A Cost Effective, High Performance, Highly Scalable, Non-RDMA NVMe Fabric

Bob Hansen, VP System Architecture
bob@apeirondata.com
Agenda

• 3rd Platform Opportunity for High Performance Storage
• Applications with enhanced user experience require:
  • High IOP performance & low-latency
    – Storage performance = $$ PROFITS
  • Scalability
  • Scale out, in-memory compute/storage architecture evolution
    – In-memory => in-box flash => external flash
• The ideal solution
• Use cases
• Apeiron’s Shared DAS™ Architecture
  • Software with HW acceleration
  • Apeiron Data Fabric™
  • System architecture
3rd Platform Opportunity

According to IDC, 3rd Platform technologies already drive 30% of ICT spending and 100% of growth and 2nd Platform will enter recession in 2015.

Enhanced User Experience Applications driving high IOP/low latency storage performance

- Customer personalization and simplified data management
- Fortune 500 companies mid-layer meta cache rapidly growing
- Kayak
  - Caching aged airline quotes to speed service
- Netflix
  - Personalization for >50M customers
- Amadeus
  - 3.7 Million Bookings per Day
Ad Tech Example

- Deliver ad
- Log results
- Match user to ad
- Bid for ad
- Log transaction
- Match request to user
- Fast Lookup (<1ms)
  - Read consumer profile
  - Update profile
- Win
- Lose
- Determine best ROI

Deliver Ad within ms

- >1 billion consumers
- >3 billion devices

Storage IOPs / latency = $$
NoSQL solution
– Scale out nodes with dataset in-memory

Scale-out in-memory goodness
- Shared nothing compute nodes scale well
- Database is “sharded” evenly across all nodes
- Data set in-memory is VERY FAST
- To scale – just add another node, shard the DB again and go

Issues
- DRAM can be VERY expensive
- Node failure = very long recovery time
  - Data at risk during recovery
- As data set grows more servers must be added
  - = higher cost and footprint
- CPU to mem ratio can not be optimized

This breaks down as you approach 100TB
Expensive DRAM? Add Internal Flash

Scale-out in-memory goodness
- Share nothing compute nodes scale well
- Database is “sharded” evenly across all nodes
- Data set in-memory is VERY FAST
- *Data in flash is FAST*
- To scale – just add another node, shard the DB again and go

Issues
- Flash size must be equal on all nodes
  - Adding storage = downtime
- Node failure = very long recovery time
  - Data at risk during recovery
- As data set grows more nodes must be added
  - = higher cost and foot print
- CPU to mem ratio can not be optimized

*Storage Management is a Pain!*
Very High Performance External Storage is the answer

Application Servers

Shared DAS Goodness
- CPU and Storage scale independently
  - Minimize cost / rack space
  - Improved CPU utilization
- Fine Grain, On-line provisioning
- Server failures don’t take out data
  - Minimize failure recovery time

Issues
- Performance
  - IOPs and Predictable Latency
- Availability
  - HA design and Replicas
- Scale
  - PBs and 100s of nodes
The Ideal Solution -
Shared Direct Attached Storage

- Best performing persistent storage media
  - *Standard NVMe SSDs* – also best cost
- Bare metal Ethernet storage network HW
  - Low cost, *industry standard networking*
- Add value where you get the best ROI
  - *HW Accelerated, Networked Data path*
  - *NVMe SSD Virtualization*
  - *High availability with no performance penalty*
- Best in class management
  - *On-line provisioning and failure recovery*
  - *Storage performance statistics / predictive modeling*

*Keep it simple!*

*Deliver raw NVMe performance to the application*
Application Use Cases

**Scale-out**
Pooled Flash For Operational Big Data

**High Bandwidth**
Data storage for applications with high bandwidth

**Fast cache**
Tiered storage acceleration

---

**Scale-Up**
Compute Complex

**Customer Personalization**
Ad Tech RTB
Fraud detection
Customer personalization

**NoSQL Cluster**
Structured data store

**Structured data store**

---

**In-memory check points requiring massive bandwidth**

---

**Accelerates response to time critical data**
Seamless scaling

---

**SAN**

---
Apeiron’s Solution - **Shared DAS™**

> Scale-out NVM storage architecture
> Intelligent software with hardware accelerated data path
> Ethernet storage fabric with <3uS round-trip latency overhead
> Seamless scaling to petabytes
Apeiron’s Software with Hardware acceleration

- **Instant Failover**
  - Reduces node rebuild time from **>10hrs to <1sec**
  - Remaps metadata to spare

- **Automatic Replication**
  - Transparent backup
  - Hardware assisted SW configured
  - Increases OP/s 40%

- **Transparent Server to Server**
  - Reduces network congestion
  - Accelerates DB manageability
  - Increases application throughput

---

**Pooled external storage at near DRAM Performance**

✓ Faster response generates more profits
Why not “PCIe on a rope”? 

A PCIe storage network is possible but faces several challenges: 

- PCIe is not a network
  - PCIe is an evolution and extension to a parallel system bus
    - Initially scoped to support a handful of devices
- PCIe was not designed to be resilient
  - Bus errors = panic
- Failure isolation is a work in progress
- There are currently no PCIe networking standards

Why re-invent PCIe as a high cost, very complex external storage fabric?
RDMA / Apeiron Data Fabric™ Comparison

Apeiron Data Fabric™

- Simple, Robust
- Optimized for Ethernet and NVMe
- *light weight* Layer 2

The standard is not tied to any particular physical layer
RDMA approach adds between 26B and 96B of headers, in addition to NVMe Encapsulation

**Flexible but adds complexity, link consumption and latency!**
Apeiron System Architecture

**Shared DAS™**

**Application Servers**

- **Deliver raw NVMe performance to the application**

- **HW Accelerated Storage Processing Offload On Application Servers**

- **40Gb Bare Metal Ethernet Storage Network**

- **Apeiron Data Fabric™**

- **Very High Performance Network to SSD interface**

- Simple, scalable architecture with better than in-box flash performance
- Highly available, shared storage using standard SSDs and networking components
- Virtualized storage, on-line provisioning, failure isolation
Apeiron Technology Delivers

> NVMe Virtualization
> Performance Density
  • 18M IOPs, 72GB/s BW
  • In a 2U form factor
> < 90 μS 4K read latency P99 (NAND flash)
  • Ready for 3D Xpoint (<3 μS Fabric Latency)
“All the simplicity and promise of DAS with the efficiency and capability of network attached storage.”