blkreplay: Experiences with Commercial vs OpenSource Storage Systems

LinuxTAG 2013 presentation by Thomas Schöbel-Theuer

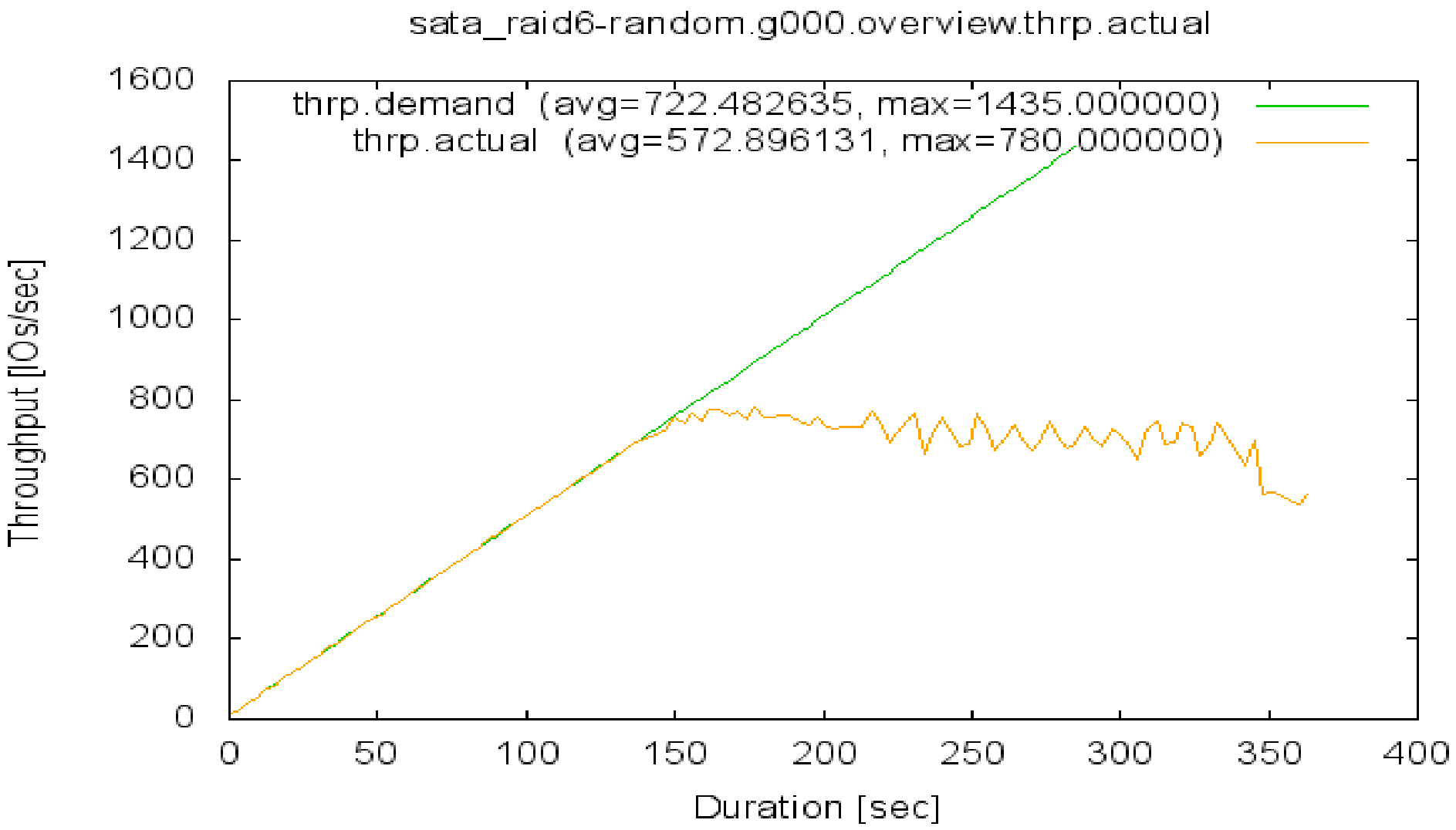
Agenda

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

