Fresh Flash Issus:
A Perspective from a Diehard DB Person

Apr. 23, 2009

Sang-Won Lee

http://icc.skku.ac.kr/~swlee

Gray's Observations

What If FLASH Disks Delivered Thousands of IO/s and Were "Big"?

My tests and those of several others suggest that FLASH disks can deliver about 3K random 8KB reads/second and with some re-engineering about 1,100 random 8KB writes per second. Indeed, it appears that a single FLASH chip could deliver nearly that performance and there are many chips inside the "box" – so the actual limit could be 4x or more. But, even the current performance would be VERY attractive for many enterprise applications. For example, in the TPC-C benchmark, has approximately equal reads and writes. Using the graphs above, and doing a weighted average of the 4-deep 8 KB random read rate (2,804 IOps), and 4-deep 8 KB sequential write rate (1233 IOps) gives *harmonic average* of 1713 (1-deep gives 1,624 IOps). TPC-C systems are configured with ~50 disks per cpu. For example the most recent Dell TPC-C system has ninety 15Krpm 36GB SCSI disks costing 45k$ (with 10k$ extra for maintenance that gets "discounted"). Those disks are 68% of the system cost. They deliver about 18,000 IO/s. That is comparable to the requests/second of ten FLASH disks. So we could replace those 90 disks with ten NSSD if the data would fit on 320GB (it does not). That would save a lot of money and a lot of power (1.3Kw of power and 1.3Kw of cooling).

The current flash disks are built with 16 Gb NAND FLASH. When, in 2012, they are built with a 1 terabit part, the device will have 2TB of capacity and will indeed be able to store the TPC-C database. So we could replace a 44k$ disk array with a few (say 10) 400$ flash disks (maybe).

If one looks at the system diagram of the Samsung NSSD there are many opportunities for innovation. It suggests interesting RAID options for fault tolerance (combining the MSR-TR-2006-176 ideas with non-volatile storage map and a block-buffer, and with writing raid-5 stripes of data across the chip array), adding a battery, adding logic for copy-on-write snapshots, and so on. These devices enable whole new approaches to file systems. They are potential gap fillers between disks and RAM and they are interesting "hot data" storage devices in their own right.

IOPS Crisis

- Access density vs. Object density

- "$ / IOPS / GB"
  - SSD winner > 6.2 IOPS/GB > HDD winner

Case, Case, Case for FlashSSDs

- FlashSSD's message is still unclear in the market
  - Ken Salem, University of Waterloo
  - J. Hellerstein, Berkeley

