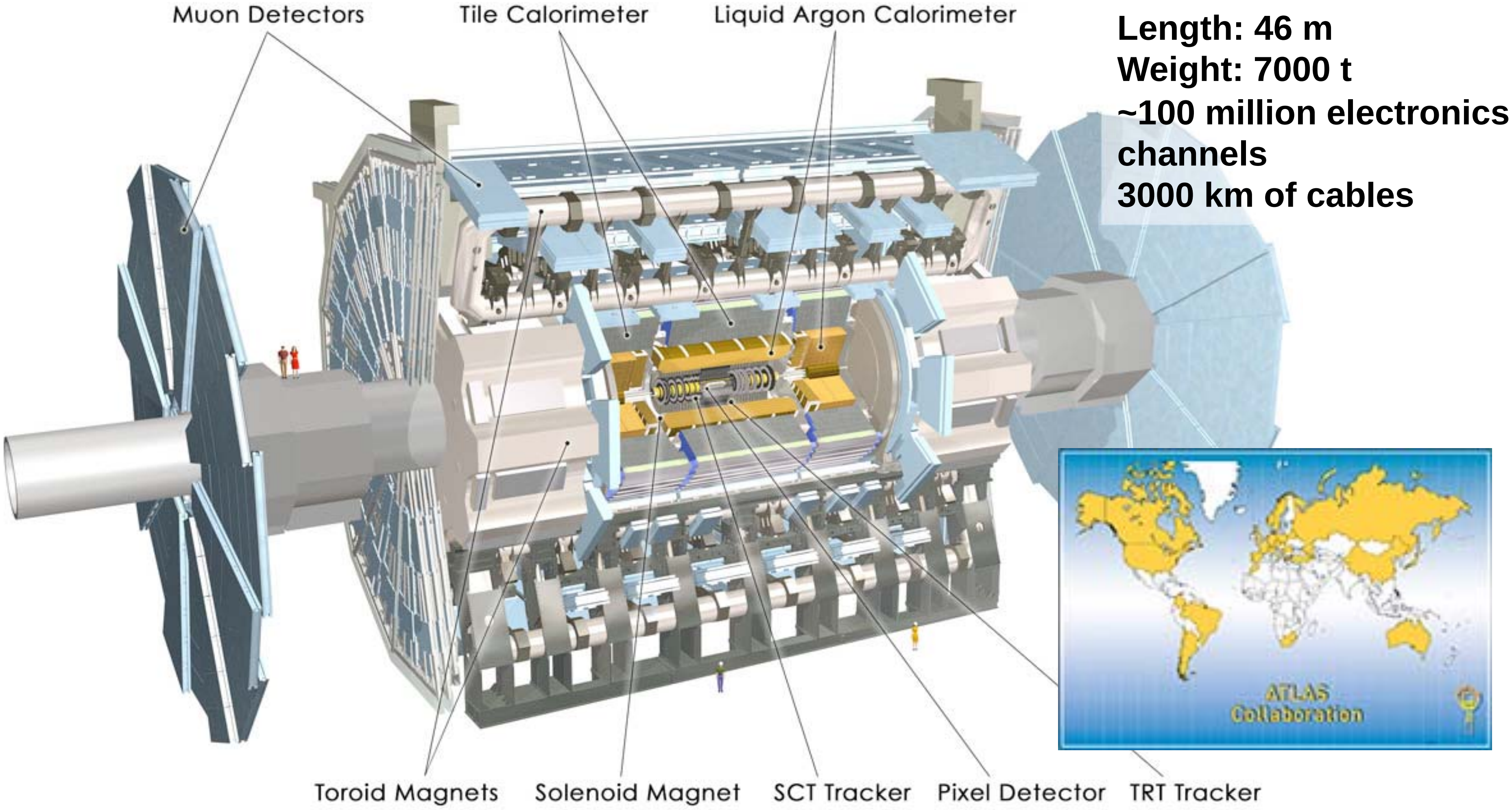
Disk Pool Manager experience and performance

