Building dense NVMe storage

Mikhail Malygin, Principal Software Engineer
Driven by demand

• Demand is changing
  • From traditional DBs to NO-SQL
  • Average NO-SQL DB size: 300TB
  • Analytics is everywhere
  • 50% of storage issues: performance
• Year 2022 forecast
  • NVMe – 80% of SSD market
Motivation

- There is clear storage paradigm shift
- Established architectures can’t cope with it
- Storage Architecture need to be revisited
- New solutions in HW and SW
- Storage should achieve balance between density, performance and availability
Clustered storage

- Hardware setup with shared drives
- Need a symmetric A+A solution
- Node synchronization required
  - Parity updates
  - Placement metadata
  - Background tasks
- Works fine on spinning drives
NVMe challenges

- PCIe bus limitation
- Hotplug
- CPU usage
Need a new architecture

- CPU is the bottleneck, back again
- Generic solutions do not work
- Software becomes complex
- Complex software is less performant
Hardware Accelerated Architecture

U.2 NVRAM Module is the patent pending solution by YADRO
Clustering revised

- Cluster with PCIe Fabric
- All members share the same bus topology
- Simultaneous access to all system devices
- Storage, IO, Acceleration, Synchronization
- SR-IOV for multi host access
Storage revised

- NVMe RAID Controller
- Dedicated CPU for EC and NVMe operations
- Aggregated drives, less PCIe devices on CPU
- Direct I/O path, from card to controller
NVMe RAID Controller Features

- Powerful multi-cores ARM A72 SoC
- PCIe Gen4 support
- Flexible protection algorithms
- SR-IOV provider for drives array
- Multiple namespaces support
- Battery-protected cache
- NVRAM support
U.2 NVRAM Module

- Industry standard form-factor
- PCIe Gen4 support
- Powerful multi-core A72 SoC for in-situ data processing
- Up to 256GB of DDR4 memory backed-up with 512GB flash
- Unlimited RAM write endurance
- NVMe mode and direct memory access mode
- External battery support
Solving Active-Active problem

- Hardware accelerated Key - Value as PCIe device
- Based on U.2 NVRAM Module
- Makes KV operations atomic
- Stores metadata and cache
- Scales up by partitioning

U.2 NVRAM Module is the patent pending solution by YADRO
PCIe Fabric NVMe Drives Module

- 16xU.2 NVMe SSDs
- Management sideband
- Dual-port drives support
- Optional NVMe RAID

PCIe switches

Mezzanine-attached NVMe RAID controllers
PCIe Fabric Controller

Up to eight unified PCIe-attached modules

Rear 16xFHHL AIC I/O module (optional, double-wide)

Side 6xHHHL AIC I/O uplinks module (4x16 or 2x16+4x8)

Root PCIe fabric switches & management
Thank you!

• Questions?
Data protection challenges

- 10+ TB per drive
- Traditional RAID does not scale good enough
  - Rebuild overhead for RAID 5/6
  - Spare overhead for RAID 60
- Use RAID with thin provision and flexible placement
  - Just place data somewhere in the enclosure and keep index
- Flexible protection scheme with erasure coding
NVMe challenges (CPU usage)
PCle Fabric Controller Features

- Redundant PCle topology
- PCle Gen4 ready
- 4kW redundant PSUs
- Redundant management
- Sideband management network
- Internal/external hosts support
- Wide (PCle 4x16) uplinks
NVMe RAID diagram
PCIe Fabric NVMe Drives Module

- 16xU.2 NVMe SSDs
- Management sideband
- Dual-port drives support
- Optional NVMe RAID
NVRAM Module Features

- Dual mode access (NVMe & direct mapped memory)
- Transactional memory with atomic operations
- SR-IOV for sharing among multiple hosts via PCIe Fabric
- Acceleration for storage applications
  - Hash calculation, compression, encryption
- Shadow replication (redundancy)
- Configuration & management via NVMe command set