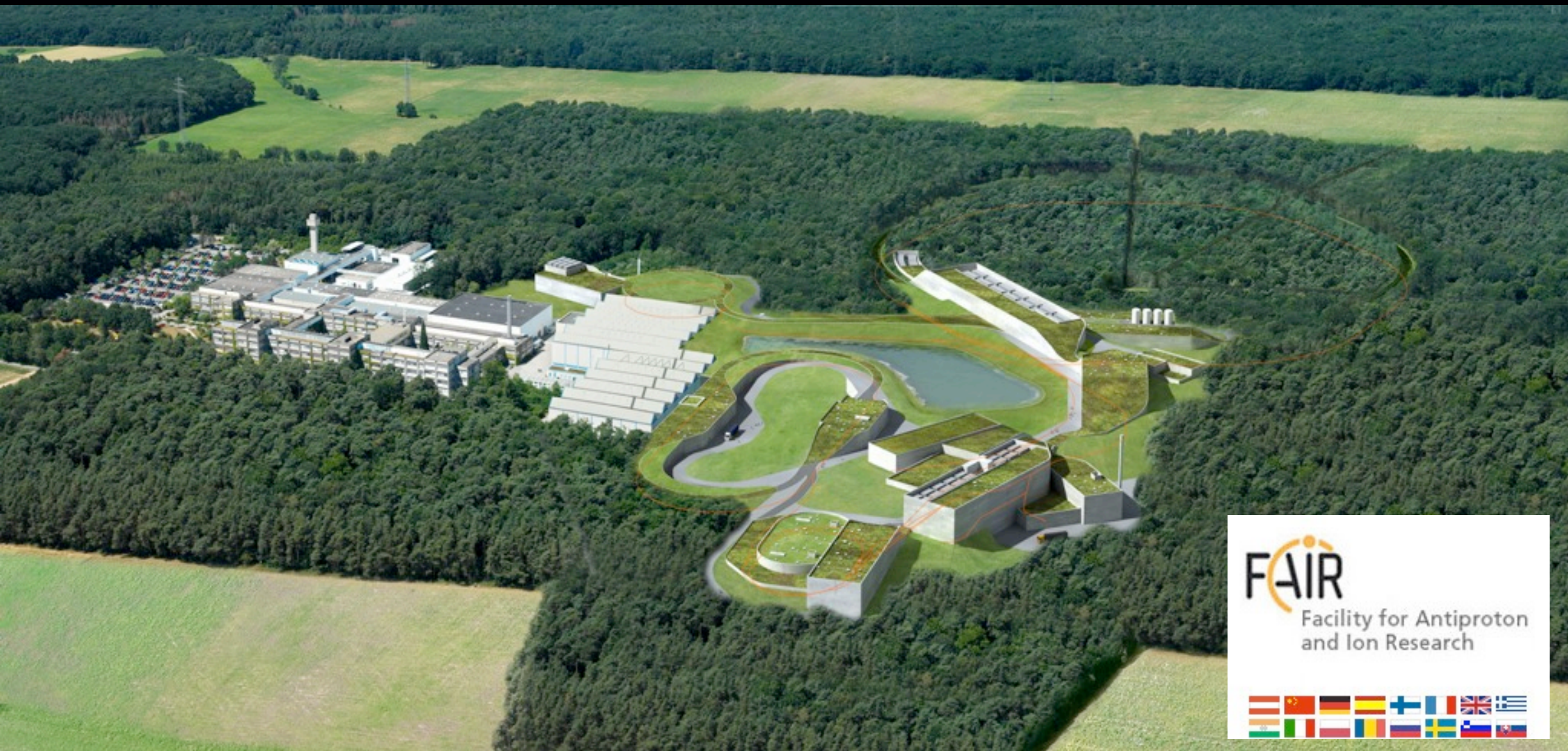# Computing On Demand: Analysis in the Cloud

Anar Manafov, GSI Darmstadt

GSI: a German National Lab for Heavy Ion Research
FAIR: Facility for Ion and Antiproton Research ~2018

FAIR
Facility for Antiproton
and Ion Research

http://www.gsi.de

FAIR http://fair-center.org/

GSI computing 2011

ALICE T2/T3
HADES
2800 (LSF) + 1800 (SGE) Cores
2.1 PByte Lustre
Topical Center LQCD

FAIR computing 2018

CBM
PANDA
NuSTAR
APPA
LQCD
300000 Cores
40 PB Disk and 40 PB Archive

Analysis Software:

dynamic or on-demand Analysis Model

Essential Properties

- portability,

- simplicity,

- confidence (trust),

- lightweightness.

Portability

How can we achieve portability of our analysis code?

