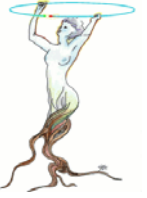




Status of PROOF

G. Ganis / CERN

Application Area meeting, 24 May 2006



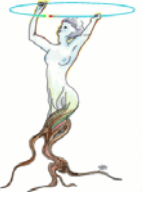
Outline

- Reminder about PROOF
- Recent developments and status
- Near future
- Testing at CAF
- Quick Demo

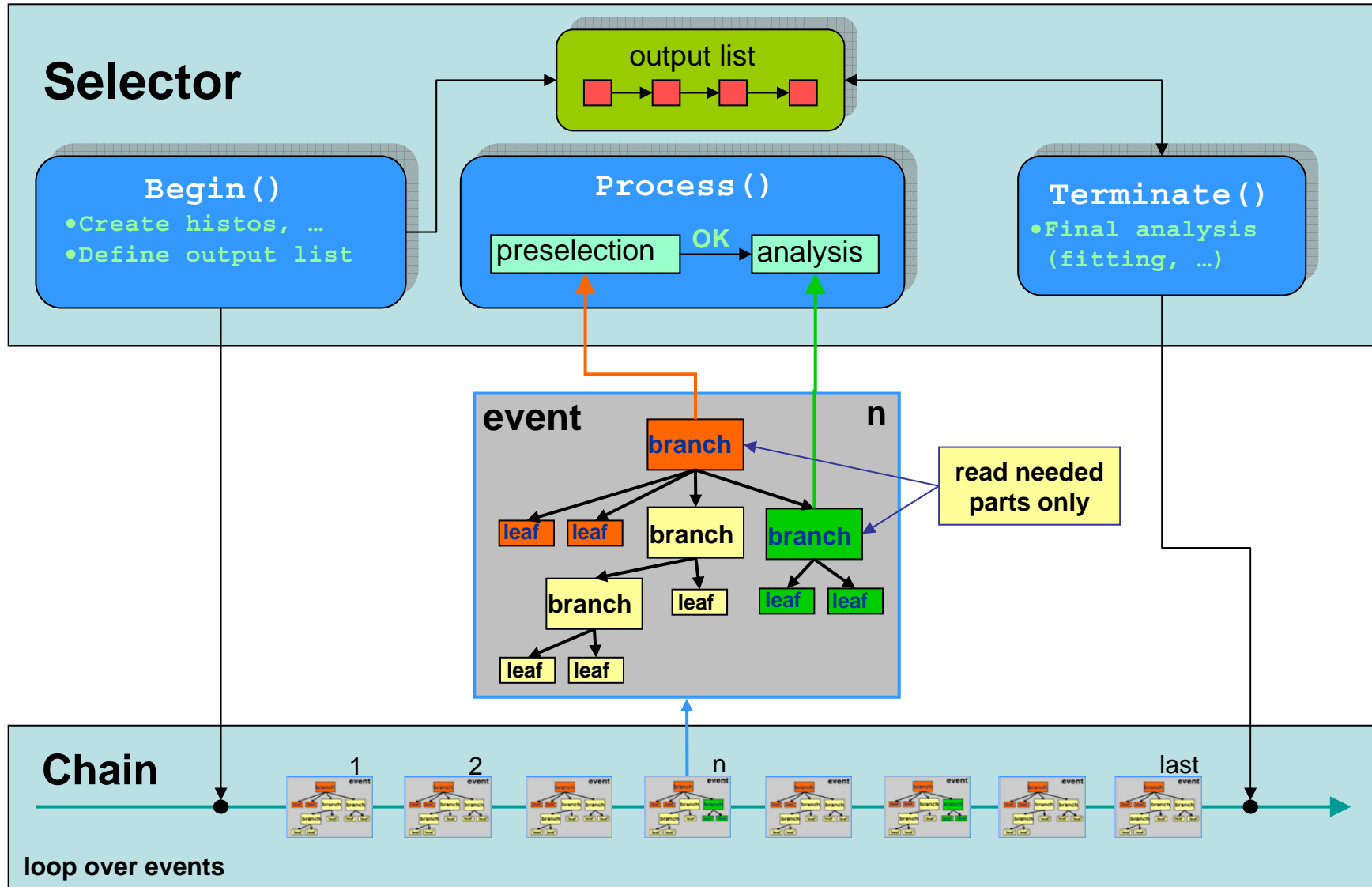


Outline

- **Reminder about PROOF**
- Recent developments and status
- Near future
- Testing at CAF
- Quick Demo