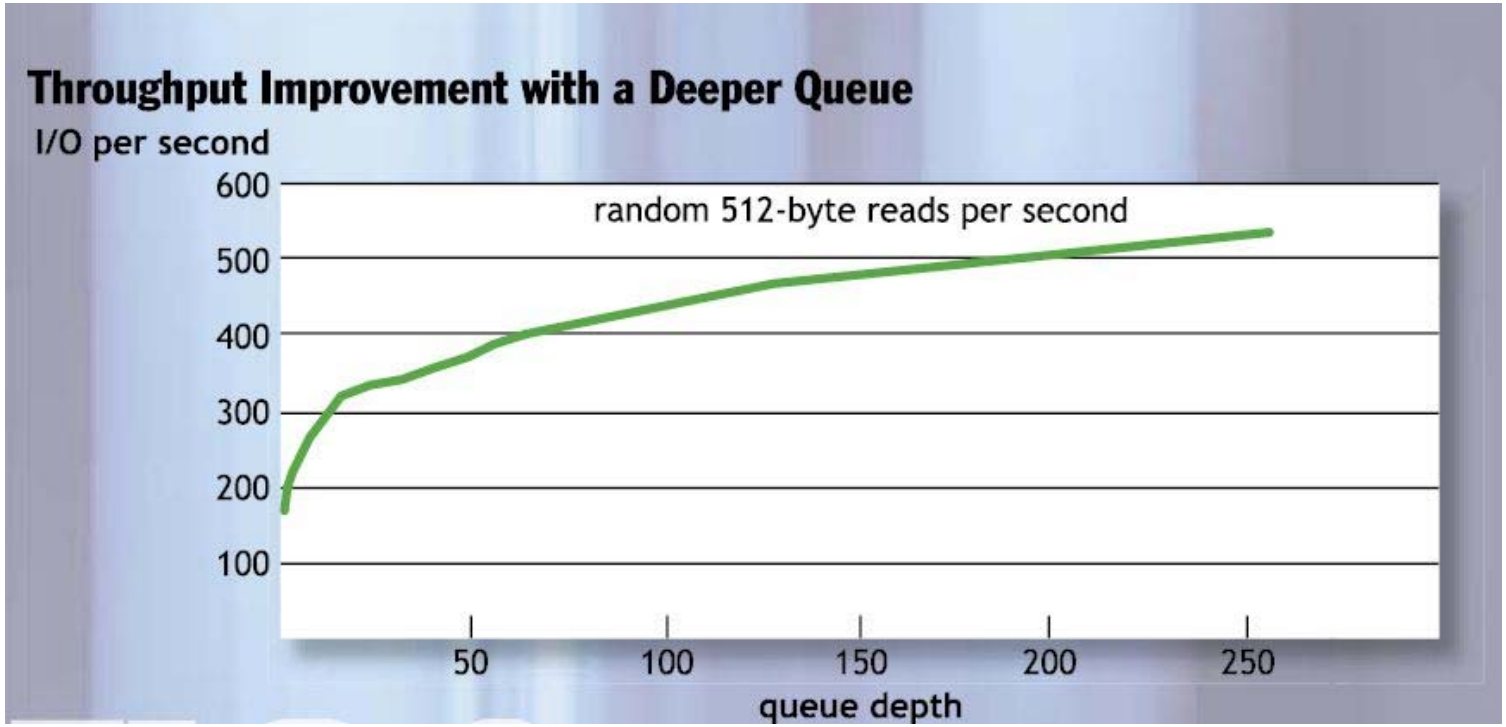
- It is urgent to develop "the case for flash memory SSD" (or killer applications) and "the right message"
- SIGMOD 08, SIGMOD 09
  - FlashLogging (SIGMOD 09, Intel)
  - Query Processing Techniques (SIGMOD 09, HP)

