Agenda

- blkreplay Features
- Why Artificial Benchmarks suck
  - Example: random-sweep comparison
- blkreplay: Real-Life Performance
  - Example continued
- Pitfall: EMPTY vs FILLED
- Chances for OSS
**blkreplay Features**

- Reproduction (at block level) of both
  - Artificial Loads
  - **Natural Loads**

- Reproduces
  - Timely behaviour
  - Positionly behaviour
  - IO parallelism
  - In future: compressibility of data

- **Test suite** for automation of large benchmarking projects, stress tests, etc
  - Extensible with plugins

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

```
sata_raid6-random.g000.overview.thrp.actual
```

Throughput [I/Os/sec]

```
thrp.demand (avg=722.482635, max=1435.000000)
thrp.actual (avg=572.896131, max=780.000000)
```

Duration [sec]
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)
Who is *really* the winner?

- Artificial random IO can be **extremely** different from real life

- **Alternative:** use [blkreplay.org](http://blkreplay.org)
  - record your real application behaviour with **blktrace**
  - or, use a published real-life load from [blkreplay.org](http://blkreplay.org)
  - exactly replay your original timely and positionally behaviour in the lab

- **Avoid AIO** [bottleneck, distortions from page cache]
  - use processes / threads

- **Does artificial ↔ natural make a difference?**
  => next slides

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

![Graph showing throughput vs. duration](image)

- **thrp.demand**: (avg = 16613.132879, max = 58379.333000)
- **thrp.actual**: (avg = 2146.161297, max = 14504.000000)
Example 2b: real-life load on EMPTY Commercial Box

```
comm1_empty.g000.overview.thrp.actual
```

- **thrp.demand** (avg=13574.519863, max=58379.333000)
- **thrp.actual** (avg=5456.227771, max=14927.000000)

Throughput [I/Os/sec]

Duration [sec]
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

physical address space

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

![Graph showing throughput vs. duration for comm1_filled_g000Overview.thrp.actual](image)

- **thrp.demand** (avg=16613.132879, max=58379.333000)
- **thrp.actual** (avg=1740.991119, max=8086.000000)
Chances for OSS Stacks

- 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 often do better & cheaper
Conclusions

- 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