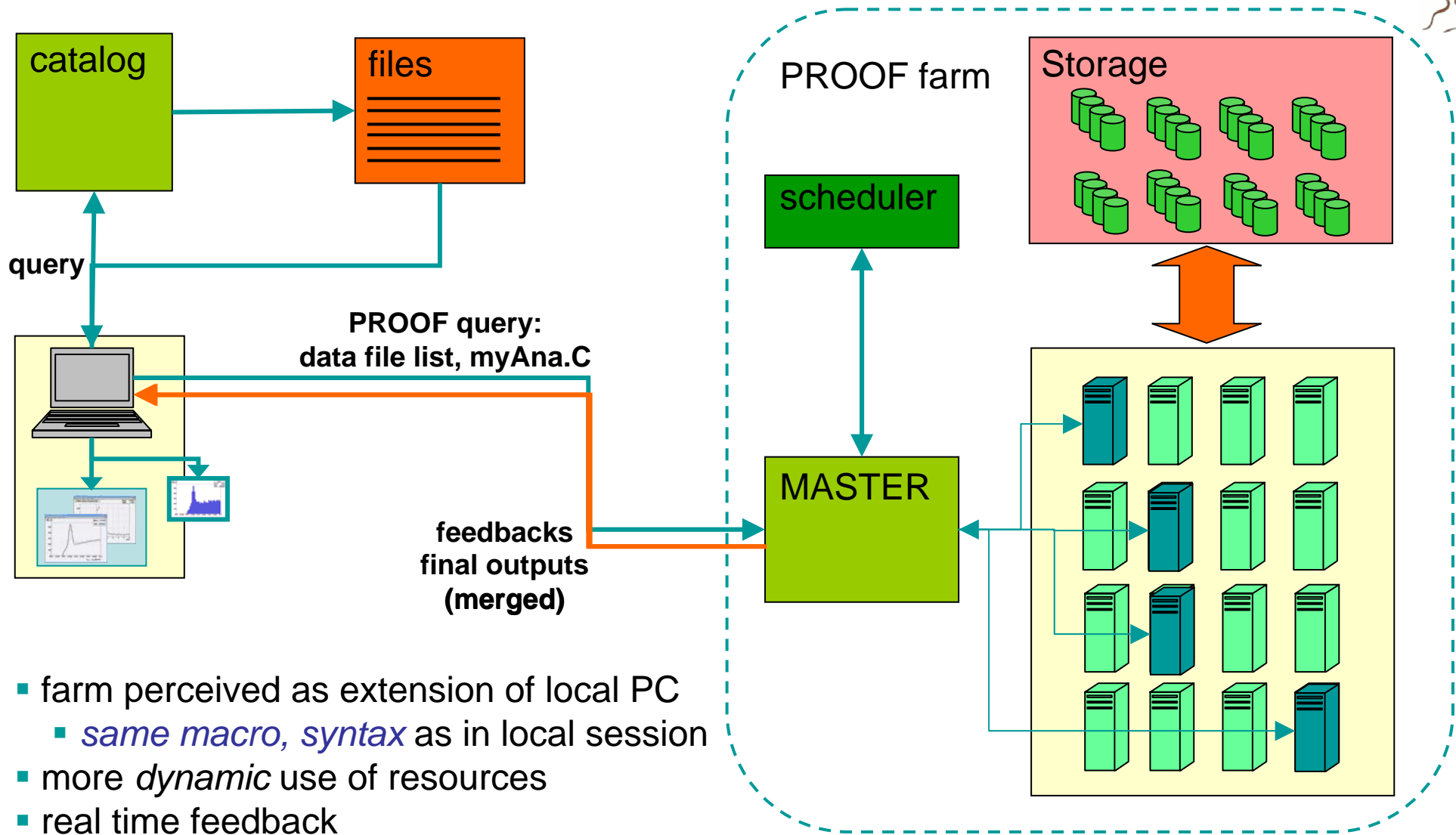


The ROOT data model: Trees & Selectors





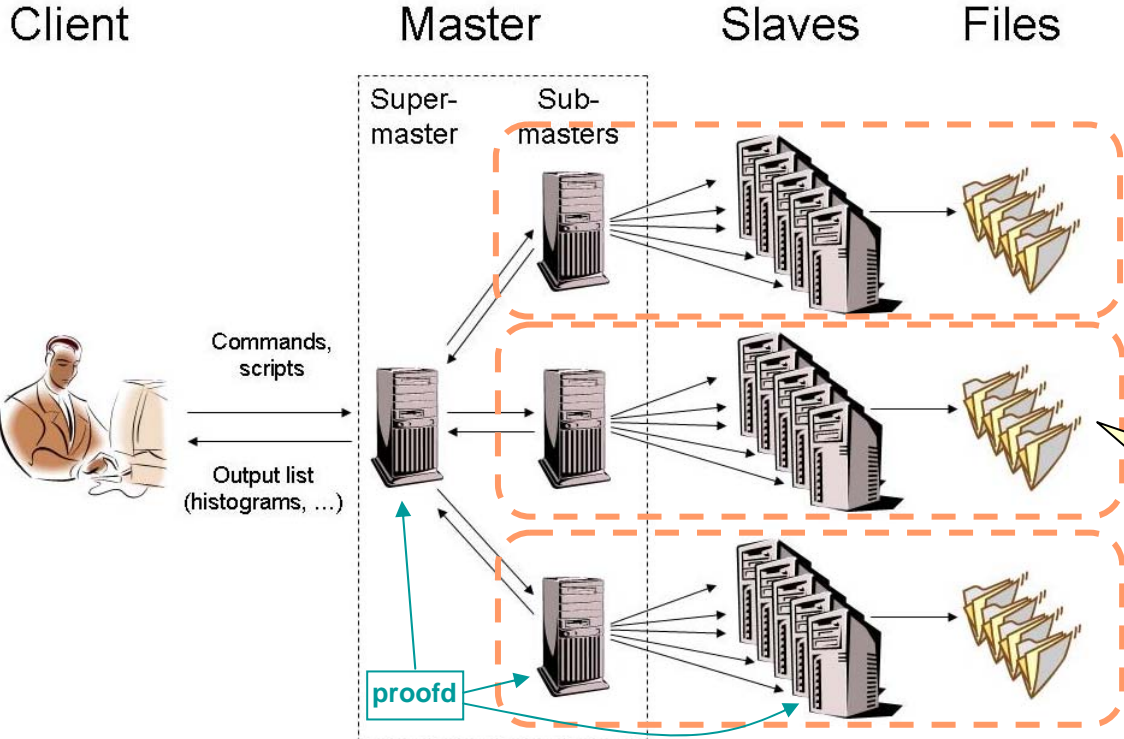
The PROOF approach



- farm perceived as extension of local PC
 - *same macro, syntax* as in local session
- more *dynamic* use of resources
- real time feedback
- automated *splitting* and merging