Ask not what flash can do for you,

Ask what you can do for flash

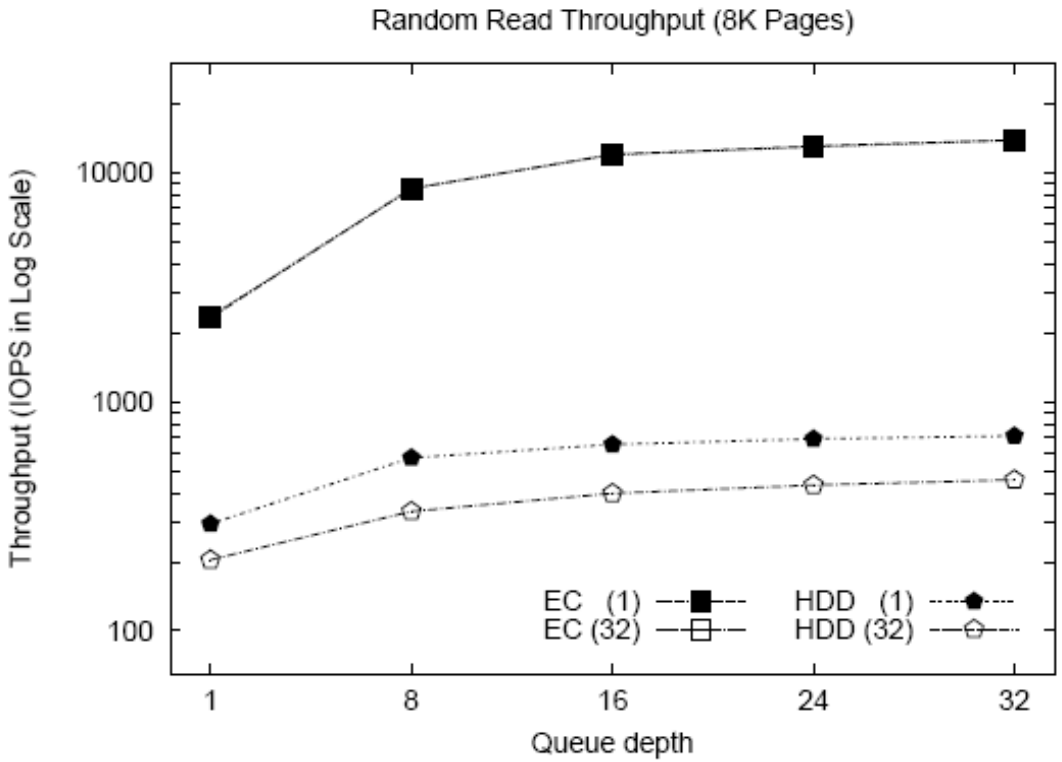
Scheduling from the past 30 Years

- Command queuing
  - Throughput improvement with deeper queue (source: You don't know jack about disks, ACM Queue 2003 June)

