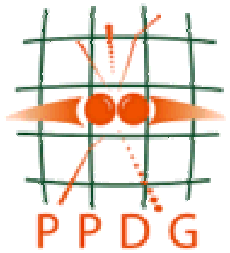


Interactive Data Analysis Tools:

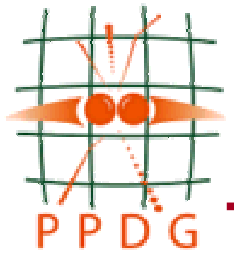
Report of PPDG CS-11 Activities

Rick Cavanaugh,

Ruth Pordes

CS-11 Goals, Scope

- Interface and Integrate Interactive Data Analysis Tools with the Grid
- Identify Common Components and Services
- Task oriented approach
  - Identify concrete tasks which can be accomplished in relatively short order (within months)
- Involves a broader HEP community than the LHC
  - ATLAS, BaBar, CMS, D0, etc
- Led by Doug Olson (LBL) and Joe Perl (SLAC)

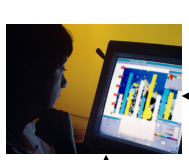


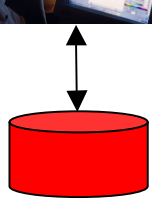
Four Models Considered for Interactive Analysis on the Grid

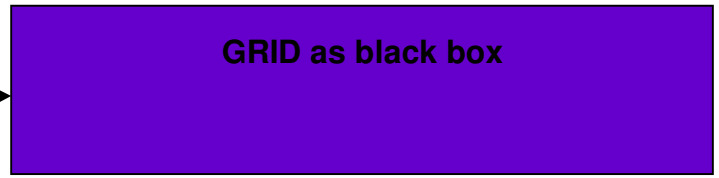
- Grid as a Black Box
- Real-time Batch
- Interactive Batch
- Pre-started Analysis Services

Grid as a Black-Box

- User performs queries for data from the grid
- Data is extracted and stored on local, non-grid resources
- Analysis performed without using the grid