PROOF – Multi-tier Architecture



adapts to cluster of clusters or wide area *virtual clusters*

Geographically separated domains; heterogenous machine types

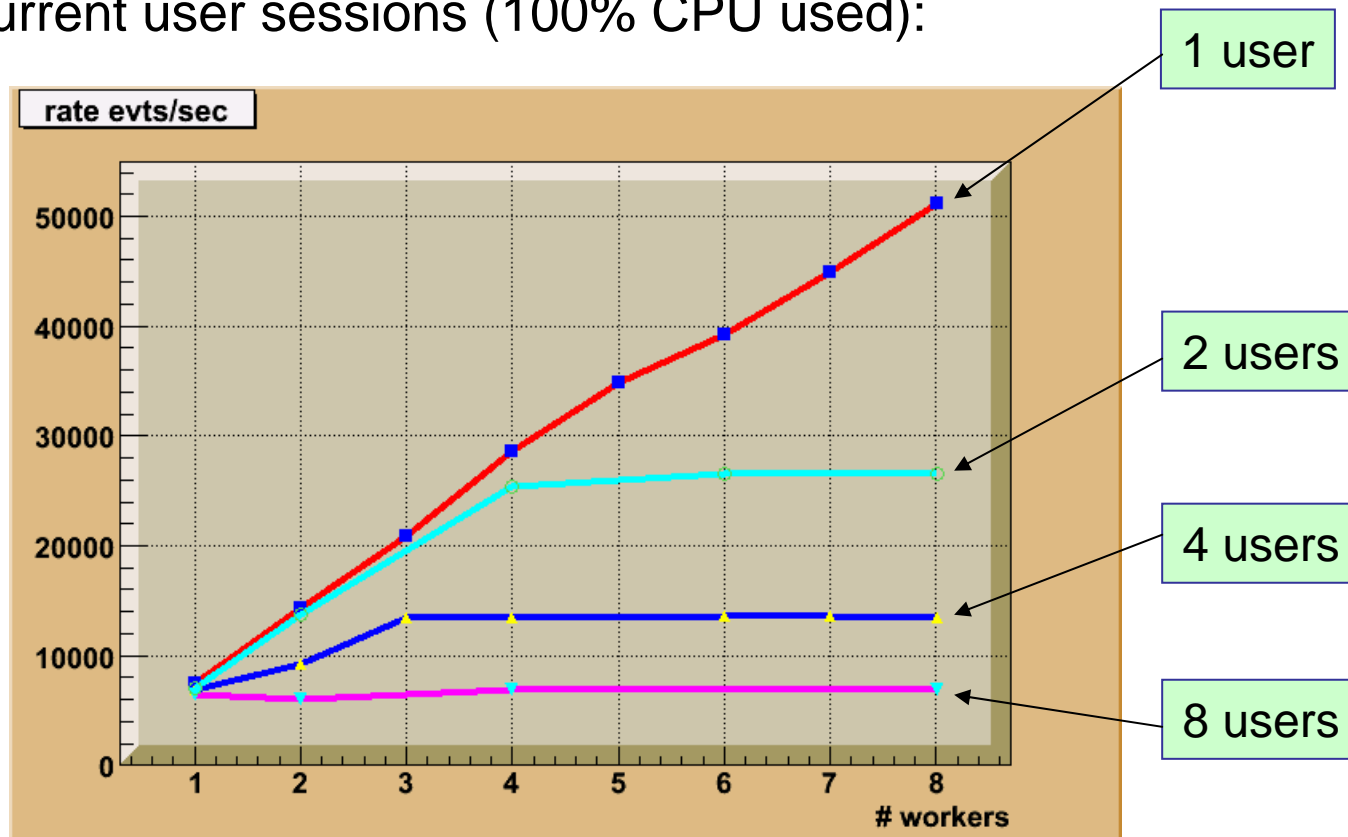
less important $\xrightarrow{\text{good connection ?}}$ VERY important

Optimize for **data locality** or efficient data server access



PROOF – Scalability

- CAF, 4 dual Xeon machines
- CMS selector, 120 MB data (290 files), distributed locally
- Strictly concurrent user sessions (100% CPU used):

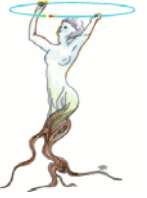


- No inefficiency introduced by PROOF internals



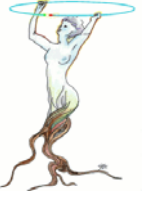
Outline

- Reminder about PROOF
- Recent developments and status
- Near future
- Testing at CAF
- Quick Demo



