SNIA DSI
DATA STORAGE INNOVATION CONFERENCE

Accelerating Real Time Big Data Applications

Bob Hansen

apeiron DATA SYSTEMS

SNIA

Apeiron Data Systems

Apeiron is developing a VERY high performance Flash storage system that alters the economics of Big Data and HPC applications. Our storage systems dramatically improve application performance while simultaneously reducing the size of the cluster.

Agenda

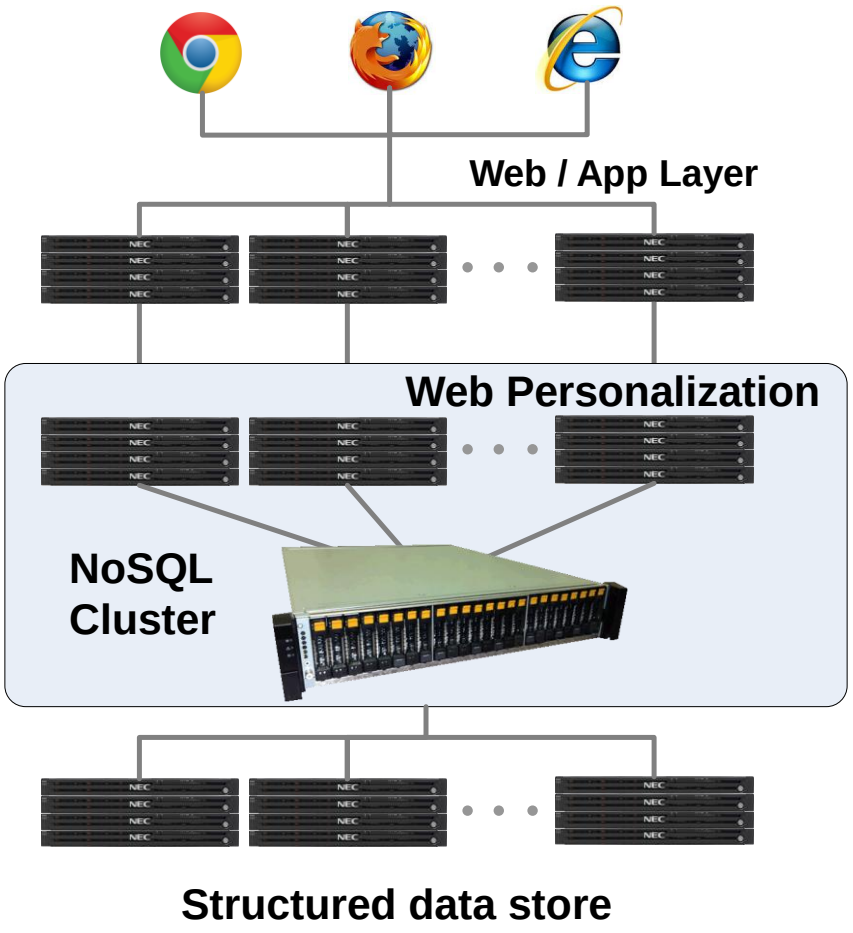
- "Real Time, Big Data" What is that?
- Enhanced user experience drives high IOP performance
  - Storage perf = $$
- Scale out, in-memory compute/storage architecture evolution
  - In-memory => in-box flash => external flash
- The Ideal, Very High Performance scale out system
- VHP System Vision
- What is possible?

