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

Bob Hansen,
VP System Architecture
bob@apeirondata.com

apeiron™
DATA SYSTEMS

Storage Developers Conference, September 2015

Agenda

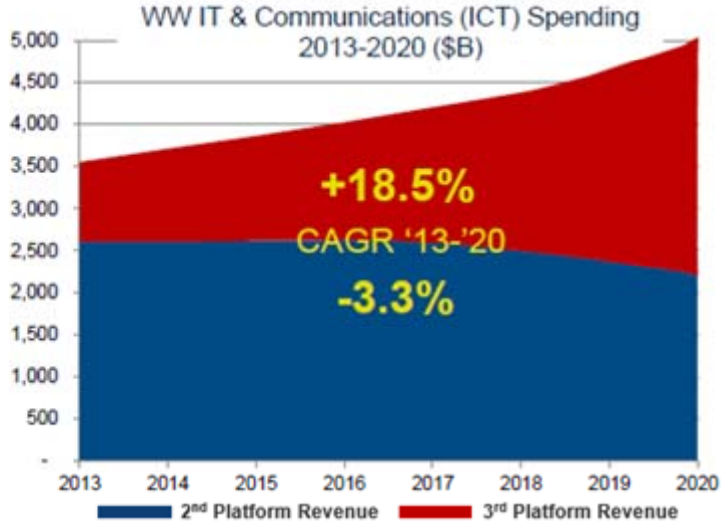
- 3rd Platform Opportunity for High Performance Storage
- Applications with enhanced user experience require:
  - High IOP performance & low-latency
    - Storage performance = $$ PROFITS
  - Scalability
- Scale out, in-memory compute/storage architecture evolution
  - In-memory => in-box flash => external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

apeiron

3rd Platform Opportunity

Source: IDC Predictions 2015: Accelerating Innovation - and Growth - on the 3rd Platform, Doc #252700, December 2014

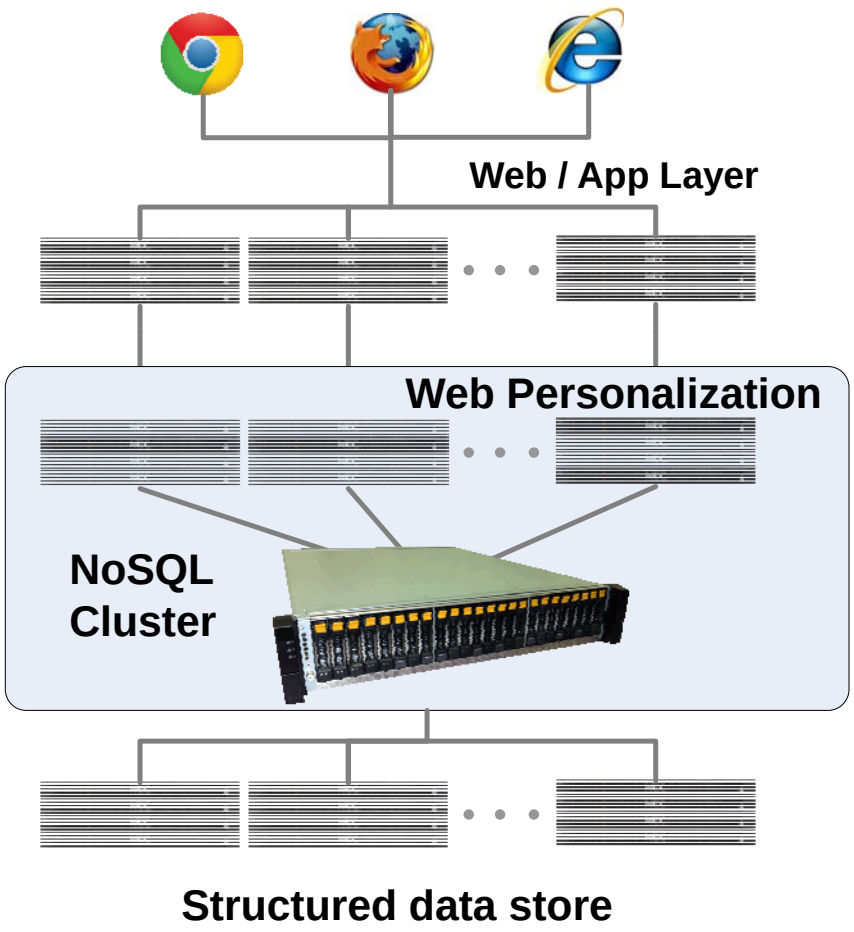
According to IDC, 3rd Platform technologies already drive 30% of ICT spending and 100% of growth and 2nd Platform will enter recession in 2015



