Storage Speed and Human Behavior

Eric Herzog
CMO and Senior VP of Business Development
Violin Memory
“I Feel the Need for Speed”

Enterprises

Software-as-a-Service

Cloud Providers
But “Houston, we have a problem”

Latency

Latency

Latency

Latency

Latency is THE ENEMY
What is Latency?

Essentially any delay or lapse in time. In general, it is the time between initiating a request in the computer and receiving the answer. Data latency may refer to the time between a query and the results arriving at the screen or the time between initiating a transaction that modifies one or more databases and its completion. Hard disk latency is the time it takes for the selected sector to be positioned under the read/write head. Channel latency is the time it takes for a computer channel to become unoccupied in order to transfer data. Network latency is the delay introduced when a packet is momentarily stored, analyzed and then forwarded.
Low latency makes a human difference

Low latency makes a business difference

Based on Google/Bing experiments on ecommerce apps
Latency Impacts Customer Satisfaction and REVENUE

- Full impact of latency not reached till 50ms
- Economic impact of latency is significant
- Human engagement depends on low latency
- Quality of interaction improves with low latency
Latency comes in Many Forms

- Compute Latency
- Software/Application Latency
- Network Latency
- Storage Latency
Solving for Latency

**Compute Latency**
- Faster CPU
- CPU Core Density
- Faster CPU-Bus Interfaces
- Faster CPU Memory
- More CPU Memory
- Virtualization

**Software/Application Latency**
- Rewrite Code
- Optimize Code
- Repeat Again
- Rewrite Code
- Optimize Code
- Repeat Again
- Rewrite Code
- Optimize Code
- Repeat Again

**Network Latency**
- Faster Network Speeds
- Hardware Switching
- Packet Compression
- Over Engineering Networks to Handle Traffic Bursts
- Transport Layer and TCP/IP Stack Optimization

**Storage Latency**
- Faster Rotational Speed – from 3600 to 5400 to 7200 to 10,000 to 15,000
- Faster RAID Controllers
- RAID Controller Caching
- HDD Short Stroking
- Short Stroked HDD Aggregation
- Additional RAID Controllers
The Storage Bottleneck is the WORST

Growing performance gap illustration

SOLUTION: THE ALL FLASH ARRAY

Server
- Multi-core
- Virtualization

Networking
- 100 GbE
- Network optimization

Storage
- Disc centric
- High latency
Flash Arrays SOLVE the LATENCY Problem

CPU Cycle with Traditional Disk Arrays:
- 80% Wait
- 20% Work

CPU Cycle with All-Flash Arrays:
- 5% Wait
- 95% Work

Lowest Latency
Billions More Operations Per Second
More Work in the Same Time

Source: Violin
Flash Delivers

Delivered a latency of .12 ms compared to latency of 8 ms with their previous storage

30X reduction in peak application I/O read latency

“The combination of low latency, high IOPS and high throughput made Violin Flash Memory Arrays a perfect fit for all workloads.”

40X reduction in application I/O latency

Random I/O latency down to .6 ms from 9 ms
Sequential I/O latency down to 2 ms from 14 ms

8X reduction in application I/O latency

270% better latency
21X more IOPs

40X reduction in application I/O latency
18X increase in IOPS

I/O latency under .15 ms
35x faster than previous tier-1 storage
Summary

- Today’s Cloud/Compute World is all about Speed
- Latency is the ENEMY
- Several Approaches to Solve the Latency Conundrum
- Storage is the WORST Latency Offender
- All-Flash Arrays Solve the Storage Latency Bottleneck