Big Data

- Big Data –
  - Structured or unstructured
  - Too big or ***too fast*** for traditional data base and SW techniques
  - Drove the development of scale-out applications using DAS
- Big Data Analytics
  - Off-line analysis of huge data sets
  - Storage requirements = high capacity and BW
- Real Time Big Data (web personalization)
  - Very fast lookups, small records, large data sets

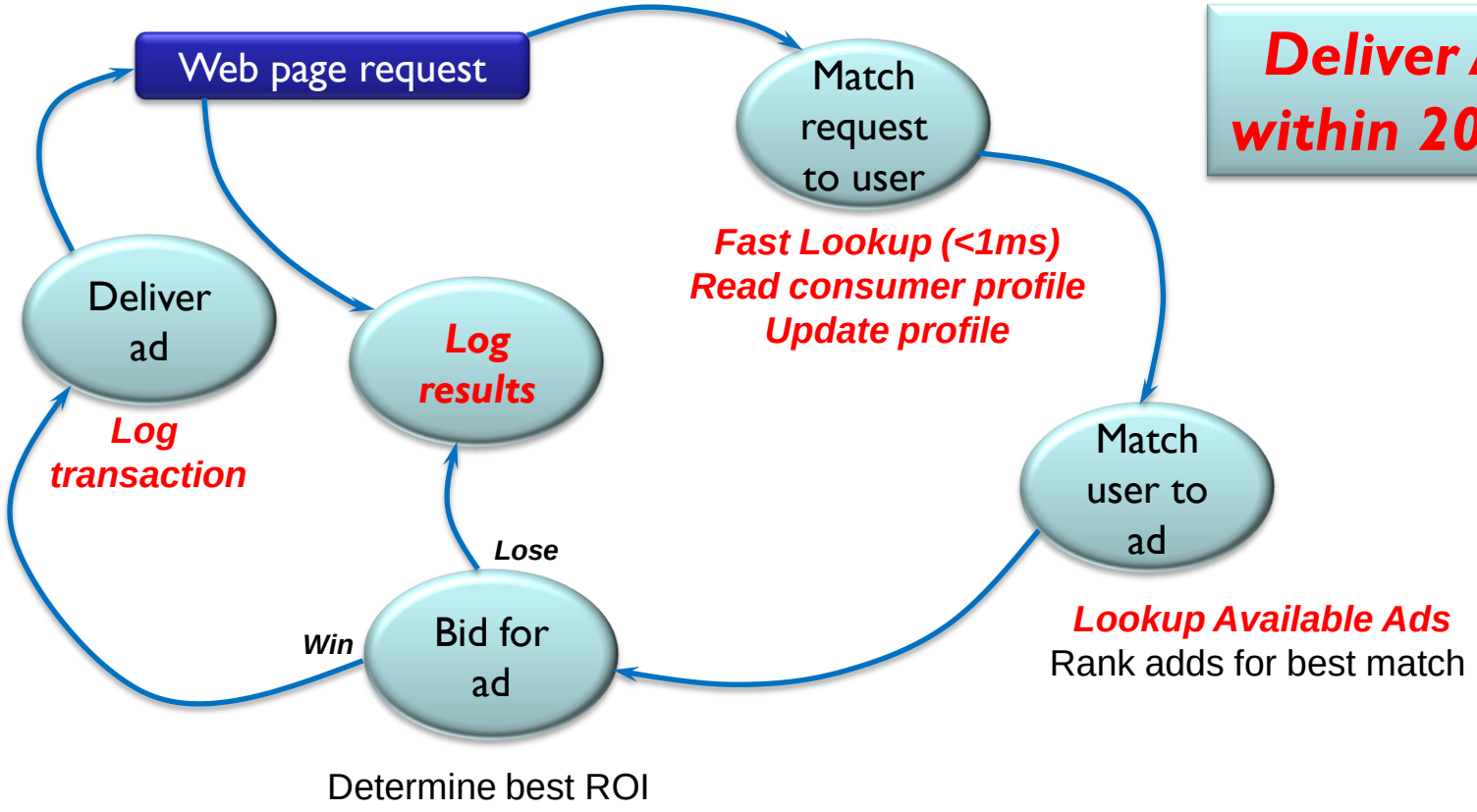
Storage requirements
Large block writes, high IOPs,
very low, predictable, consistent latency
with linear performance scaling