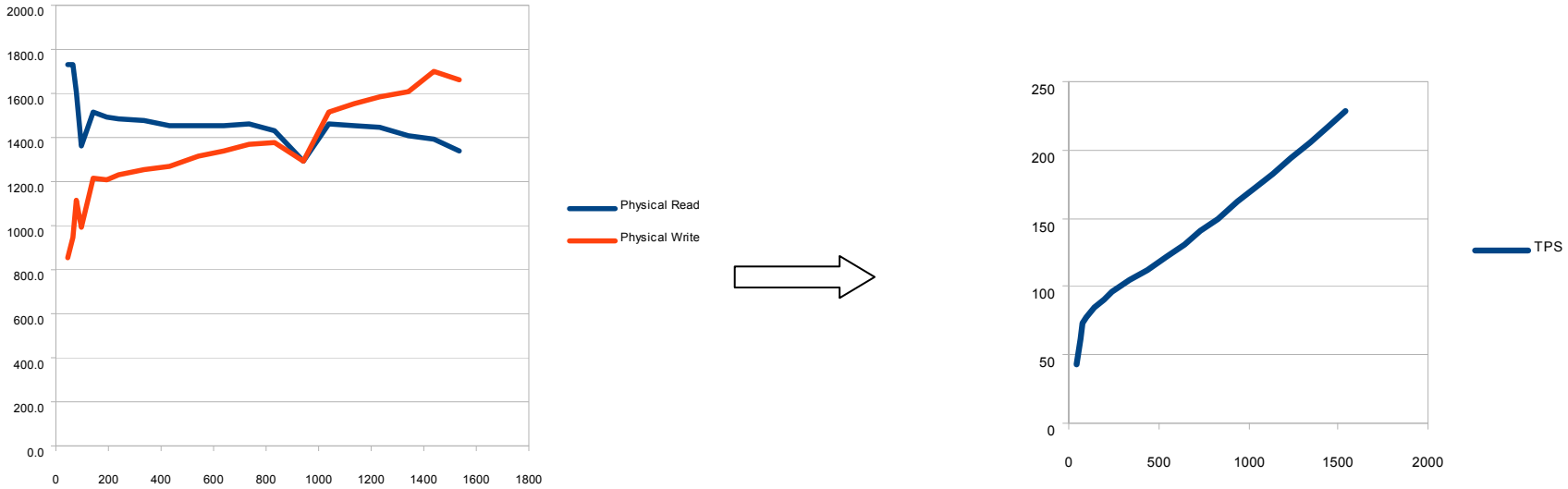


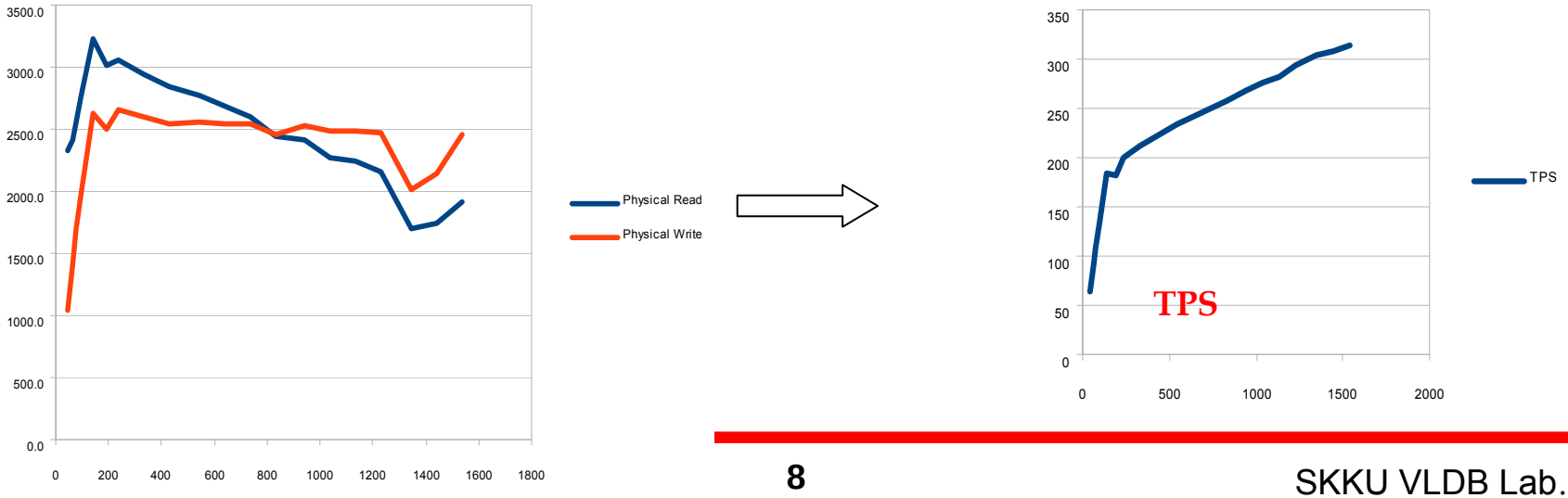
Scheduling for the coming 30 Years

- Random read vs. NCQ queue size

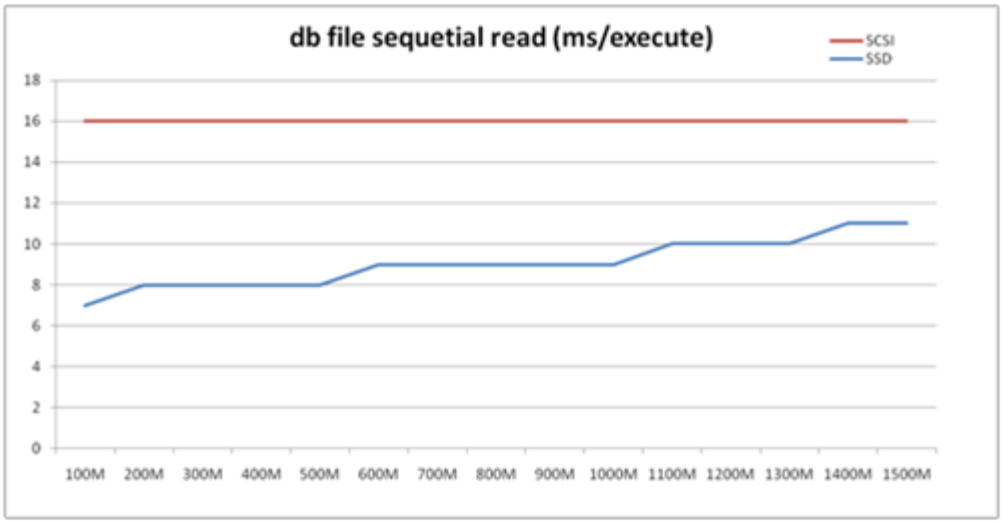