Develop portable analysis frameworks:

- don't be irresponsible! Just use ROOT,

- constantly revise all dependencies.

Use open standards, even more - be a part of standardization committees...
For example, HEP is a big C++ customer. Are we presented well in the C++ standards committee?

Simplicity

Keep it simple!

- Keep your code supportable.

- Provide simple interfaces.

- Maintain clear and safe deployment model.

Confidence

Trust your code!

- unit/functional tests.

  We must write testable code.
  Code must fit for tests and not the other
  way around - "a Test-driven development"

Lightweightness

Minimal dependency of business logic code to infrastructural code

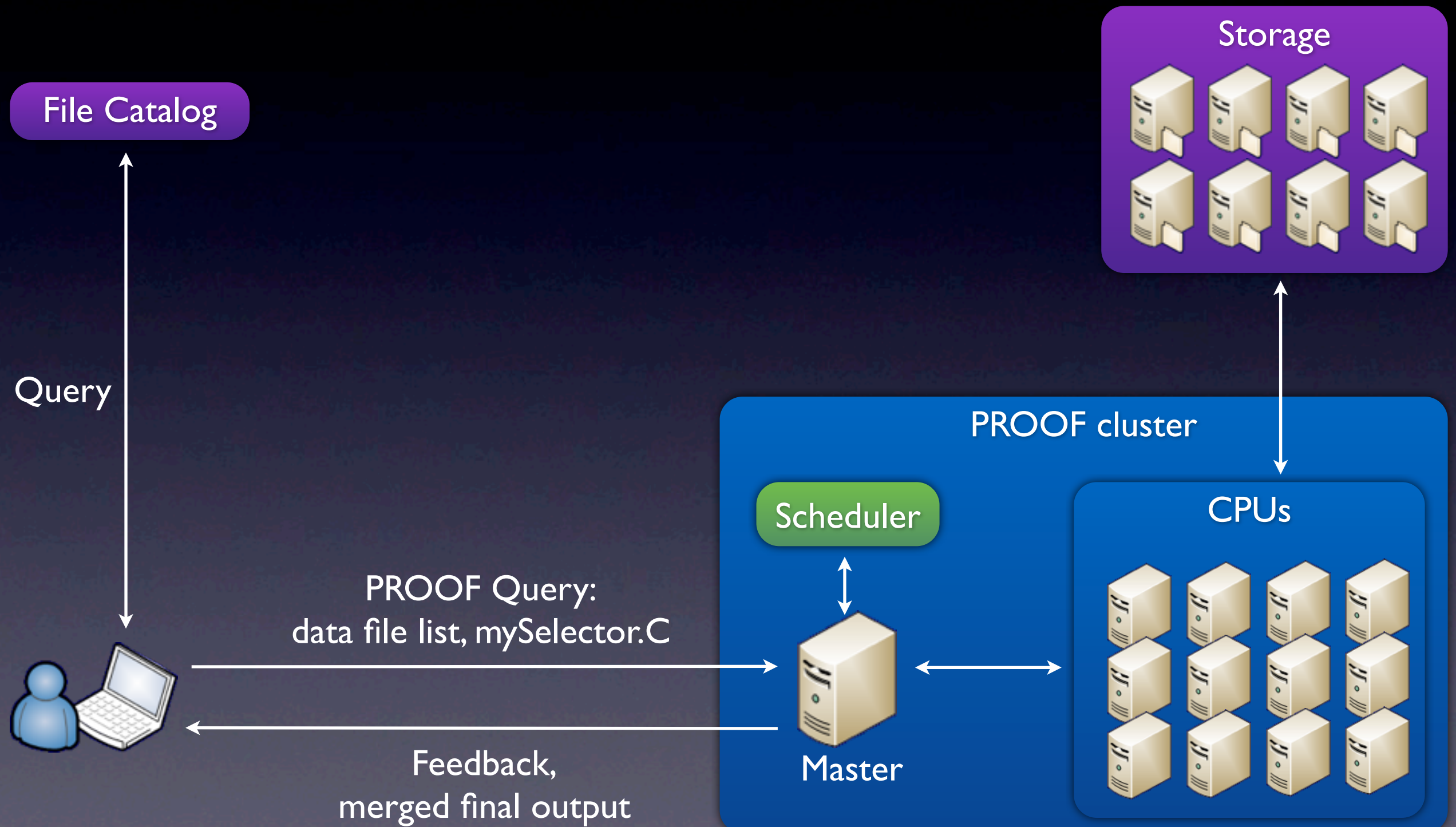(aka separation of business logic and infrastructural stuff)

Tools

HEP Data Analysis

Implement algorithm → Run over data set → Make improvements

Typical HEP analysis needs a continuous algorithm refinement cycle

PROOF

File Catalog

Storage

Query

PROOF cluster

PROOF Query:
data file list, mySelector.C

Scheduler

CPUs

Master

Feedback,
merged final output

13

Motivation

cons of a "static"/pre-installed PROOF cluster

- One user can disturb other users.