Source: IDC, Accelerating Innovation on the 3rd Platform, doc #DR2015_GS1_FG, March 2015

apeiron

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

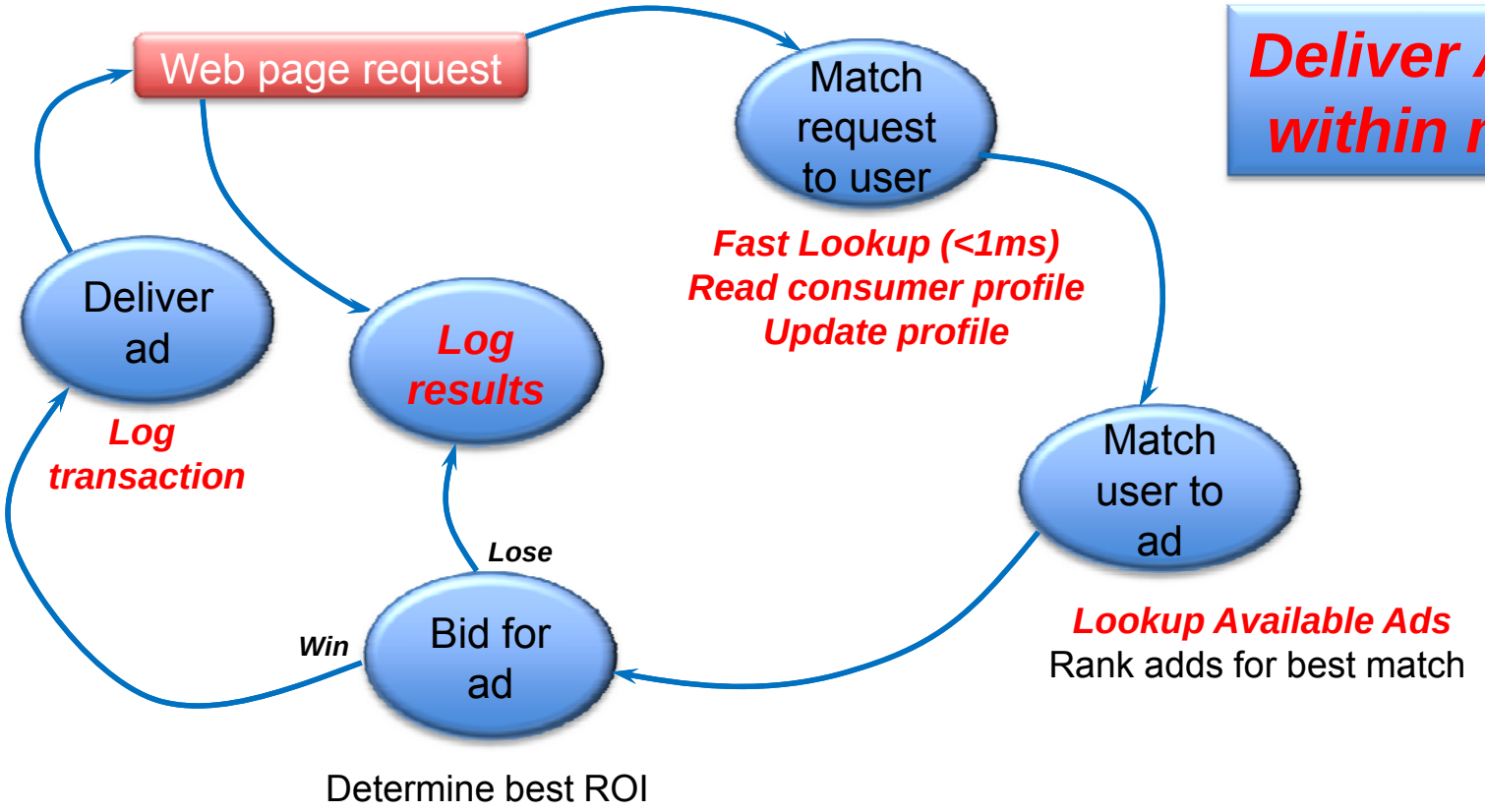
AEROSPIKE

DATASTAX

Couchbase

apeiron

Ad Tech Example



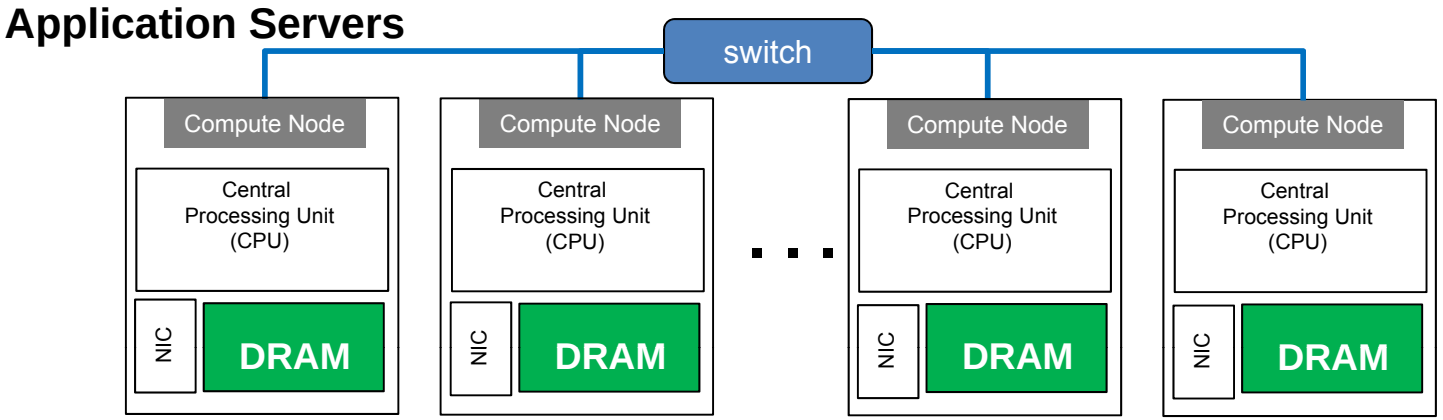