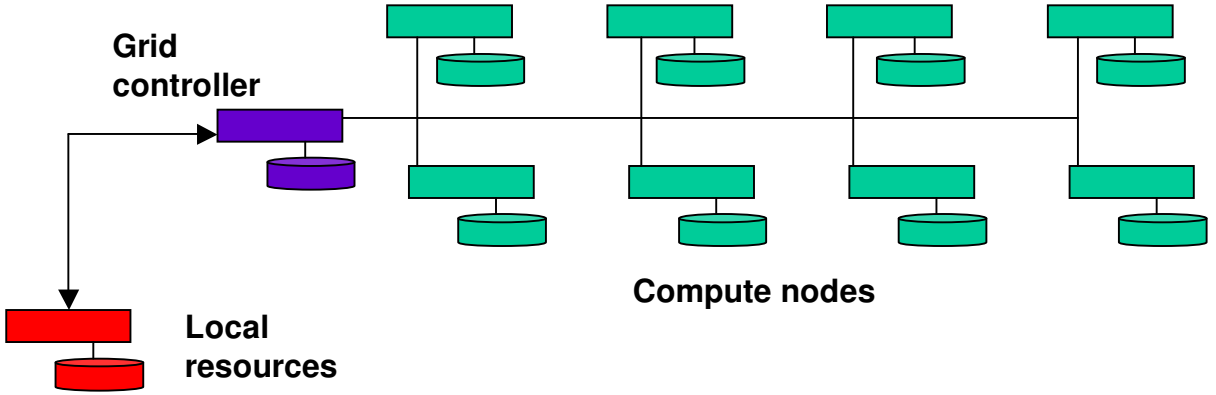


Select and extract data from grid,

Analyze using local resources without Further interaction with grid

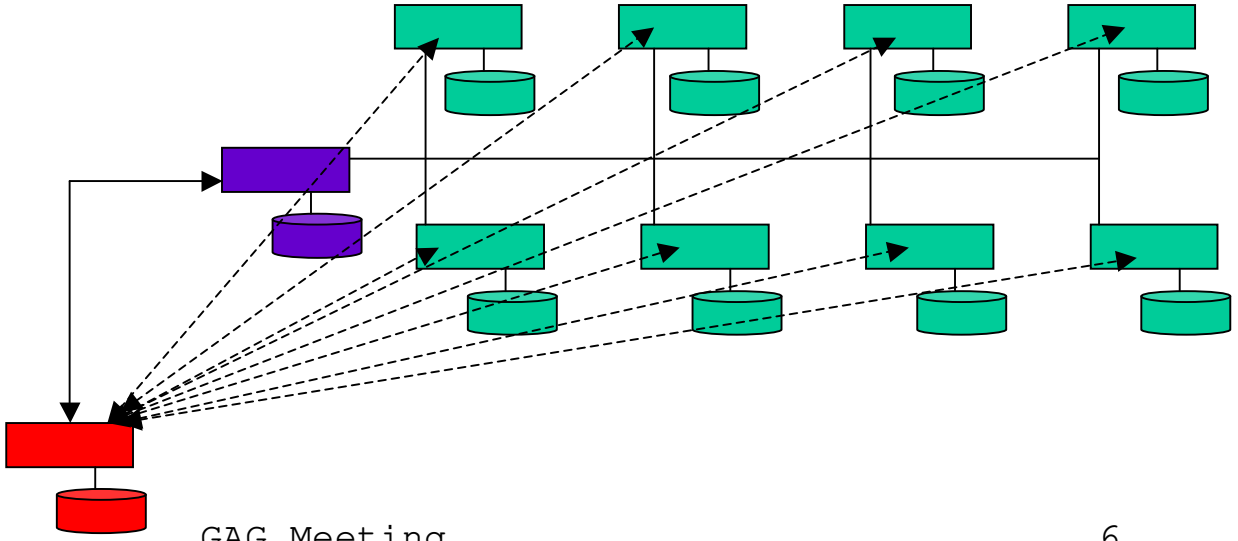
Real-time Batch

- Work is submitted to a grid scheduler which distributes it across grid nodes
- Intermediate results from the individual batch jobs are returned to the user
- No interaction with individual batch jobs



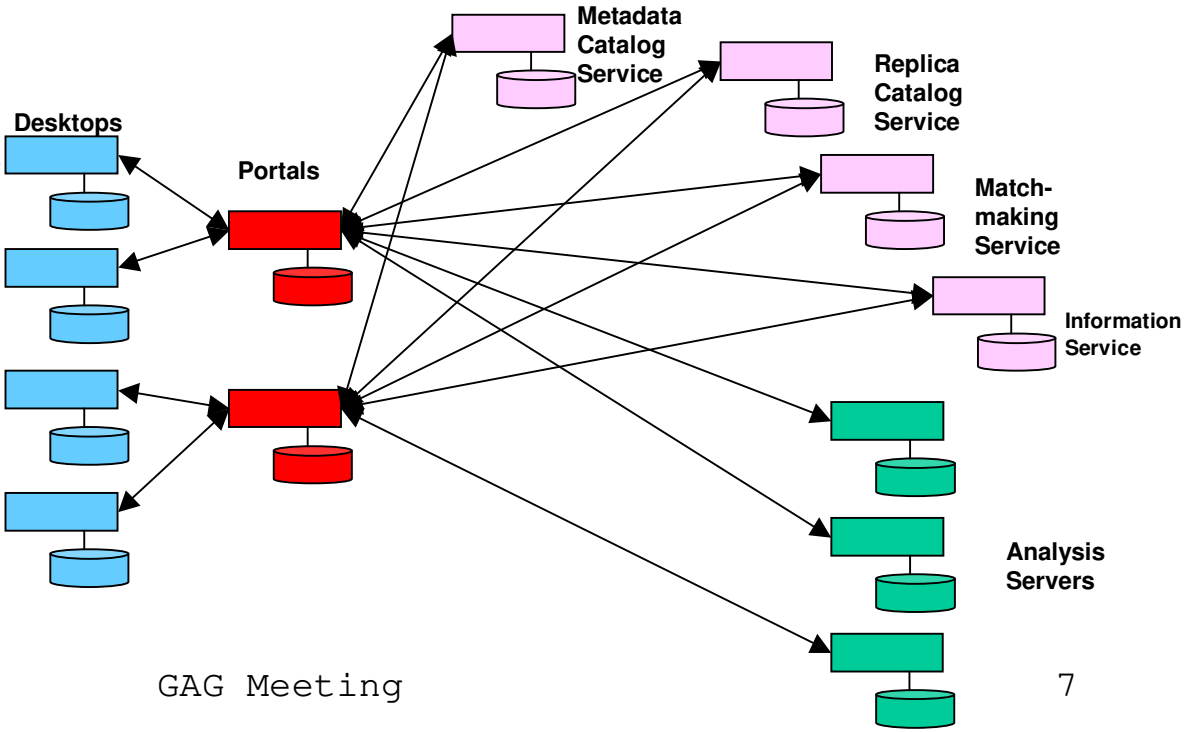
Interactive Batch

- Similar to real-time batch model with the addition of control channels between user and analysis jobs
- User can modify analysis jobs as they run
- User may have a rich desktop environment providing fine grained control and feedback



