PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin

E. Vilucchi[1], A. De Salvo[2], R. Di Nardo[1], C. Di Donato[3], A. Doria[3], G. Ganis[4], A. Manafov[5], G. Mancini[1], S. Mazza[6], A. Sanchez Pineda[3], F. Prelz[6], D. Rebatto[6], A. Salvucci[7]

on behalf of the ATLAS Collaboration

15/10/2013

CHEP 2013, Amsterdam

1. INFN-Laboratori Nazionali di Frascati, Frascati (RM), Italy
2. INFN-Roma1, Roma (RM), Italy
3. INFN-Napoli, Napoli (NA), Italy
4. CERN, Geneva, Switzerland
5. GSI, 64291 Darmstadt, Germany
6. INFN-Milano, Milano (MI), Italy
7. Radboud University Nijmegen and Nikhef, Netherlands

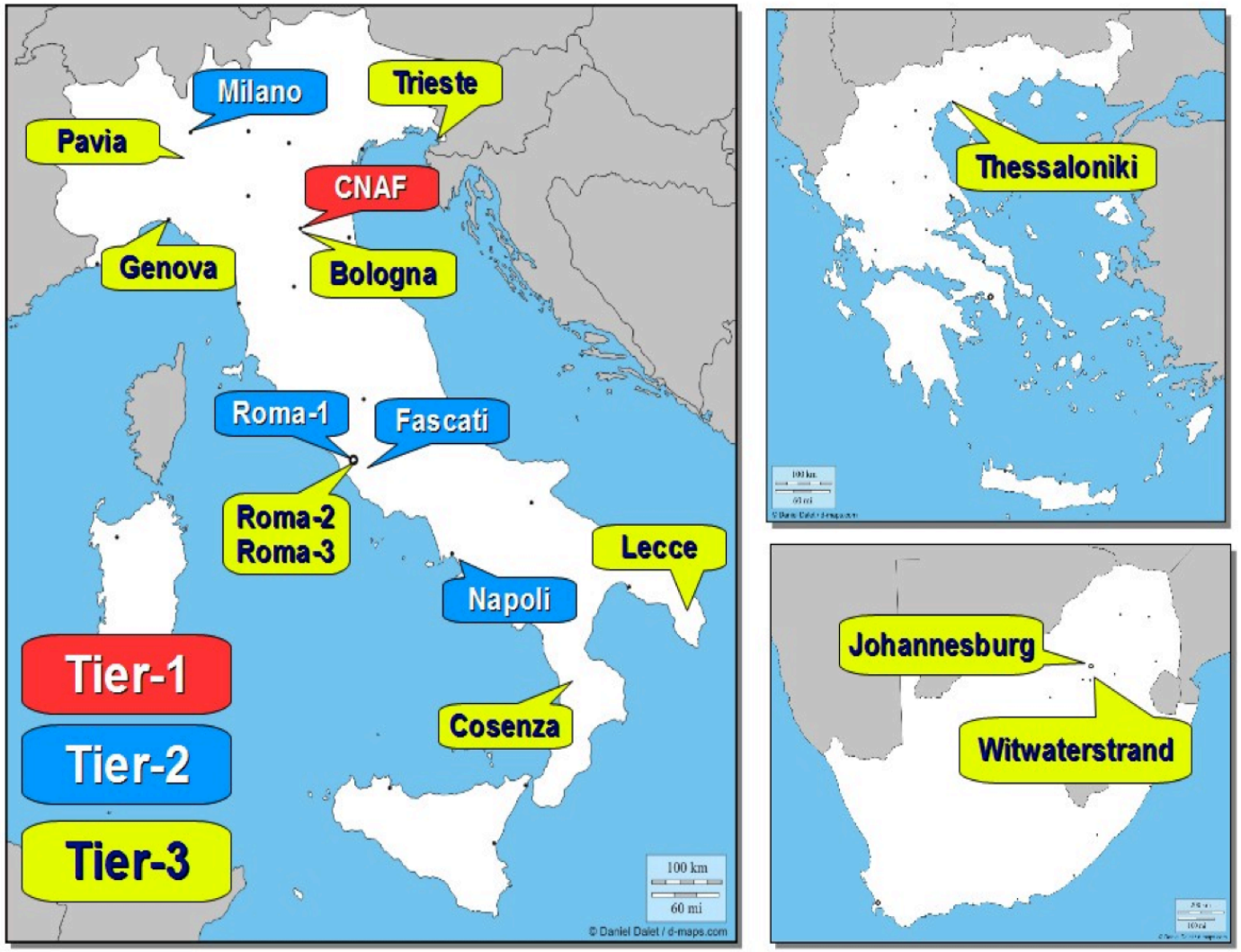
Outline

- The ATLAS Italian cloud
- CHEP 2012: tests of PROOF on Demand (PoD) usage with LRMS and gLite WMS
- CHEP 2013: tests of real analyses run in the ATLAS Italian cloud, Tier-2s/1, and at CERN submitting jobs with PoD on Panda system:
  - PoD workflow and usage examples
  - New startup latency tests
  - Calibration tests
  - Analysis tests
- Conclusions and future developments.

PoD: PROOF on Demand

- PROOF on Demand is a tool-set which dynamically sets up a PROOF cluster, at a user's request, on any resource management system (RMS).
- We successfully run real PROOF-based analysis examples, on non-dedicated resources of an ATLAS Italian Tier, at the same time of usual activities, without doing additional installation.
  - This investigation started in 2012, and the first PoD development was a plug-in based system able to use different job submission front-ends to enable a PROOF cluster on a batch system (LSF, PBS, Condor) and a Grid cluster, submitting the request to the gLite WMS.
  - At CHEP 2012 we presented a poster about our experience with PoD plug-in for gLite WMS. Now we are going to present our experience with PoD plug-in for Panda system.

