# High Performance, Highly Scalable Storage Architecture Using NVMe

Bob Hansen,
VP System Architecture
bob@apeirondata.com

August 2015

# External, Virtualized NVMe Storage Apeiron Data Fabric™

Apeiron’s *Shared DAS™* virtualization platform delivers industry leading latency and bandwidth, accelerating Real Time Big Data analytics, while optimizing scale-out cluster efficiency.

Apeiron's Data Fabric delivers seamless scalability and easy manageability.

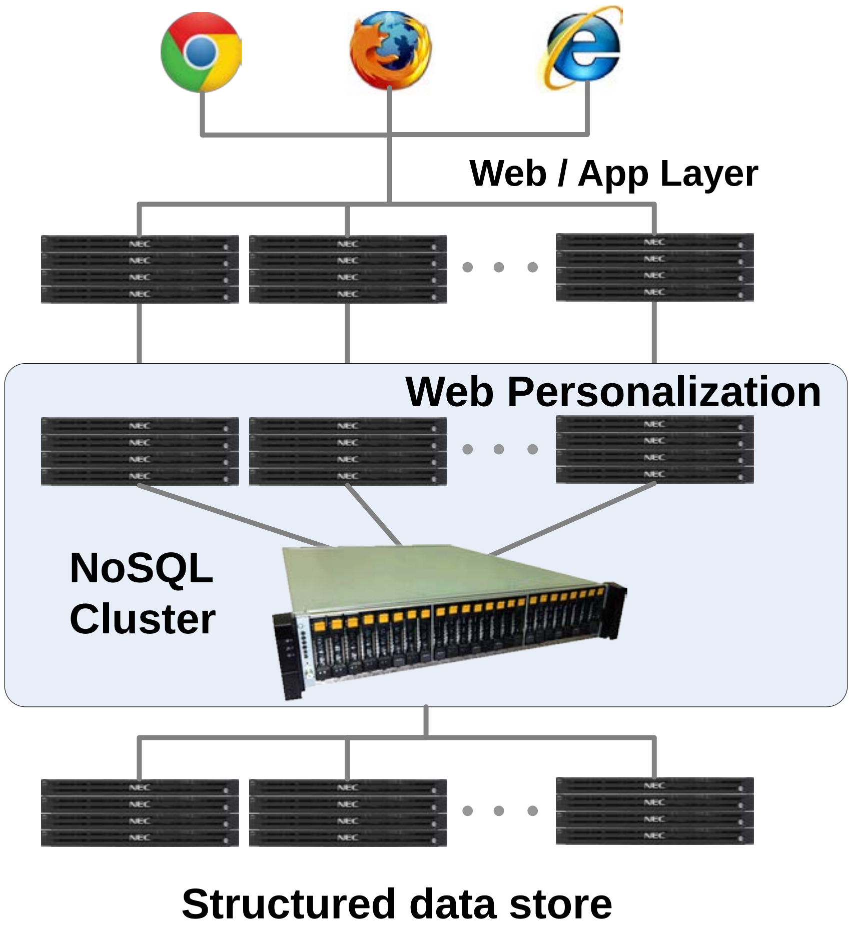
***Come visit us at Booth 819***

apeiron™

# Agenda

- “Real Time, Big Data”  What is that?
- Applications with enhanced user experience requires
  - 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, Very High Performance scale out system
- Apeiron’s Shared DAS™ Architecture

