Storage Systems

NPTEL Course
Jan 2012
(Lecture 12)

K. Gopinath
Indian Institute of Science

Disk Systems

- Direct Attached Storage (DAS)
- Network Attached Storage (NAS)
- Storage Area Network Storage (SAN)
- Distributed Storage

Disk

- Electromechanical component. Example disk (late 2010)
  - 5400 RPM, 12 ms seek avg, full stroke 20 ms
  - 375 Gb/sq in max
  - 875 Mb/s buffer to/from media; 3 Gb/s max
- Block based device
  - Typ sector size 512B, block size 4KB
  - Some newer devices 4KB sectors
    - Reduces ECC overhead and intersector gap overhead
  - (logical) Heads, Tracks, Cylinders: 16, 63, 16383.
    - Actually only 2 physical heads!
  - Non uniform number of sectors across disk. Eg.
    - Zone 0: 0-7153 cylinders: 1920 sectors (960KB)
    - Zone 11: 80850-91139 cylinders: 1440 sectors (720KB)
    - Zone 23: 170226-172675 cylinders: 912 sectors (456KB)

Tape?

- Only BW matters!
- Latency high (minutes!)
- Now used only for backup
- Controller cost very high

Flash/Disk/Tape

- Flash: semiconductor persistent memory
  - Capacity low, access time low, cost high
    - 50 microsecs for 4KB (say, 1 page)
    - 20,000 random 4KB transfers per sec (IOPs) for one channel
    - Easy to scale IOPs by adding more interfaces, planes,...
      - IOPs cheap (and very high IOPs reqd only in most demanding large appl)
      - But increases controller cost
    - Cost per page higher than cost per IOP
- Disk/tape: magnetic persistent memory
  - Capacity high, latency time high, cost low
    - 5-10ms for 4KB, most of it seek/rotational latency
    - Atmost 100-200 random 4KB transfers per sec (IOP limited)
    - Easy to scale capacity by adding more disks/tapes
      - Capacity cheap
    - Cost proportional to reqd IOPs

Tiering

- Memory, SSD, Disk, Tape, ...
  - Cache imp stuff and evict to slower layer
  - Migrate up (in stages?) if necessary
- How do we decide cost/benefit?
- Suppose we have written a new 4KB page. Either we keep it in mem, or after some time k, write to disk
  - Currently, 4GB DDR3 approx Rs 1200
  - 4KB approx Rs 1200/10^6 = Rs 1.2/10^3 (c_m)
  - Disk is Rs 5000/1TB for 100 IOPs/s (Rs 50/IOP/s, c_IOP/s)
    - For IOP/k secs, cost is Rs 50/k (c_IOP/k)
- Break even when 50/k=1.2/10^3 (k=c_IOP/c_m)
  - k approx 4*10^4 secs (approx 11 hours) for 4KB
  - 8KB approx 5.5 hours

Add eSSD

- Since we have to keep data in memory for long periods, need LARGE memory
  - But in mid 80's, interval 5 min (“5 minute rule”)
    - Disk $2,000 / IOP/s; RAM $5 / KB
    - 1 KB breakeven = 400 seconds
  - Smaller memory sufficed
- Now if we add eSSD to our current systems?
  - Rs 12,500 per 20,000 IOPs/s (Rs 0.62/IOP/s)
  - Breakeven betw DRAM/eSSD: 500 secs (approx 8 min)
  - Need much smaller amount of DRAM memory (8 min vs 5.5 hrs)
- Between eSSD and HDD
  - SSD Rs 50/GB, HDD Rs 50/IOP
  - Breakeven betw eSSD/HDD about 70 hours (for 4K) and 35h (8K)!
  - Can add one more layer!

Benchmarked Configs: SPECsfs2008

- Use Flash as cache (writethru)
- 64TB (224 FC drives) vs 16 TB (56 FC drives)

with flash cache

  - About the same perf (throughput ops/s vs response time ms)
  - Half the cost
  - 2/3 power savings
  - 2/3 space savings

Summary

- Tiering an important trend in storage systems
- Need to be careful wrt sizing