Hardware evaluation 2012

Jiří Horký

horky@fzu.cz

25/4/2012

Outline

Evaluations on borrowed hardware

- scope of tests limited

- Deduplication solution Fujitsu CS800 S2
- Disk performance scaling on Dell C6145
- Intel Sandy Bridge performance
  - already covered in details by previous speakers!
  - addition: performance/Watt numbers almost 50% better than the previous generation of Intel processors

Deduplication

- Big hype about this technology
  - deduplication factors of 1:20+ often advertised
  - some marketing materials even claiming it as a general purpose storage technology
- Performance very data dependent
  - but how much?

Deduplication - Hardware

- System - Fujitsu Eternus CS800 S2
  - 1x IntelE5620 @ 2.40GHz
  - 24GB RAM
  - 10 x 500GB 7200RPM 2.5” SAS drives in RAID6
  - LSI MegaSAS 9260 (rev 05)

- Software:
  - RPM based Linux with restricted shell access
  - Running Quantum software internally (StoreNext filesystem)

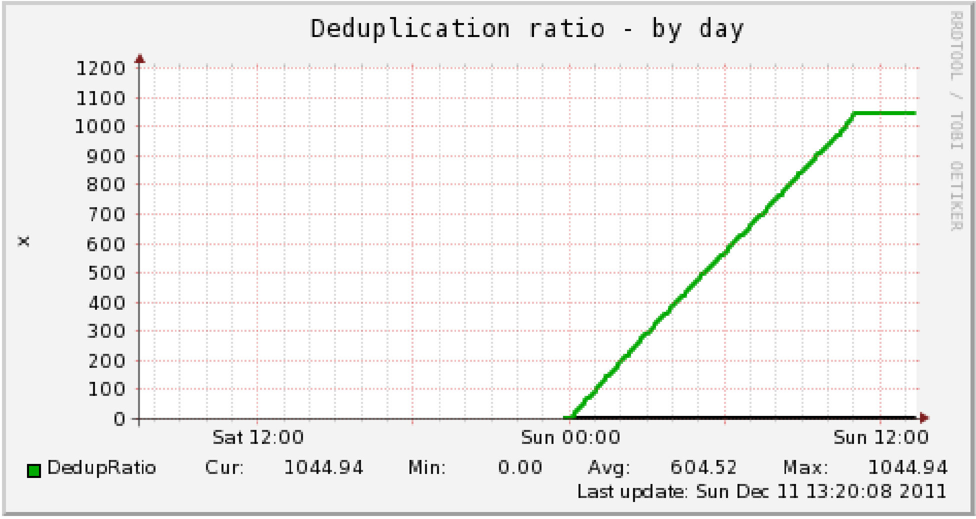
Deduplication – test methodology

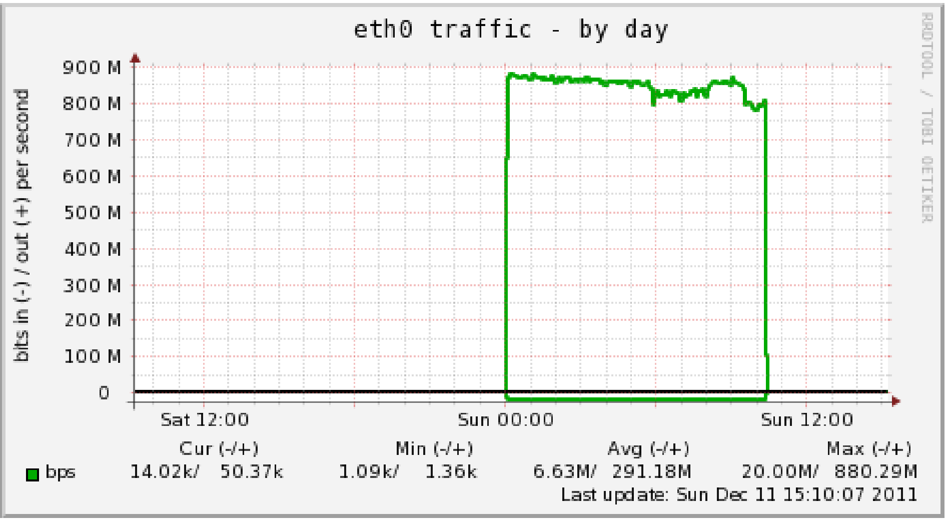
- /dev/zero
  - 1GB files, 3.8TB in total, pregenerated
- /dev/urandom
  - 1GB files, 3.8TB in total, pregenerated
- ATLAS data sets
  - 10MB-1GB .root files, 4TB in total
- regular backups
  - several full + incremental backups of SQL servers, /home directories, Linux servers..., 2.8TB in total
- virtual machines snapshots
  - daily snapshots of virtual servers, 800GB in total
  - 7x 115GB image file