Recent developments

- Goals:
 - support for interactive-batch mode
 - stateless connection
 - multi-sessions
 - user-friendliness
 - management tools
 - GUI



Sample of analysis activity

AQ1: 1s query produces a local histogram
AQ2: a 10mn query submitted to PROOF1
AQ3->AQ7: short queries
AQ8: a 10h query submitted to PROOF2

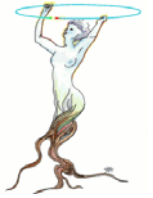
Monday at 10h15
ROOT session
on my laptop

BQ1: browse results of AQ2
BQ2: browse temporary results of AQ8
BQ3->BQ6: submit 4 10mn queries to PROOF1

Monday at 16h25
ROOT session
on my laptop

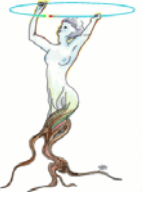
CQ1: Browse results of AQ8, BQ3->BQ6

Wednesday at 8h40
Browse from any
web browser



New connection layer based on XROOTD

- Interactive batch requires a **coordinator** on the server side
- Candidate: XROOTD
 - light weight top component (networking, protocol handler)
 - new protocol implemented as a plug-in to launch and control PROOF server sessions
- Non-destructive disconnections handled naturally
 - stateless connection
- Xrd/olbd control network can be exploited to circulate information
- Can use same daemon for data serving and PROOF serving



PROOF management tools

- Data sets
 - distribution of data files on local worker pools
 - by direct upload
 - by staging out from a mass storage (e.g. CASTOR)

- Query results
 - classification and handling tools
 - retrieve, archive

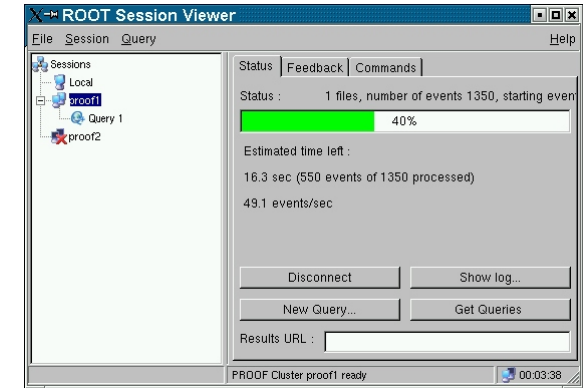
- Packages
 - optimized upload of additional libraries needed by the analysis

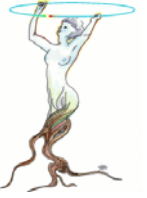
GUI controller



Allows full *on-click* control on everything

- define a new session
- submit a query, execute a command
- query editor
 - execute macro to define or pick up a **TChain**
 - browse directories with selectors
- online monitoring of feedback histograms
- browse folders with results of query
- retrieve, delete, archive functionality
- **start viewer** for fast **TChain** browsing





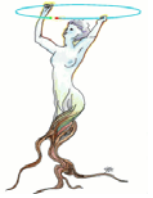
Outline

- Reminder about PROOF
- Recent developments and status
- **Near future**
- Testing at CAF
- Quick Demo



Near Future Plans

- Data access (see next)
- Multi-user scheduling (see next)
- **Packetizer** optimizations
 - re-assignment of being-processed packets to fast idle workers
- **Dynamic cluster configuration**
 - come-and-go functionality for worker nodes
 - *olbd* network to get info about the load on the cluster
- **Improve handling of error conditions**
 - identify cases hanging the system, improve error logging, ...
 - exploit *olbd* control network for better overview of the cluster
- **Testing and consolidation**
- **Monitoring of cluster behaviour**
 - MonAlisa: allows definition of *ad hoc* parameters, e.g. I/O / node / query



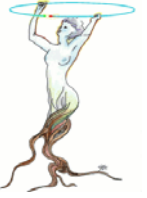
PROOF: data access issues

- **Low latency in data access is essential**
- **File opening overhead**
 - minimized using asynchronous open techniques
- **Data retrieval**
 - caching, asynchronous pre-fetching of data segments to be analyzed
- Asynchronous features supported by XROOTD / XrdClient
- Plan of work to fully exploit this in TFile / TSelector using the knowledge available in TTree (L. Franco)



PROOF: multi-user scheduling issues

- New *scheduler* component being developed to control the use of available resources in multi-user environments (J. Iwaszkiewicz)
- Decisions taken on per query base following a metric based on:
 - load of the cluster
 - resources need by the query
 - user history and priorities
 - ...
- Requires support for dynamic re-configuration of worker assignments (prototype being tested)
- Generic interface to external schedulers planned
 - Condor, LSF, ...



