High Performance Storage in Today’s Critical Applications

March 23, 2014

Andy Walls, IBM Fellow, CTO and Chief Architect
IBM Flash Systems
Hard Disk Drive History

- RAMAC was the first hard disk drive!
  - One of the top technological inventions... EVER!!

- 5MB across 50 HUGE platters

- After 50 years, the capacity increase is incredible.

- As are the reliability increases... . . .

- Performance limited by the rate at which it can spin.
  - 15K RPM

- Has not kept up with the speed of CPUs

RAMAC Prototype
**Hard Disk Drive History**

**Areal Density**

- **HDD growth focus:** areal density for 50 years
- **Data rate has just topped 100MB/sec. But RPM not increasing. New increases will come from linear density improvement**

**Access Latency**

- **SO:** With HDDs, Performance improvements have been gained by scaling out high speed disks and only using a portion

**Data Rate**

- **HDD access latency:** <10% / y for most of that period
Hard Disk Drive Technology Has Not Kept Up With Advances in CPUs or CPU Scaling

As you can see from this database example, which uses rotating disk drives, even well-tuned databases have the opportunity to improve performance and reduce hardware resources.

Reducing I/O wait time can allow for higher server utilization

Source: Internal IBM performance lab testing
IT Infrastructure Challenges

CPU performance up 10x this last decade
Storage has grown capacity but unable to keep up in performance
Systems are now Latency & IO bound resulting in significant performance gap

From 1980 to 2010, CPU performance has grown 60% per year*
...and yet, disk performance has grown ~5% per year during that same period**
Flash is a powerful accelerator for today’s critical applications

- Big Data – Hadoop, MongoDB, Cassandra
- High Performance Cloud
- Business Analytics
- OLTP
- HPC
How Flash Accelerates Today’s Most Critical Applications

• Latency
  – Inherent read latency
  – Systems employ DRAM for buffering so write latency can be very fast

• IOPS
  – Very high IOPS
  – More importantly, high IOPS with low average response time under load.
  – More consistent performance - can handle temporary workload spikes

• High Throughput
  – Reduced table scan times
  – Reduced time for clones and snapshots
  – Reduced time for backup coalescence

• Reduction in batch windows
The Impact of Low Latency on CPU Performance

**MicroLatency**
deliver microseconds response time to accelerate critical applications to achieve competitive advantages

- Faster decision making
- Increase revenue
- Accelerate cost savings
- Eliminate wait time
- Scale performance with capacity

100 microseconds : 1 second :: 1 second : 2.78 hours
The Value of Performance

Extreme Performance enables businesses to unleash the power of performance, scale, and insight to drive services and products to market faster.

- Improved end-user experience
- Faster insights into critical applications

A 1-SECOND DELAY IN PAGE LOAD TIME = 7% LOSS IN CONVERSIONS, 11% FEWER PAGE VIEWS, 16% DECREASE IN CUSTOMER SATISFACTION

In dollar terms, this means that if your site typically earns $100,000 a day, this year you could lose $2.5 million in sales.

Source: Aberdeen Group

CCBCC cut data processing time by 75% without replacing a single server.

Source: Coca-Cola Bottling Co. Consolidated case study
Much has Changed Around Flash Enabling Technology

• Given the right controller technology, one really does not have to worry about endurance any more
  – IBM is a Leader in enabling MLC for enterprise applications

• Well designed all flash arrays can be designed with excellent write performance

• Flash has excellent sequential throughput characteristics
  – Not just good random IOPs
  – Most workloads have some attributes of each and Flash excels
Flash Offers Other Significant Advantages

• Power reductions
  – A key consideration in driving Internet data centers to Flash
  – Can be the main driver in internet data centers and Big Data

• Density
  – Incredible densities per rack unit possible with Flash
  – Saves rack space, floor space

• Form Factors and Flexibility
  – Can be placed in many parts of the infrastructure
  – Can go on DIMMs, PCIE slots, attached directly via cables, unique form factors, etc.

4TB Custom Flash Module
High Performance Networked Flash Storage Architectures

• Inside Traditional Storage Systems
  – Hybrid or pure storage

• All Flash Arrays
  – SAN Attached
  – RDMA SAN
    • IB SRP, iSER, RoCE
  – SAN “Less”
    • Ethernet, iSCSI
  – Building blocks for scale out storage.

• Advantages
  – Shared!
  – High Availability built in
  – Advanced storage function like Disaster Recovery
  – All flash array is flash optimized from ground up

• Perceived weaknesses
  – Network latencies
  – Further away from CPU
World Class and Consistent Performance!

IBM FlashSystem 840
Random 4K Read/Write Performance

- 100% rr
- 90% rr-- 10% rw
- 80% rr-- 20% rw
- 70% rr-- 30% rw
- 60% rr-- 40% rw
- 50% rr-- 50% rw
- 40% rr-- 60% rw
- 30% rr-- 70% rw
- 20% rr-- 80% rw
- 10%rr -- 90% rw
- 100% rw
High Performance Direct Attached Flash Storage Architectures

- **PCIe Drawers**
  - Dense and can be attached to 2 servers

- **PCIe Cards**

- **Flash DIMMs**

- **Advantages**
  - Attached to lowest latency buses
  - Memory bus is snooped
  - Uses existing infrastructure for power/cooling

- **Perceived weaknesses**
  - No inherent high availability
  - Mirroring more expensive than RAID
  - No advanced DR or storage functionality
Bottlenecks in Flash Storage

- **RAID Controllers**
  - Flash Optimized RAID controllers with hardware assists now exist

- **Network HBAs**
  - Reductions in latency
  - RDMA protocols

**OS and Stack Latency!**

- Standard driver model adds significant latency and reduces IOPS per core by an order of magnitude
- Fusion-io Atomic Writes
- sNVM and SCSi
- IBM Power CAPI

- **Many Legacy Applications written around HDDs**
  - Added path length to coalesce, avoid store, etc.
CAPI Attached Flash Value

**Concept**
- Attach FlashSystem to POWER8 via CAPI coherent attach
- CAPI flash controller operates in user space to eliminate 97% of instruction path length
- Lowest achievable overhead and latency memory to flash.
- **Saves up to 10-12 cores per 1M IOPs**
Workload Optimized Systems and Flash

• Analytics
  – Very fast table scans
  – Tremendous IOPS capability to identify patterns and relationships

• OLTP
  – Credit card, travel reservation, other
  – Can share without sacrificing IOPs
  – But low response time is key

• Cloud and Big Data
  – Either inside servers as hyper converged or
  – Linear scale out with QoS for Grid Scale.