Deliver Ad within ms

- >1 billion consumers
- >3 billion devices

Storage IOPs / latency = $$

apeiron

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

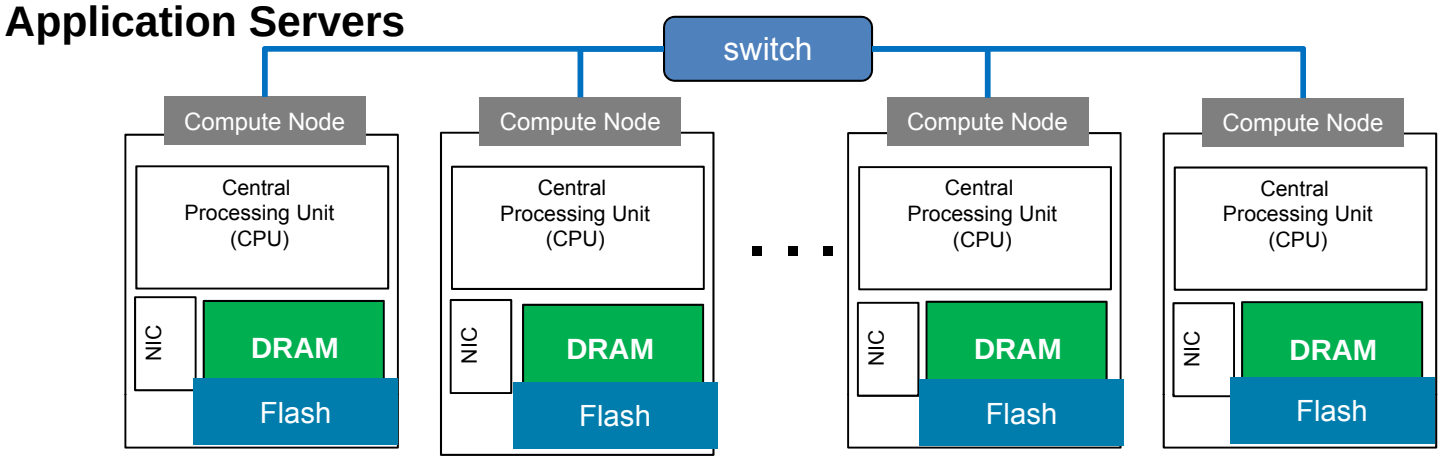
Issues

- 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 foot print
- CPU to mem ratio can not be optimized

