type
NV_PCI_MLBAR_RTE	PCI	0x10	0	0	0x0	Resource Type Indicator
NV_PCI_MLBAR_HI	PCI	0x14	31	0	0x0	Memory Register Base Addr; upper 32 bits
NV_PCI_IDBAR	PCI	0x18	31	0	0x0	Index/Data Pair Register Base Addr
NV_PCI_IDBAR_BA	PCI	0x18	31	3	0x0	Base Address
NV_PCI_IDBAR_RTE	PCI	0x18	0	0	0x0	Resource Type Indicator
NV_PCI_INTR	PCI	0x3c	31	0	0x10b	Interrupt Information
NV_PCI_INTR_IPIN	PCI	0x3c	15	8	0x1	Interrupt Pin
NV_PCI_INTR_ILINE	PCI	0x3c	7	0	0xb	Interrupt Line
NV_PCI_MGNT_GNT	PCI	0x3e	7	0	0x0	Grant

NVME 0600

State: NOT CAPTURING   Save/Load Trace Progress Bar:

View Context: ● All ○ IT Nexus ○ ITL Nexus ○ IO Context

Index	Direction	Time	Frame Type	Decoded Frame	LBA	Sector Size	IO Len (Bl...	IO Thread	IO Tag	Queue ID	Name Spa...	IO Duration
NVME 0	To Tgt	2.718959s	Command	Identify				0	IO Tag 0x0	Sub Q: Admin		
NVME 1	To Ini	2.729089s	Data					0	IO Tag 0x0			
NVME 2	To Ini	2.729090s	Response	Status: OK				0	IO Tag 0x0	Comp Q ID: 0x0		.010130970
NVME 3	To Tgt	2.729118s	Command	Set Features				0	IO Tag 0x1	Sub Q: Admin		
NVME 4	To Ini	2.729138s	Response	Status: OK				0	IO Tag 0x1	Comp Q ID: 0x0		.000020124
NVME 5	To Tgt	2.729149s	Command	Create I/O Completion Queue				0	IO Tag 0x2	Sub Q: Admin		
NVME 6	To Ini	2.729189s	Response	Status: OK				0	IO Tag 0x2	Comp Q ID: 0x0		.000039702
NVME 7	To Tgt	2.729197s	Command	Create I/O Submission Queue				0	IO Tag 0x3	Sub Q: Admin		
NVME 8	To Ini	2.729227s	Response	Status: OK				0	IO Tag 0x3	Comp Q ID: 0x0		.000029873
NVME 9	To Tgt	2.919027s	Command	Identify				0	IO Tag 0x4	Sub Q: Admin		
NVME 10	To Ini	2.919143s	Data					0	IO Tag 0x4			
NVME 11	To Ini	2.919143s	Response	Status: OK				0	IO Tag 0x4	Comp Q ID: 0x0		.000116597

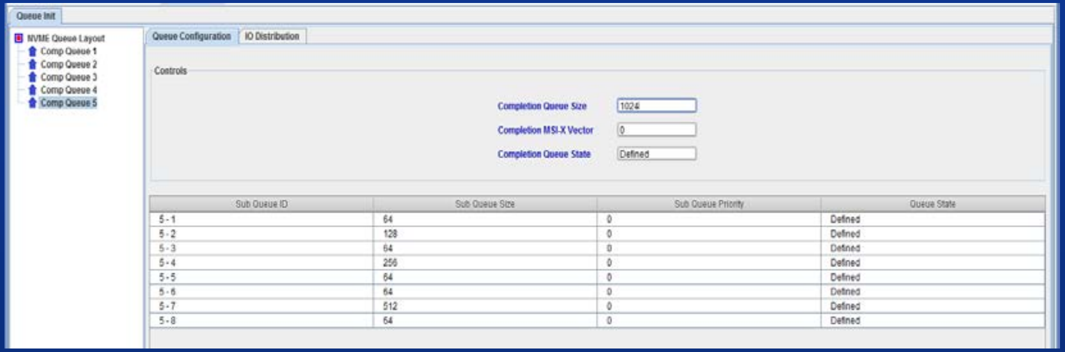
Buf Fill 0 %   Bw 1 0 MB/Sec   Bw 2 0 MB/Sec   IOPS 0 IOs/Sec