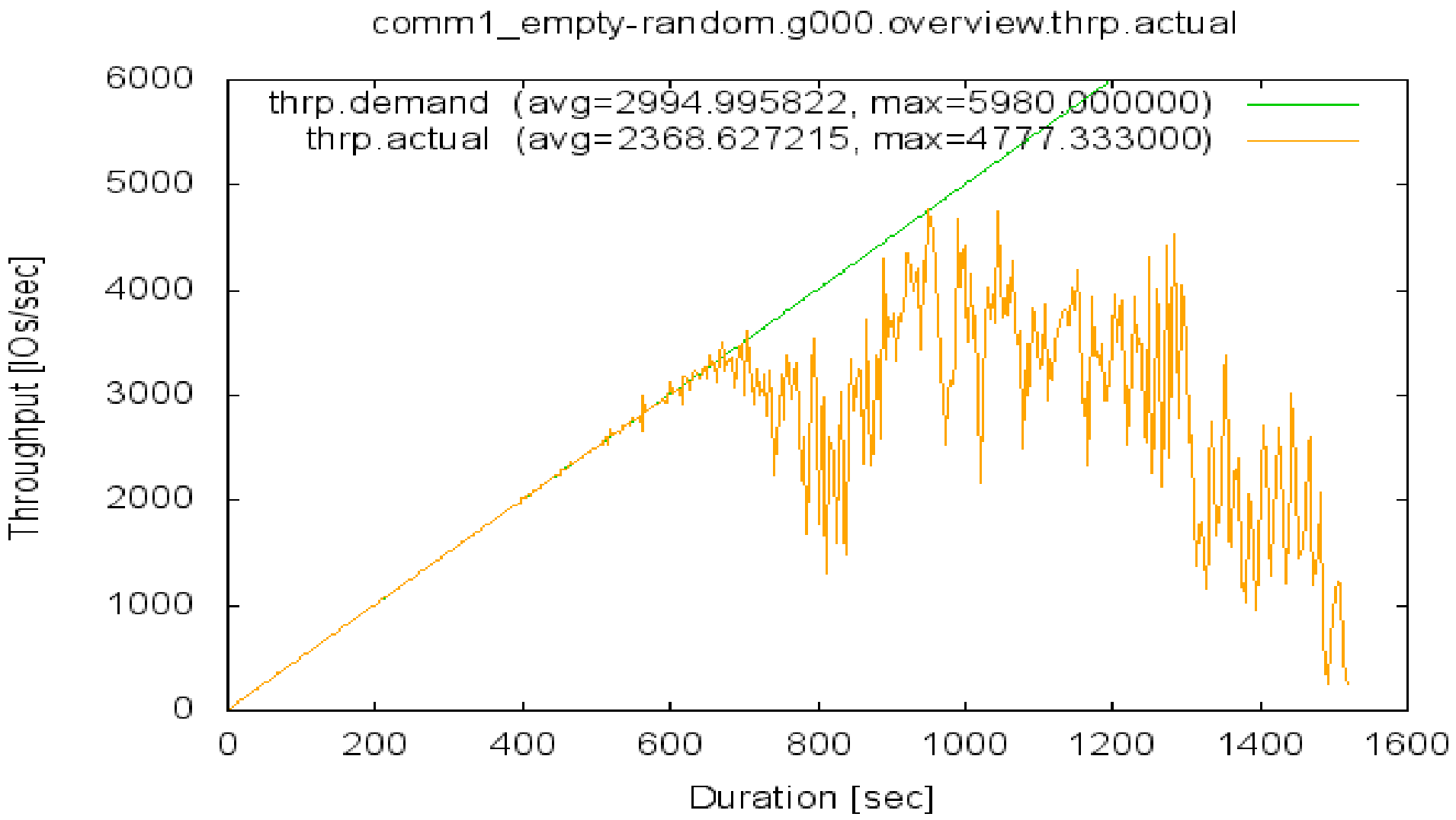
blkreplay Features

- 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



Example 1b: random sweep on Commercial Box



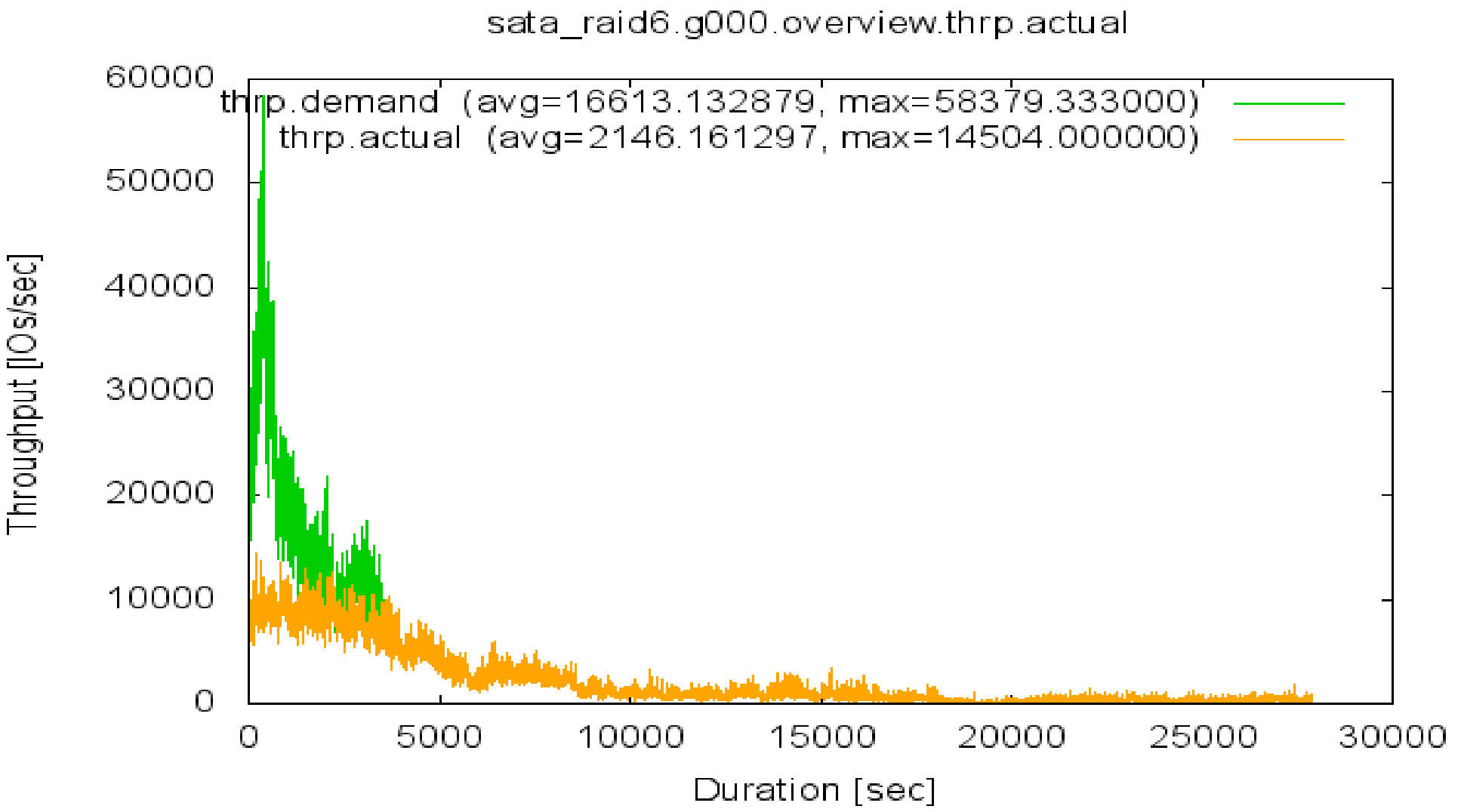
Who is *really* the winner?

- 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 in the lab
- Avoid AIO [bottleneck, distortions from page cache]
  - use processes / threads
- Does artificial↔natural make a difference?