- NFSv3 used to access the storage

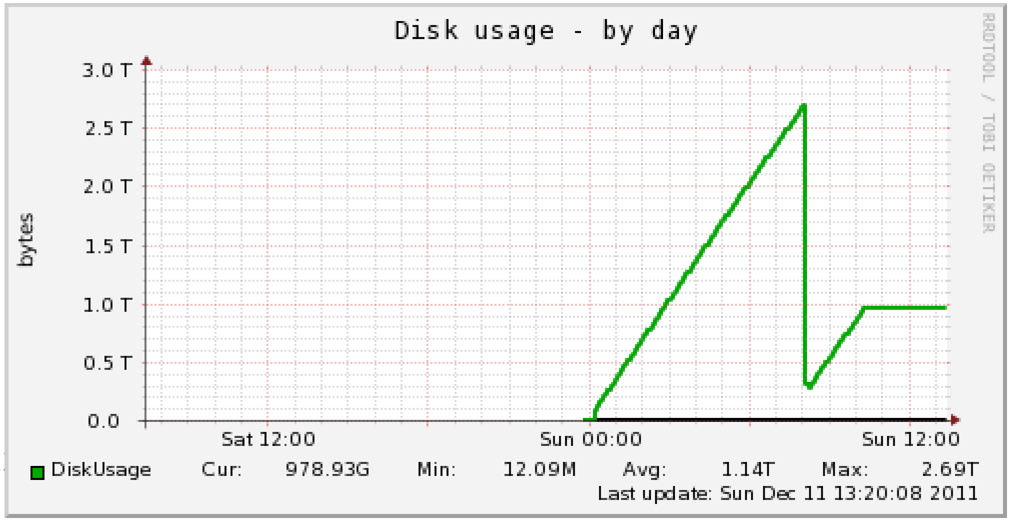
Deduplication – results I

- all zeros (/dev/zero) - “best case”