This breaks down as you approach 100TB

apeiron

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

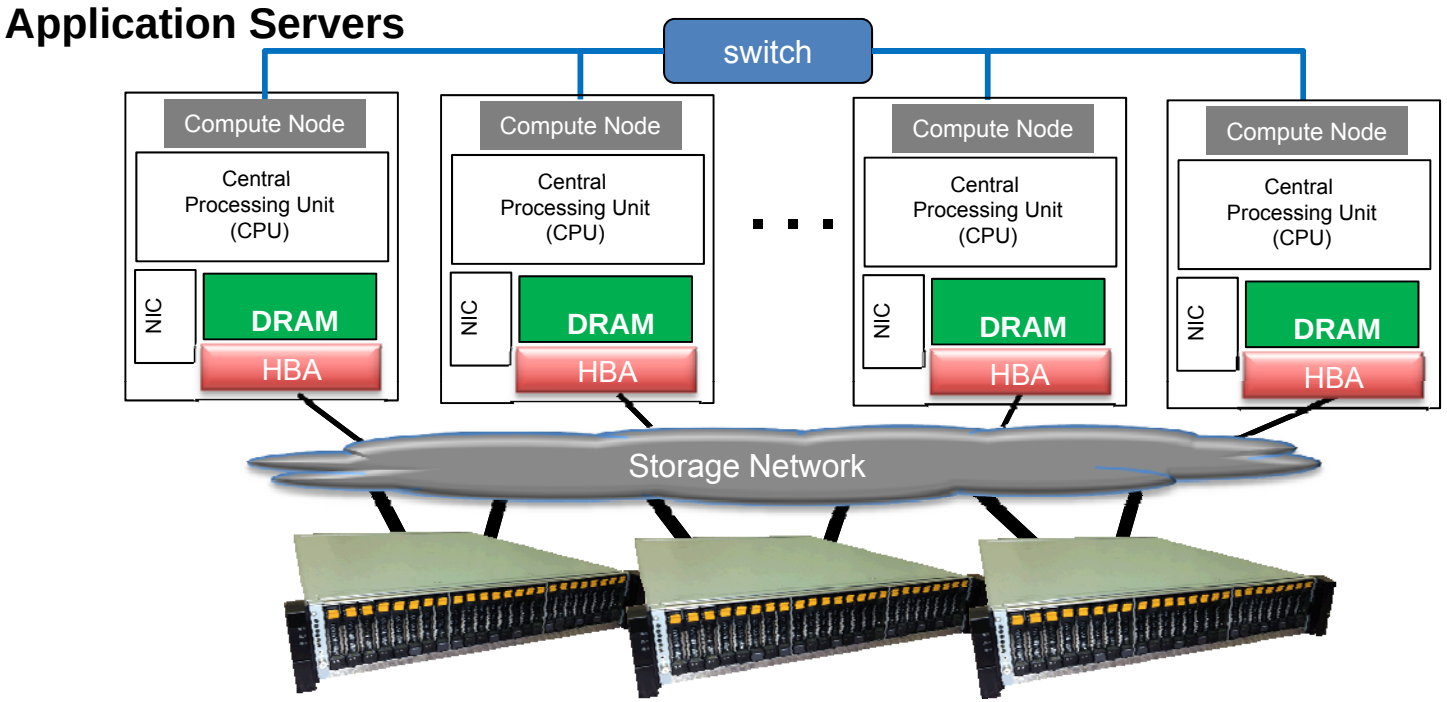
Issues

- 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

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!

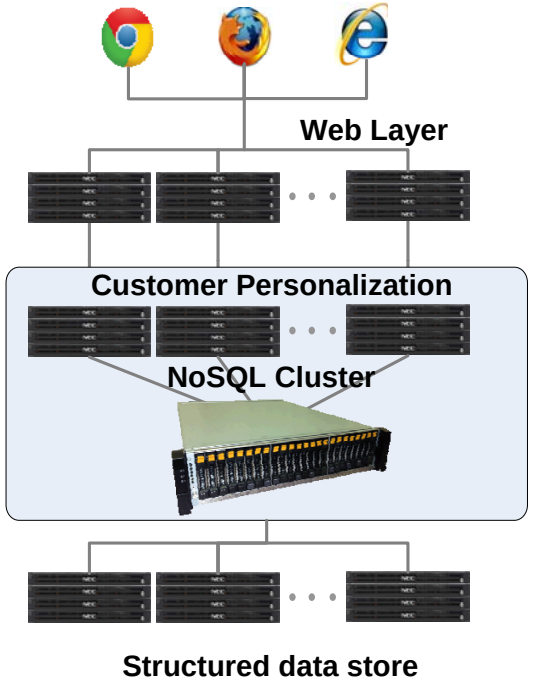
Deliver raw NVMe performance to the application