- NVME Submission
  - Command: Identify
  - Op Code: 0x6
  - Fused Operation: 0x0
  - Cmd Identifier: 0x0
  - Name Space Identifier: 0x00000000
  - CDW 2: 0x00000000
  - CDW 3: 0x00000000
  - Meta Data Ptr: 0x0000000000000000
  - PRP 1: 0x000000006432e9000
  - PRP 2: 0x0000000000000000
  - CDW 10: 0x00000001
  - CDW 11: 0x00000000
  - CDW 12: 0x00000000
  - CDW 13: 0x00000000
  - CDW 14: 0x00000000

Scsi CMD	Buffered Frames	Seen Frames
ADM Create Sub Q	1	1
ADM Create Comp Q	1	1
ADM Identify	2	2
ADM Set Features	1	1

Queue Configuration

- Multiple Sub/Comp Queues
- Definable Sub/Comp Queue Size
- Sub Queue IO Loading
- Multi-CPU Queue Distribution
- Definable Interrupt Support
  - MSI-X
  - MSI
  - Line



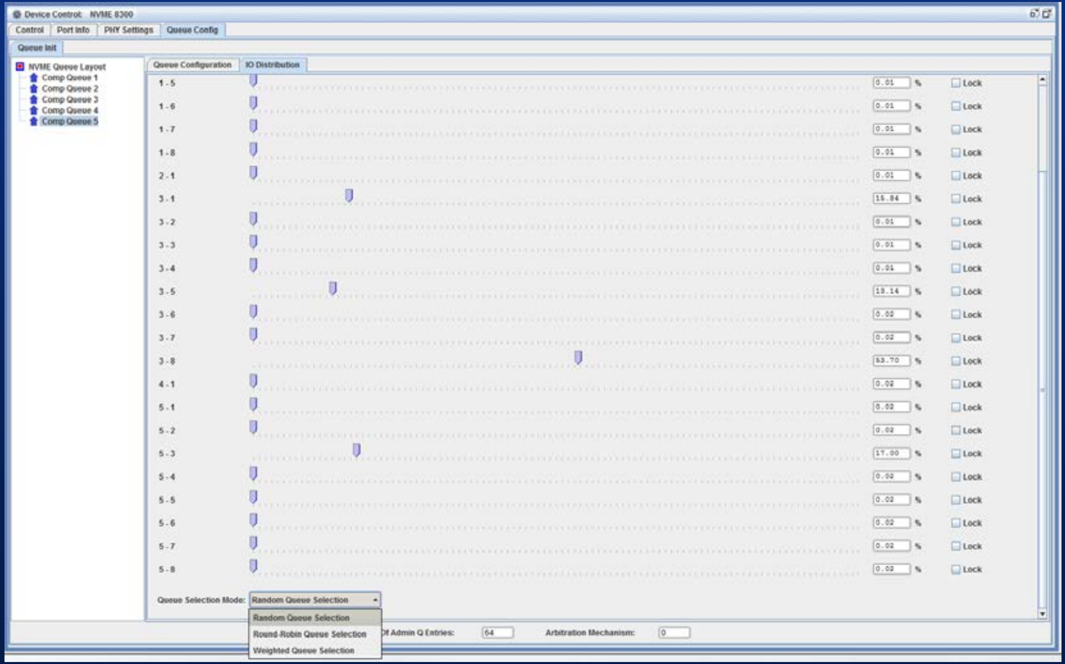
Controls

Completion Queue Size: 1024

Completion MSI-X Vector: 0

Completion Queue State: Defined

Sub Queue ID	Sub Queue Size	Sub Queue Priority	Queue State
5 - 1	64	0	Defined
5 - 2	128	0	Defined
5 - 3	64	0	Defined
5 - 4	256	0	Defined
5 - 5	64	0	Defined
5 - 6	64	0	Defined
5 - 7	512	0	Defined
5 - 8	64	0	Defined



