The Need For Higher-Level Software for Flash

Adam Leventhal
Fishworks Flash Architect
Sun Microsystems

Flash in the Data Center (2008)

- Beginning: direct replacement for HDDs
- Lower power / foot print for high IOPS applications
- No application changes required
- Problems
  - > Prohibitively expensive in many cases
  - > Limited to niche applications

slides available here: http://blogs.sun.com/ahl/resource/fms09_leventhal_software.pdf

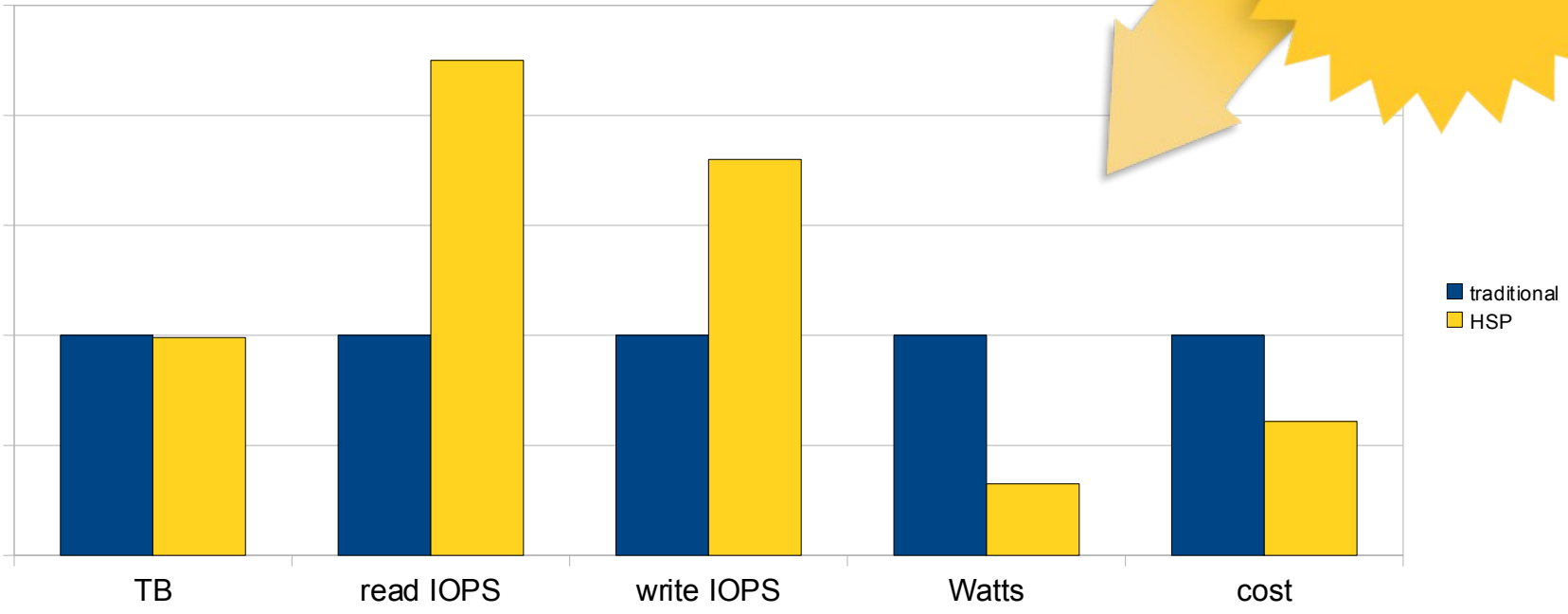
Flash in the Data Center (2009)

- Flash as a new tier in the storage hierarchy
- Keep hard drives
  - Great throughput, capacity per $ or watt
- Use flash to accelerate read/write IOPS
- Change the software to be aware of flash

Example: Hybrid Storage Pool

- 216 x 450GB 15K RPM HDD
- 48 x 1TB 7200 RPM HDD
  6 x read-optmized SSD
  2 x write-optimized SSD

60% the cost
32% the power



Hybrid Storage Pool

- Optimize the ZFS filesystem for flash
- Write IOPS accelerated with write-optimized SSDs
  - Intent-log stored on dedicated device
- Read IOPS accelerated with commodity SSDs
  - Second-level flash cache behind primary DRAM cache
  - Designed with flash performance in mind
  - Populated via “evict-ahead” algorithm
- Independent scaling ($$, TB, MB/s, IOPS, watts)
- Enables use of lower-power, lower-cost HDDs

Flash in the Data Center (2010)

- NAND flash on “Lithography Death March” – Michael Cornwell in his keynote
  - Focus on commodity space and $/GB
  - Higher latency than HDDs in two generations
  - Rapidly dropping reliability
  - The enterprise is not the design center for NAND
- New opportunities
- New innovation required

Flash Interface

- HDD form factors were convenient
- IOPS and visibility limited by FC/SAS/SATA
- Need new interface for flash in the data center
  - > PCIe, NVMHCI
- Software needs to know it's talking to flash!

Further Specialization

- NVDRAM for write IOPS
- Use of MLC flash
  - Ride the commodity trends
  - Cache **huge** datasets
  - Errors should affect performance not correctness
  - Software to manage reliability

Flash Translation Layer

- Software for managing flash
  - Wear-leveling, remapping, etc.
- Scope limited to one device
  - Wear-leveling across RAID-0?
- Move parts of FTL into higher level software
- Enable further commoditization
- Smarter software → Dumber devices → Cheaper systems

Questions?

Adam Leventhal
http://blogs.sun.com/ahl