- From time to time admin. interventions are needed.

- There is only one ROOT/xrootd version for PROOF services.

- There is a master node limitation.

PoD

PoD vs "Static" approach

User

- can entirely control his/her dynamic cluster,

- can setup and use it on demand,

- can dynamically change an amount of workers,

- can select a preferable master host,

- doesn't need admins to take an action,

- doesn't disturb other users,

- is free to choose a ROOT version for services.

PoD vs VM based "static"

PoD: a "Green-PC" concept

Data location

For the dynamic analysis model is the best to keep data remote from WNs.

Interesting study:

ISGC2011, "Investigation of storage options for scientific computing on Grid and Cloud facilities" (Fermi National Accelerator Laboratory, US)

http://www2.twgrid.org/event/isgc2011/slides/GridsandClouds/2/Storage-Evaluation-ISGC-2011.pdf
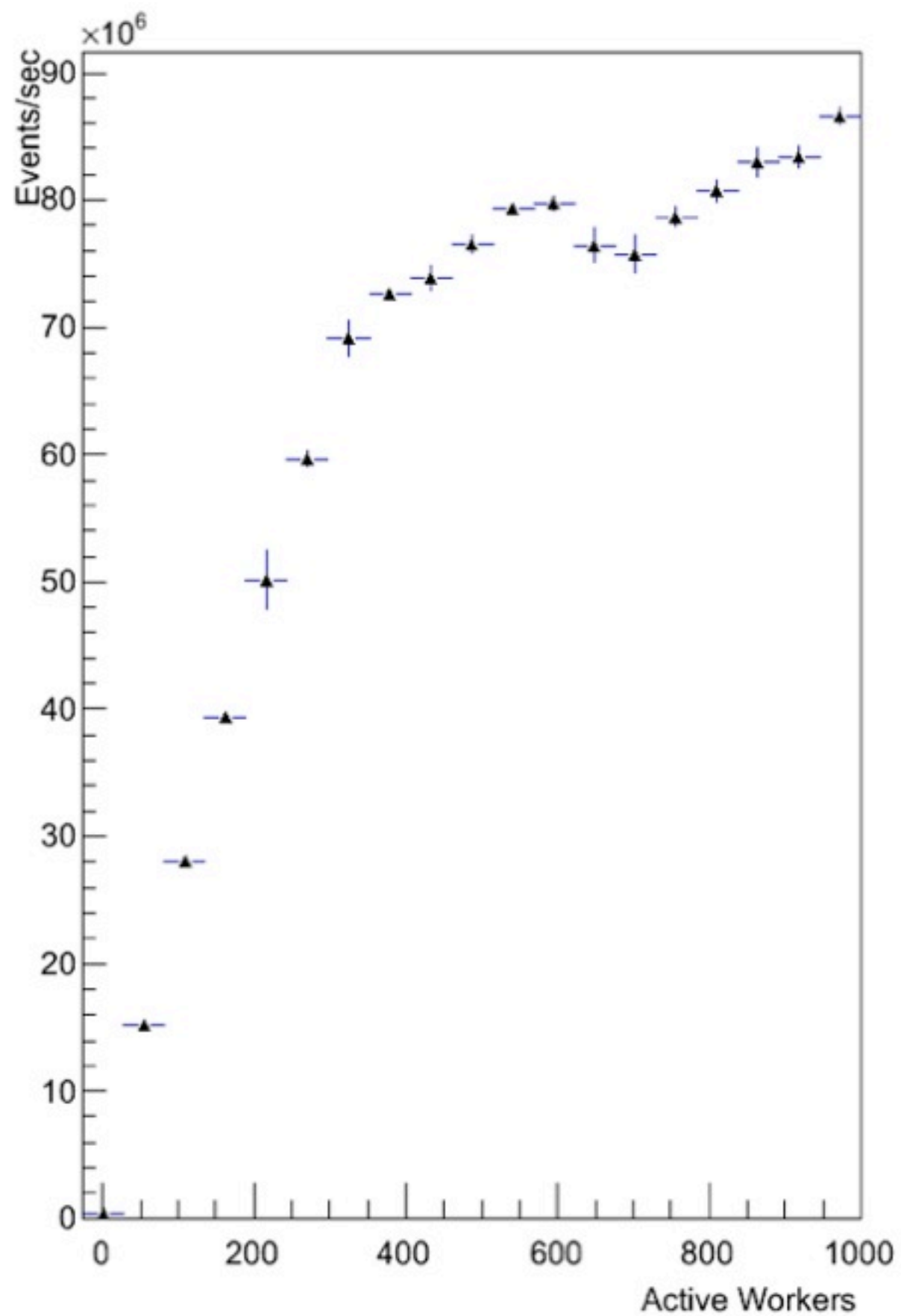
Some examples