Outline

- Reminder about PROOF
- Recent developments and status
- Near future
- **Testing at CAF**
- Quick Demo



Testing at CAF

- CAF
 - 40 dual Xeon 2.8 GHz machines, 4 GB RAM, GB/s Ethernet
 - 215 GB disk pool
 - SLC4
- `ixcafdev`: 5 machines
 - testing developments
- `ixcaf`: 35 machine tested by ALICE (J.F. Grosse-Oetringhaus)
 - stress testing functionality
 - performance tests



Quick demo

- CMS test analysis selector
- 290 files (115 MB) distributed on lxcafdev

- Local run accessing files from lxcafdev
- PROOF run with 8 workers



A real PROOF session - query definition and running

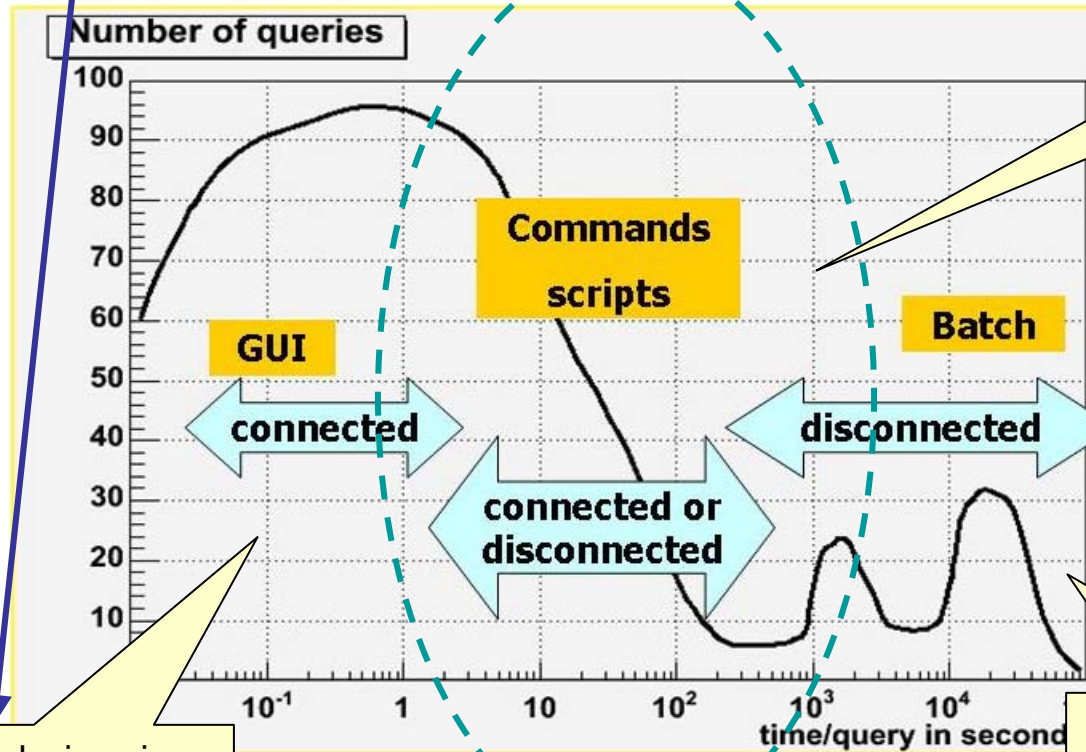
The screenshot displays the ROOT Session Viewer interface with several callouts:

- name**: Points to the 'Query Name' field containing 'Query 1' in a 'Query Dialog' window.
- Select chain**: Points to a 'Chain Selection' dialog window.
- Execute to create chain**: Points to a button in the 'Chain Selection' dialog.
- Feedback histograms**: Points to a bar chart titled 'Events processed per Slave' showing four bars with values approximately 21000, 17000, 17000, and 21000.
- Choose selector**: Points to a 'Choose selector' dialog window.
- Processing information**: Points to a progress bar and text indicating '31%' completion, 'Estimated time left : 13.1 sec', and 'Processing Rate : 14821.3 events/sec'.
- Open/Cancel buttons**: A red box highlights the 'Open' and 'Cancel' buttons in a dialog window.

The main ROOT Session Viewer window shows a tree view with 'Query 1' selected under the 'lxb0130-4' session. The status bar at the bottom indicates 'PROOF Cluster lxb0130-4 ready' and a timer at '00:06:41'.

Typical end-user job-length distribution

Goal: bring these to the same level of perception



Medium term jobs, e.g. analysis design and development using also non-local resources

Interactive analysis using local resources, e.g.
- end-analysis calculations
- visualization

Analysis jobs with well defined algorithms (e.g. production of personal trees)



A real PROOF session - connection

Define new session

Predefined session

```
Session E
pcepsft4
pcepsft43:~ $
pcepsft43:~ $
pcepsft43:~ $
pcepsft43:~ $
pcepsft43:~ $ root -l
root [0] TProof::Open()
```

Session startup progress bar

ROOT Session Viewer

Sessions

- Local
- lxb0130-4

ready

ROOT Session Viewer

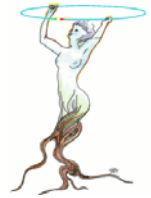
Status | Commands | Packages | Options

*** Connected to lxb0130.cern.ch (parallel mode, 4 workers) ***

Port number :	11093
User :	ganis
Client protocol version :	9
Remote protocol version :	9
Log level :	0
Session unique tag :	0-lxb0130-1138917966-12630
Total MB's processed :	0.00
Total real time used (s) :	0.070
Total CPU time used (s) :	0.000

New Query... Get Queries Show log...

