Analysis model for $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$ analysis

Valerio Ippolito

Università di Roma “La Sapienza”

INFN Sezione di Roma / CERN

Outline

- $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$: a challenging analysis
  - what we want and what we have
- HiggsAnalysis / Higgs4lepAnalysis
  - code structure and analysis model
- dealing with the GRID
- performance and improvements
- conclusions

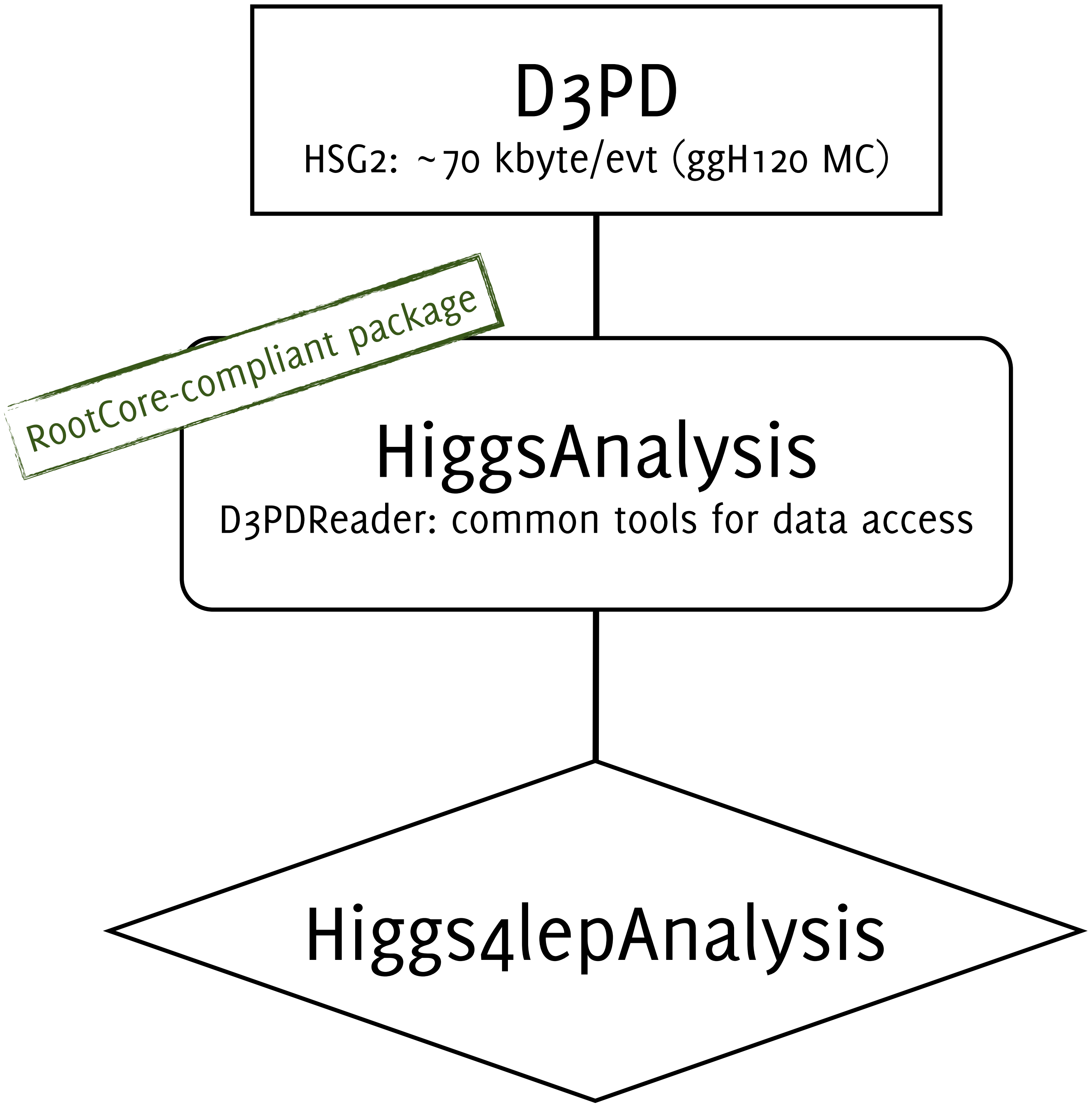
What we want to do

- $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$ is one of the leading analyses for 2012
  - we need to have an immediate feedback as soon as new data is collected
- many common tasks need to be performed
  - candidate reconstruction ("main analysis")
  - reducible background estimation from control regions ("relaxed selection")
- data analysis must be quick
  - try to build an ntuple reliable for both main and relaxed selections
  - make sure a quick common framework is nevertheless available for detailed studies