The ATLAS Italian Cloud

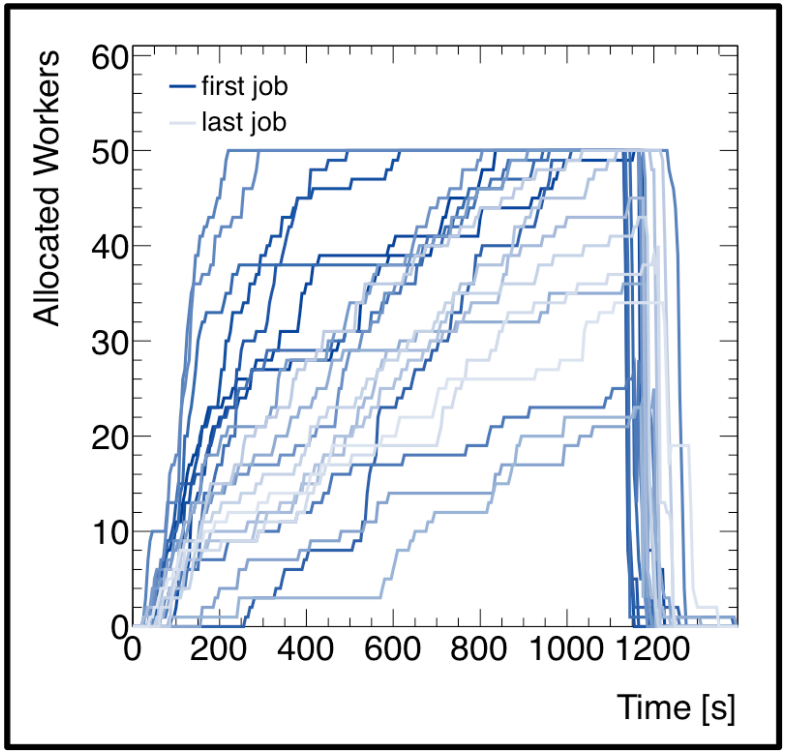


SRM adopted in the ATLAS Italian cloud

- Disk Pool Manager (DPM) is adopted in Frascati, Napoli and Roma-1.
  - DPM natively offers the XRootD protocol
    - Full support of X509-cert/proxy;
  - DPM sites are part of *FAX*: Federated ATLAS storage systems using XRootD
    - *FAX* brings Tier-1 and Tier-2s storage resources together into a common namespace: storage systems of all Tiers are viewed as a large single system;
    - *FAX* provides data access via a single entry point using XRootD's redirection technology;
  - In a recent DPM version HTTPS/WebDAV access is supported.
- StoRM, with GPFS underlying file system, is adopted in Milano and at CNAF
  - Installing GPFS on the WNs, data access with standard POSIX I/O function calls is allowed;
  - Recently XRootD access was enabled for GPFS at CNAF. Then CNAF SRM is part of *FAX*.

CHEP 2012: PoD for gLite WMS

- Startup latency: time necessary to allocate a certain number of nodes with PoD before running PROOF analysis.
  - Startup latency tested in Frascati (available workers as a function of time). The latency depends on the share allocated in the scheduler and many other parameters, e.g. cluster load. The color scale is proportional to the job submission time (from darker to lighter). The results of the test are in agreement with the expectations from the cluster load and fair-shares.
- Aggregated read-out rate with direct XRootD access over LAN
  - Input rate in MB/s using the PROOF statistics tools as a function of the number of workers (plot in backup slides).



Startup latency tested in Frascati: available workers as a function of time

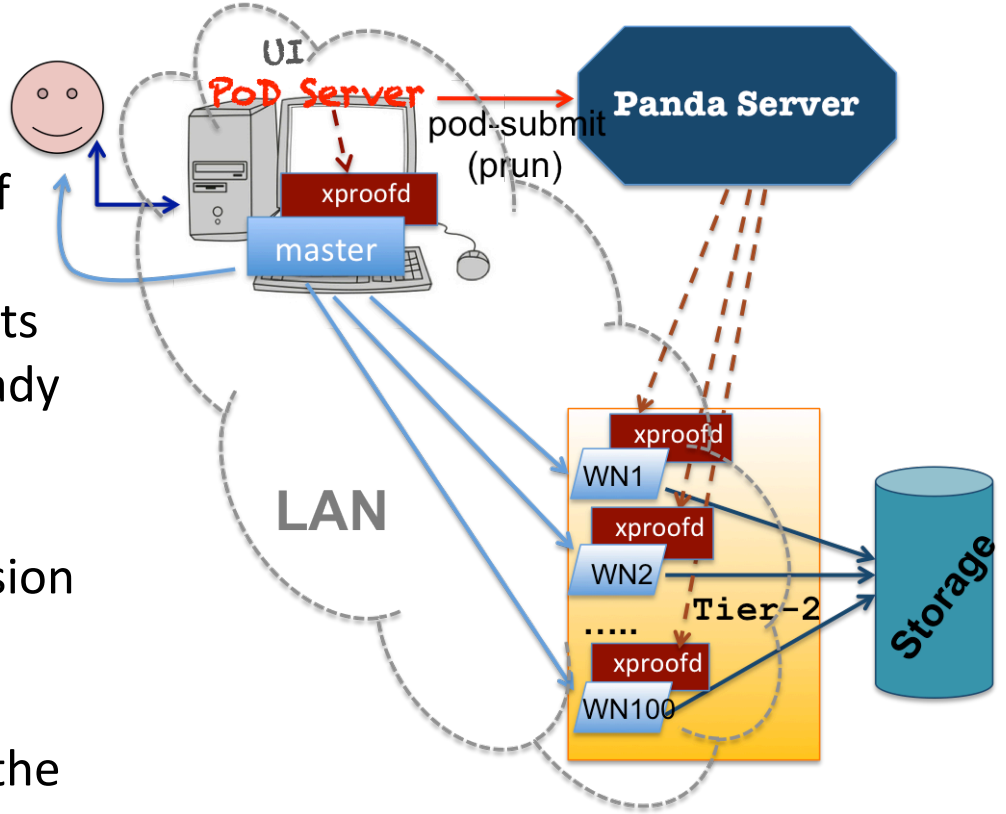
R. Di Nardo, G. Ganis, E. Vilucchi, A. Annovi, M. Antonelli, G. Carlino, A. De Salvo, A. Doria, A. Manafov, A. Martini, M. Testa, *"Enabling data analysis à la PROOF on the Italian ATLAS Tier-2s using PoD"*, Proc. Conf. 183 for Computing in High-Energy and Nuclear Physics (CHEP 2012) (New York, USA).

CHEP 2013: PoD on Panda System

- WMS PoD plug-in was successfully tested in 2012, but ATLAS does not support WMS submission anymore.
- Now we tested PoD with the new PoD plug-in for the Panda system.
  - Panda: the ATLAS data-driven workload management system for production and distributed analysis.
- With PoD Panda plug-in, users can enable a PROOF cluster on a Panda site.
  - The user connects to a UI, starts the PoD server and sends to Panda the request of a PROOF cluster with *N* workers.
  - Once the pilot jobs have picked up the PoD request from the Panda server and workers became available, user can send his jobs directly to the workers.

PoD workflow in detail

- PoD server starts *xproofd* daemon on the UI and uses prun to make a bulk submission of the required number of workers to PanDa.
- When the jobs run, PanDa server starts an xproofd daemon on the nodes, ready to accept PROOF connections from the master.
- Now the user can open a PROOF session on the master and the xrpoofd daemons run ROOT processes inter-connected between the master and the workers. The final merging is made from the master.



- In this test phase, users must submit the requests to the site the UI belongs to, otherwise the workers will not be able to connect back to the PoD server running on the master machine.
  - A solution allowing request submission to any site will be available in the next release.