PROOF Cluster lxb0130-4 ready 00:01:18



A real PROOF session - package manager

The screenshot displays the ROOT Session Viewer interface. A yellow callout box labeled "Package tab" points to the "Packages" tab in the top window. The bottom window shows a file listing for the package cache:

```
total 4
drwxr-xr-x  3 ganis  sf          4096 Nov 10 15:34 AliESD
*** Package cache lxb0103.cern.ch:/home/ganis/proof/packages ***
total 124
drwxr-xr-x  3 ganis  sf          4096 Nov 10 15:33 AliESD
-rw-r--r--  1 ganis  sf       114953 Nov 10 15:32 AliESD.par
```

At the bottom of the interface, there are several control buttons: "Upload", "Enable", "Disable", "Clear", "Show packages", and "Show Enabled". The "Enable" button is highlighted with a red box and an arrow. Below these buttons is a checkbox labeled "Enable at session startup".

PAR (Proof ARchive)

- ROOT-INF directory, BUILD.sh, SETUP.C
 - Control setup of each worker



A real PROOF session: query browsing and finalization

The screenshot displays the ROOT Session Viewer interface. The main window shows a query result for 'Query No : 1' with the following details:

- Ref : "session-0-lxb0130-1138917966-12630"
- Selector : h1analysis
- Status : completed
- Started : Thu Feb 2 23:12:54 2006
- Real time : 21 sec (CPU time: 49.4 sec)
- Events : 283813 events (size: 70.694 MB)
- Rate : 13514.9 evts/sec
- Objects : 9 objects

A callout box labeled 'finalization' points to the 'Finalize' button in the bottom right of the main window. Below the main window, there are status bars indicating 'Query Result Ready for session-0-lxb0130-1' and 'PROOF Cluster lxb0130-4 ready'.

Two plots are overlaid on the interface:

- tauD0 plot:** Titled 'Fitted value of par[1]=p1', showing a scatter plot of data points with error bars. The y-axis ranges from 100 to 200. The plot includes a statistics box with Mean = 0.4503 and RMS = 1.011.
- h1analysis analysis plot:** Titled 'dm_d', showing a histogram of the invariant mass difference $m_{K\pi\pi} - m_{K\pi}$ in GeV/c^2 . The x-axis ranges from 0.13 to 0.17. The plot shows a sharp peak at approximately 0.145 GeV/c^2 . The statistics box indicates Mean = 0.1551 and RMS = 0.008494.



A real PROOF session: disconnection / reconnection

- Running sessions kept alive by server side coordinator
- Reconnection is much faster: no process to fork

The screenshot displays the ROOT Session Viewer application. The main window shows a tree view of sessions and queries. A callout box points to the 'Status' field in the query details, which reads 'completed'. Another callout box points to the 'Status' field, stating 'The query is now terminated'. The interface includes a menu bar (File, Session, Query, Options, Help), a toolbar, and a status bar at the bottom showing 'User : Gerri - sf' and '00:00:00'.

reconn

disc

The query is now terminated

Query No : 2
Ref : "session-0-pcepsft43-1138910936-27977:q"
Selector : hlanalysis
Status : completed

Started : Thu Feb 2 21:29:02 2006
Real time : 34 sec (CPU time: 17.7 sec)
Processed : 283813 events (size: 46.126 MBs)
Rate : 8347.4 evts/sec

Results : <PROOF SandBox>/queries/session-0-pcepsft43-1138910936-27977:q
Outlist : 9 objects

Retrieve Finalize Show Log

Query Result Ready for session-0-pcepsft43- PROOF Cluster lxb0130-4 ready 00:01:54