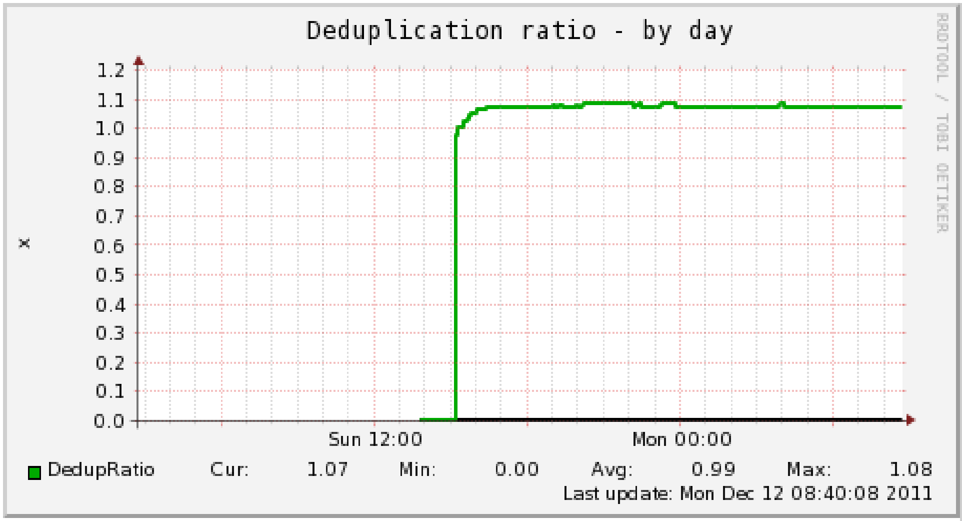


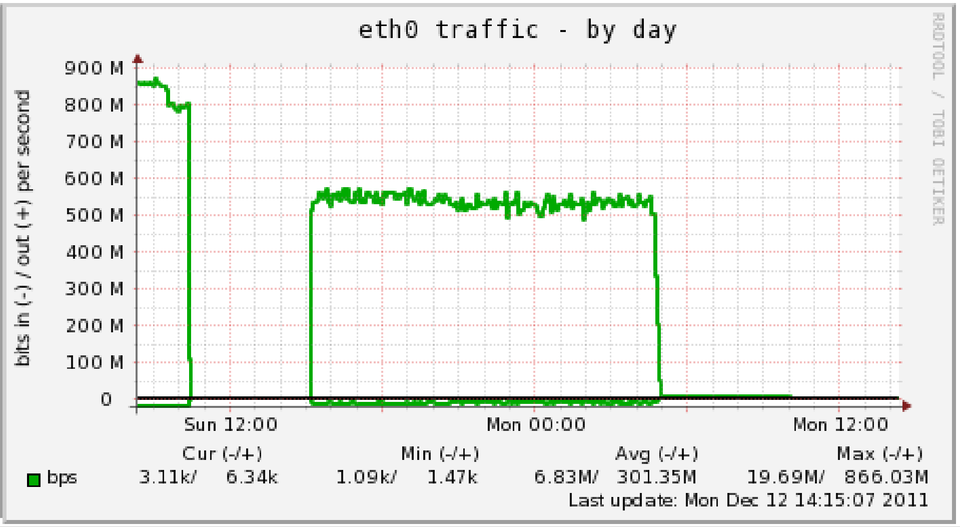
- avg. load around 10
- network bound
- dedup ratio 1:**1045**



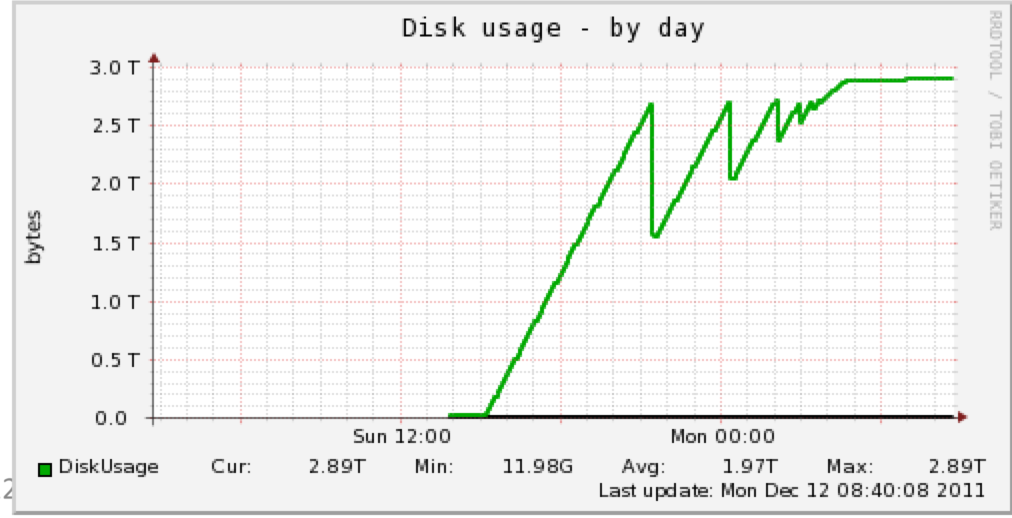
Deduplication – results II

- random data (/dev/urandom) - “worst case”