1 - 5	0.01	%	□ Lock
1 - 6	0.01	%	□ Lock
1 - 7	0.01	%	□ Lock
1 - 8	0.01	%	□ Lock
2 - 1	0.01	%	□ Lock
3 - 1	15.84	%	□ Lock
3 - 2	0.01	%	□ Lock
3 - 3	0.01	%	□ Lock
3 - 4	0.01	%	□ Lock
3 - 5	13.14	%	□ Lock
3 - 6	0.02	%	□ Lock
3 - 7	0.02	%	□ Lock
3 - 8	53.70	%	□ Lock
4 - 1	0.02	%	□ Lock
5 - 1	0.02	%	□ Lock
5 - 2	0.02	%	□ Lock
5 - 3	17.00	%	□ Lock
5 - 4	0.02	%	□ Lock
5 - 5	0.02	%	□ Lock
5 - 6	0.02	%	□ Lock
5 - 7	0.02	%	□ Lock
5 - 8	0.02	%	□ Lock

Queue Selection Mode: Random Queue Selection

Of Admin Q Entries: 64

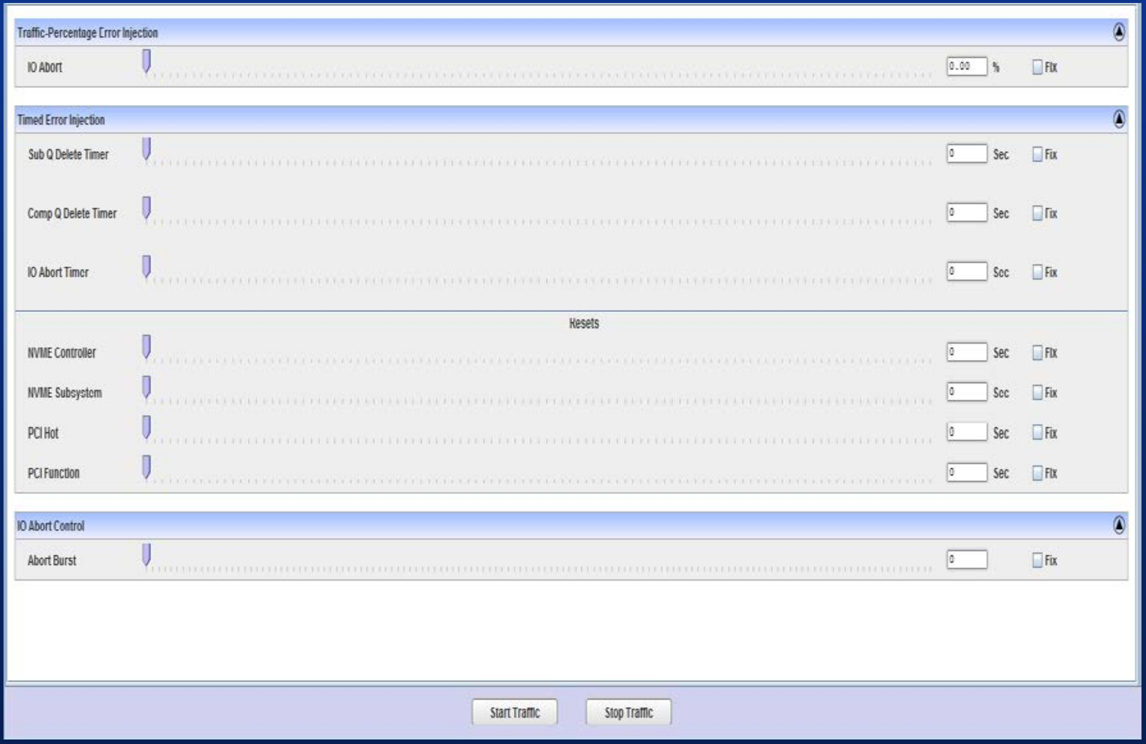
Arbitration Mechanism: 0

Hot Power Loss/Recovery

- Host Independent Power Control
  - Ability to control power to each device
  - Measure power of each device
- Surprise Power Loss/Recovery
  - Detection of device being powered off
  - Auto re-discovery and drive state verification
  - No Host reboot required
- Auto Queue Management
  - Re-instantiate all Queues
  - Clean up of outstanding commands
- Auto Traffic Control
  - Continue with traffic workload
  - Check for data corruption