User : Gerri - sf 00:00:00

PROOF Cluster lxb0130-4 ready 00:06:41



A real PROOF session: chain viewer

Right-click

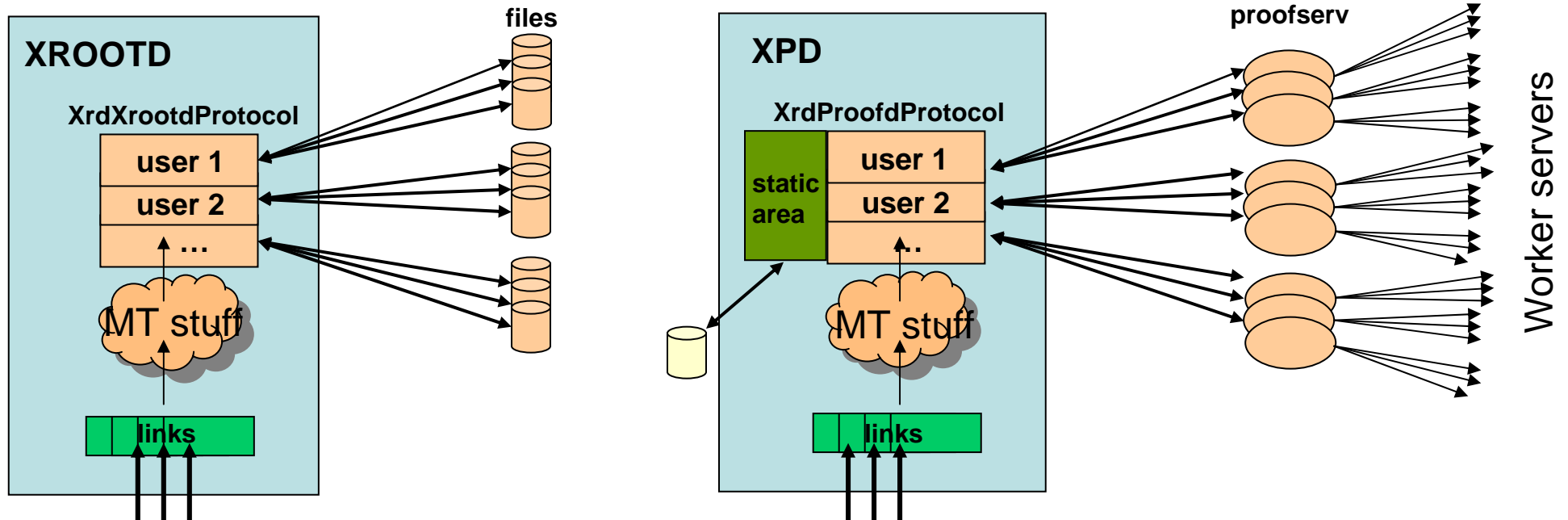
The screenshot shows the PROOF software interface. In the background, the 'PROOF Session Viewer' window displays a tree view of sessions, with 'Query 1' selected. A context menu is open over 'Query 1', showing options: Edit, Submit, Start Viewer (highlighted), and Delete. In the foreground, the 'Chain viewer' window is open, showing a tree structure of data files. The 'Chain viewer' window has a menu bar (File, Edit, Run, Options, Help) and a toolbar (Command, Option, Histogram, htemp, Hist, Scan, Rec). The main area is divided into two panes: 'Current Folder' and 'Content Tree : bg_filtered_Wra'. The 'Current Folder' pane shows a tree structure with 'TreeList' and 'bg filtered_Wrapped'. The 'Content Tree' pane shows a list of files with columns for X, Y, Z, and file names. The files listed are: X: -empty-, Y: -empty-, Z: -empty-, -empty-, Scan box, E: -empty-, E: -empty-, E: -empty-, E: -empty-, E: -empty-, E: -empty-, E: -empty-, E: -empty-, E: -empty-. The status bar at the bottom of the 'Chain viewer' window shows '0%' and 'First entry : 0 Last entry : 1349'. The status bar at the bottom of the 'PROOF Session Viewer' window shows 'Query Result Ready for session-0-lxb0130-1', 'PROOF Cluster lxb0130-4 ready', and '00:07:29'.