Pre-started Analysis Services

- Persistent analysis services run on pre-determined servers
- Capable of discovering grid resources and executing analysis tasks for users
- Early examples of such analysis services include
  - PROOF
  - Distributed JAS
- Does not necessarily have the view of a batch model

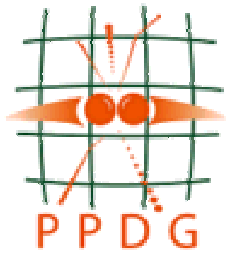


Interactive Data Analysis Use-cases and Requirements

- 23 (so far) identified use-cases for Grid Services
- Detailed in a PPDG CS-11 document

Current List of Requirements for Grid Services

- Select data
- Select subset of data
- Inspect data
- Move data
- Choose version of code
- Run mini-analysis
- Retrieve results
- Estimate resources
- Negotiate availability
- Run full-analysis
- Check status
- View results
- Suspend/resume analysis
- Abort analysis
- View results
- Display events
- Add refined data
- Share refined data
- Add tag data
- Compare results
- Calculate cross sections
- Maintain audit trail
- Security and access control

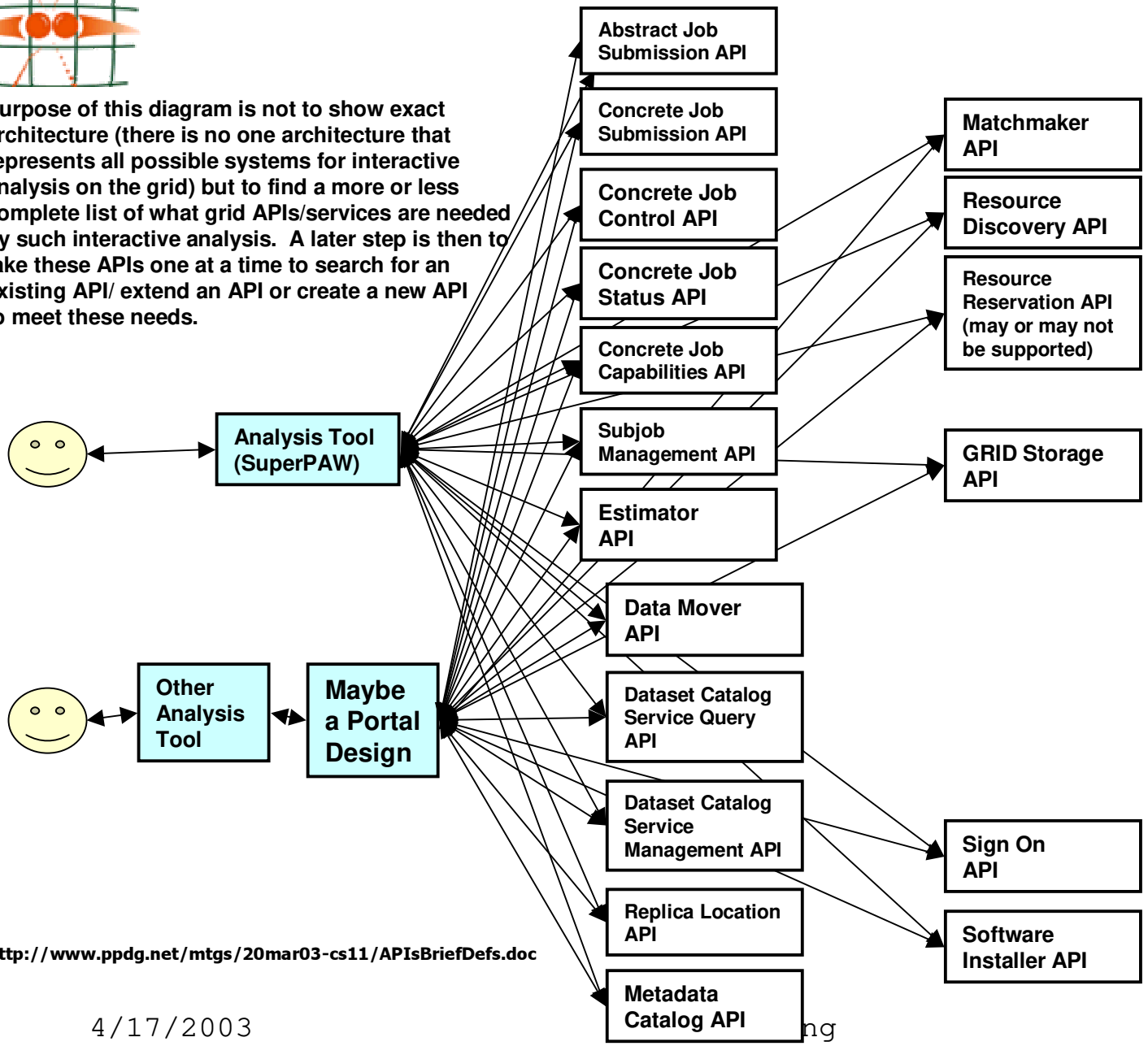


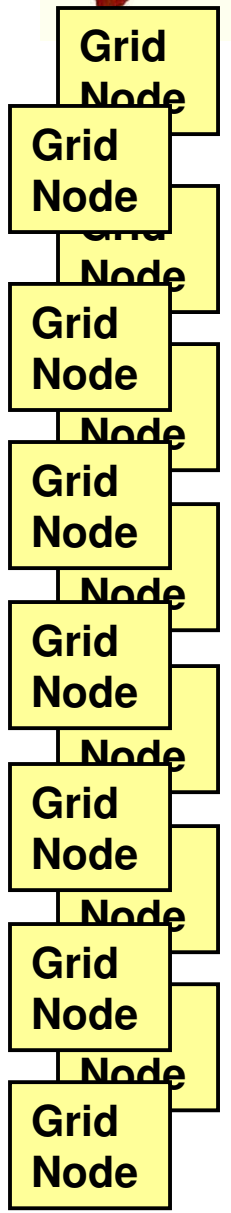
Definition of Interactive Data Analysis APIs

- 18 (so far) APIs identified
  - Abstract Job submission
  - Concrete Job Interaction, Control, Status, Capabilities
  - Data Mover
  - Etc.

APIs for Interactive Data Analysis using the GRID

Purpose of this diagram is not to show exact architecture (there is no one architecture that represents all possible systems for interactive analysis on the grid) but to find a more or less complete list of what grid APIs/services are needed by such interactive analysis. A later step is then to take these APIs one at a time to search for an existing API/ extend an API or create a new API to meet these needs.





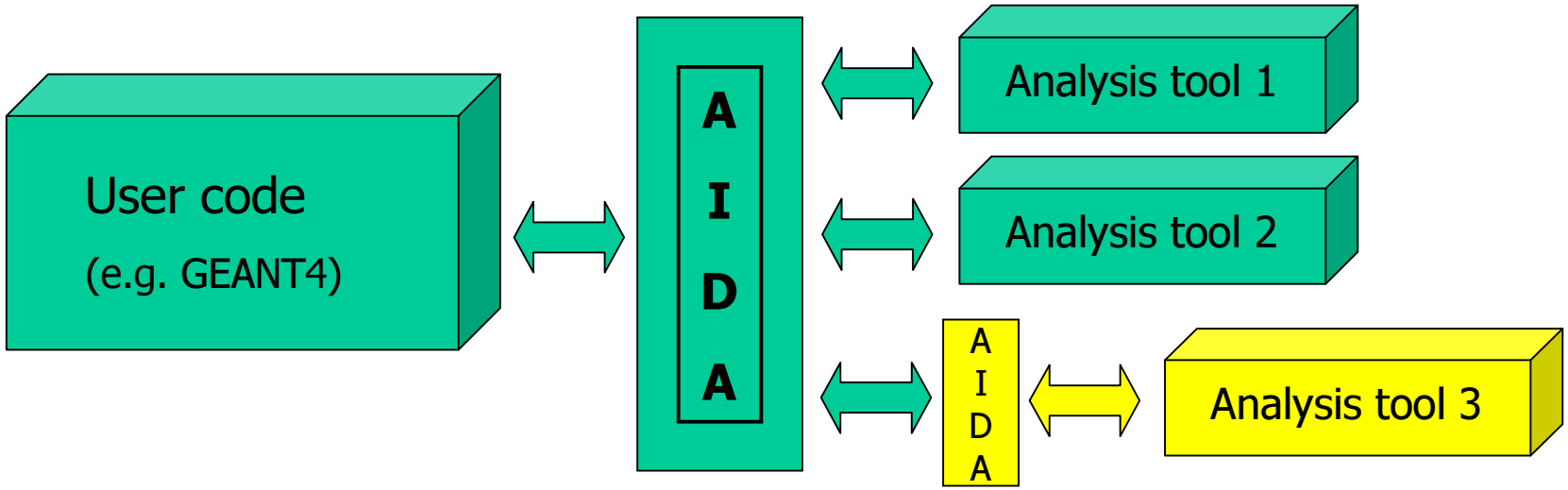
http://www.ppdg.net/mtgs/20mar03-cs11/APIsBriefDefs.doc

Joseph Perl and 20-21 March 2003 CS11 Workshop Participants

AIDA End User View

Abstract Interfaces for Data Analysis

- Use same code with any AIDA-compliant analysis tool.



