Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006
Tape Is Dead
Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: \(~\$400 \rightarrow \sim \$80\)

- But: \sim 5..15 \text{ hours to read (sequential)}
- \sim 15..150 \text{ days to read (random)}

- Need to treat most of disk as Cold-storage archive
FLASH Storage?

• 1995 16 Mb NAND flash chips  
  2005 16 Gb NAND flash  
  Doubled each year since 1995

• Market driven by Phones, Cameras, iPod,…  
  Low entry-cost,  
  ~$30/chip → ~$3/chip

• 2012  1 Tb NAND flash  
  == 128 GB chip  
  == 1TB or 2TB “disk”  
  for ~$400  
  or 128GB disk for $40  
  or  32GB disk for $5

Samsung prediction
FLASH Some Parameters

5,000 IO/s per chip!

- Chip read ~ 20 MB/s
  write ~ 10 MB/s

  N chips have N x bandwidth

- Latency ~ 25 μs to start read,
  ~ 100 μs to read a “2K page”
  ~ 2,000 μs to erase
  ~ 200 μs to write a “2K page”

- Power ~ 1W for 8 chips and controller
What’s Wrong With FLASH?

• Expensive: $/GB
  – 50x more than disk today
  – Ratio may drop to 10x in 2012

• Limited lifetime
  – ~100k to 1M writes / page
  – requires “wear leveling”
    but, if you have 1B pages,
    then 15,000 years to “use” ½ the pages.

• Slow to write
  you can only write 0’s,
  so erase (set all 1) then write.
Obvious Uses For Flash

• PDAs, cameras, iPod, ....
• Laptop disks
  – power, rugged, quiet, big enough, ...

• Not so obvious use:
  – ARCHIVE for photo/music/.. because it’s simple to understand.
  – Enterprise drives (lots of IO/s per $ per watt per liter )
One Could Make a Flash Disk (or a Flash File System)

- 6K random reads/sec, 3K random writes/sec
- The IO capacity of 30..45 disks
- Uses 1 W vs 500W...
- Less space, ...

- See “A Design for High-Performance Flash Disks” Birrell, Isard, Thacker, Wobber

MSR-TR-2005-176
We Are Not There Yet

• Current FLASH disks could do much better on writes (100x better (!))
  Algorithms are known but…

• This changes many ratios
  Access time is 20x less (~200us)
  IOps is 100x more

• Re-evaluate page sizes MSR-TR-2006-168
  FlashDB: Dynamic Self-tuning Database for NAND Flash, Suman Nath, Aman Kansal
RAM Locality is King

- The cpu mostly waits for RAM
- Flash / Disk are 100,000 … 1,000,000 clocks away from cpu
- RAM is ~100 clocks away unless you have locality (cache).
- If you want 1CPI (clock per instruction) you have to have the data in cache (program cache is “easy”)
- This requires cache conscious data-structures and algorithms sequential (or predictable) access patterns
- Main Memory DB is going to be common.
Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006