blkreplay: Experiences with Commercial vs OpenSource Storage Systems

LCA2013 presentation by Thomas Schöbel-Theuer
Agenda

- blkreplay **Features**
- Why Artificial Benchmarks suck
  - Example: random-sweep comparison
- blkreplay: Real-Life Performance
  - Example continued
- Pitfall: EMTPY vs FILLED
- Chances for OSS
Reproduction of both artificial and **natural loads** (block level)
- Positionly behaviour
- Timely behaviour
- IO parallelism

**Test suite for automation** of large benchmarking projects / stress-testing, etc
- extensible with plugins

**Large database** (>70GB) with natural loads from 1&1 datacenters on blkreplay.org
- contributions welcome!
Example 1a: random sweep on Linux SATA RAID-6

![Graph showing throughput over duration](image)

- thrp.demand (avg=722.482635, max=1435.000000)
- thrp.actual (avg=572.896131, max=780.000000)
Example 1b: random sweep on Commercial Box

comm1_empty-random.g000.overview.thrp.actual

- thrp.demand (avg=2994.995822, max=5980.000000)
- thrp.actual (avg=2368.627215, max=4777.333000)
Artificial random IO can be extremely different from real life

Alternative: use blkreplay.org

- Record your real application behaviour with blktrace

- Or, use a published real-life load from blkreplay.org

- Exactly replay your original timely and positionly behaviour, degree of IO parallelism, etc

- Don't use AIO [bottleneck, distortions from page cache]

- Use processes / threads

Okay, does it make a difference?
=> next slides

25 VMs (XenServer) in parallel, iSCSI over 10GbEth
Example 2a: real-life load on Linux SATA RAID-6
Example 2b: real-life load on EMPTY Commercial Box

comm1_empty.g000.overview.thrp.actual

thrpdemand (avg=13574.519863, max=58379.333000)

thrp.actual (avg=5456.227771, max=14927.000000)
Commercial black-boxes / SSDs / etc often implement **Storage Virtualization**

Translation from **logical block addresses** to **physical block addresses**

Problem: benchmarks touch only a **tiny fraction**!

**Pitfall: Filled vs Empty Logical Volumes**

- *(sparse) logical address space*

- 50 TB

- 1 GB

- Solution: pre-fill the whole LV with random data
Example 2c: real-life load on FILLED Commercial Box

```
comm1_filled.g000.overview.thrp.actual

thrp.demand  (avg=16613.132879, max=58379.333000)
thrp.actual   (avg=1740.991119, max=8086.000000)
```

Throughput [IoS/sec] vs Duration [sec]
Mass Data: > 1 PB total
   → price/TB matters

Admins know what they are doing

Management often believes sales personnel from commercial storage vendors
   → find out the TRUTH
      prejudices can be HARD

   → evaluation projects

   ✅ Automated by the blkreplay test suite

   → convince your management that OSS can
do often better & cheaper
Never trust any claim / benchmark from sales!

Always check yourself, e.g. with natural loads from blkreplay.org

OSS Performance often better

OSS Price / Performance even more often better