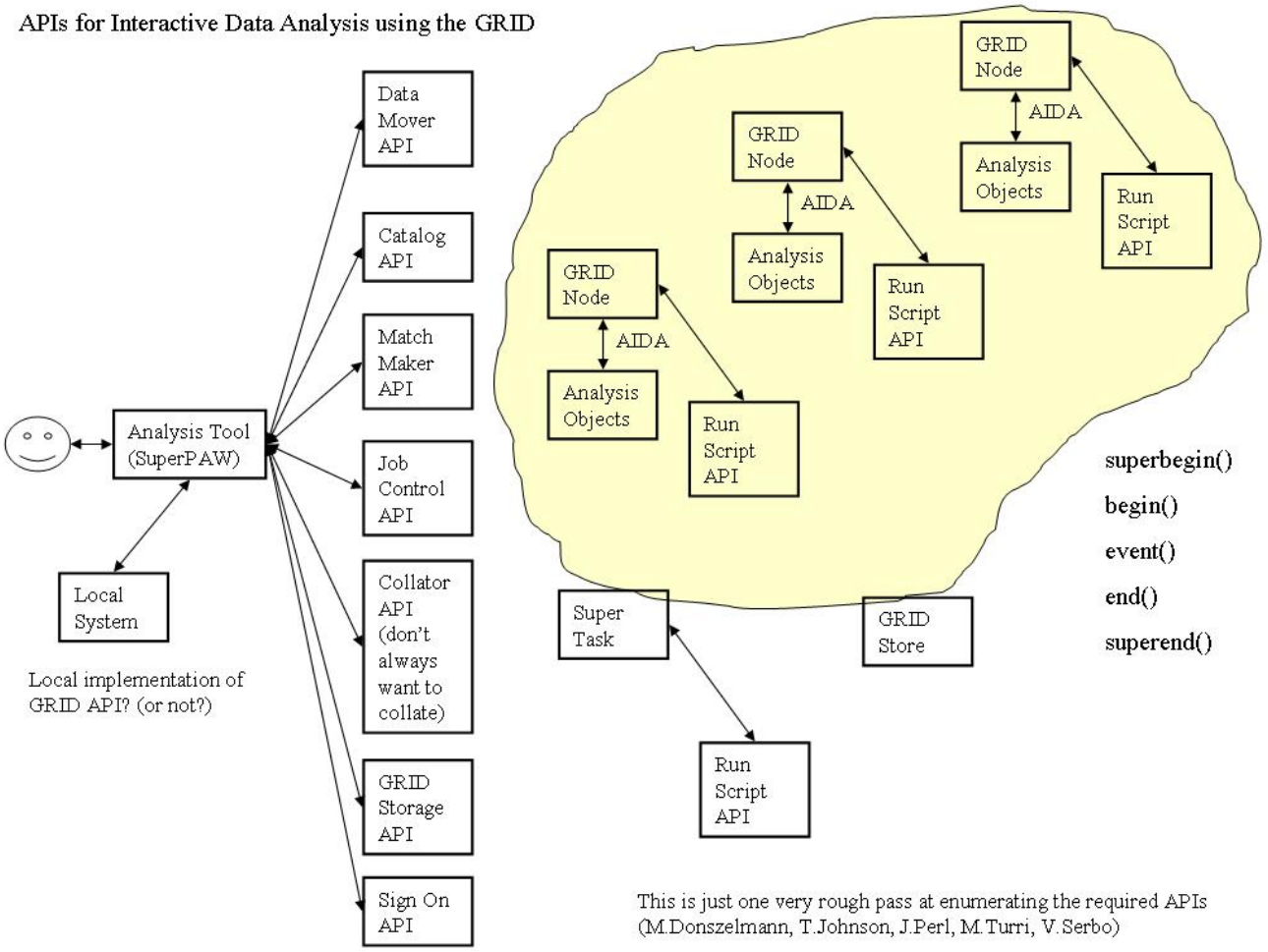
- User only has to learn one set of terminology
- Can use best tool for the job
- Can migrate to new tools as they become available
- Can interchange data between tools

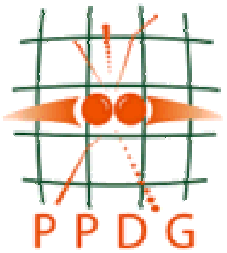
AIDA Developer View

Abstract Interfaces for Data Analysis

- Started with 3 separate teams with different goals, constraints and terminology
  - By collaborating on interfaces the AIDA group were able to
    - Separate design of user-interface from implementation
    - Harness combined experience of many people, while leaving individual teams to produce implementations that target their own needs
    - Generate tools that are better than any one team would have produced alone
    - Share IO formats, test suites, utilities, and ultimately components