Usage examples

```
$ pod-server start
```

start PoD server

submit to Panda the request of a PROOF cluster with 100 cores in the analysis queue at INFN-FRASCATI

```
$ pod-submit –r panda –q ANALY_INFN-FRASCATI –n 100
```

```
$ pod-info –nl
8
```

query the PoD server to know how many workers are available

```
worker pilatlas009@atlaswn050.lnf.infn.it:21001 (…)
startup: 79s (…)
worker pilatlas009@atlaswn046.lnf.infn.it:21003 (…)
startup: 82s (…)
….
```

Once enough workers are available, the PROOF analysis can be started

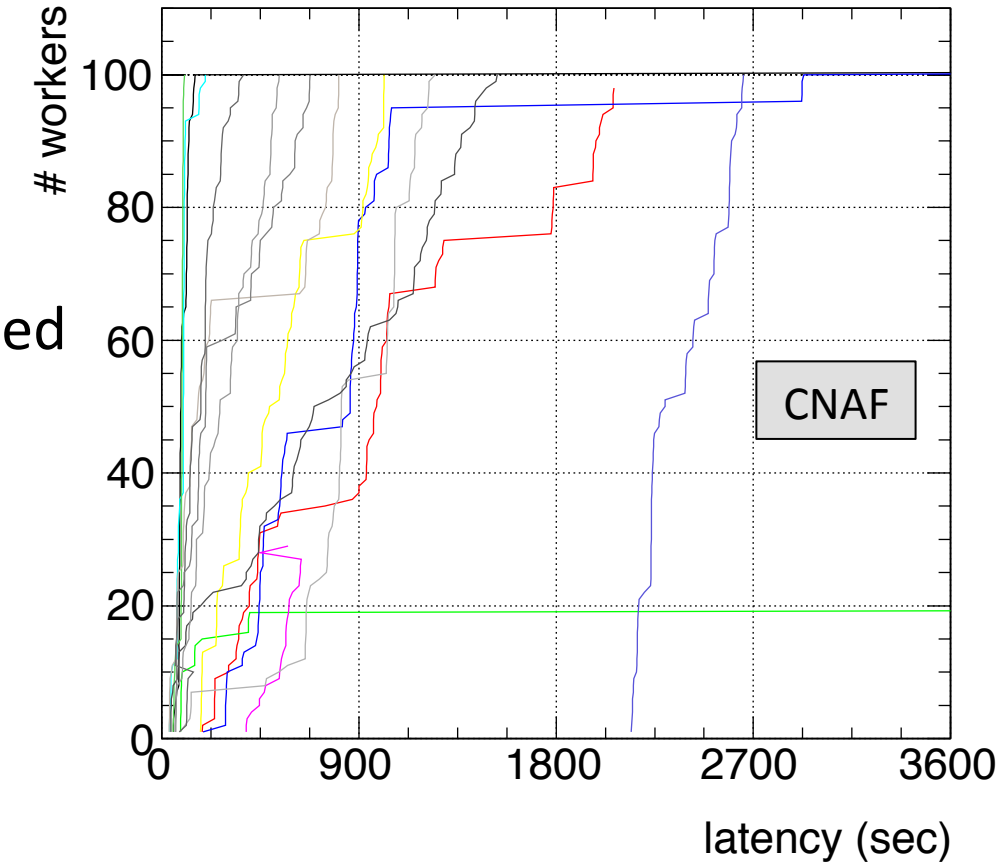
```
root [0] Tproof *proof=TProof::Open("pod://”)
```

Startup Latency

- Single test consists in:
  - Submitting 100 jobs (for convenience of the tests)
  - Retrieving startup time using "`pod-info -nl`"
- Time to obtain requested workers, depending on many factors:
  - Site status:
    - available job slots;
    - site occupancy rate:
      - Running/Activated/Defined jobs.
  - Number of requested job slots;
    - 100 for our tests
  - User priority in Panda.
- Many tests were run by different users in different sites.

Startup latency tests: CNAF, as an example

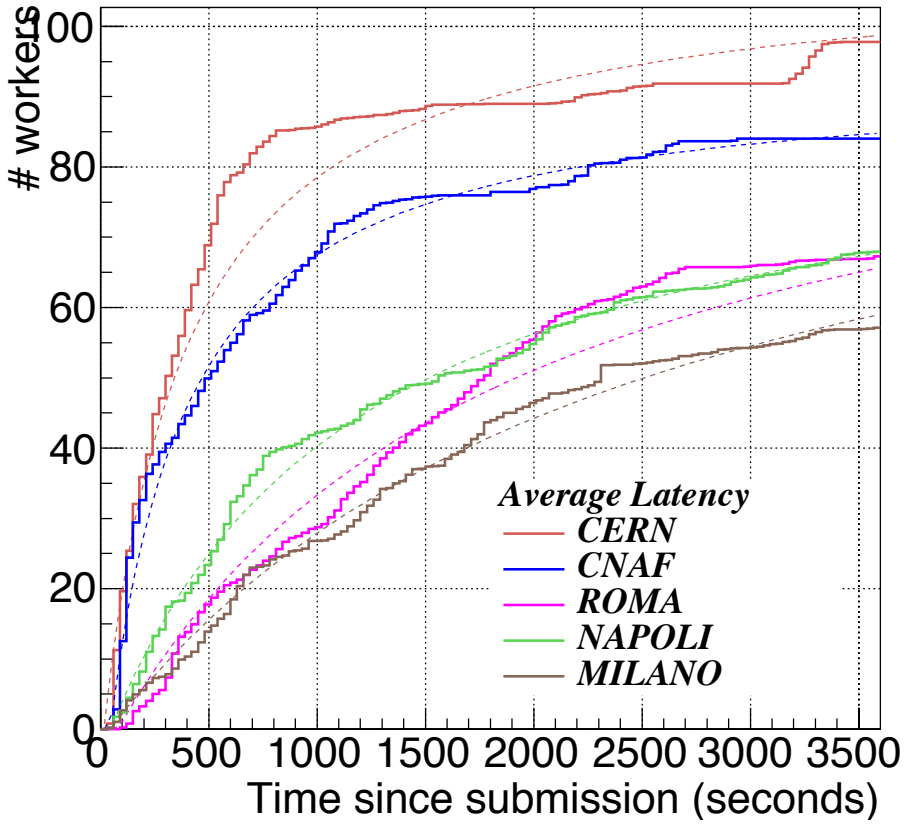
- Waiting time for the first job: order of a few minutes.
- Once the request has got attention: assignment rate s.t. 100 job slots are typically allocated in the order of 10 minutes.
- Time distributions are broad
  - varying load conditions of the site.
- The other sites show similar behaviours
  - the rates at which the jobs are assigned and the queuing time may vary due to the different available site resources.