# High IOP Application
# Enhanced User Experience



- 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

AEROSPIKE

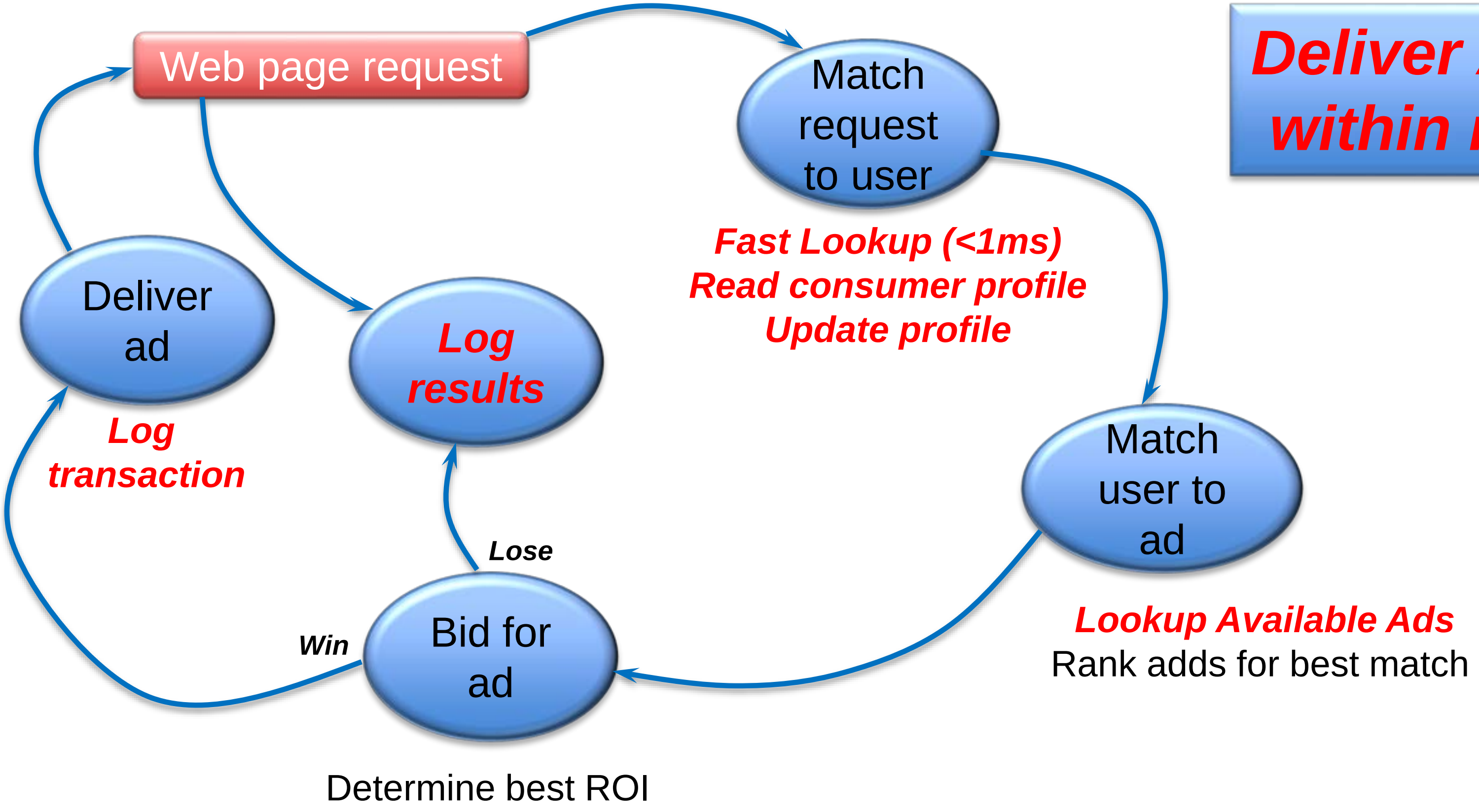
DATASTAX

Couchbase

apeiron

# Ad Tech Example

**Deliver Ad within ms**



Web page request

Match request to user

**Fast Lookup (<1ms)**
**Read consumer profile**
**Update profile**

Match user to ad

**Lookup Available Ads**
Rank adds for best match

Bid for ad

Determine best ROI

*Win*

*Lose*

Deliver ad

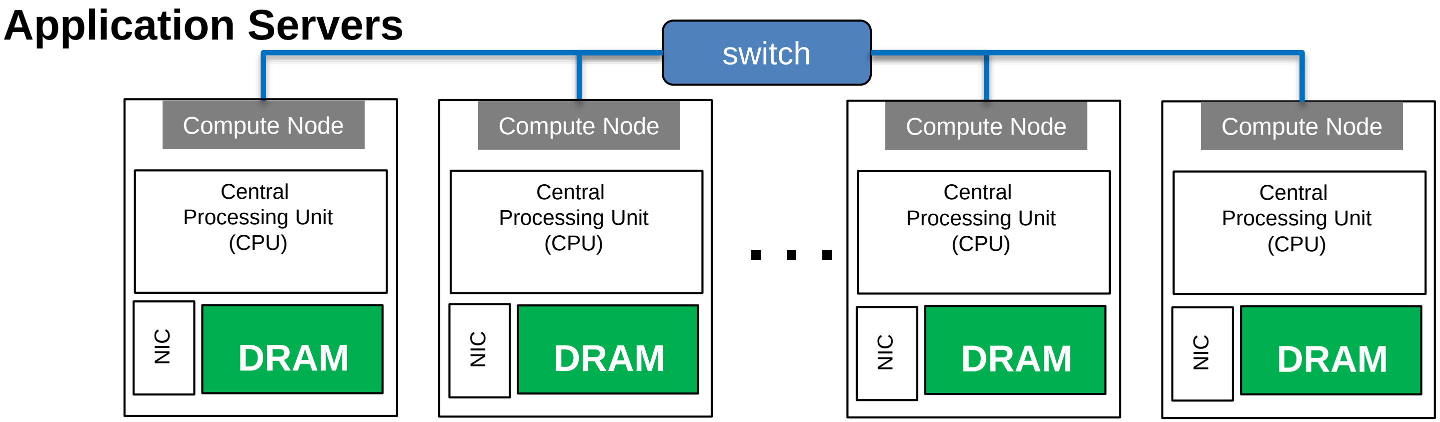
**Log transaction**

**Log results**

- >1 billion consumers
- >3 billion devices

**Storage IOPs / latency = $$**

# NoSQL solution – Scale out nodes with dataset in-memory


Application Servers