Resets/Queues/Abort

- IO Aborts
- PCIe Resets
  - Function Resets
  - Hot Resets
  - Controller
  - Subsystem
- Auto Re-Discovery
- Queue Management & Recovery



Traffic-Percentage Error Injection

IO Abort 0.00 % □ Fix

Timed Error Injection

Sub Q Delete Timer 0 Sec □ Fix

Comp Q Delete Timer 0 Sec □ Fix

IO Abort Timer 0 Sec □ Fix

Resets

NVME Controller 0 Sec □ Fix

NVME Subsystem 0 Sec □ Fix

PCI Hot 0 Sec □ Fix

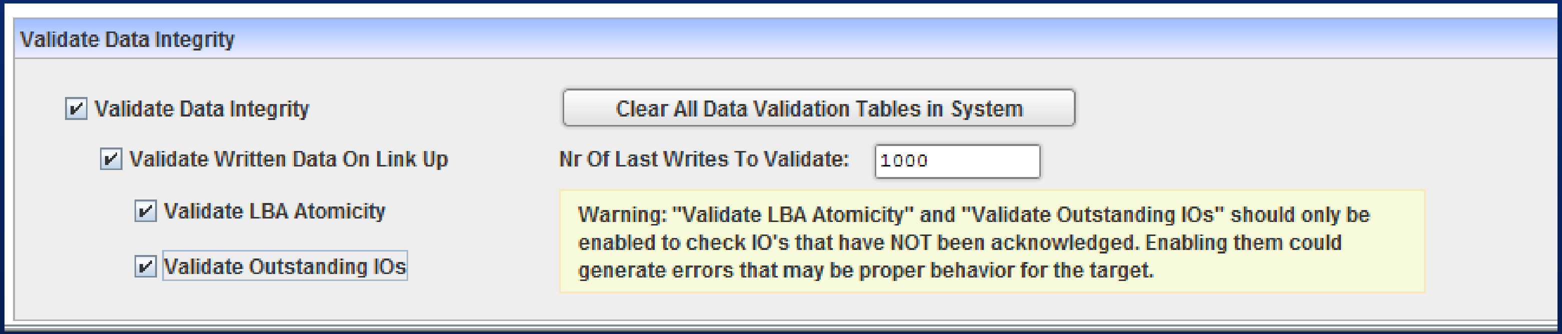
PCI Function 0 Sec □ Fix

IO Abort Control

Abort Burst 0 □ Fix

Data Retention/Corruption

- Validate Pre-existing Data
- Validate ‘Acknowledged’ Writes
- Evaluate 'Outstanding' Writes
  - Verify Atomic Write Functionality
  - Single LBA corruption
  - Visibility to Partial Writes