How we do it

- we run over HSG2 D3PDs
  - previously we produced an intermediate object-oriented ntuple (ZNtuple)
  - it was not affordable anymore (processing speed, data transfer)
- D3PDs are subscribed to PHYS-HIGGS disks in ROMA1 and NAPOLI
  - running via grid we profit from resources reserved to IT-cloud users
- D3PDs are "light AODs"
  - all informations we need are there
  - we build interesting objects (e.g. quadrileptons) and we use them in the analysis
  - in this way plotting and downstream studies are quick (we run on tiny minimal ntuples)

Code structure



- HiggsAnalysis
  - common framework to read D3PDs
  - uses D3PDReader to improve reading performance
  - robust against different D3PD tags (e.g. missing variables)
- Higgs4lepAnalysis
  - inherits from HiggsAnalysis
  - implements Summer2012 selection
  - produces a common light ntuple covering most studies

Data flow

NTUP_HSG2	GRID	DaTRI
(IT cloud, PHYS-HIGGS)	(PanDA)	(ROMA1 / NAPOLI LOCALGROUPDISK)

MC

mc11c

~ 3 Tb

data

Muons

Egamma

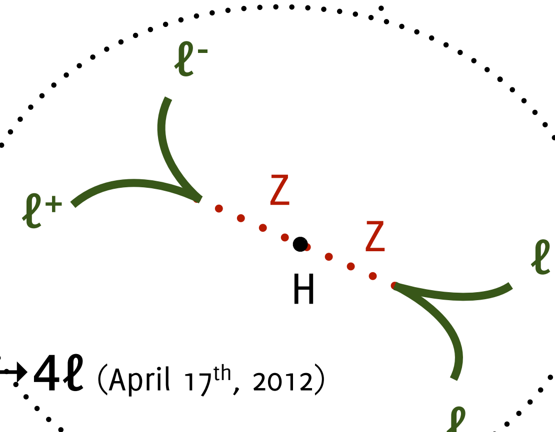
~ 2×2 Tb / fb^{-1}

Higgs4lepAnalysis
[prun]

flat H4l TTree

~0.2 Gb / fb^{-1}

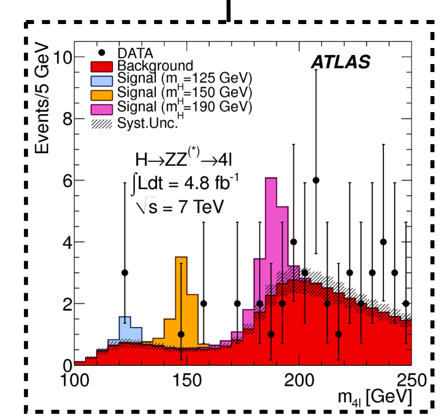
(~ 3 Mb / 50k Higgs @120GeV)



one entry per quadrilepton

(no charge requirement for background studies)

TSelector/macros for each study
[prun/local]



Grid submission

We use PanDA via prun (launch jobs from a full Athena+RootCore setup via CVMFS):

prepare a working RootCore environment and launch analysis

```
prun
     --writeInputToTxt IN:input.txt
     --exec "cat input.txt; Higgs4lepAnalysis/bin/test"
     --useAthenaPackages
     --useRootCore

     --inDS  mc11_7TeV.116761.PowHegPythia_ggH110_ZZ4lep.merge.NTUP_HSG2.
             e873_s1310_s1300_r3043_r2993_p869/
     --outDS user.vippolit.H4l000000.116761.PowHegPythia_ggH110_ZZ4lep.p869.fullproc.01

     --outputs output_test.root
     --extFile=Higgs4lepAnalysis/packages/files/pileup/*root*,Higgs4lepAnalysis/packages/
              files/ggFHiggsPtWeight/*root*,TrigMuonEfficiency/share/*root*,
              egammaAnalysisUtils/share/*root*,Higgs4lepAnalysis/files/*root*

     --nGBPerJob=10
     --cloud=IT

     --inTarBall tarball.tar
     --tmpDir=/tmp

     --destSE INFN-ROMA1_LOCALGROUPDISK
```

input/output datasets

input/output files

use IT cloud (all data and most MC samples are in Italy)

speed-up multiple submissions (first job must be submitted twice, with `--outTarBall tarball.tar --noSubmit`, to create the tarball)

data transfer from SCRATCHDISK to LOCALGROUPDISK

Job monitoring on PanDa monitor website

(or try `python /afs/cern.ch/user/j/jha/public/Atlas/Panda/latest/panda_task.py --server --showPandaID --outDS=XXX`)

Performance

We performed a test run over all the available mc11c samples and the full 2011 data sample

(both data and MC samples are fully replicated in the IT cloud)

Monte Carlo

* running over 106 signal and background samples
* ~300 jobs [just 1 subjob per sample, except Z+jet, JF17]
* submission takes ~2 h
* each subjob takes ~5' to run (locally we process signal at > 1 kHz)

→ overall MC processing time is 2h30'

data

* two big jobs (~900 subjobs each) on Egamma and Muons streams
* submission takes ~ 1h
* each subjob takes less than 30' to run

→ overall processing time is 4h30'

(after MC! priority for data is from 830 to 630...)