Chain viewer



XrdProofd basics

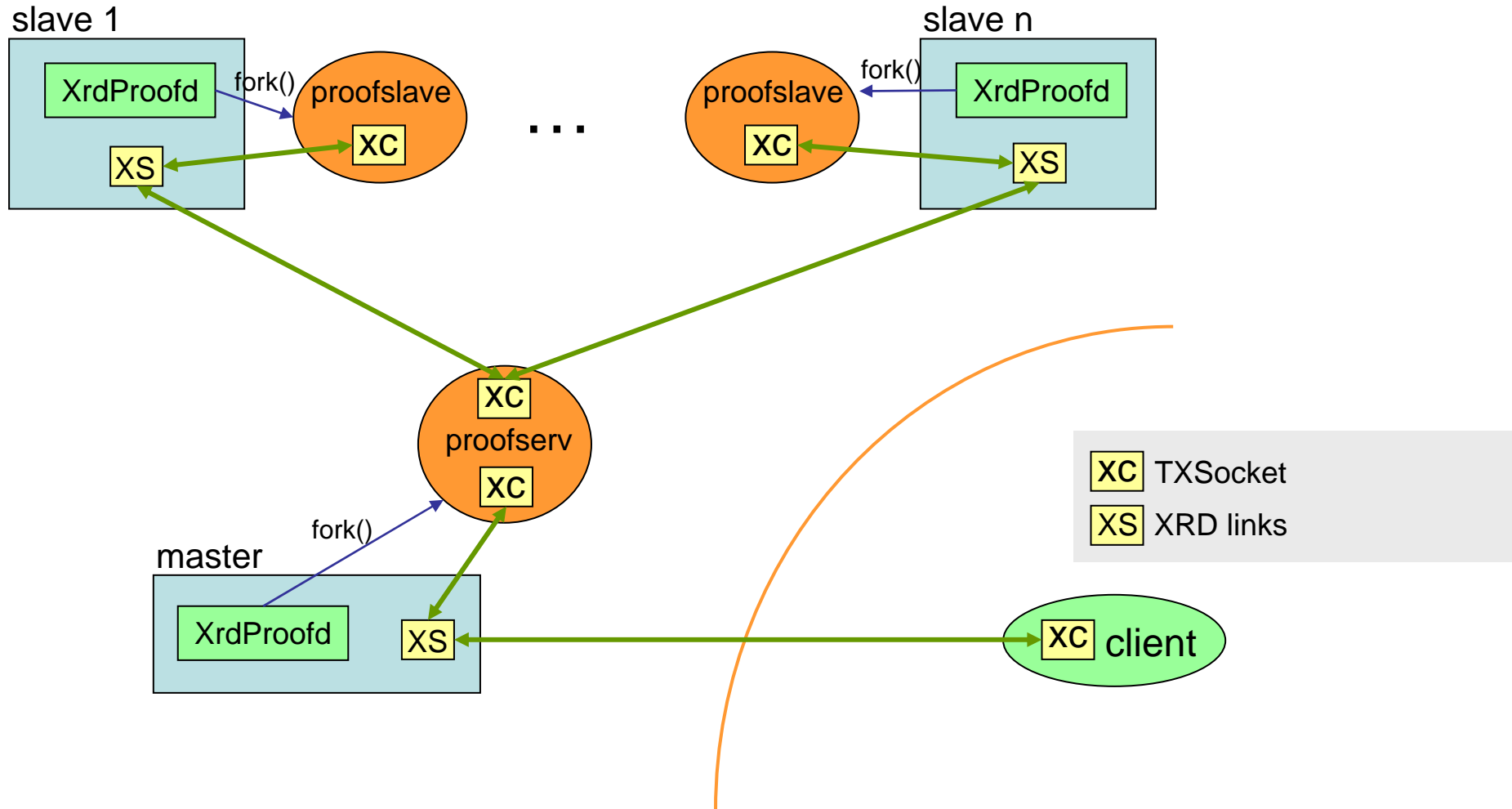
- Prototype based on XROOTD



- XrdProofdProtocol: client gateway to proofserv
- static area for all client information and its activities

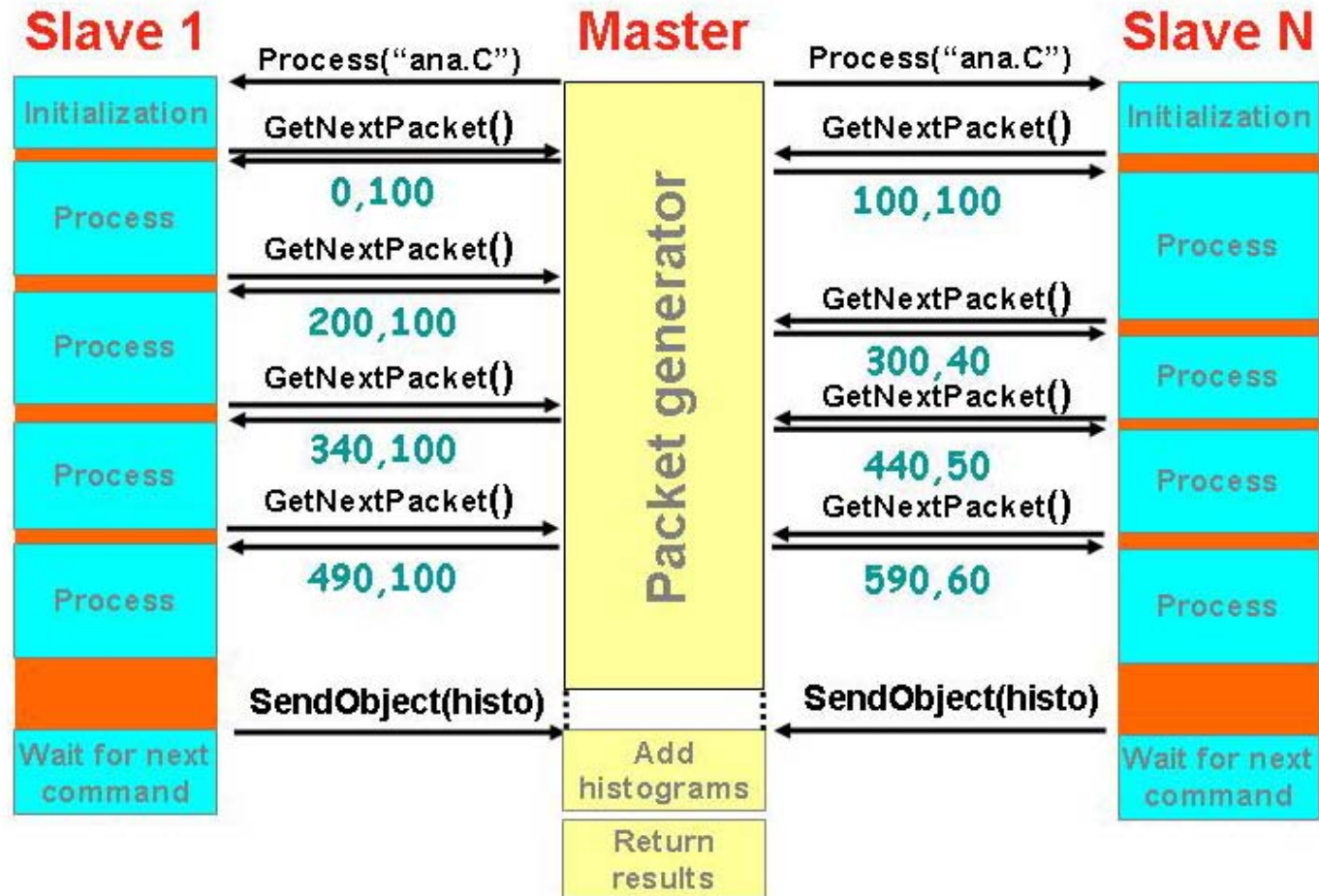


XrdProofd communication layer





Workflow: Pull Architecture



dynamic load balancing naturally achieved



PROOF @ AliEn: command-line session

- **TGrid**: abstract interface for all services

```
// Connect
TGrid *alien = TGrid::Connect("alien://");

// Query
TString path= "/alice/cern.ch/user/p/peters/analysis/miniesd/";
TGridResult *res = alien->Query(path, "*.root");

// Create chain from list of files
TChain chain("Events", "session", res->GetFileInfoList());

// Open a PROOF session
TProof *proof = TProof::Open("proofmaster");

// Process your query
chain.Process("selector.C");
```



PROOF @ GRID