Szymon Gadomski, HPC-CH, October 2010

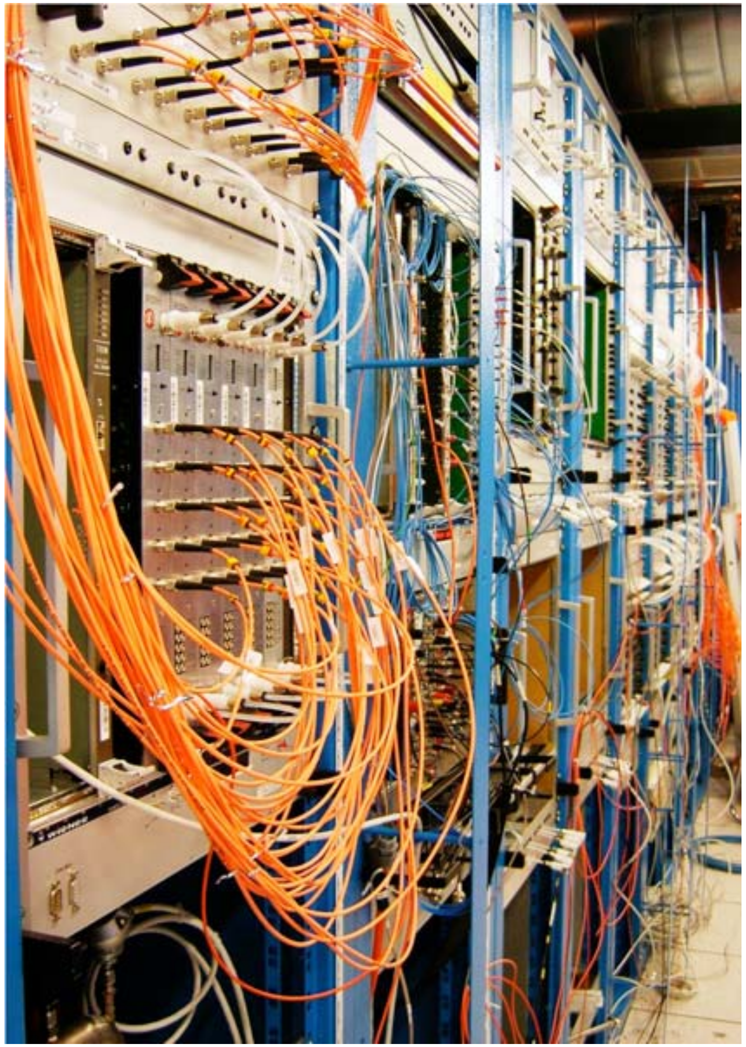
- our challenge in computing
- the Disk Pool Manager
- measured performance
- real life experience

The ATLAS detector

Diameter: 25 m
Length: 46 m
Weight: 7000 t
~100 million electronics channels
3000 km of cables