Colored curves represent the different startup latency tests performed at CNAF

Average startup latency in different sites

- Average number of allocated workers from each test in bins of 30 seconds
  - At any fixed time, average of allocated workers in the site.
  - One hour window
- Ramp-up slope is steeper for CNAF and CERN than for Tier-2s
  - Larger amount of resources.
- On average, in all cases the first worker starts within O(1') while the ramp time depends on the size of the site.
- We tried to write a formula to represent the average curves.
  - The dotted lines show the result of an unweighted fit to a formula based on simple assumptions for job slot assignment.

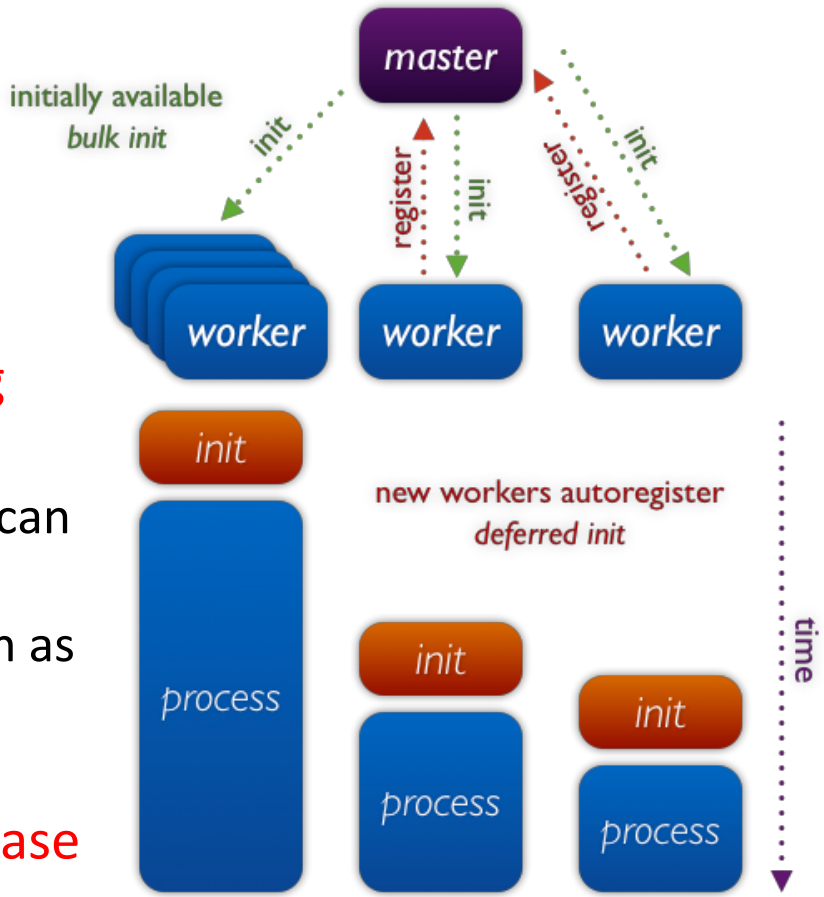


Continuous line: average startup latency
Dashed line: fit to a formula

Dynamic addiction of workers

Dario Berzano
CHEP 2013, Amsterdam
"PROOF as a Service on the Cloud: a Virtual Analysis Facility based on the CernVM ecosystem"

- New workers join and offload an existing analysis
  - Startup latency impact on the analyses can be significantly reduced.
  - Users can perform their analysis as soon as they have run the *"pod submission"* in PanDa.
- Implementation ongoing in the next release of PROOF



Thanks to D. Berzano for the contribution

Disk access and analysis tests

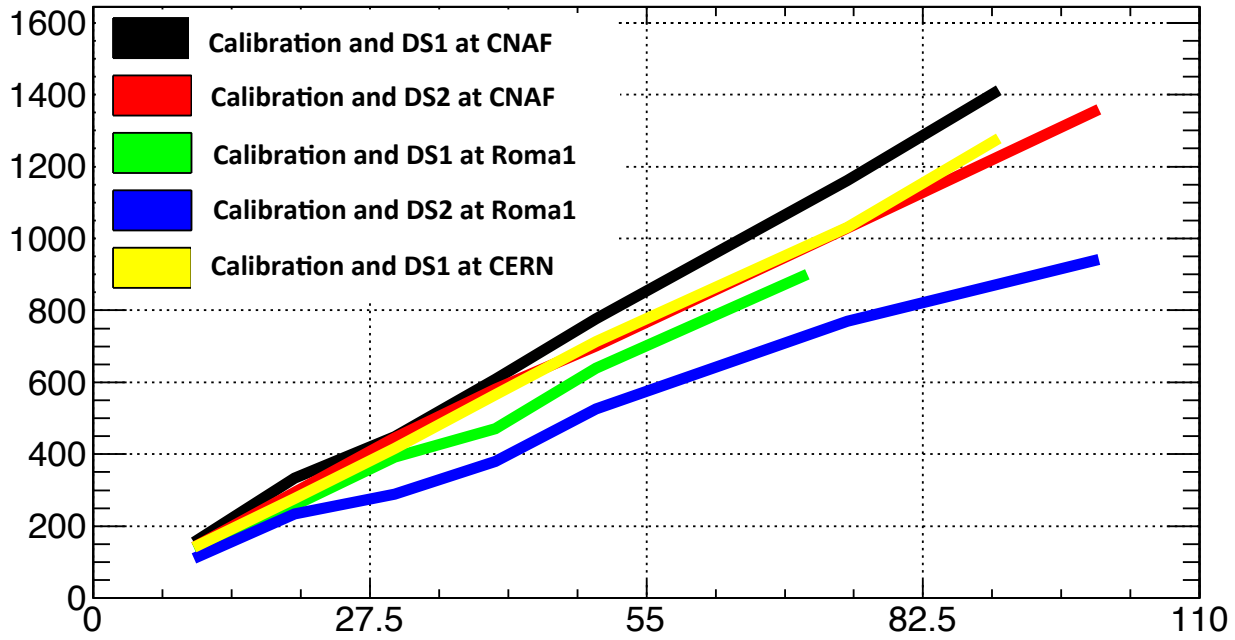
- Analysis tests were performed in any site of the ATLAS Italian cloud and at CERN.
- Before and after any analysis test we performed disk access test (calibration).
  - Tests were performed with 100 workers.
- Analysis tests and calibration tests where performed on any available SRM interface
  - on local storage system (workers and storage on the same site);
  - on remote storage systems (workers and storage on different sites);
    - XRootD direct access and trough the FAX service.

Disk access tests

- Disk access tests were performed to:
  - check the site efficiency
  - obtain an upper limit on the performance in terms of MB/s as a function of number of workers.
- Different storage systems or different access protocols
  - XRootD on DPM over LAN
  - XRootD on DPM over WAN
  - File protocol over StoRM with GPFS
  - XRootD on Storm/GPFS over WAN
  - XRootD on EOS
  - FAX infrastructure was tested for XRootD access over LAN and WAN
  - Few tests of *root* access with HTTP for DPM.