switch
Compute Node
Central Processing Unit (CPU)
NIC
DRAM
Compute Node
Central Processing Unit (CPU)
NIC
DRAM
. . .
Compute Node
Central Processing Unit (CPU)
NIC
DRAM
Compute Node
Central Processing Unit (CPU)
NIC
DRAM

## Scale-out in-memory goodness

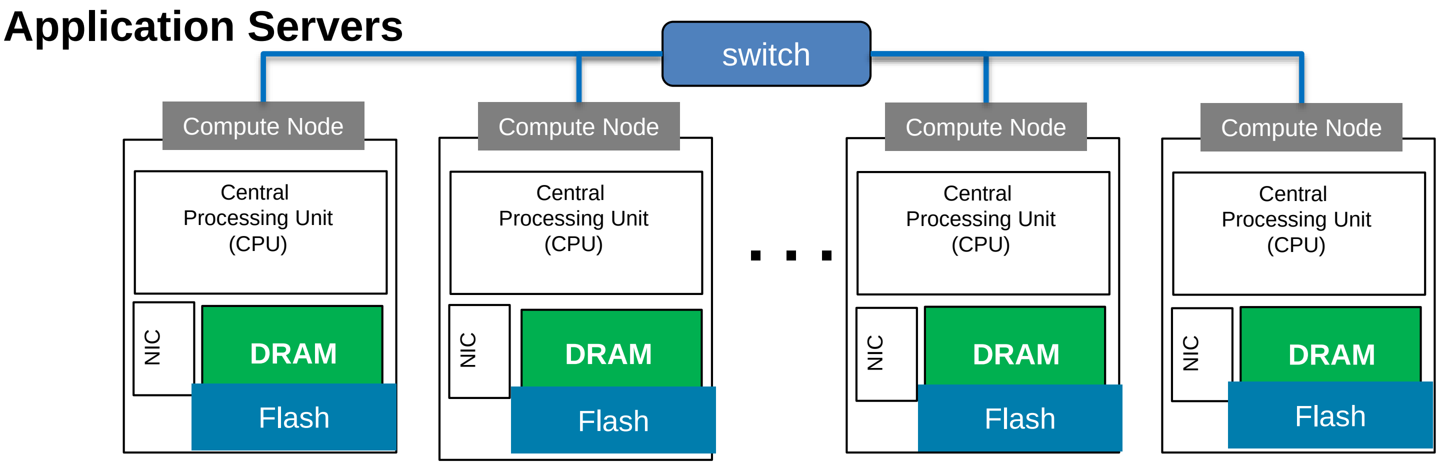
- Shared nothing compute nodes scale well
- Database is "sharded" evenly across all nodes
- Data set in-memory is VERY FAST
- To scale – just add another node, shard the DB again and go

## 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 foot print
- 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**