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

Ad Tech Example



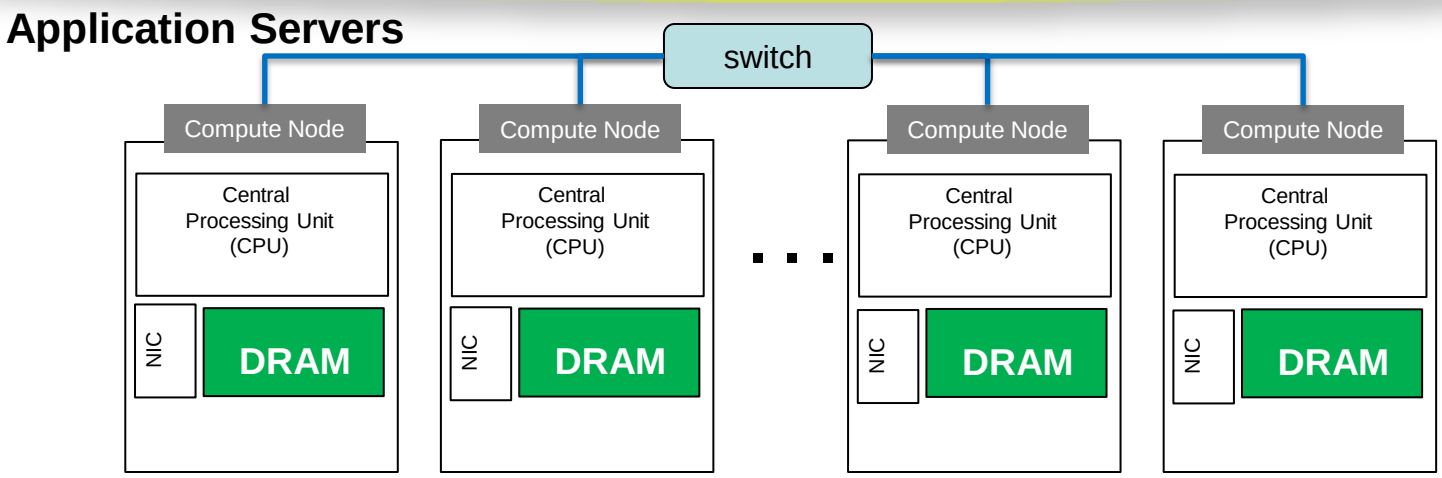
- >1 billion consumers
- >3 billion devices

Storage IOPs / latency = $$

NoSQL and other “in-memory” data applications

High Performance (IOPs) Compute / Storage Architecture Evolution

NoSQL Solution – Scale Out nodes with the data set In-memory



Scale-out in-memory goodness

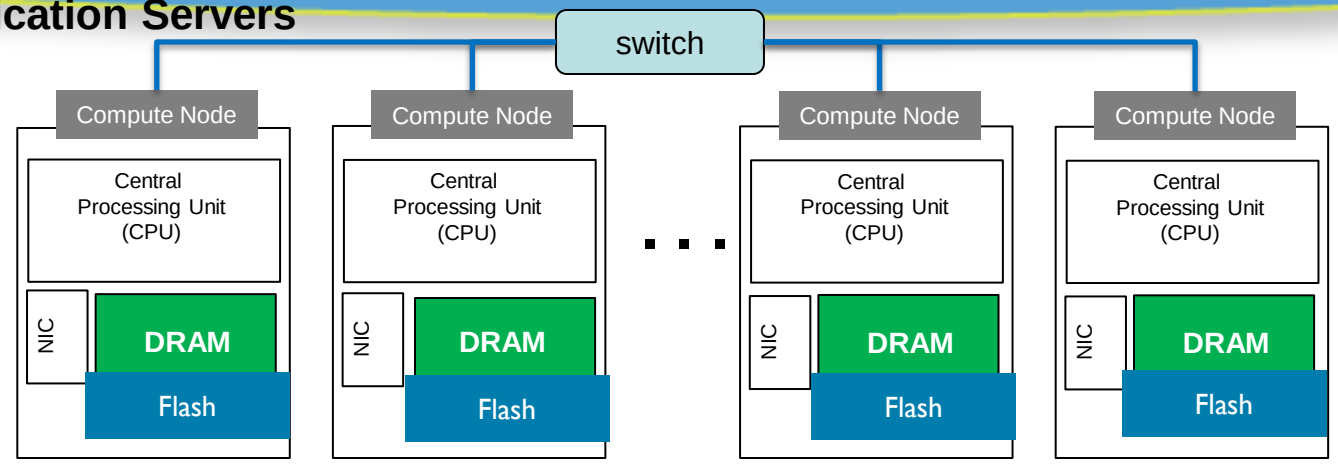
- Shared nothing compute nodes scale well
- Data base is "sharded" evenly across all nodes
- Data set in-memory is VERY FAST
- To scale – just add another node, shard the DB again and go