apeiron

Application Use Cases

Scale-out

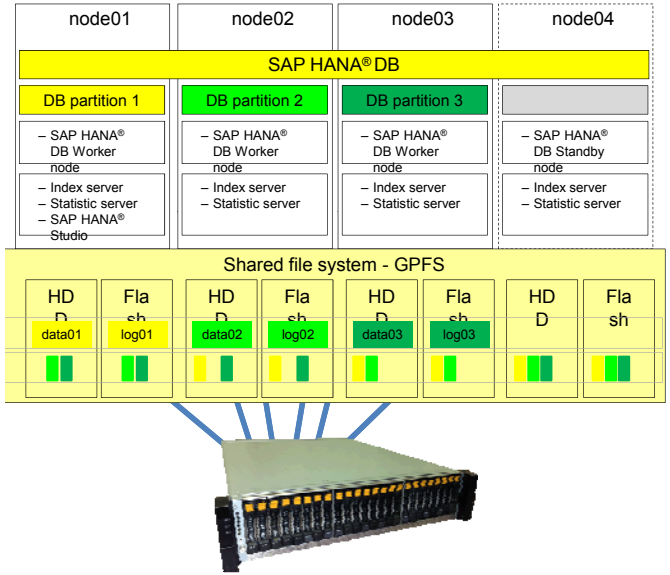
Pooled Flash For Operational Big Data



Ad Tech RTB
Fraud detection
Customer personalization

High Bandwidth

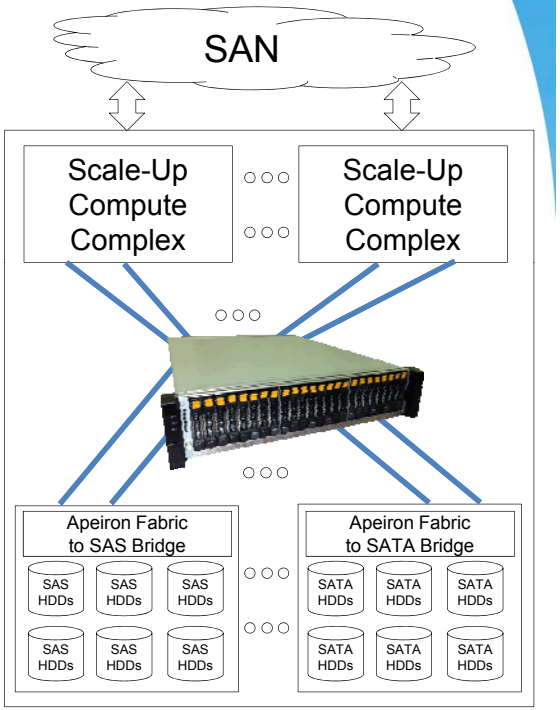
Data storage for applications with high bandwidth