~7 h for a single user to process everything!

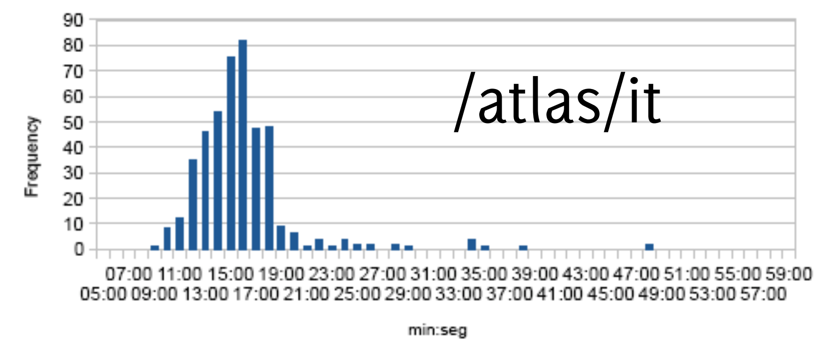
Finalizing the analysis

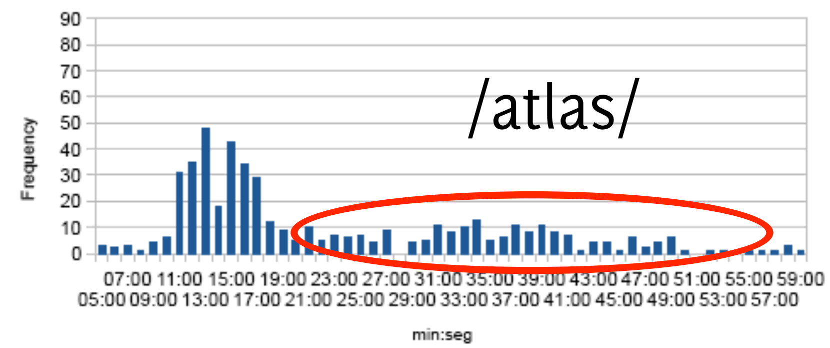
- subntuple production is done once for all
  - integrate with new collected data
  - re-run if changes in main analysis (e.g. selections, corrections/calibrations) are introduced
- downstream studies rely on this output
  - candidate reconstruction and limit (“the final plot”)
  - selection optimizations (cut-based, MVA, mass resolution studies)
  - background studies (Z+μμ)
- subntuple size is ~2.5 Gb (0.5 Gb for MC)
  - run on the grid with simple TSelector via prun (or locally)

What did we gain?

- we switched from ZNtuple to D3PD
  - direct access to all needed variables
  - no need to transfer huge ntuples (1/10 of D3PD size: ~7 days for 5 fb^{-1}...)
  - D3PDReader allows a quick analysis reading only branches actually needed (>1 kHz, even more with TTreeCache)
- we have PHYS-HIGGS disks in ROMA1 and NAPOLI
  - GRID jobs profit from resources reserved to `/atlas/it` users (average time-to-start is ~30% shorter with much less tails)
  - in ~7 h a single user can analyze 5 fb^{-1} of data plus MC

job time-to-start





What did we gain / 2

	before	after
data/MC location	worldwide	IT cloud
ntuple size	~ Tb	~ Gb
signal MC processing time	10 min / 50k evts	50 s / 50k evts
GRID time (MC⊕2011 data)	~ 1 day	~ 1/2 day
data transfer to IT cloud	~ 1 week	already on SCRATCHDISK

from a physics point of view:

- almost instantaneous signal MC studies
  (acceptance challenge, resolution, optimizations...)
- quick framework to analyze real data
  (data-driven background studies - Z+μμ, Z+ee, ttbar)
- constantly run on new data in 2012
  (monitor D3PDs replication to stay tuned with production)

Conclusions

- $H \to ZZ^{(*)} \to 4\ell$ is a demanding analysis
  - results need to be updated on daily basis
  - this is possible only with a robust and responsive analysis framework
- HiggsAnalysis+Higgs4lepAnalysis model does the job
  - common structure using D3PDReader to read D3PDs
  - well structured code: final analysis relies on small candidate-wise ntuples reliable for background studies as well
- D3PDs are replicated to PHYS-HIGGS disk in the IT cloud
  - shorter job time-to-start, access reserved resources
  - full 2011 analysis can be performed in almost half a day
  - of course we can't have every space token in Italy, but replicating jobs' output to localgroupdisks grants high GRID performances!