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 all breaks before you approach 100TB

Expensive DRAM? Add Internal Flash



Scale-out in-memory goodness

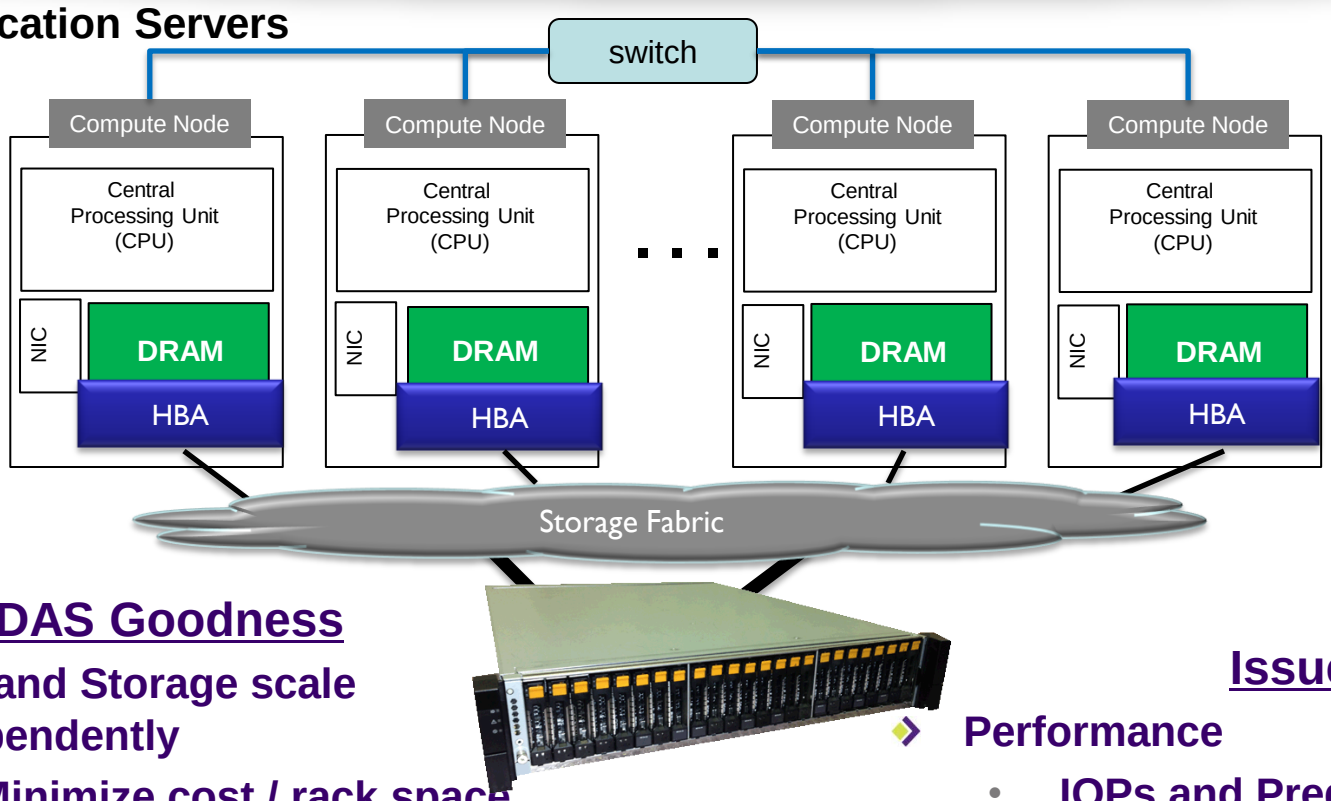
- Shared nothing compute nodes scale well
- Data base is "sharded" evenly across all nodes
- Data set in-memory is VERY FAST
- *Data in flash is FAST*
- To scale – just add another node, shard the DB again and go