Results of local disk access tests

- Sample datasets (DS) of real analysis used with PoD
  - *DS1* (1M events), *DS2* (3M events): standard *D3PD* dataset (few files many information, ~100KB per event).
- Calibration tool:
  - a PROOF job running a dedicated TSelector reading the whole content of each entry of the TTrees from the files.
- Results show that
  - Up to 100 nodes performance scale linearly;
  - GPFS at CNAF gives best I/O performance than XRootD on DPM at Roma1;
  - CNAF and CERN have comparable performance.



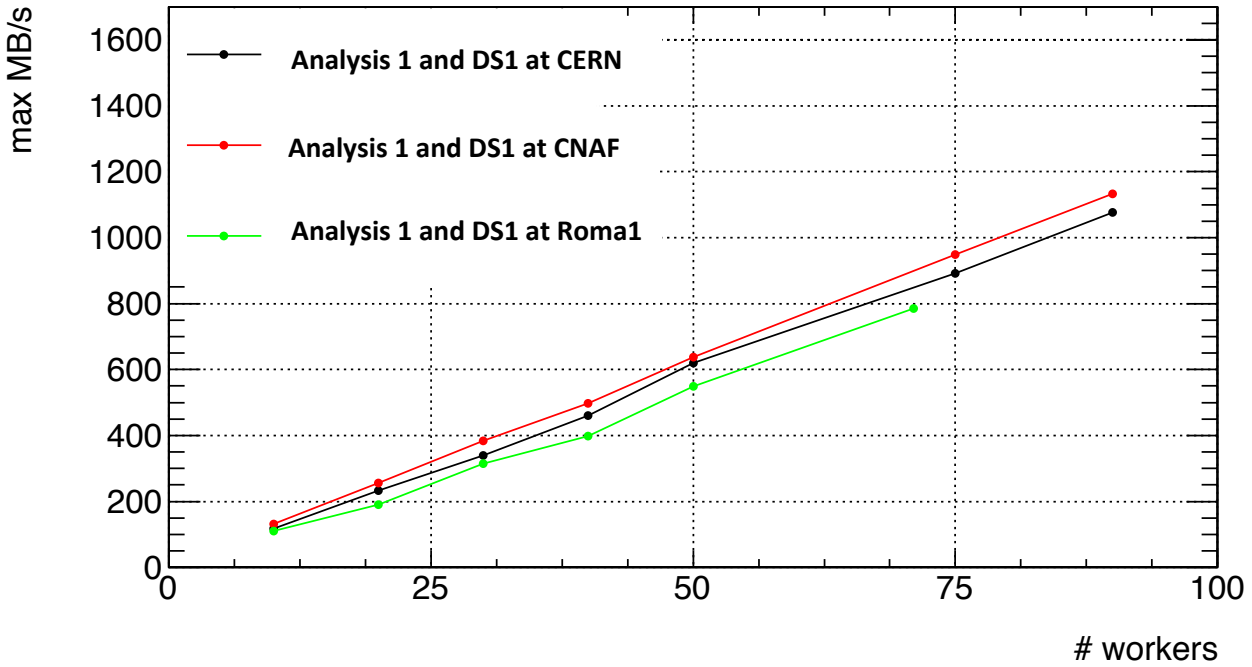
Max MB/s in function of number of workers

Analysis Tests

- We performed analysis tests with three different real analyses
  - Analysis 1: standalone C++ code without configuring the whole ATLAS software framework and RootCore.
  - Analysis 2: RootCore based with the EventLoop and D3PDReader packages.
  - Analysis 3: performed with a simple C++ code (based on Tselector), it uses the PROOF cluster to run over user-defined ntuples produced (with a "standard" Grid resources usage) in a previous analysis step.
- Input datasets of analysis 1 and 2 are the ones used for the calibration tests

Analysis 1 test results

- I/O scalability performance at CERN, CNAF, Roma1
- Comparing the results of the calibration with *DS1* (see plot in slide 16), we see that the weight of the additional event processing is about:
  - ~18% at CNAF
  - ~16% at CERN
  - ~16% at Roma1

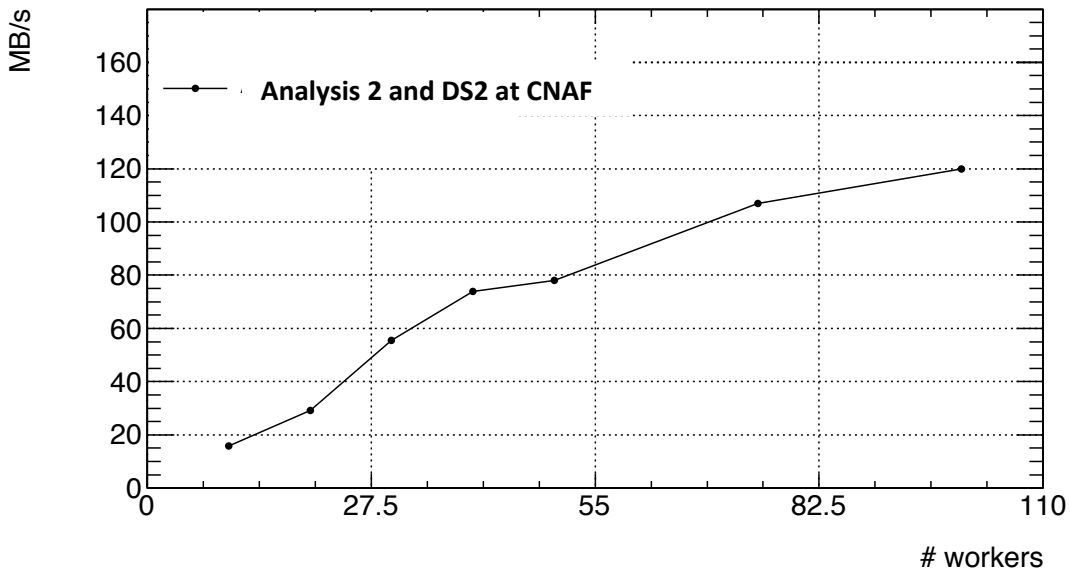


Max MB/s versus number of worker running analysis 1

Analysis 2 test results

- Average MB/s running analysis 2 with PoD at CNAF with *DS2.*
- Analysis 2 has an optimized disk access allowing higher processing rate
  - The first "standalone" analysis has a higher -not optimized- read-from-file rate (up to 1 GB/s) and around 7500 evts/s analyzed with 75 workers (reading ~500 branches).
  - Analysis 2, based on the D3PDReader package, has an optimized file access (around 120 MB/s reading ~100 branches) analyzing around 30000 evts/s with 75 cores.

- In Napoli site it was also possible to use the http protocol. Some preliminary tests gave:
  - 100 workers: 90 MB/s, 25567 Evts/s in average. Slightly lower performance but comparable with XRootD.
  - No https access (due to the installed DPM version).