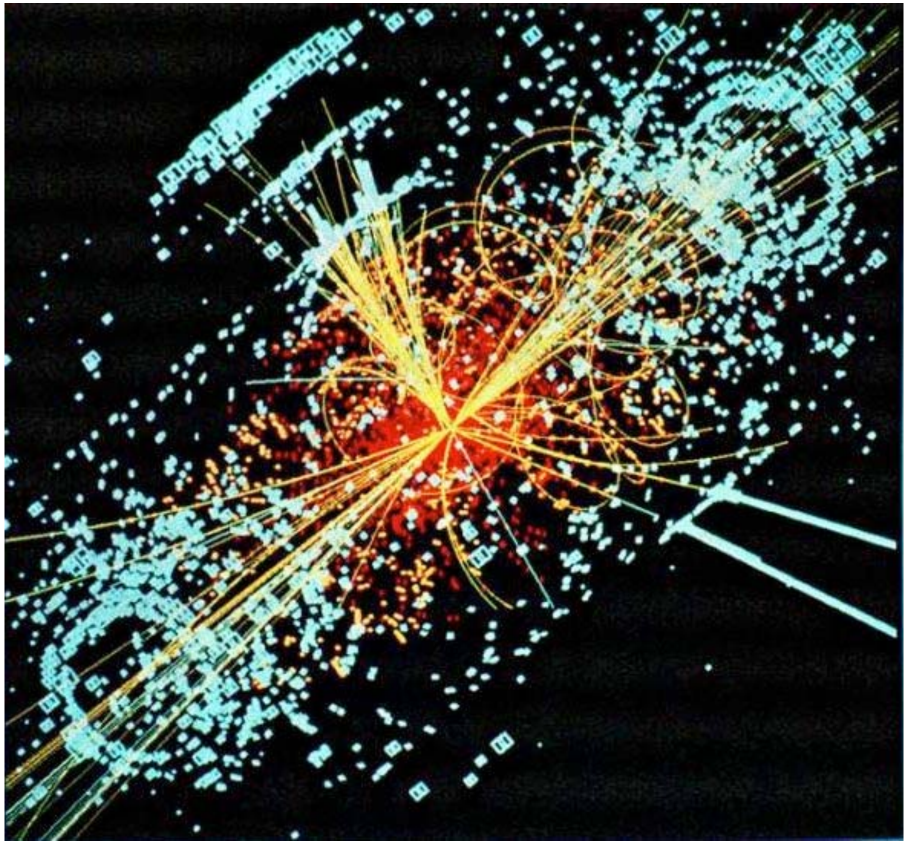
Online selection of data



recording collisions at 200 Hz (1 in ~200’000)

Recorded data

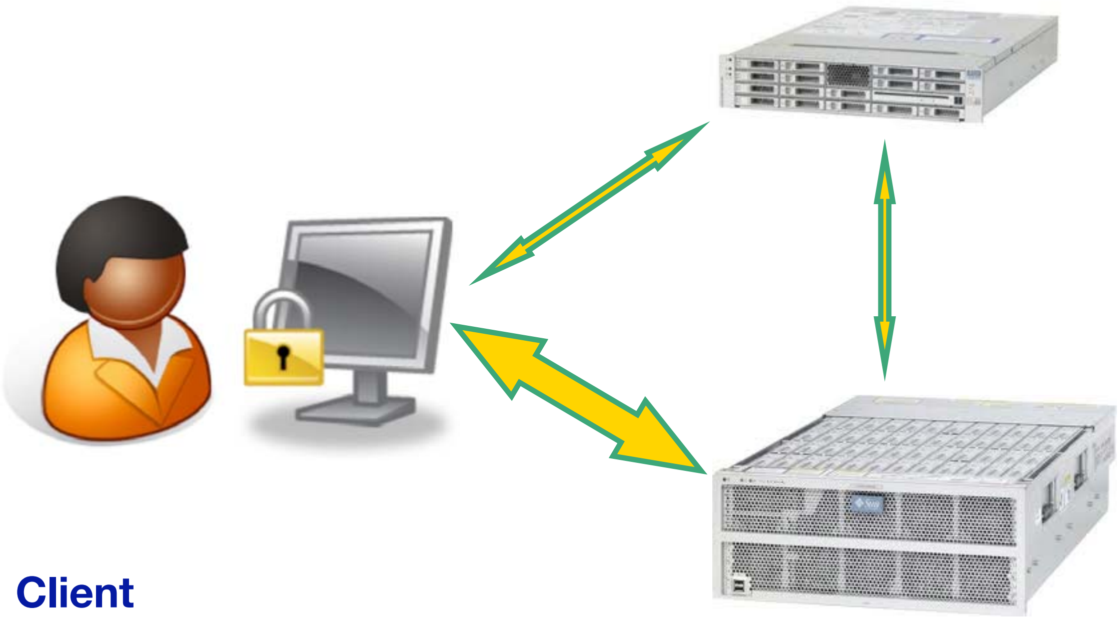
- 3 PB per year of raw data from one experiment
- up to 15 PB per year for the four experiments, (counting derived formats)
- ~25 pp collisions per “event”



ATLAS computing in Geneva

- 268 CPU cores
- 180 TB for data
  - 68 in a Storage Element
  - 110 on NFS
- special features:
  - direct line to CERN at 10 Gb/s
  - latest software via AFS
  - data channels from CERN Tier 0 and from the NDGF Tier 1
- the data analysis facility for Geneva group
- Trigger development, validation, commissioning
- grid batch production for ATLAS

Disk Pool Manager (simplified)



Head node (one)

- the grid interface (SRM)
- name space (directories and files) independent of file systems used
- map of logical to physical files in a database (MySQL)

Disk servers (4 now + 5 planned)

- THE DATA (4*17 = 68 TB)
- processes sending or receiving data (gridftp, rfio)

Client

- user authentication with grid proxy (X.509)
- command line client interface (rfdir, rfcp, rfrm)
- API in C (rfio_fopen(*f.,…))

The data are distributed between physical servers file by file, in a round-robin way.