These kinds of exercise you can't easily do without PoD!
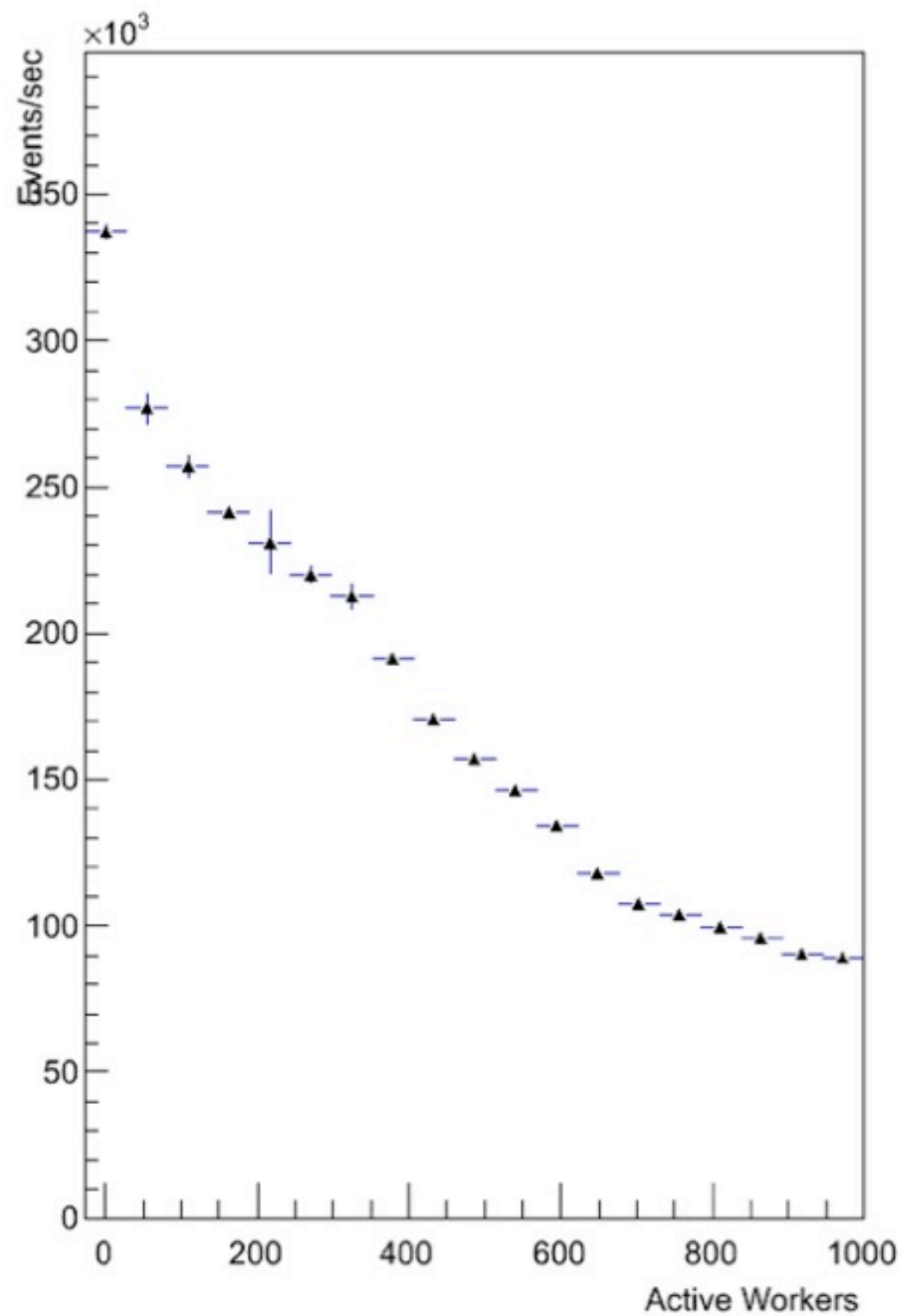
Some examples

For tests we use the new PROOF Benchmark Suite.

BTW, the suite is a very good example of everything I said above about software.

Profile CPU QueryResult Event - TSelHist_Hist1D

Profile CPU Normalized QueryResult Event - TSelHist_Hist1D

Summary