In-memory check points requiring massive bandwidth

Fast cache

Tiered storage acceleration

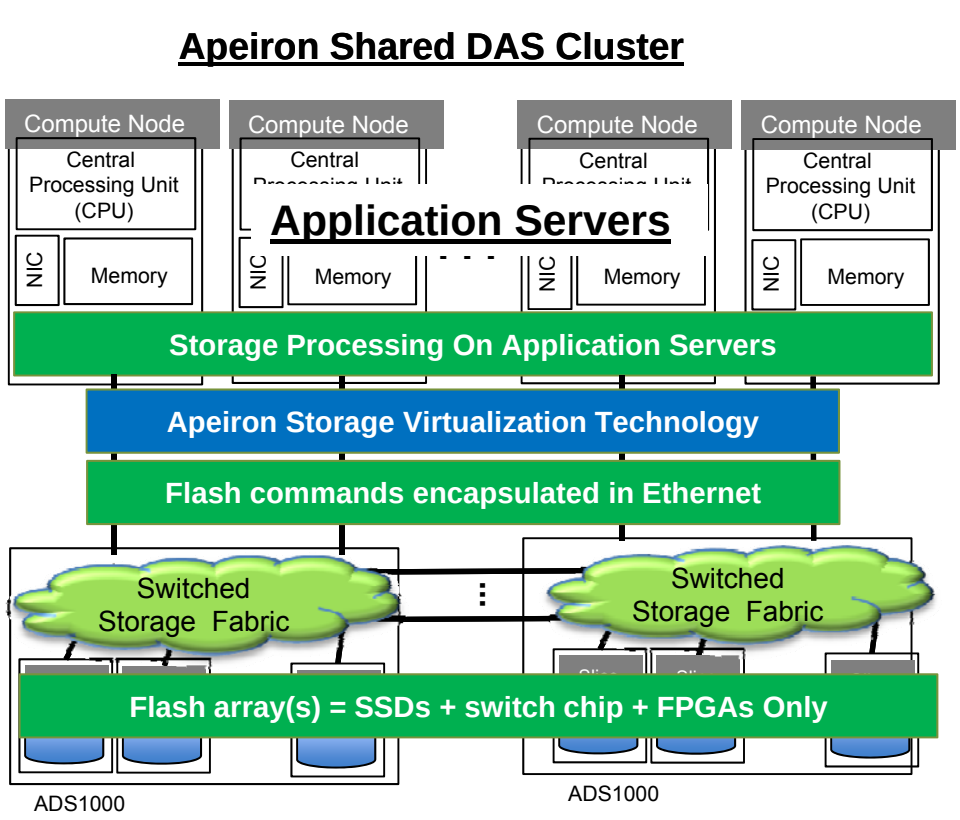


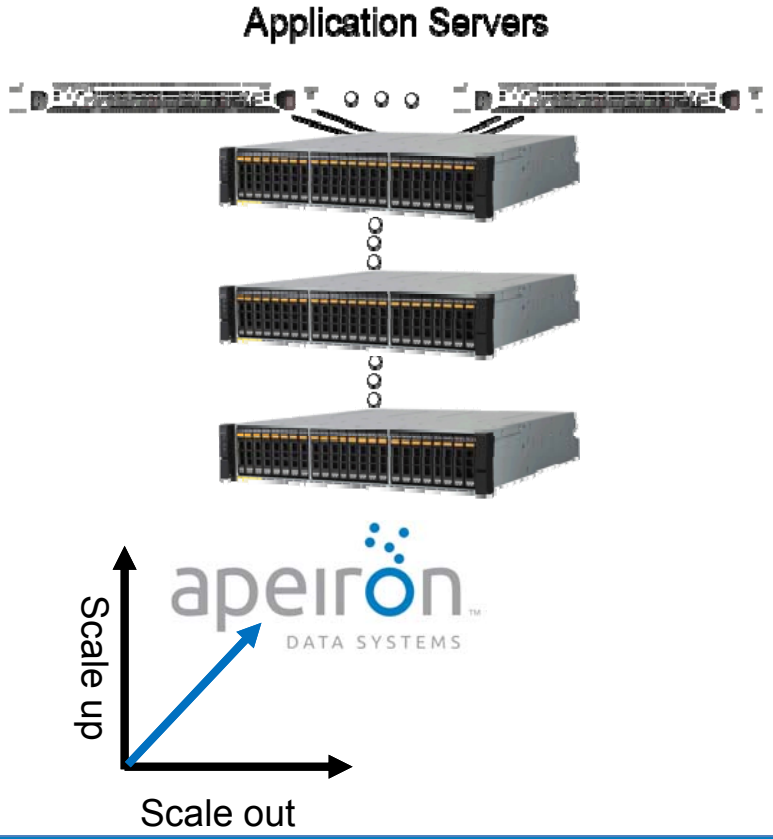
Accelerates response to time critical data
Seamless scaling

apeiron

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

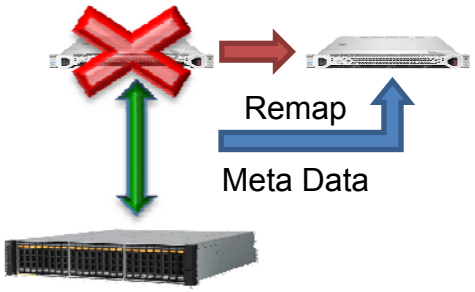
- Scale-out NVM storage architecture
- Intelligent software with hardware accelerated data path
- Ethernet storage fabric with <3uS round-trip latency overhead
- Seamless scaling to petabytes





Apeiron's Software with Hardware acceleration

Instant Failover



Reduces node rebuild time from ***>10hrs to <1sec***

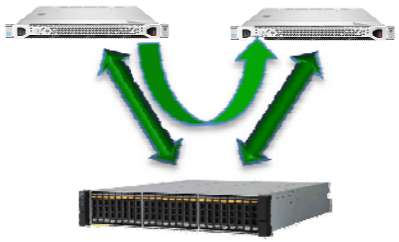
Remaps metadata to spare

Automatic Replication

Transparent backup
Hardware assisted SW configured

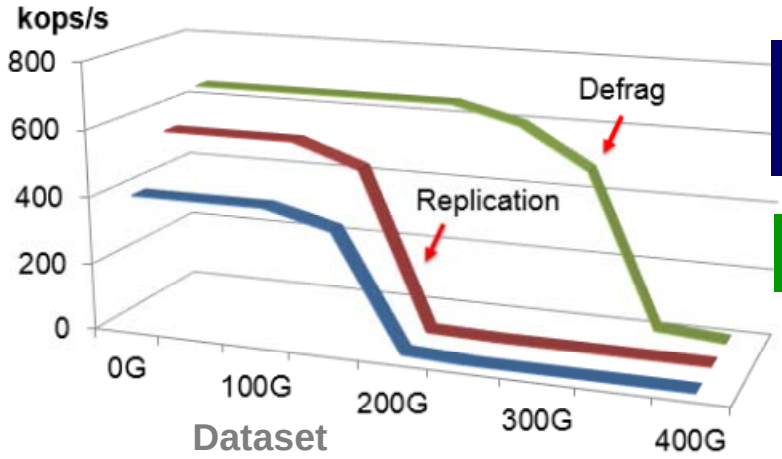
Increases OP/s 40%

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

apeiron

Why not “PCIe on a rope”?