More information: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm

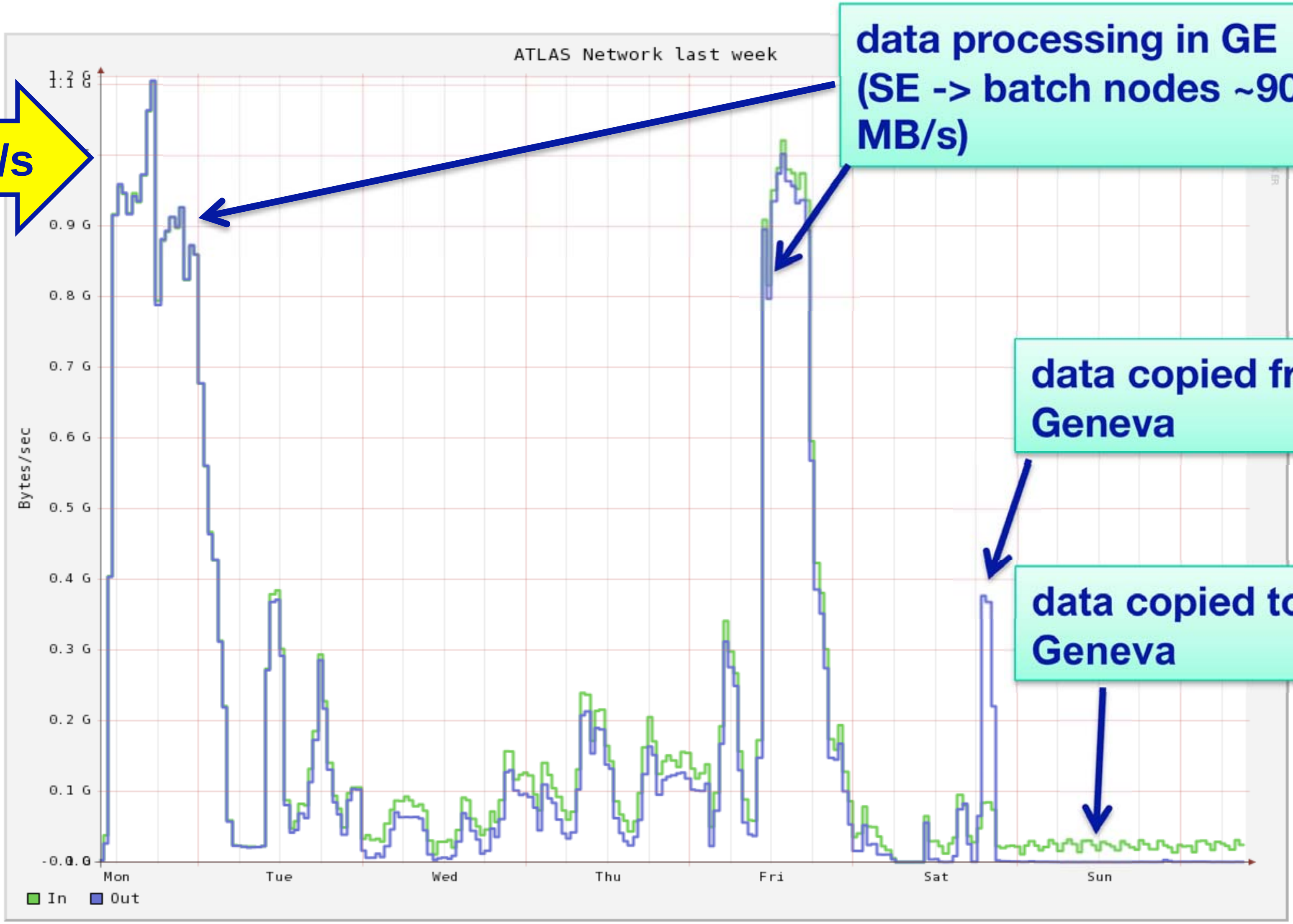
Tested performance of our storage

Data rates internal to the Cluster

tested with 100 batch jobs, 5 GB/job

Storage system	direction	max rate [MB/s]
NFS 3, 1 server	read	300
	write	200
DPM SE, 4 servers	read	800
	write	210