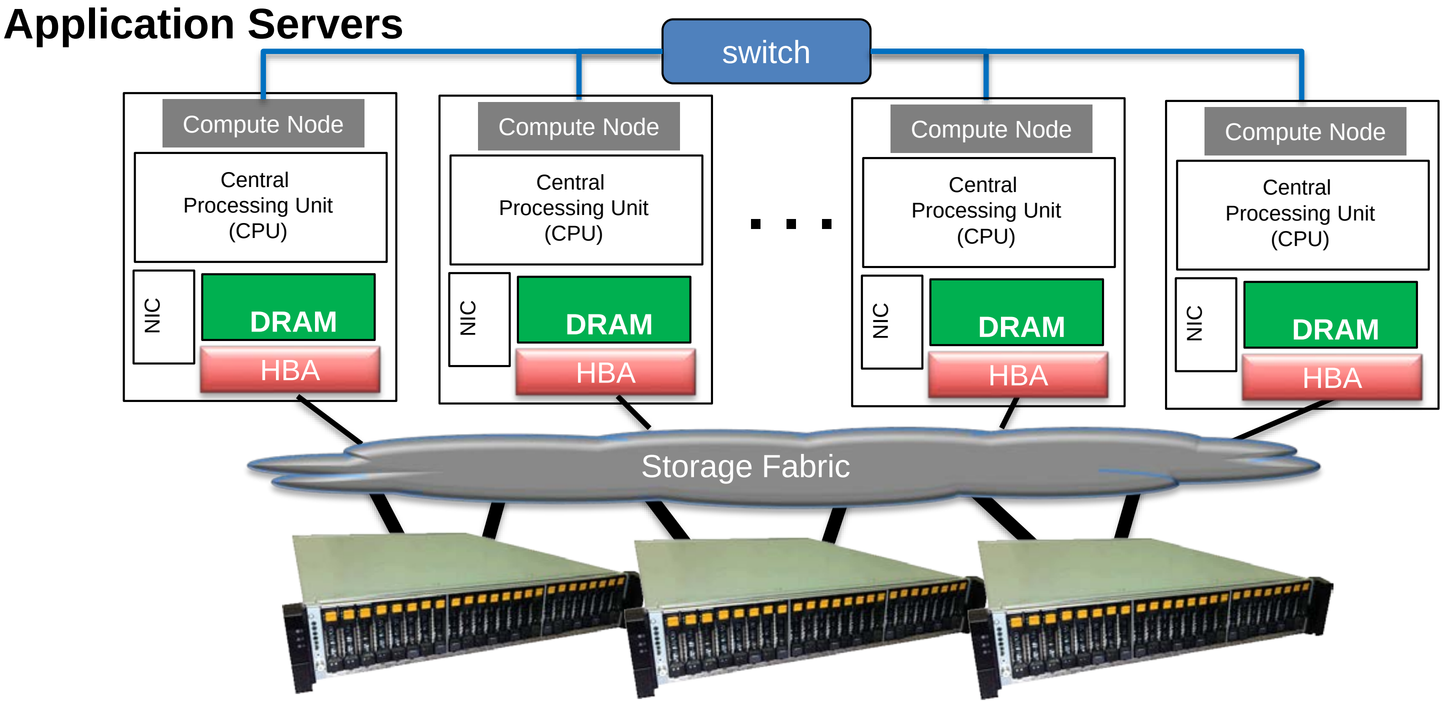
## Issues

- **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!***

apeiron™

# Very High Performance External Storage is the answer



## 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

apeiron™

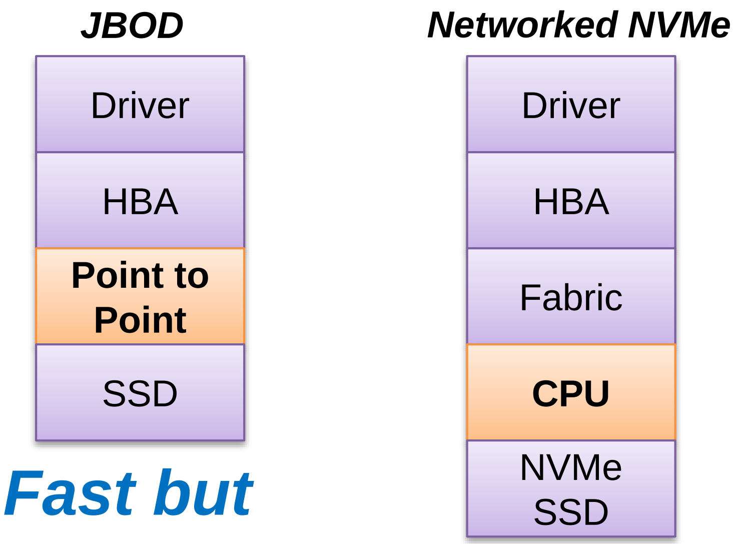
# Storage technology choices
# IOPs / latency performance