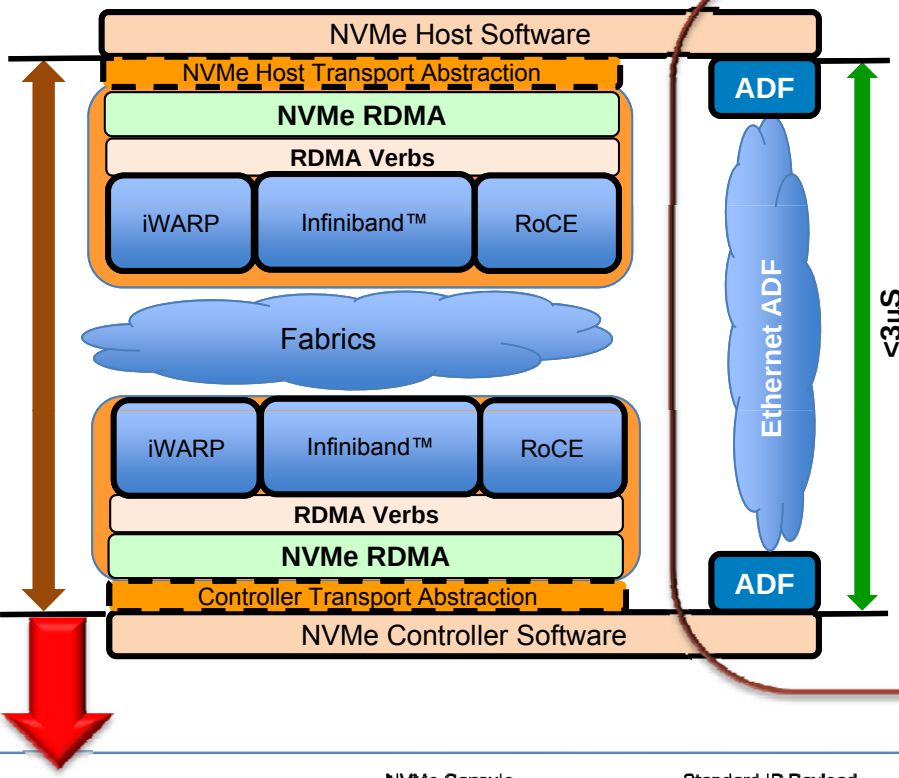
A PCIe storage network is possible but faces several challenges -

- 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?