Monitoring of the network



1 GB/s

data processing in GE (SE -> batch nodes ~900 MB/s)

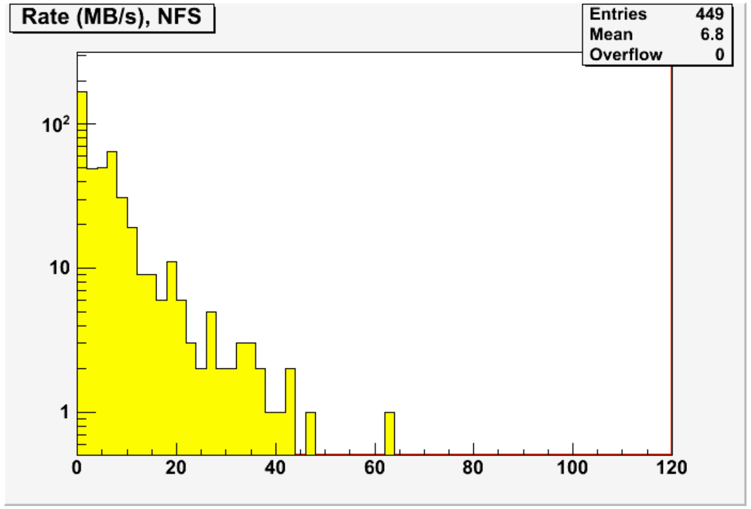
data copied from Geneva

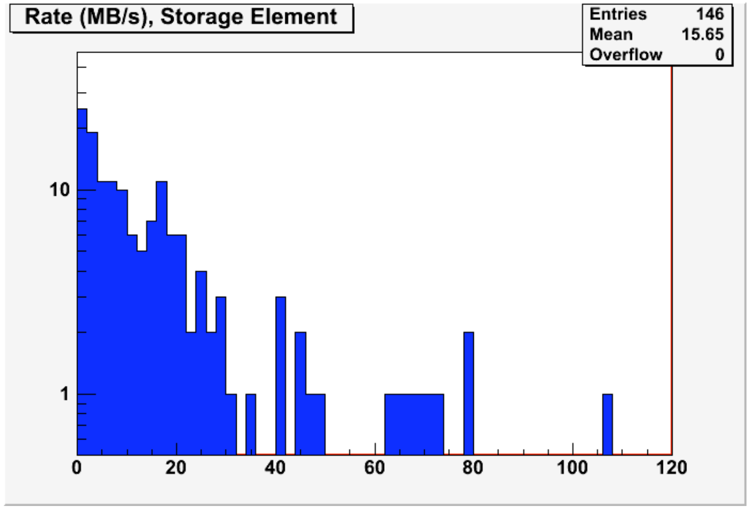
data copied to Geneva

Reliability issues

- copy of data from the SE to a local disk may fail
  - seems related to load on the file servers
  - otherwise at random
  - rate between 0.5 and 3.0%, depending on the run
  - just repeating the copy is always enough
- needs to be understood and fixed
  - we have more monitoring in place now (number of copies going on)
- having more file servers is likely to help

Data rates to Geneva





Method	MB/s	GB/(24h)
dq2-get	6.8	570
transfer to the SE	15.7	1300

Data transfer rates from other sites need an improvement.
Our hardware + network would allow ~800 MB/s!

Summary

- The LHC experiments produce data in PB per year. A global grid has been developed to process and to analyze the data. The development includes storage solutions.
- The Disk Pool Manager is a light-weight and simple(r) Storage Element, designed for smaller sites. There exist ~250 installations, up to 1 PB of disk.
- At the University of Geneva:
  - 68 TB of disk space
  - reading data at 800 to 900 MB/s
- Reading performance scales well with the number of servers.
- Some reliability issues to understand, but in production since Aug 2009.
- For long-range data transfers the rates are far below limits of our HW and network. Room for improvement!