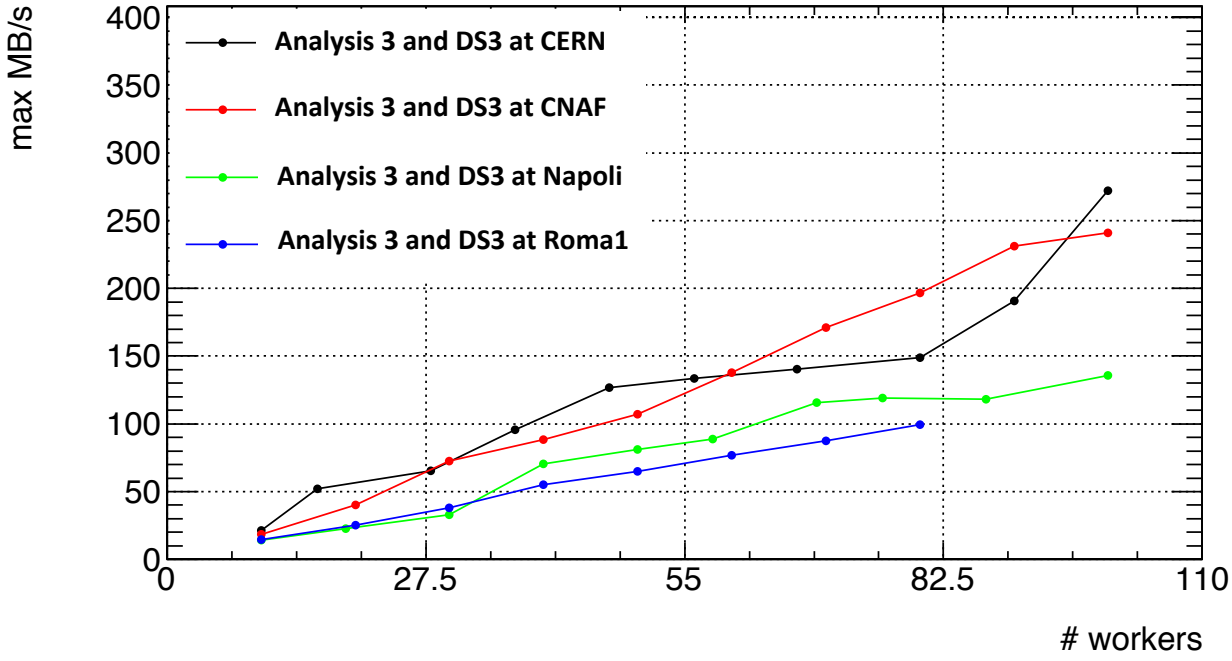


Avg MB/s versus number of worker running analysis 2

Analysis 3 test results

- Plots shows I/O scalability performance at CERN, CNAF, Roma1 and Napoli.
- Input dataset, *DS3,* is completely different from *DS1* and *DS2*. Files are root n-tuples: many files with little information, ~0.2 KB per event (140M of events with a rate of 1200K evt/s).
  - Performance is a consequence of the files structure and size.

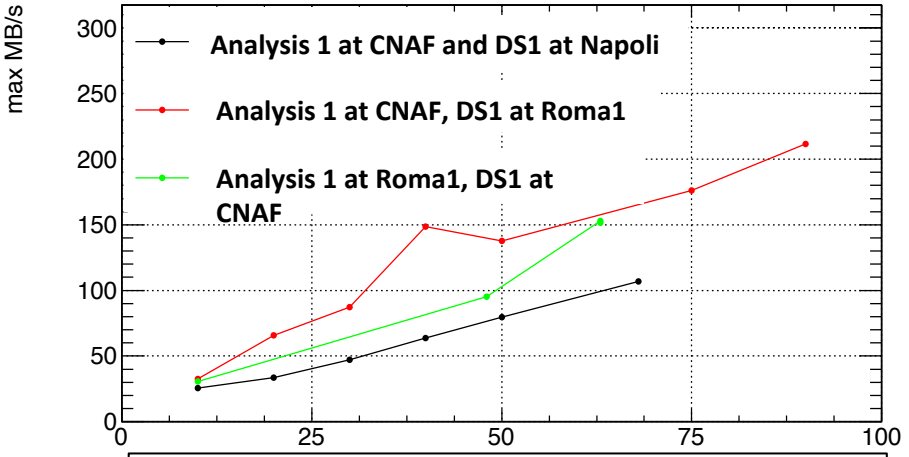
- Performance scale in agreement with expectations based on the number of workers.
- This example shows how PoD could be successfully used for a different use case.



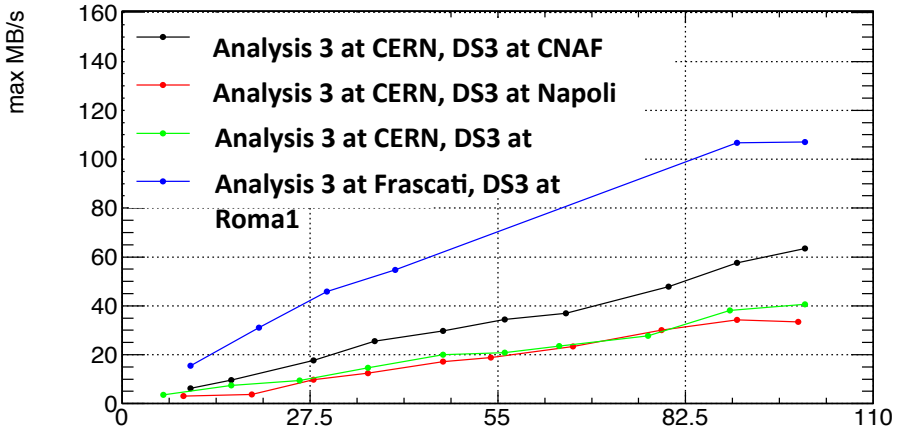
Max MB/s versus number of worker running analysis

Remote XRootD access and *FAX* infrastructure

- We run analyses with PROOF cluster and input datasets in different sites.
  - Direct XRootD access (without *FAX*);
  - *FAX* access.
- *FAX* works smoothly. Accessing, through a central *FAX* redirector, to an input container distributed across different sites, the request is forwarded to the appropriate storage elements, selecting the local SRM when possible, and the mechanism is transparent for the user.
- Performance depends on network connections.