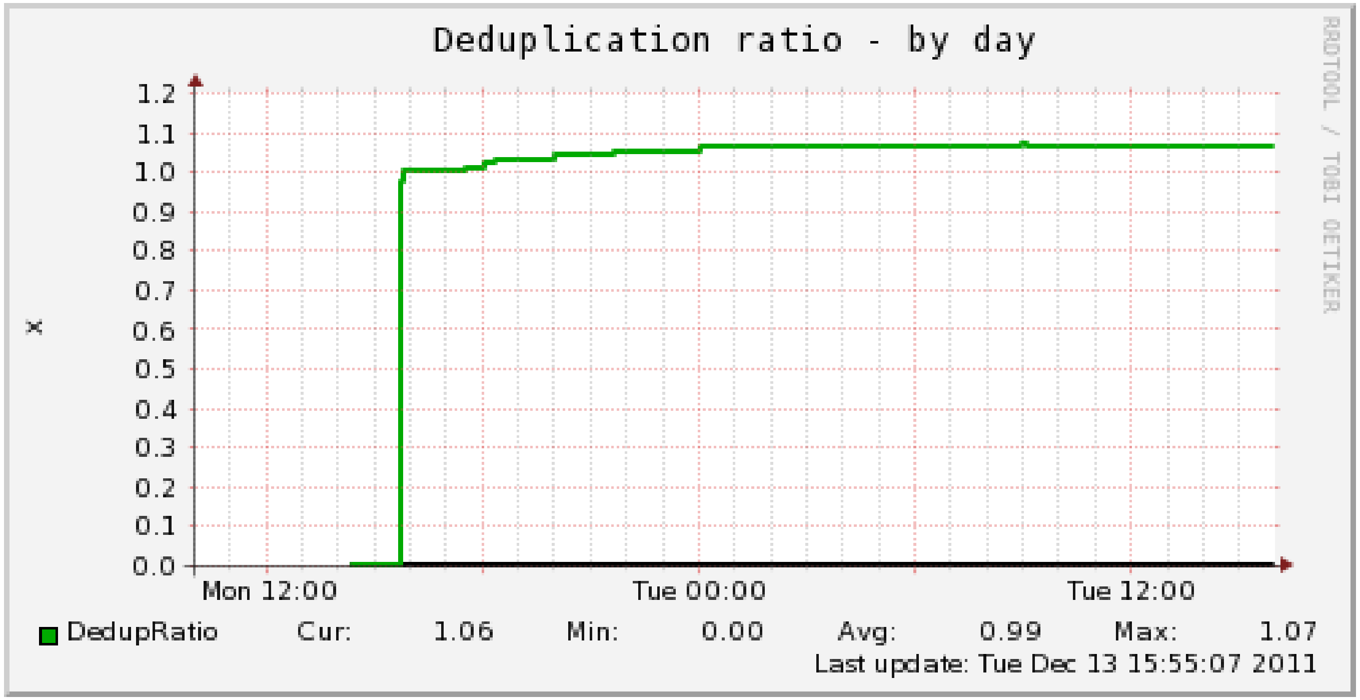


- avg. load around 25
- CPU bound
- dedup ratio 1:**1.08**



Deduplication – results III

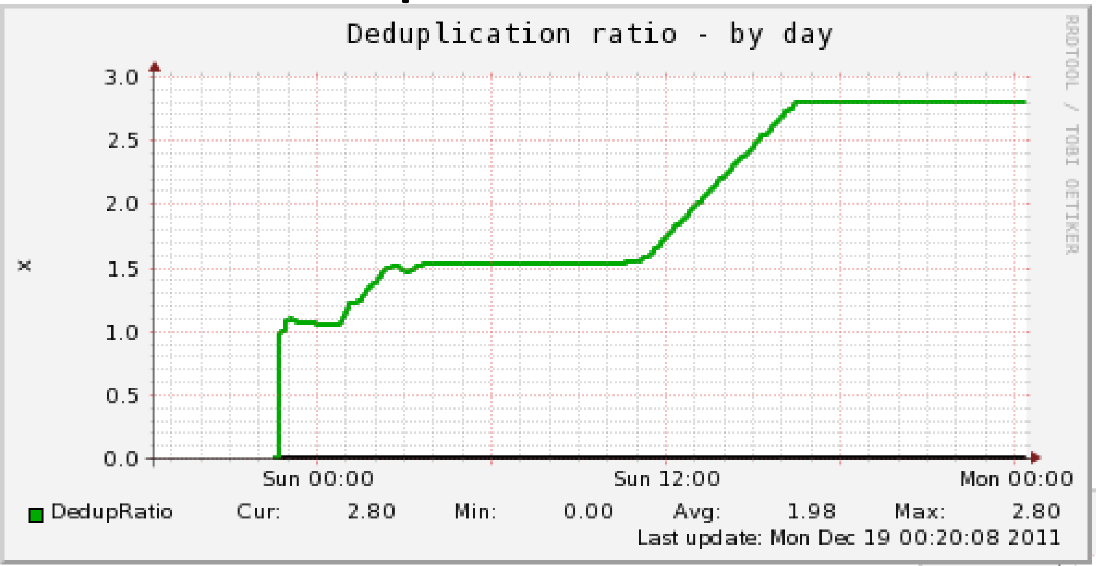
- ATLAS data sets



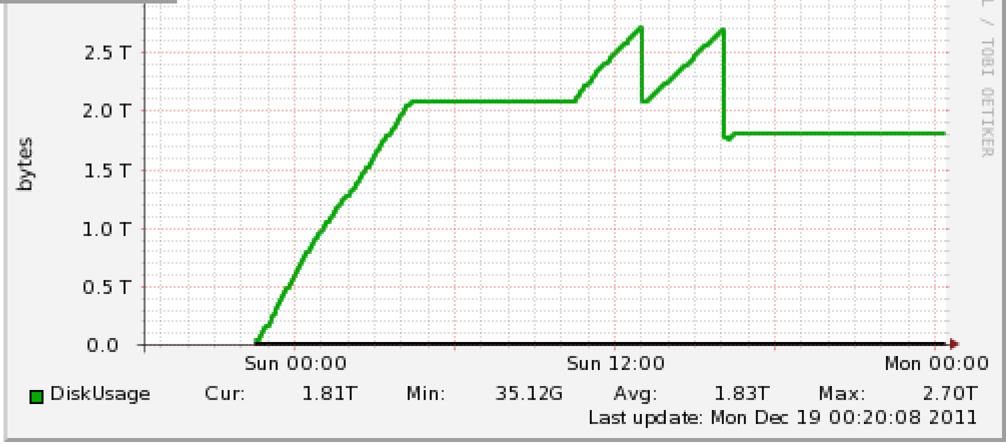
- dedup ratio 1:**1.07**

Deduplication – results IV

- Backup – real data from Networker