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 (VHP) 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

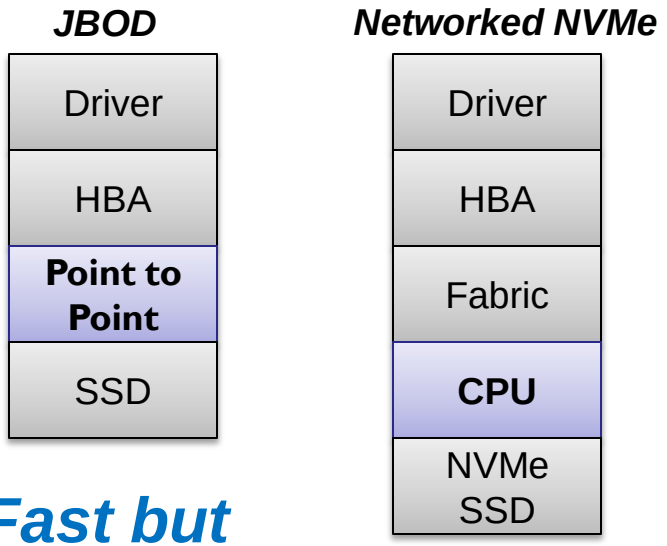
Storage technology choices
IOPs / latency performance

SSD Performance

- 15K HDD – 210 IOPs
- 6Gb SATA SSD – 90K IOPs*
- 12Gb SAS SSD – 155K IOPs*
- NVMe SSD >> ***600K*** IOPs*

SATA and SAS can't cut it!

Get Out of the Box!



Fast but Doesn't Scale

CPU is the Bottleneck

Kills Performance or Adds Cost$$

Objectives

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

* Typical 4K Random Reads

The ideal, VHP persistent storage solution

Shared Direct Attached Storage

- Best performing persistent storage media
  - Standard NVMe SSDs – also best cost
- Bare metal, commodity storage network HW
  - Low cost, industry standard networking (Enet) – also best cost
- Add value where you get best ROI
  - Data path optimization
  - SSD Virtualization
  - High availability
- Best in class management
  - On-line provisioning and failure recovery
  - Storage performance statistics / predictive modeling