Possible interfaces for Grid Data Analysis

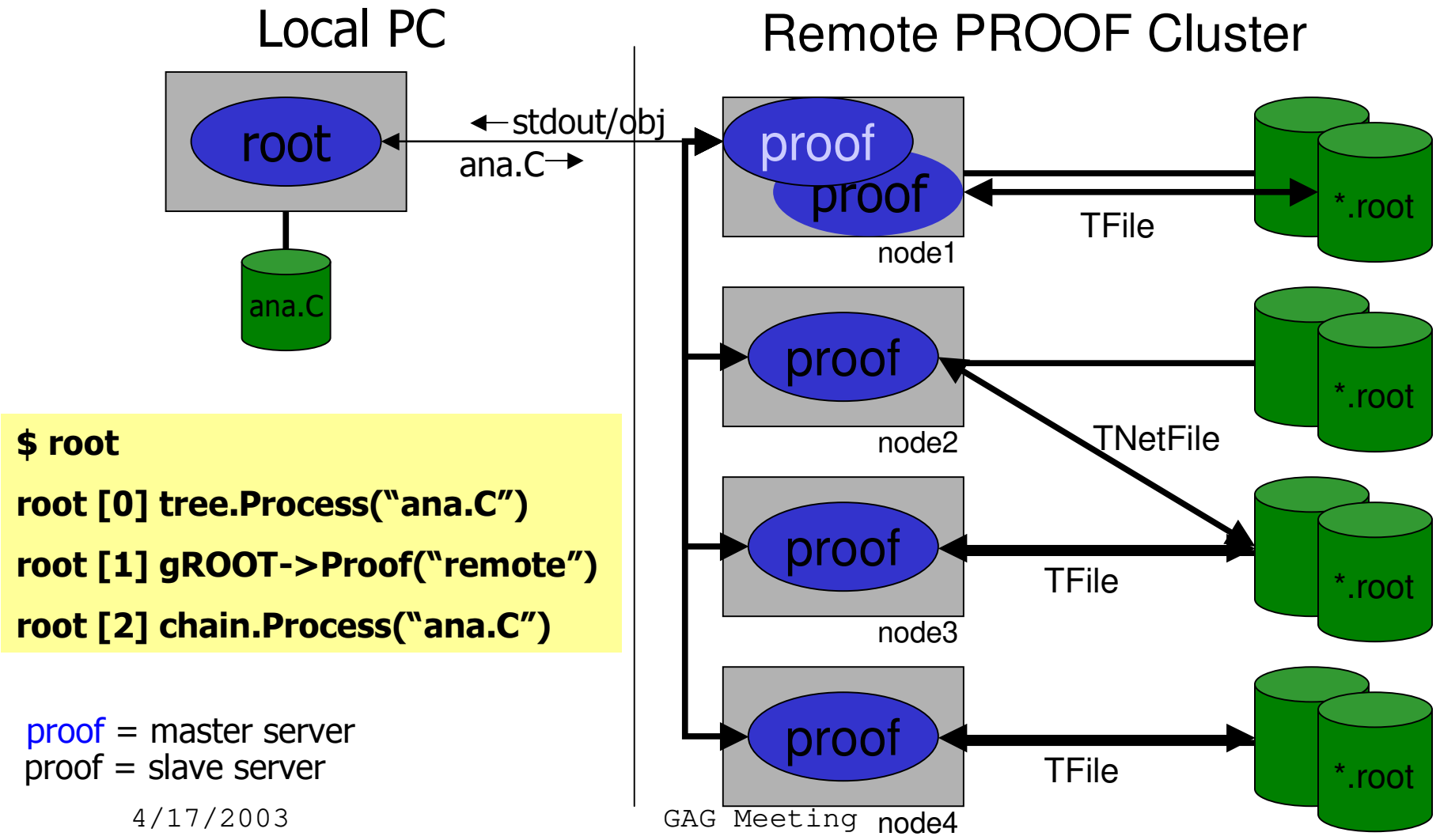




Examples of Participants in CS-11

- Package/Team
  - ROOT and PROOF
  - JAS and JAS w/ COG
  - Clarens and CMS Grid Enabled Analysis
  - Chimera
  - Grappa and Ganga
  - Dial
  - SAM
  - several others…
- Represents a diverse set of ideas and implementations

PROOF (CERN & MIT)



$ root

root [0] tree.Process("ana.C")

root [1] gROOT->Proof("remote")