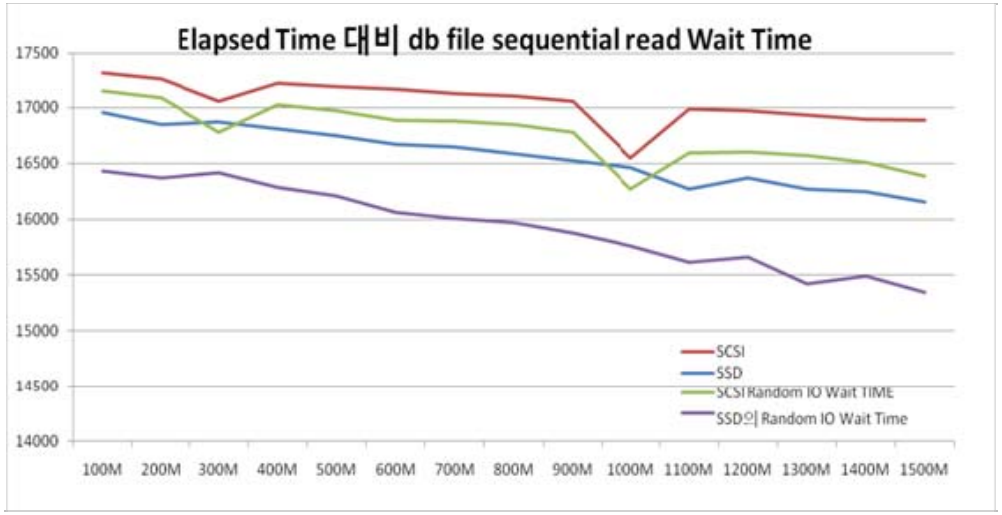
Beyond the CFLRU



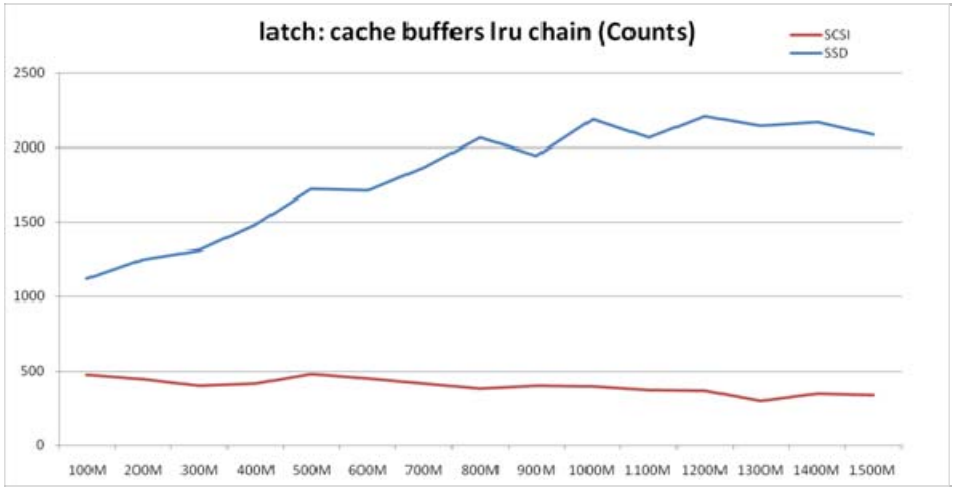


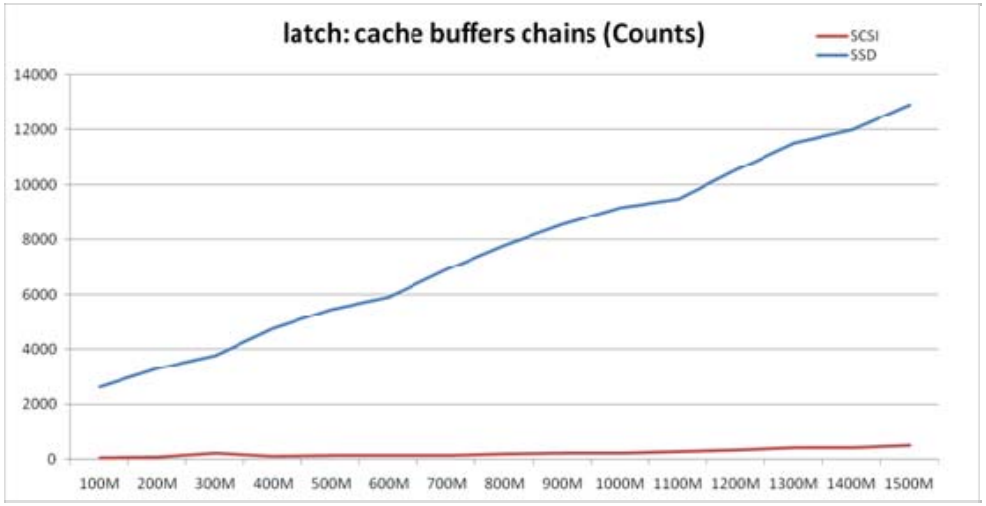
Buffer Latch Contention (?)