Deliver raw NVMe performance to the application

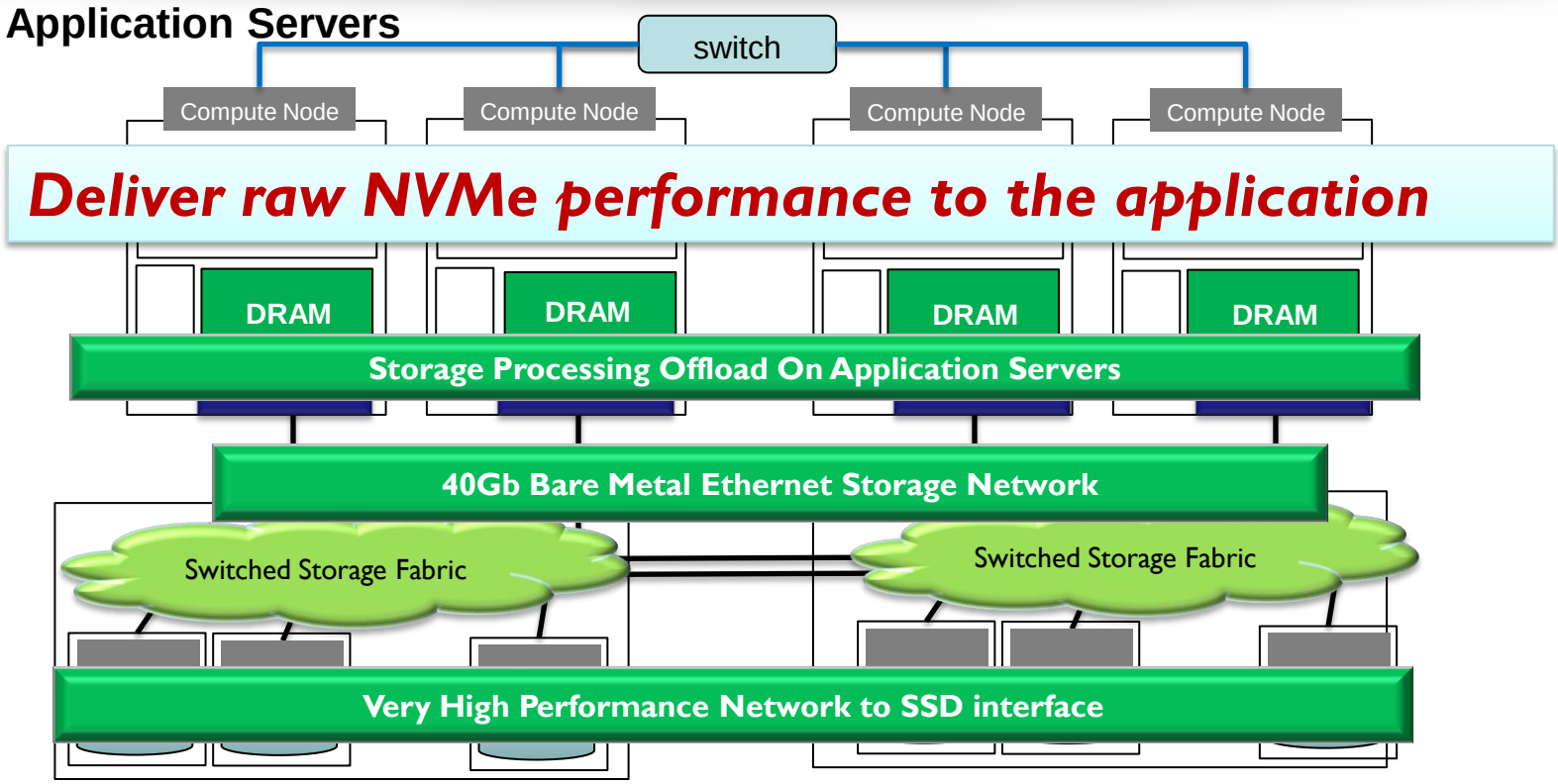
Why not PCIe on a rope?

A native 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 or components
  - Yes, there are switches designed for in-box PCI extensions

Why re-invent PCIe as a high cost, very complex, non-standard fabric?

VHP System Architecture Vision



- 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

What is possible?

- Industry leading IOPs and BW performance
- Industry leading performance density
- Linear scaling to 100s of servers and Petabytes of storage
- Dramatically lower solution cost
  - CapEx
  - OpEx
- An easy to manage, extensible solution – ready for the next gen flash, networking and PCIe technology

"Watch this Space!"

apeiron™ DATA SYSTEMS

Thank You!

www.apeirondata.com

bob@apeirondata.com