Hybrid SSD/HDD Storage: A new Tier?

George Symons, COO
Xiotech Corporation

Hybrid SSD/HDD Storage: A New Tier

- Hybrids today
- SSDs and Arrays
- HSM – Hierarchical Storage Management and SSDs
- MLC SSD enablement – technology and architecture
- Hybrid…how might it work?
- Solution Cost to Customer

Hybrids Today

- Hybrids exist today in multiple areas, even though in many cases we have come to accept them as just n

Servers (memory, HDDs)

HDDs

Automobile engines

Some for economy…e.g. prius

Some for performance…Lexus

- Why do all these things exist vs being just one type of technology: EFFICIENCY, which relates to COST

SSDs and Array Controllers

- Array Controllers are consolidators or aggregators of storage devices
  - The first arrays made physical to virtual transformations of HDDs blocks
  - Later arrays became places to add data protection, replication, and every feature you can name
  - One thing is clear, arrays handle multiple devices to abstract the physical to the virtual
- SSDs are just single controller array controllers with multiple banks with multiple flash devices on each bank. Given the sizes of chips today, we are at the point where SSDs are 2011's equivalent of a 1995 SCSI array controller.

SSDs and Array Controllers

- So why the connection of SSDs to Array Controllers?
- Because Arrays now of SSDs or including SSDs are out there…how best to use them to make the whole greater than the sum of the parts
  - SSD as cache
  - SSD as HSM component (tier)
  - SSDs in totality (SSD Array Controllers)
  - SSDs as something else…hybrid of some sort?

HSM – Hierarchical Storage Management and SSDs

- HSM or ILM is the method to store data at different levels based on usage patterns or age of data, location, etc.
- SSD as cache, however, is what is basically done in most arrays today, being used as a somewhat automated HSM. SSDs are basically written to after internal DRAM cache and then down to lower levels and this is monitored by rules about activity to decide when to move things (or lock things) in one tier or another.
- What's wrong with this?...Tons

HSM – Hierarchical Storage

- Using SSD as a cache means it is not its own tier, along with the DRAM in the array, the enterprise HDD, and the nearline HDD.
  - This means more DRAM to represent the SSD blocks in data structures
  - This means after DRAM, the next place, like a L2 cache, is the SSD, meaning writes go through the SSD all the time, on their way down to HDDs etc.
  - This explains the wholesale use of SLC flash SSDs in the big arrays and even the PCI cards being used as look-aside caches in servers today.
  - Most of the big arrays also don't have enough horsepower left with all the features that suck away performance to drive to many SSDs all the time
  - All of this is the reason many have moved to full SSD solutions, and why customers note the high cost of HSM tiering solutions…but is there possibly a more effective solution? A Hybrid Tier?

SSD enablement – technology & architecture

- The cost of flash is well known between SLC and MLC. MLC is much less but
- MLC can be erased in general about 10x less than SLC (eMLC is there and new techniques for longer lasting flash is coming)
- MLC can't handle as well as SLC on write
- SLC and MLC are still great on reads…
- Many people want to enable MLC in the enterprise and have in some situations. Add to that the efforts to make MLC last longer (LDPC, eMLC, etc)…but it still has issues that need an array to manage
  - SSD wear out
  - SSD failures (same as SLC)
- If the good parts of MLC can be focused on, while the less optimal parts mitigated, we all win

Hybrid tier…how might it work?

- First, assuming a hybrid of SSD and HDD makes sense..
  - Are flash costs still more than HDD? Yes, today. If the answer was no, then all HDD vendors would be out of business; Enterprise HDDs make up the vast majority of server and storage solutions today for applications, virtualization, and VDI.
  - How much cost savings can be gotten by such an approach? Does it save just drive costs or other?
- In an ideal hybrid solution, the HDD and Flash/SSD just work together (e.g. Seagate Momentus). For arrays, it has to be done in an enclosure or group of enclosures.
- The hybrid has to make more monetary sense than just HSM tiering
  - Eventual locking of hot application data in SSD
  - The hybrid also must not use more DRAM as HSM tiering does to point at everything.
- Essentially it has to become a tier by itself that the array manages as such to get capacity and performance for the right I/O density, lifetime, and end-user cost

Why build a hybrid

- SSD price/GB is ~10x that of enterprise HDD
- SSD random access read perf is ~100x HDD
- SSD random-access write perf is 3-10x HDD
- SSD sequential access is 1-3x HDD
- Because of access locality in applications, a hybrid SSD/HDD system could be a price/performance winner - IF you figure out what data goes on SSD…table below relative as all costs go down

20TB subsys, $4K fixed cost	All HDD	All SSD	Hybrid 5-10% SSD
Build Cost	X	6X	1.3-1.6X
Relative Perf	1	15	4 - 8

Using SSD effectively in a Hybrid

- Users must be able to create “pure” SSD volumes
  - Because some users know best, and others think they do
  - Some low-IOPS applications are excessively latency sensitive
- Users must see gains in performance without tuning or rules for manual entry…needs to be automatic
  - Most applications are not excessively latency sensitive, but need better I/O per GB to utilize all the storage they have with multiple workloads running from applications to dense storage

Managing SSD as a Storage Tier

- Goal is to put hot data on SSD, colder data on HDD
  - Rule of thumb /old folklore: IO obeys an "80-20 rule"
    - SPC1 random-access IO obeys a "90-8 rule"
  - User does not know, or can't express, where hot spots are
  - Apps may know where their hot spots are, but not how they rank with respect to another app's hot spots…but apps hints can help too since it is generating I/O
  - Only the controller really knows; it sees the traffic realtime
- A hybrid system migrates data to/from SSD based on "heat"
  - Management algorithm in a nutshell: "Count back-end IO's to every chunk of disk space; periodically migrate hot HDD chunks to SSD and cold SSD chunks to HDD"…
  - **Allows cache to work as cache, weeding out things that might cause thrashing of either HDD or SSD in ways they don't work as well, as well as sequentials for SSD as HDD works great there.**

Summary

- We feel the case for Hybrid is clear:
  - HSM is not cost effective with SSD as cache, both for number of devices, but cost of additional DRAM to make the system work right.
  - Full SSD arrays are still expensive for capacity density
  - Arrays make more sense than PCI cards because they have shared methodology, HA, etc
  - Customer's applications are so huge they need more storage, just to use what they buy more effectively, which means getting I/O to all the GB/TB they purchased for the least cost
  - MLC enablement is a key to proliferating flash storage and working with HDDs while costs go down to support more homogenous environments
- Xiotech has built the world's first hybrid storage subsystem