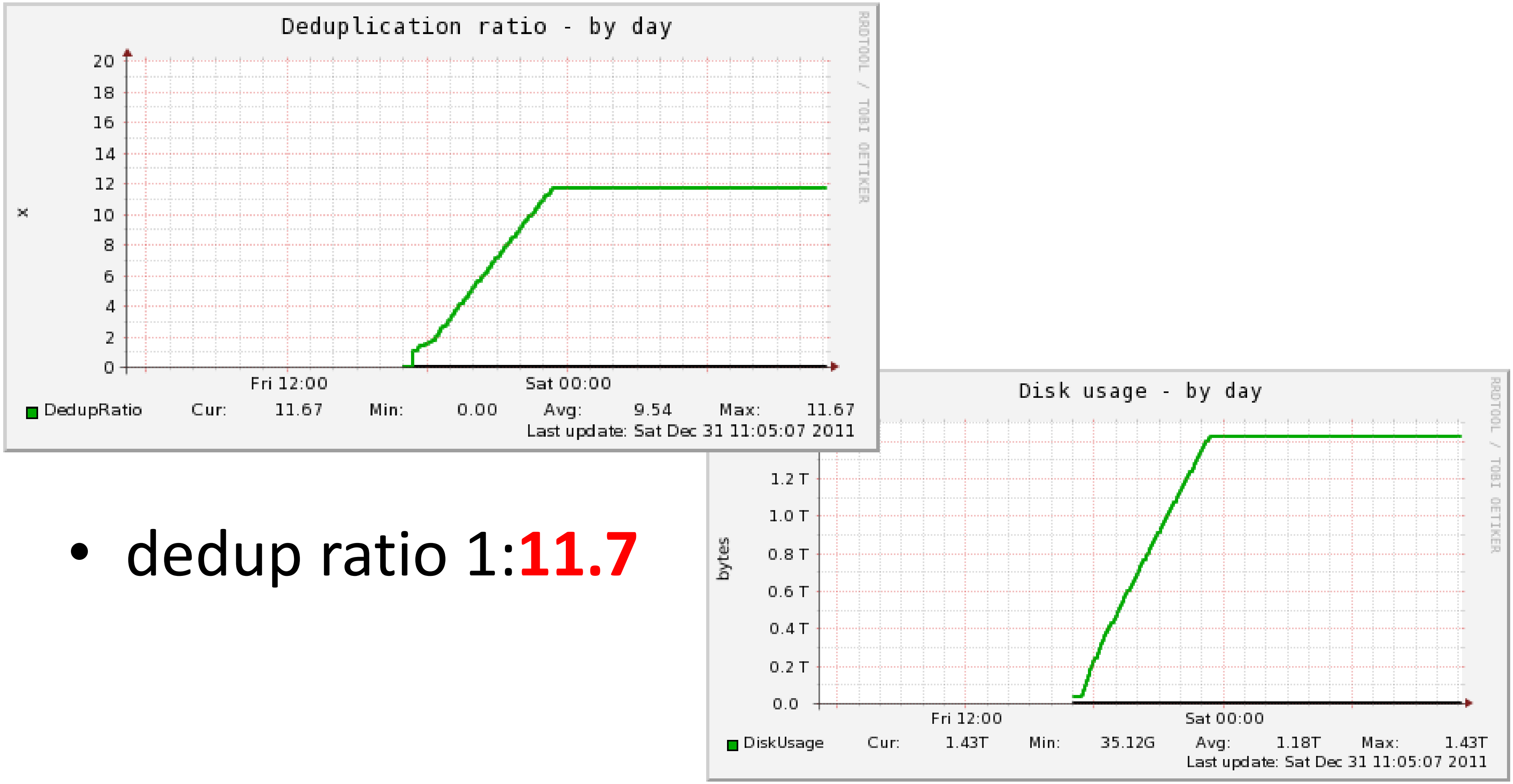


- dedup ratio 1:**2.8**



Deduplication – results V

- Backup – snapshots of virtual machines



- dedup ratio 1:**11.7**

Deduplication – conclusion

- Testing on your real data is a must!
- Performance could be better
  - maybe just an entry level configuration?
- Hardly suitable for anything else than backups?