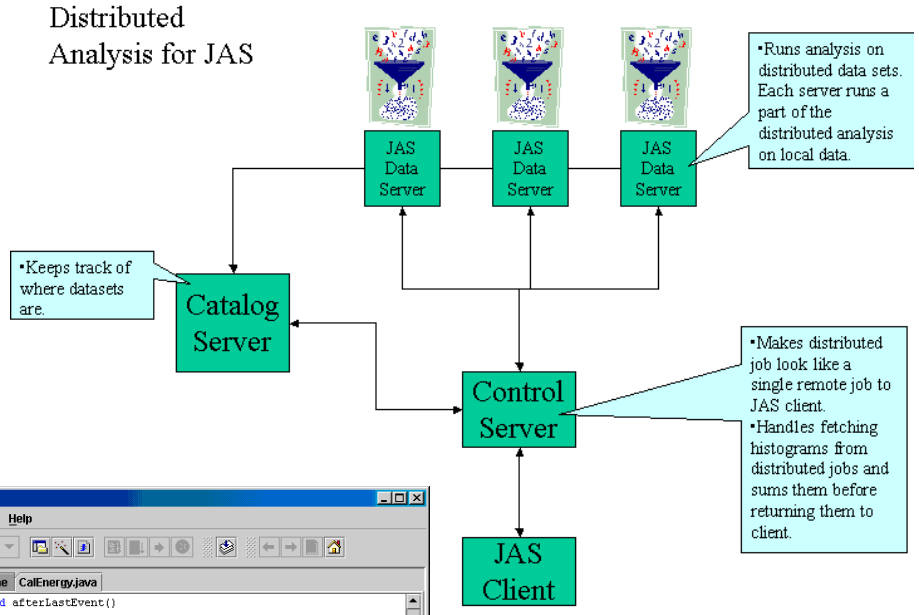
root [2] chain.Process("ana.C")

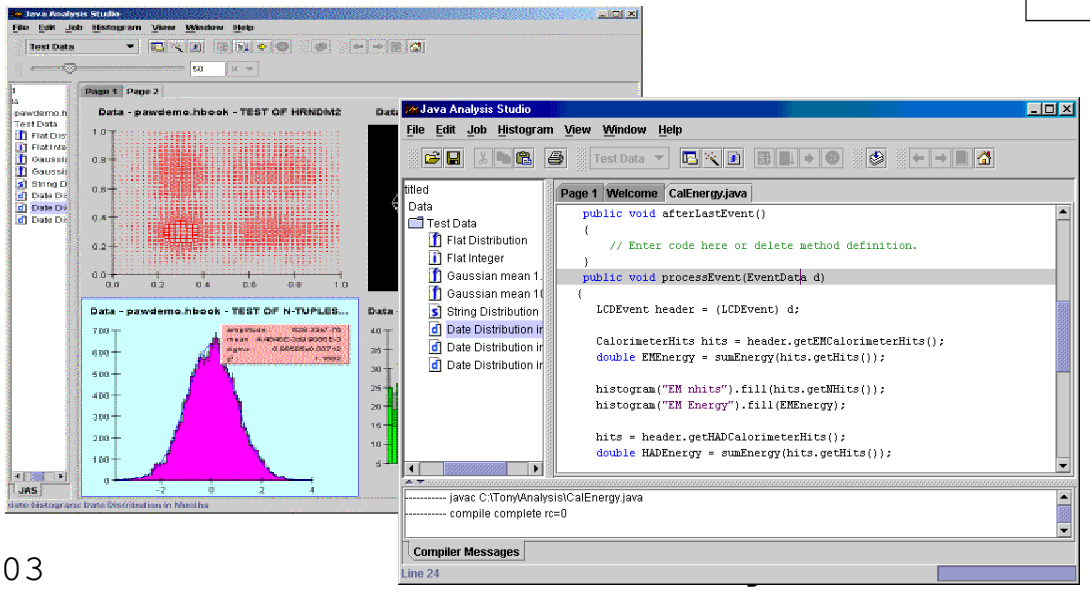
proof = master server
proof = slave server

This slide and next: Maarten Ballintijn and Fons Rademakers. Collaboration between ROOT and the MIT Heavy Ion Group

Distributed Java Analysis Studio (JAS)