Max MB/s in function of # of WNs for analysis 1 with DS1 over WAN



Max MB/s in function of # of WNs for analysis 3 with DS3 over WAN

Conclusions and future works

- In this study we used all the main components of the ATLAS analysis chain to evaluate the potential of PROOF-based analysis on non-dedicated resources.
- We successfully run real analysis examples at the same time of usual activities of an ATLAS Tier-2, competing with other users for analysis resources.
- In terms of startup latency all studied sites showed good responsiveness and its impact on the analysis will be reduced when the currently experimental feature of adding workers while processing will be consolidated (D. Berzano).
- The storage systems showed good scalability for the typical ranges of workers per session (100).
- We showed that FAX works smoothly and made some initial tests using the HTTP protocol.
- For the future work we propose to evaluate PROOF with dynamic workers addition and we will also evaluate data access with the HTTPS/WebDav protocol.

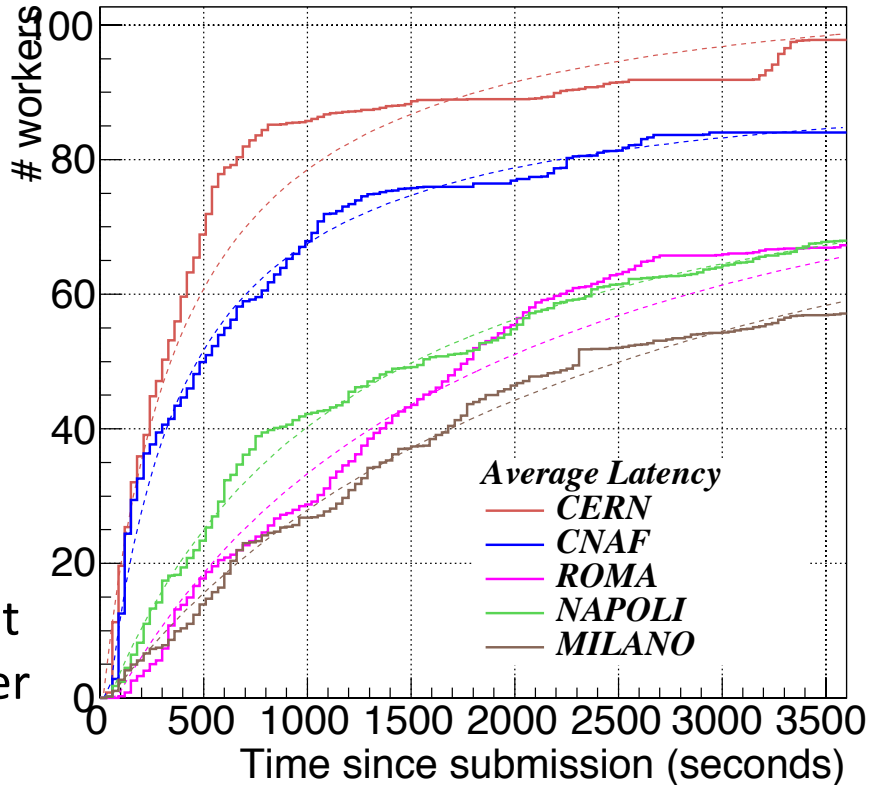
Backup

Startup Latency tests in different sites

- Assuming that
  - the rate at which job slots become available is proportional to the number of successful jobs;
  - the priority of assignment decreases with the number of jobs already started, the curves can be described by:

$$n(t) = \frac{p_0 \cdot (t - t_0)}{1 + p_1 \cdot (t - t_0)}$$

  - t_0 representing the queuing time to obtain the first worker, p_0 being related to the rate of available slots at $t=t_0$, and the ratio p_0/p_1 to the number of requested slots.
- The dotted lines show the result of an un-weighted fit to this formula.
  - The results of the fits are given the table on slide 25.



Continuous line: average startup latency
Dashed line: fit to the formula

Startup Latency tests in different sites

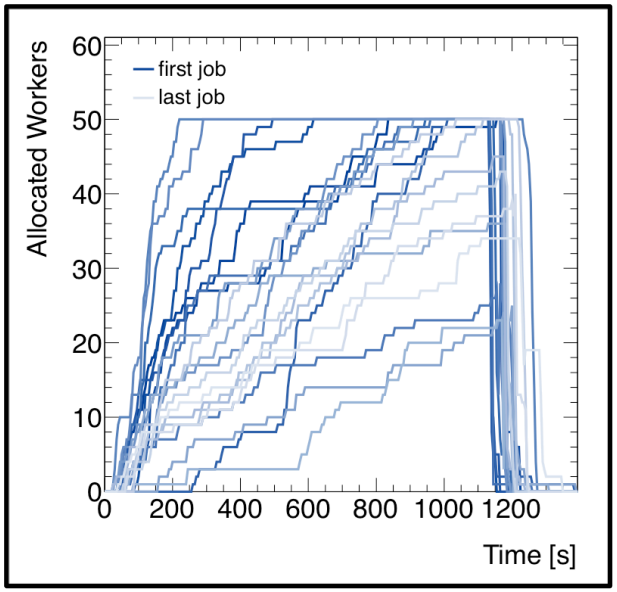
- The fact that the simple formula at slide 24 adapts reasonably well to the measured curves can also be explained with an use, in the sites, of the computing resources on the base of a fair-share between different activities (production and analysis).
- The column $N_{jobs,p0}$ is the number of jobs per day derived from p_0.
- $N_{jobs,PanDA}$ is the number of jobs per day obtained from the PanDA monitoring system.

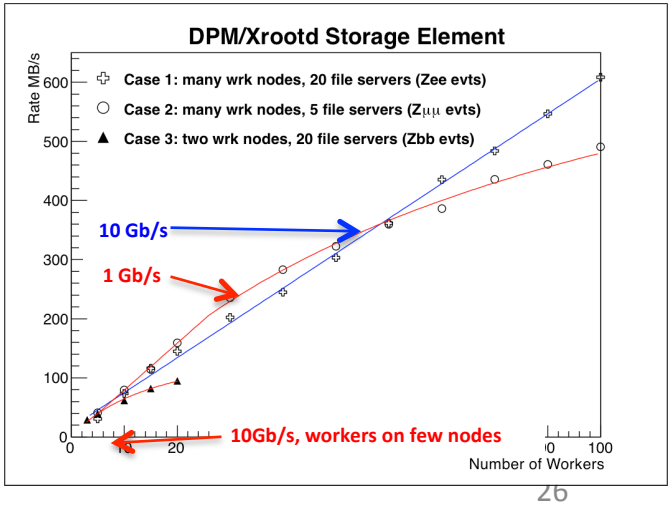
Site	t_0 [s]	p_0 [s^{-1}]	$N_{jobs,p0}$ [day]	$N_{jobs,PanDA}$ [day]
CERN	15.4	0.282	24365	30321
CNAF	44.9	0.253	21859	18957
MILANO	45.7	0.041	3542	3728
NAPOLI	46.3	0.076	6566	6971
ROMA	79.0	0.054	4697	5630

- On average, in all cases the first worker starts within O(1') while the ramp time depends on the size of the site and we get separately similar values for Tier 2 and Tier 1-like sites. The number of estimated successful jobs per day is in good agreement with the ones retrieved from the PanDA monitoring.