Disk performance with 64 cores

How many disks are needed for 64 core worker nodes?

- – especially for analysis

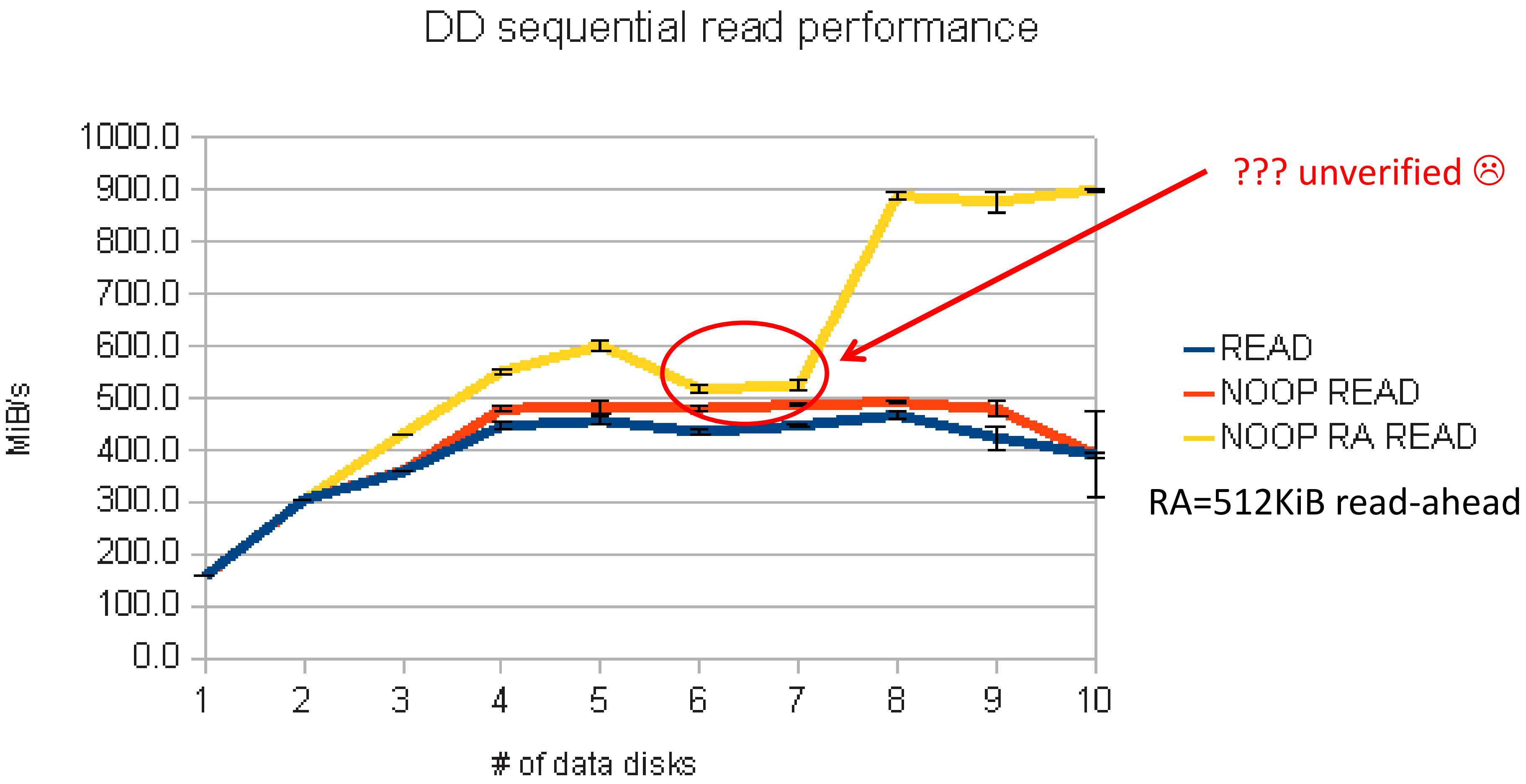
- Hardware:
  - – Dell C6145
  - – 2x AMD Opteron 6276 (64 cores)
  - – 128GB RAM
  - – LSI SAS2008 PCI-Express Fusion-MPT SAS-2
  - – 1-10 SAS 2.5” 300GB 10kRPM drives in RAID0