## SSD Performance

- 15K HDD – 210 IOPs
- 6Gb SATA SSD – 90K IOPs*
- 12Gb SAS SSD – 155K IOPs*
- NVMe SSD >> ***700K*** IOPs*

***SATA and SAS can't cut it!***

## Get Out of the Box!



***JBOD***

Driver

HBA

**Point to Point**

SSD

***Networked NVMe***

Driver

HBA

Fabric

**CPU**

NVMe SSD

***Fast but Doesn't Scale***

***CPU is the Bottleneck***

**Objectives**

- **Performance**
- **Availability**
- **Scale**

***Kills Performance or Adds Cost$$***

* Typical 4K Random Reads

# 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 best ROI
  - Data path optimization
  - 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***

# Why not "PCIe on a rope"?

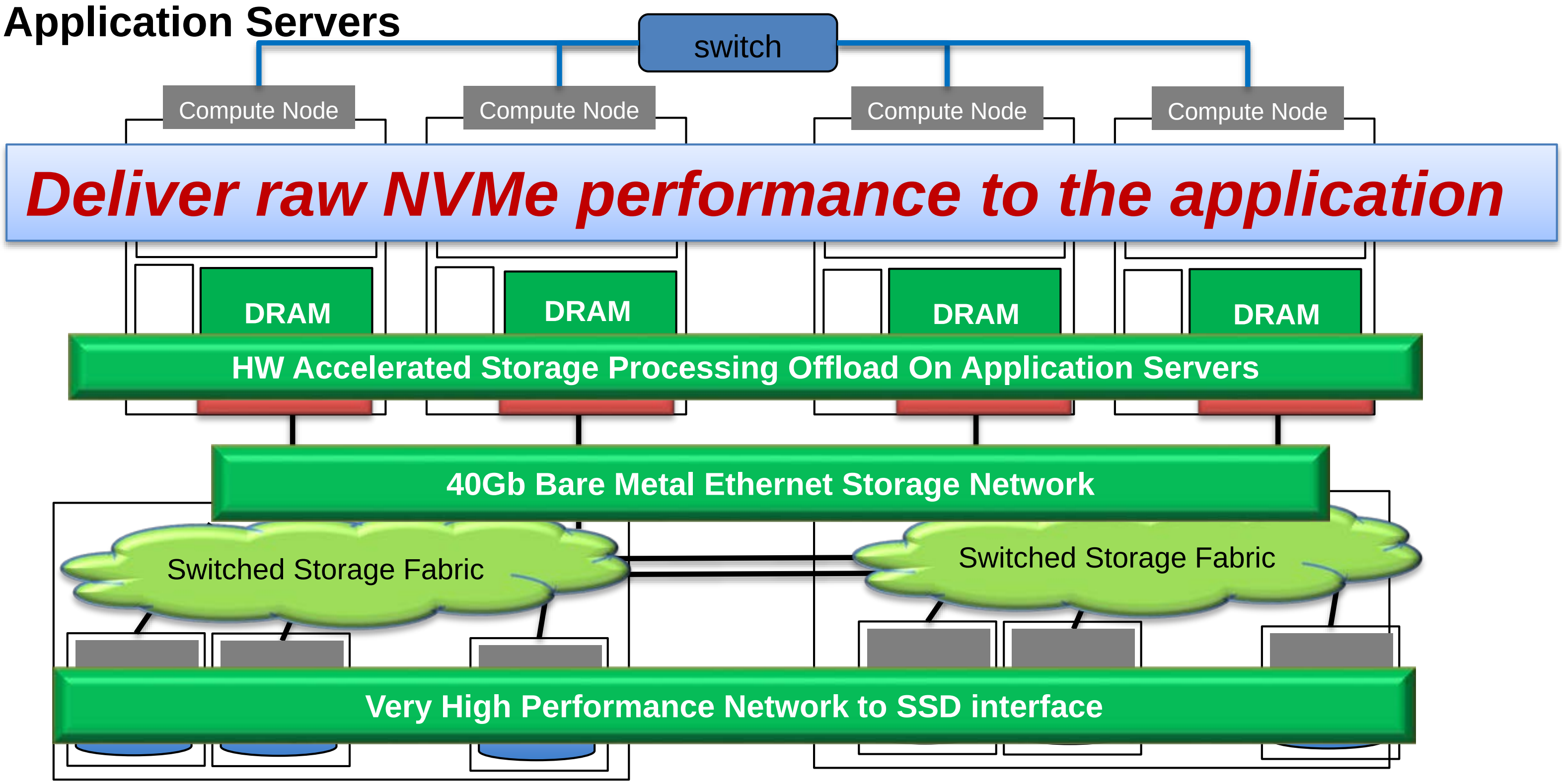
***A PCIe storage network is possible
but faces several challenges -***