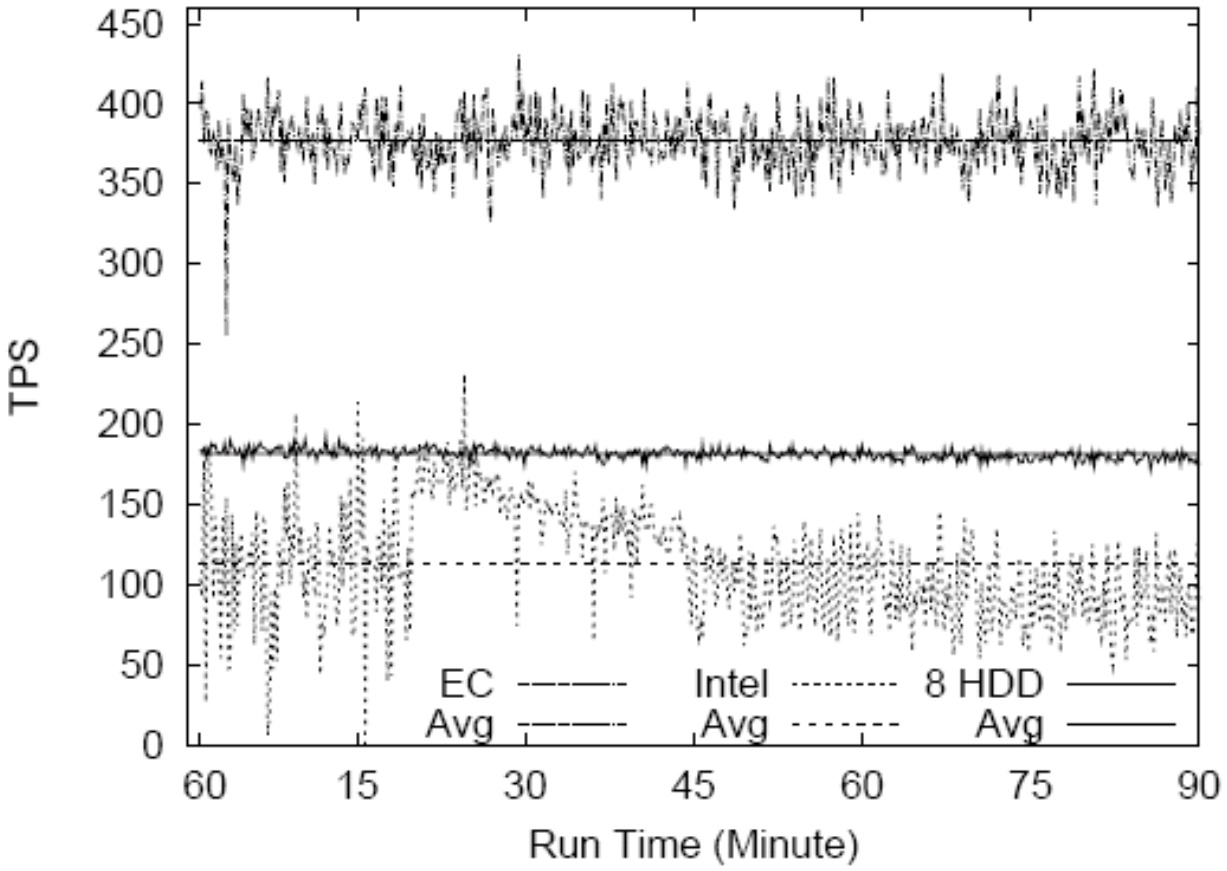
Buffer Latch Contention (?)





To page-map or not: This is the question

- TPS change over time



It’s NOT the capacity, NOT the bandwidth, Stupid!

- Flash page size vs. DB page size
  - IPL, Delta compression
  - Delta = bitwise XOR of old and new data
  - Physical delta vs. logical delta (e.g. IPL: timestamp, version)
  - 2K page size and fine-grain NOP are essential

Tucson Summit

- One SSD can beat Ten Harddisks
- “One Server + One SSD” can beat “Ten Server + 100 Harddisks”
- We are witnessing “the fittest survival in second storage”