Methodology

- Software:
  - Scientific Linux 5.7, kernel 2.6.18-274.17.1.el5
  - XFS filesystem
  - IO Scheduler: CFQ (default) or NOOP (marked **NOOP** in plots)
  - Kernel read ahead: 128KiB (default) or 512 KiB (marked **RA** in plots)
- Test cases
  - sequential read/write using dd (bs=512K)
  - simulation of 64 parallel ATLAS analysis job with "new" ordered data files using IOreplay:

  http://code.google.com/p/ioapps/

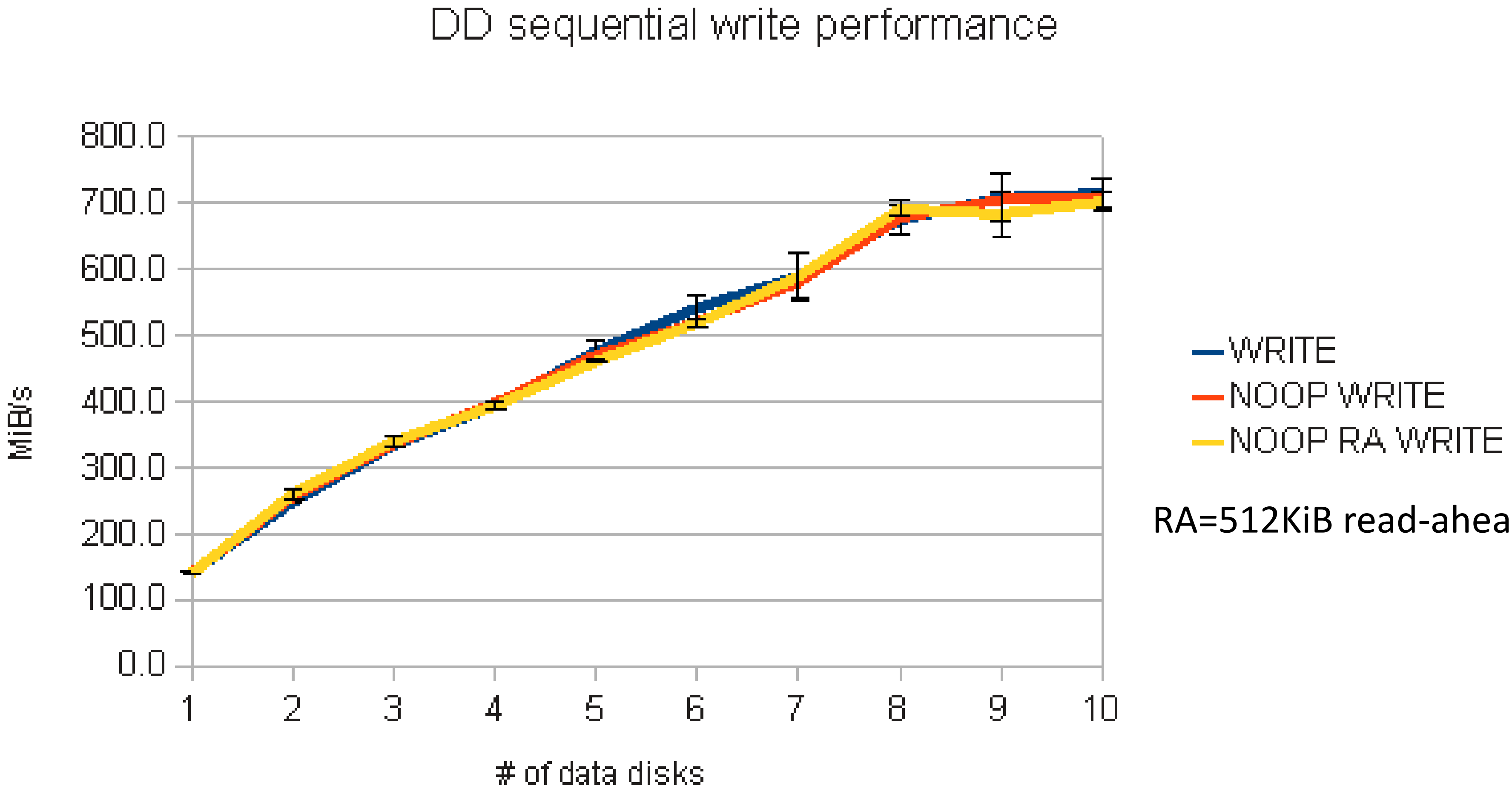
Results – dd sequential read



??? unverified ☹

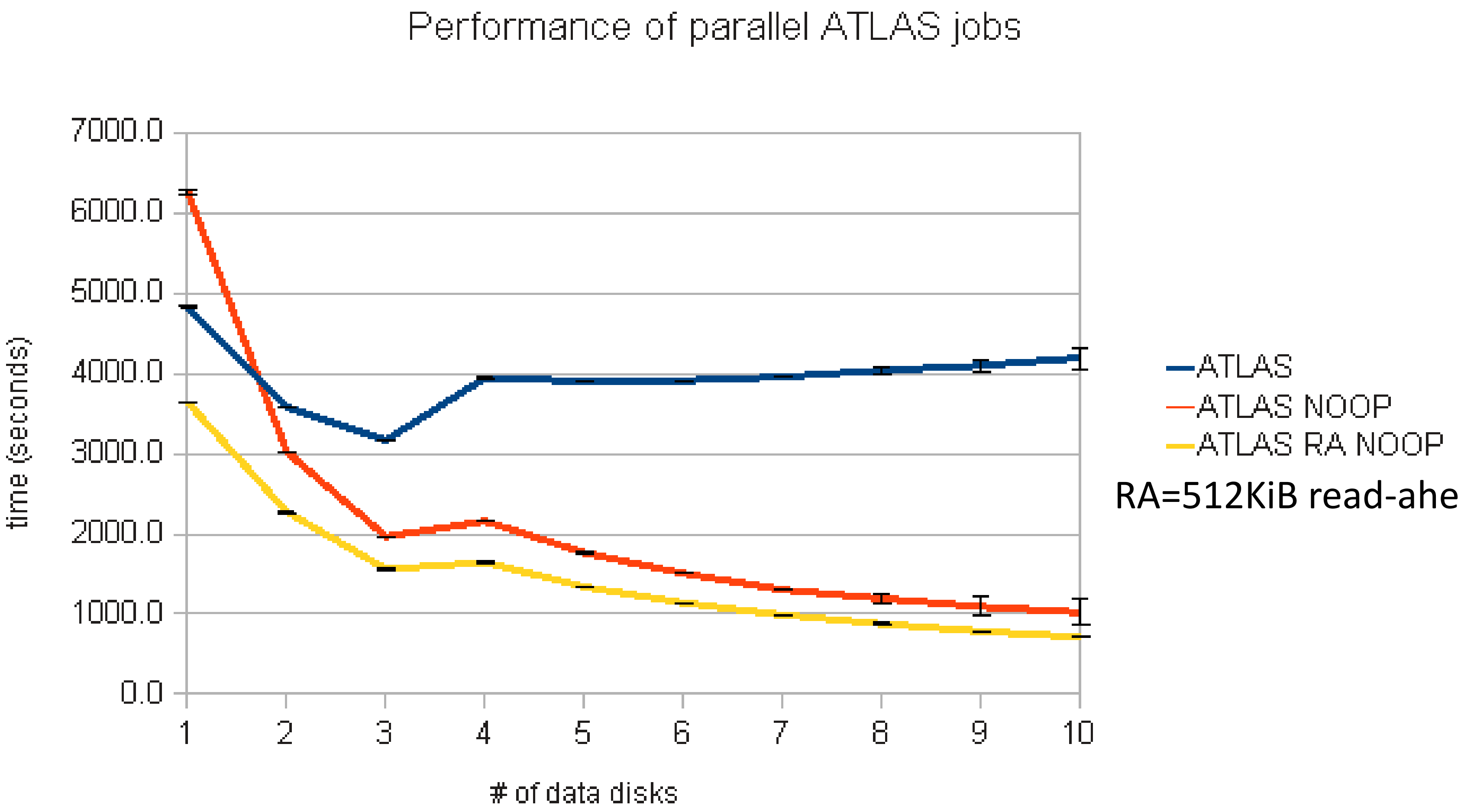
RA=512KiB read-ahead

Results – dd sequential write



RA=512KiB read-ahead

Results – parallel ATLAS jobs



RA=512KiB read-ahead

Conclusion

- IO scheduler settings can have BIG impact on IO performance
  - – RAID controllers may have much better knowledge on how to schedule the load
  - – but there is no clear winner for all IO patterns!
- Adding more than 8 drives does not help
- 3-4 drives should be sufficient
- What will we do with 128 cores machines?
  - – RAID 0 with 8 drives is probably not the right option

Thank you for your attention.

Questions?