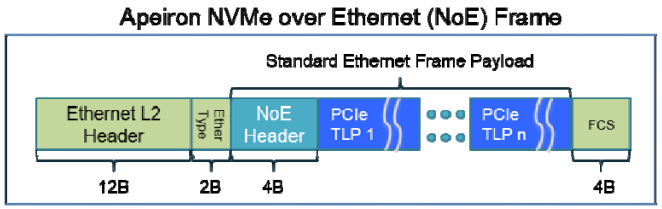
apeiron

RDMA / Apeiron Data Fabric™ Comparison



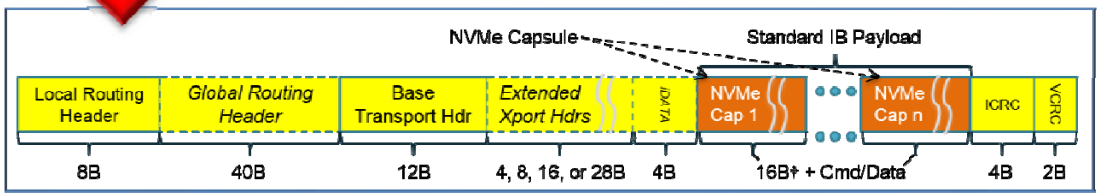
Apeiron Data Fabric™

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



22B Fixed Overhead

L2 Ethernet Headers + 4B NoE Header/encapsulated PCIe TLPs



42B ~
> 112B
Overhead

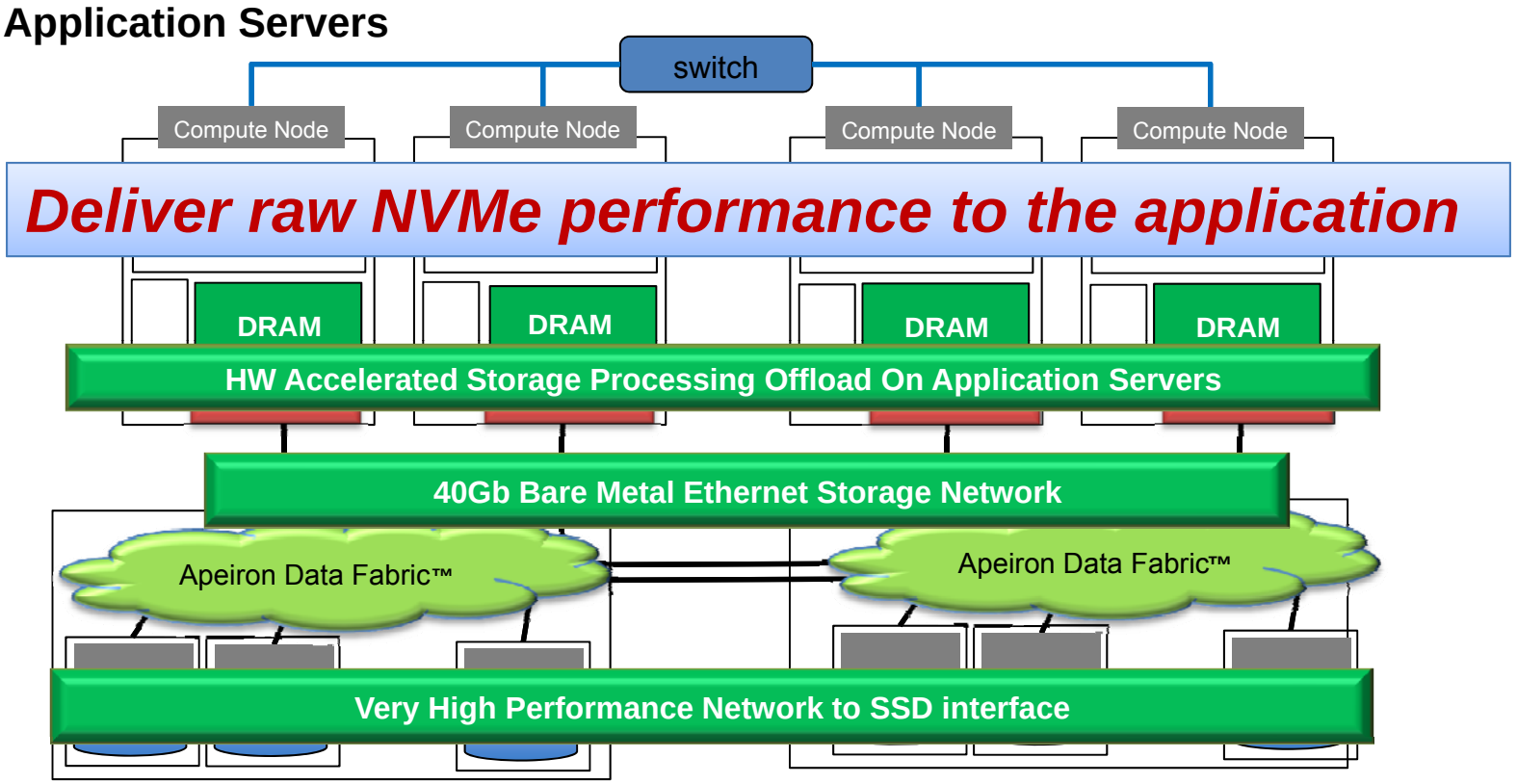
Optional in some configurations

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

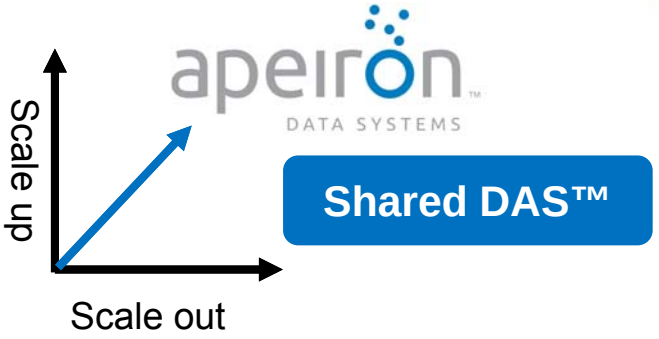
Apeiron System Architecture
Shared DAS™

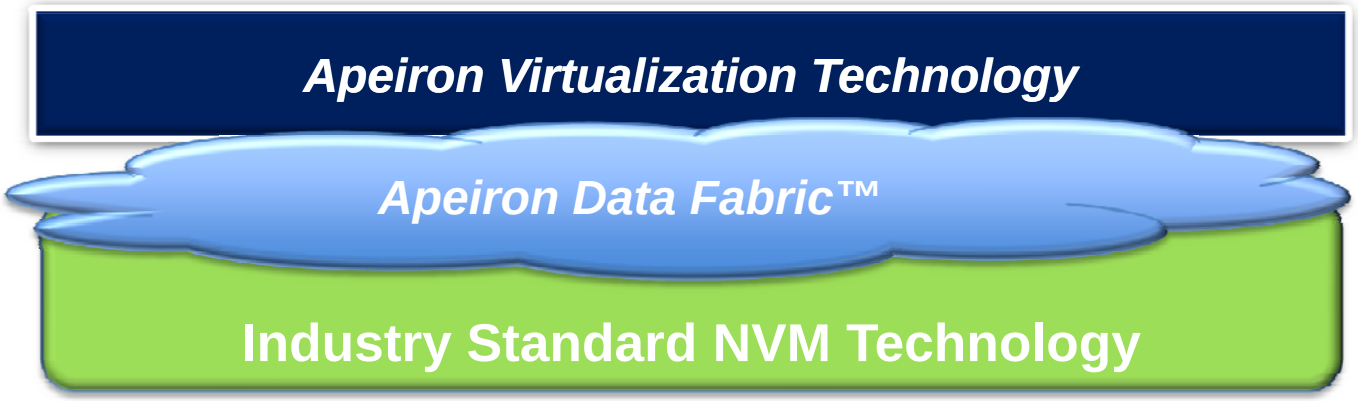


- Simple, scalable architecture with better than in-box flash performance
- Highly available, shared storage using standard SSDs and networking components
- Virtualized storage, on-line provisioning, failure isolation

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)





apeiron


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