CHEP 2012: PoD for gLite WMS

- Startup latency tested in Frascati (available workers as a function of the time).
  - Latency depends on the share allocated in the scheduler and many other parameters, e.g. cluster load. The color scale is proportional to the job submission time (from darker to lighter).
- The results of the test are in agreement with the expectations from the cluster load and fair-shares.
  - First submissions suffer from resource competition with PanDA analysis of generic ATLAS user jobs, whose computing fair-share was 25%.
- Readout performance.
  - Input rate in MBytes/second using the PROOF statistics tools as a function of the number of workers. This quantity is derived from the number of bytes effectively read out from the files by the active workers divided by the total processing time; the latter includes event decompression and construction of the event information in memory, which is the only CPU load in this simple analysis, giving a negligible effect to our study.





Datasets

- *DS2* and *DS1* are D3PD
- *DS1* is a dataset of ZZ background MC sample, containing 90 files, corresponding approximately to 1 million events and with a total size of almost 100GB.
- *DS2* is a MC sample of Standard Model Higgs boson decaying into photon pairs. The dataset contains 3 million 100KB-events composed of approximately 300 files totalizing 300 GB.
- *DS3*, is a collection of small root ntuples (~400 files produced previously running another analysis on the Grid) with an average size of 150MB/file and 0.2 KB per event. The number of events can grow up to hundreds of millions of very small size events.

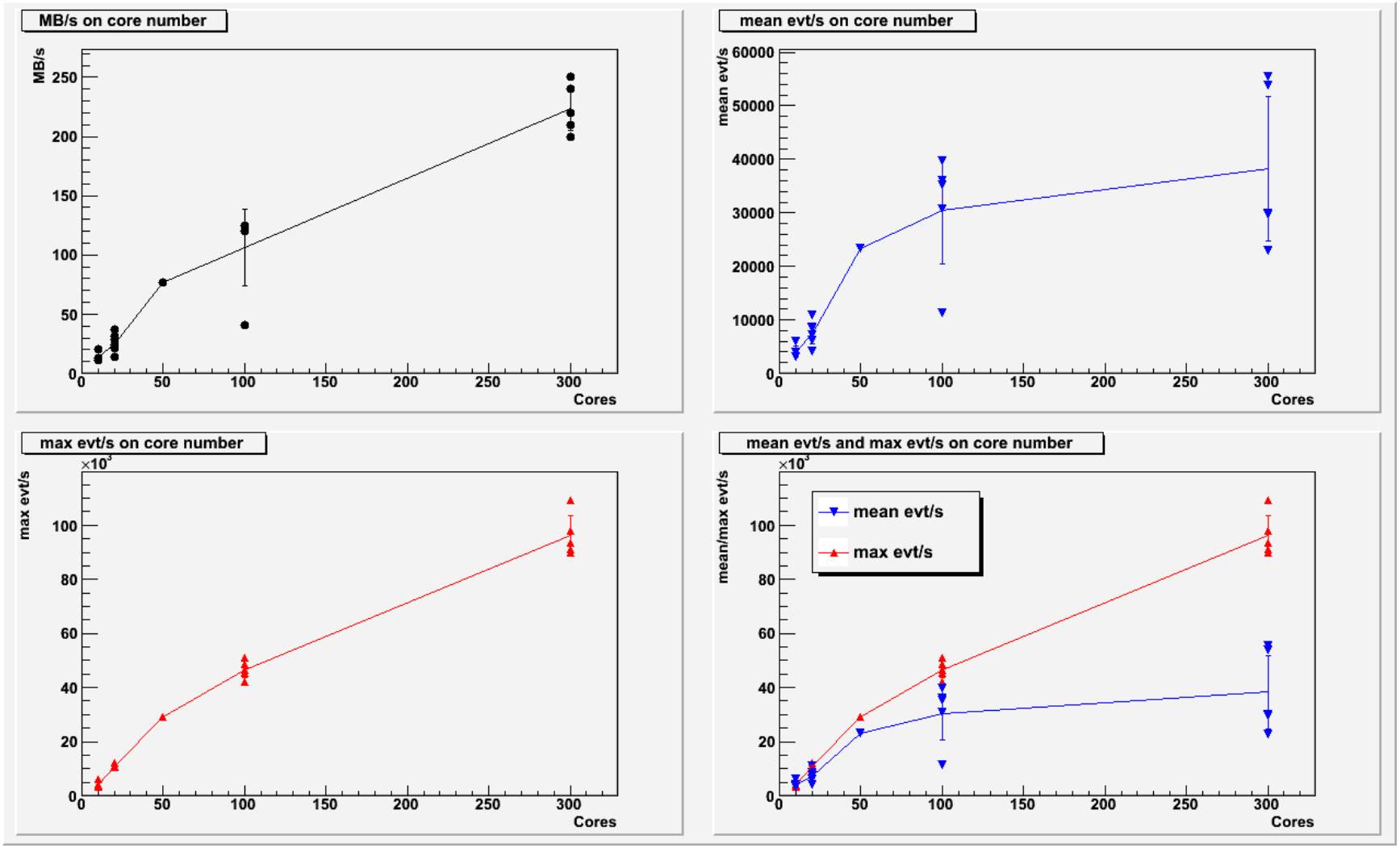
Analysis 1

- Analysis 1 is the Higgs boson selection in the four lepton channel ($H \rightarrow ZZ^* \rightarrow 4l$). It is organized as a standalone C++ code without configuring the whole ATLAS software framework and RootCore. Since the number of branches contained in a D3PD TTree is huge (around 7500) and most of them are not used in this analysis, the code loads only the needed ones (around 500) in order to speed and optimize the analysis process; in this way ROOT will not allocate memory for inactive branches and the cache and the reading process of the event is faster.

Analysis 2

- The second analysis is the selection of a Higgs decaying into two photons and it is RootCore based with the EventLoop and D3DPReader packages. It has a more efficient event access, because only the branches of the TTree used in the analysis are read from disk, in an optimized way with respect to what is done in the standard root TSelector. This reduces the I/O between the analysis and the data read from disk, enhancing greatly the number of processed events per second. For this particular case, the second analysis allows to gain in speed almost a factor of 10 in terms of events processed per second, with respect to the stand-alone version of the same analysis.

Analysis 2


MB/s on core number
MB/s
Cores
mean evt/s on core number
mean evt/s
Cores
max evt/s on core number
max evt/s
Cores
mean evt/s and max evt/s on core number
mean/max evt/s
mean evt/s
max evt/s
Cores

Analysis 3

- The third analysis selects Higgs decays in two leptons and two jets (H->ZZ->llqq). Analysis 3 uses ROOT intermediate ntuples to perform some final cuts and calculations to generate the sets of plots to be merged into the usual data-MC comparison.