- 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?***

apeiron™

# Apeiron System Architecture *Shared DAS™*


Application Servers
switch
Compute Node
Compute Node
Compute Node
Compute Node
Deliver raw NVMe performance to the application
DRAM
DRAM
DRAM
DRAM
HW Accelerated Storage Processing Offload On Application Servers
40Gb Bare Metal Ethernet Storage Network
Switched Storage Fabric
Switched Storage 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

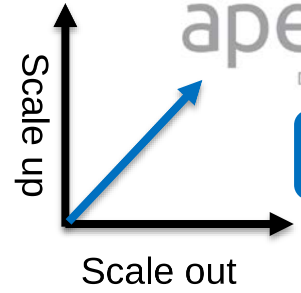
11/02/2014 Apeiron Proprietary and Confidential 12

# Apeiron Technology Delivers

> NVMe Virtualization

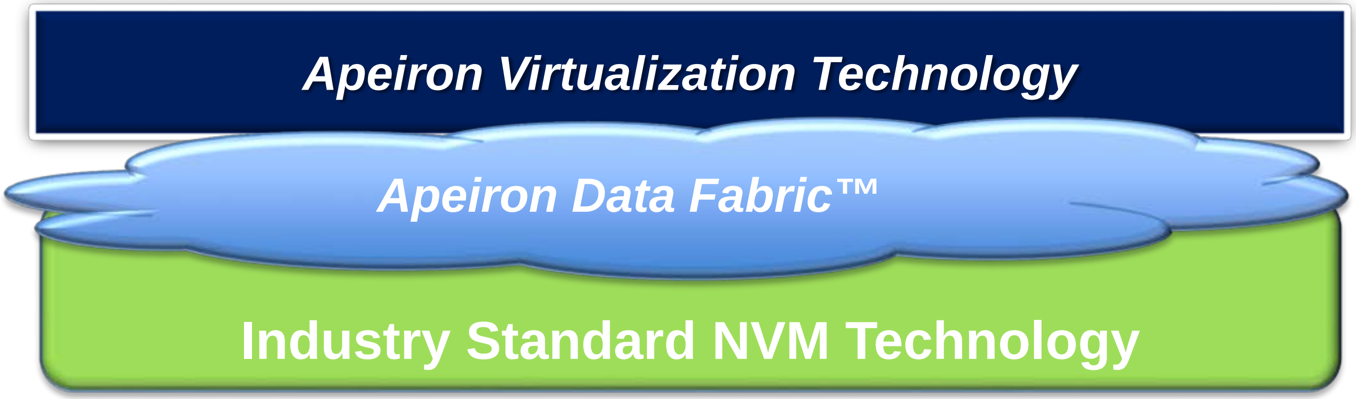
> Performance Density
  - 18M IOPs, 72GB/s BW
  - In a 2U form factor

> < 90 µS 4K read latency P99
  - Ready for Next Gen NVM (<3 µS Fabric Latency)



apeiron DATA SYSTEMS

**Shared DAS™**


Apeiron Virtualization Technology
Apeiron Data Fabric™
Industry Standard NVM Technology

***Come visit us at Booth 819***

**“All the simplicity and promise of DAS with the efficiency and capability of network attached storage.”**

apeiron™
DATA SYSTEMS