Validate Data Integrity

☑ Validate Data Integrity

☑ Validate Written Data On Link Up

☑ Validate LBA Atomicity

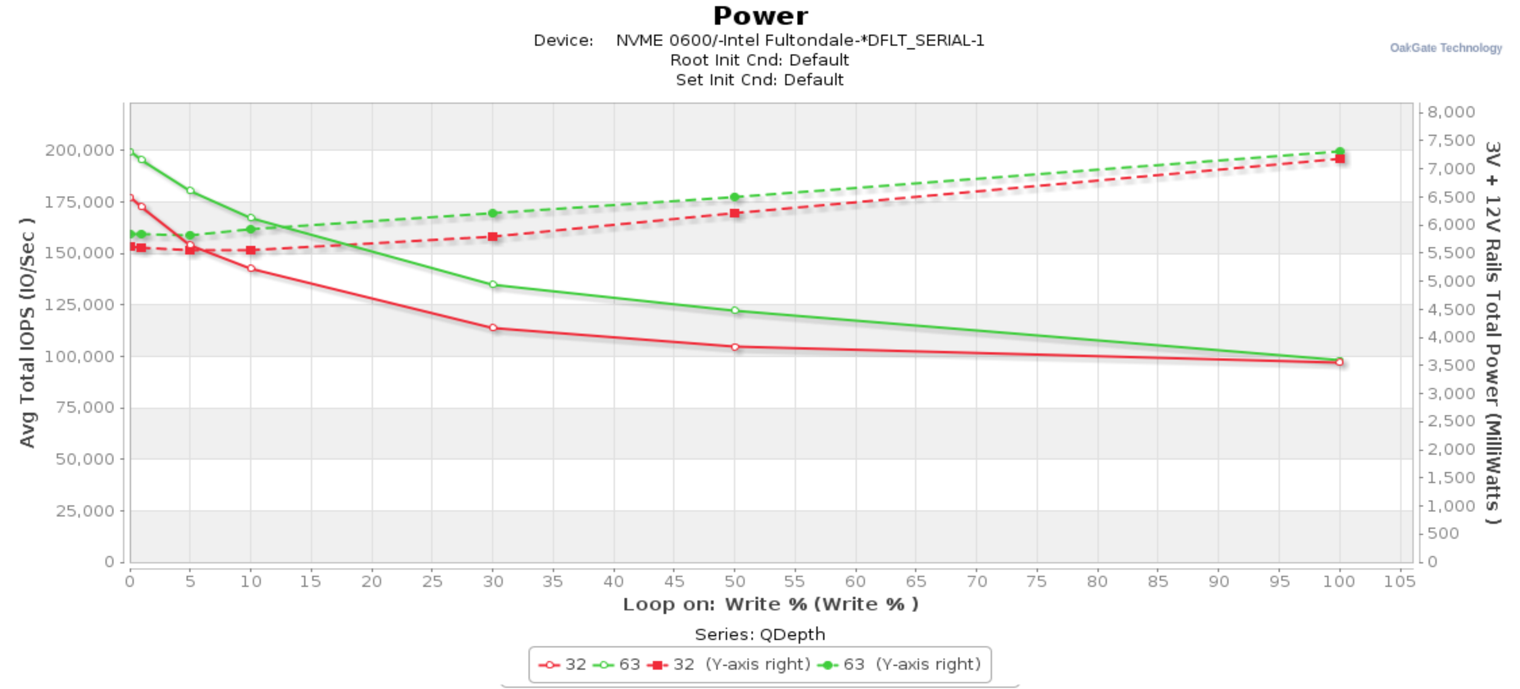
☑ Validate Outstanding IOs

Nr Of Last Writes To Validate: 1000

Warning: "Validate LBA Atomicity" and "Validate Outstanding IOs" should only be enabled to check IO's that have NOT been acknowledged. Enabling them could generate errors that may be proper behavior for the target.

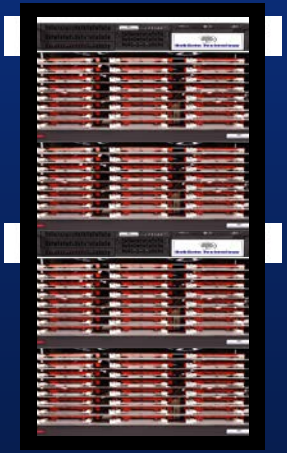
Power Assessment

- Real-time Power Measurement
  - Power vs Performance
  - Power vs Workload



Infrastructure Scale

- Improved Space Utilization
- Better Power/DUT
- Enhanced Management



SAS/SATA

44U Rack
96 SATA/SFF-8639 SSDs
OakGate

PCIe

44U Rack
256 PCIe SSDs
OakGate

Thermal

10.5x5.5 Box
1024 SATA SSDs
512 PCIe x4 SSDs
OakGate/Tanisys

Summary

- NVMe has momentum in the Market
  - Products exists from a several major suppliers
- Test/Validation Tools exist
  - With Power Control and Power Measurement
  - Detailed PCIe Register/Queue access
- PCIe Ecosystem is evolving
  - SFF-8639, M.2 etc becoming available
  - Dual Port Support
- NVMe Spec is evolving rapidly
  - Requires Validation Tools that stay current

Thank You

Questions/Follow-up:
bob.weisickle@oakgatetech.com
www.oakgatetech.com