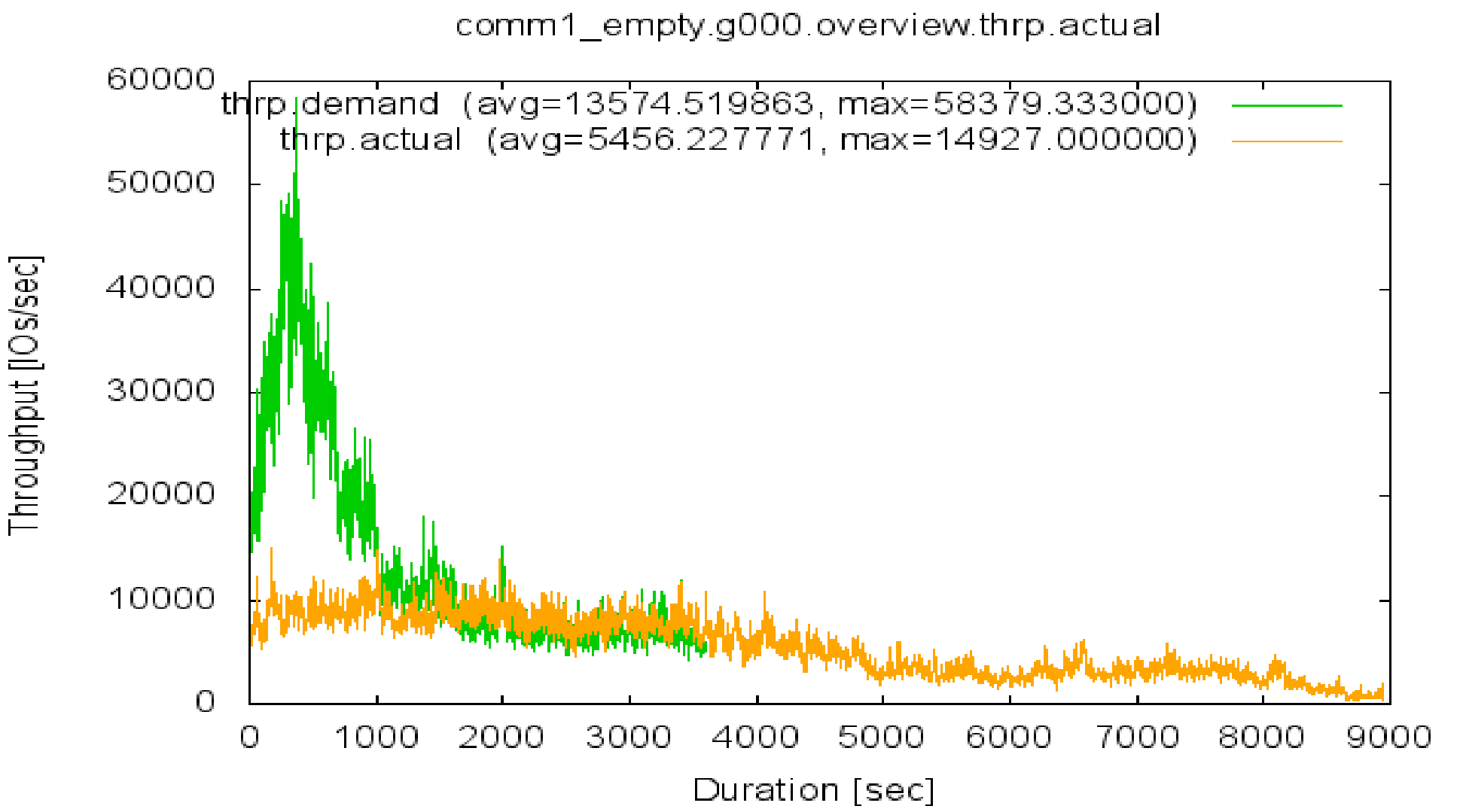
  => next slides

25 VMs (XenServer) in parallel, iSCSI over 10GbEth

Example 2a: real-life load on Linux SATA RAID-6

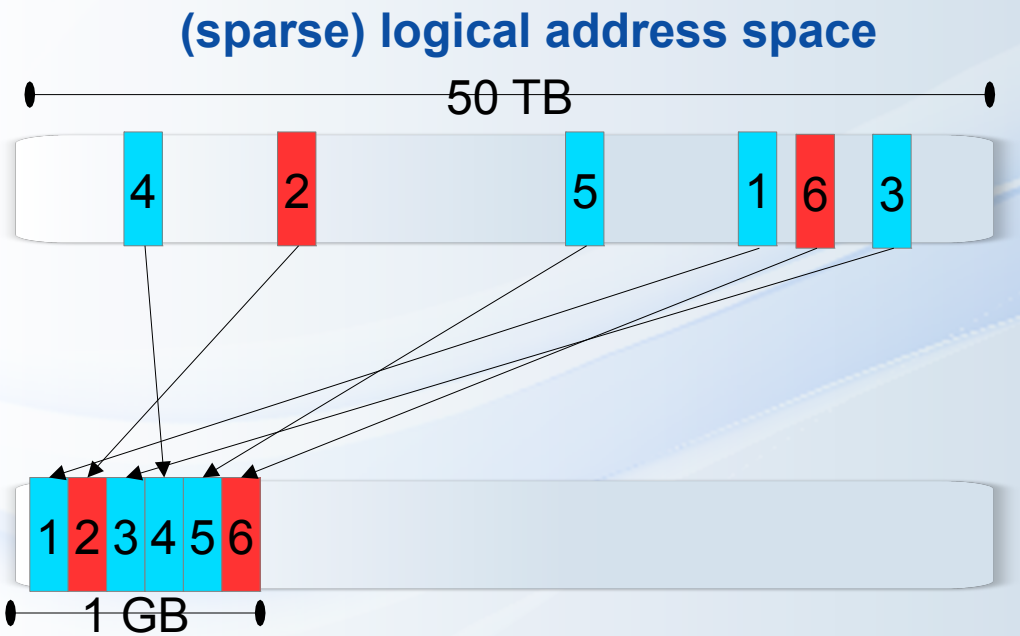


Example 2b: real-life load on EMPTY Commercial Box



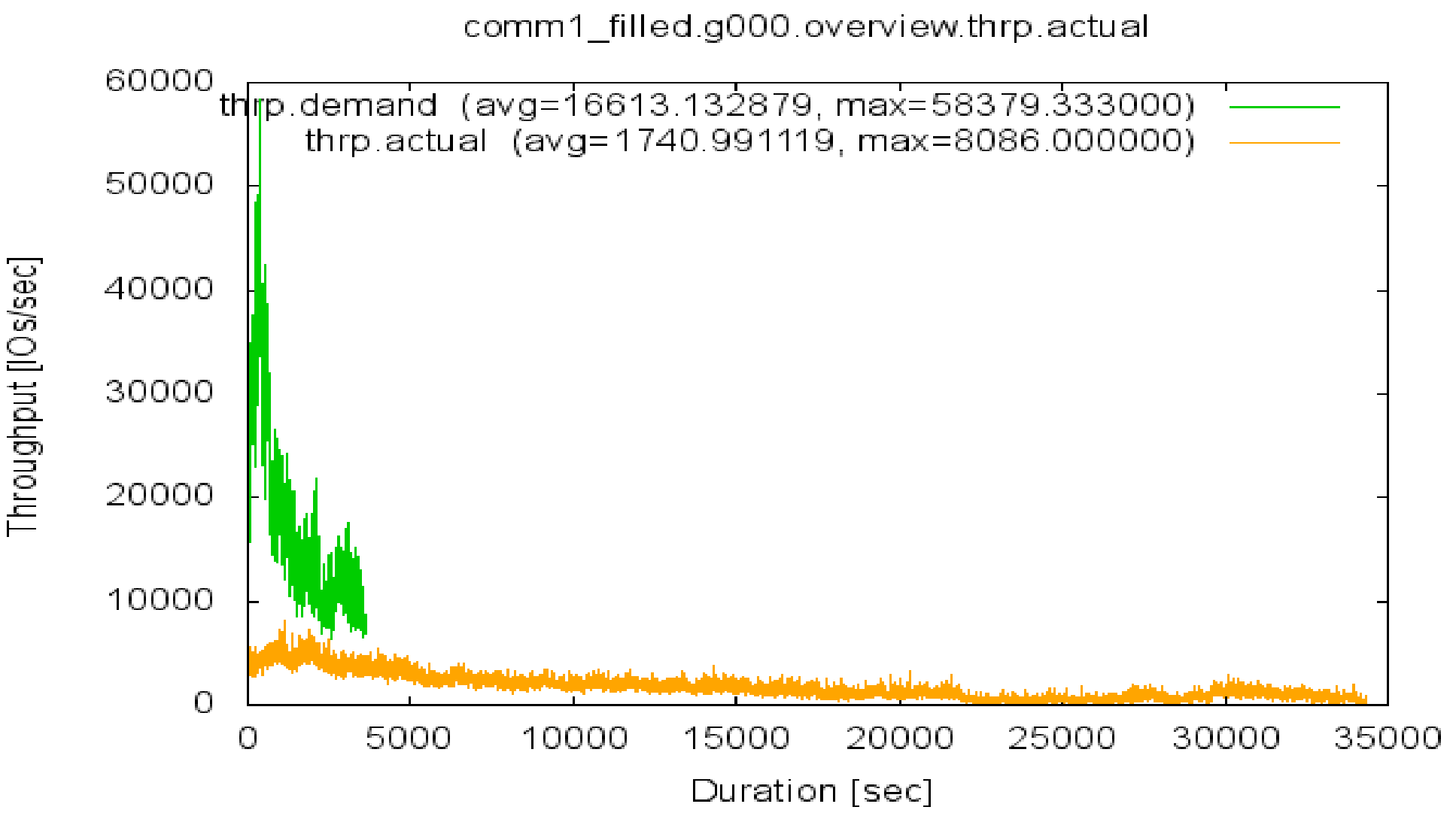
Pitfall: Filled vs Empty Logical Volumes

- 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!**



Solution: pre-fill the whole LV with random data

Example 2c: real-life load on FILLED Commercial Box



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