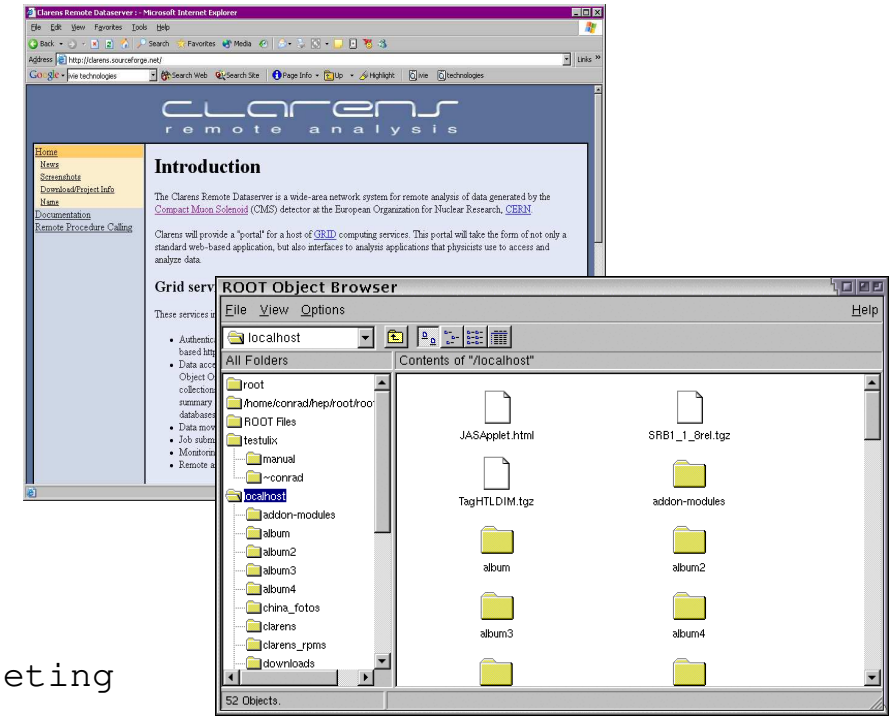
- Goal: clustered deployment, launch, & federation
- Minimal prerequisites:
  - Bare grid: Globus, Java, nothing else
  - Heterogeneous cluster
  - Off-grid (or not) client, data, codebase
  - Clients don't need to be superusers
- Optional background deployment
- Single sign on



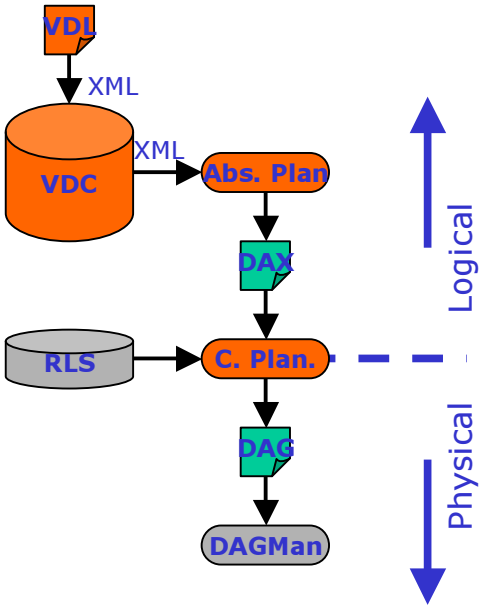


Clarens

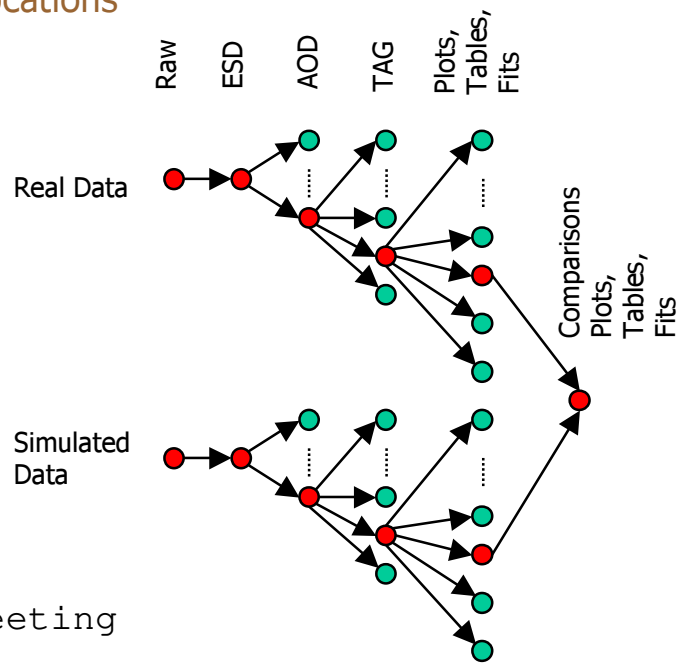
- The Clarens Remote Dataserver: a WAN system for remote analysis of data
- Clarens servers are deployed at Caltech, Florida, UCLA, UCSD, FNAL
- SRB now installed as Clarens service on Caltech Tier2 (Oracle backend)



The Chimera Virtual Data System



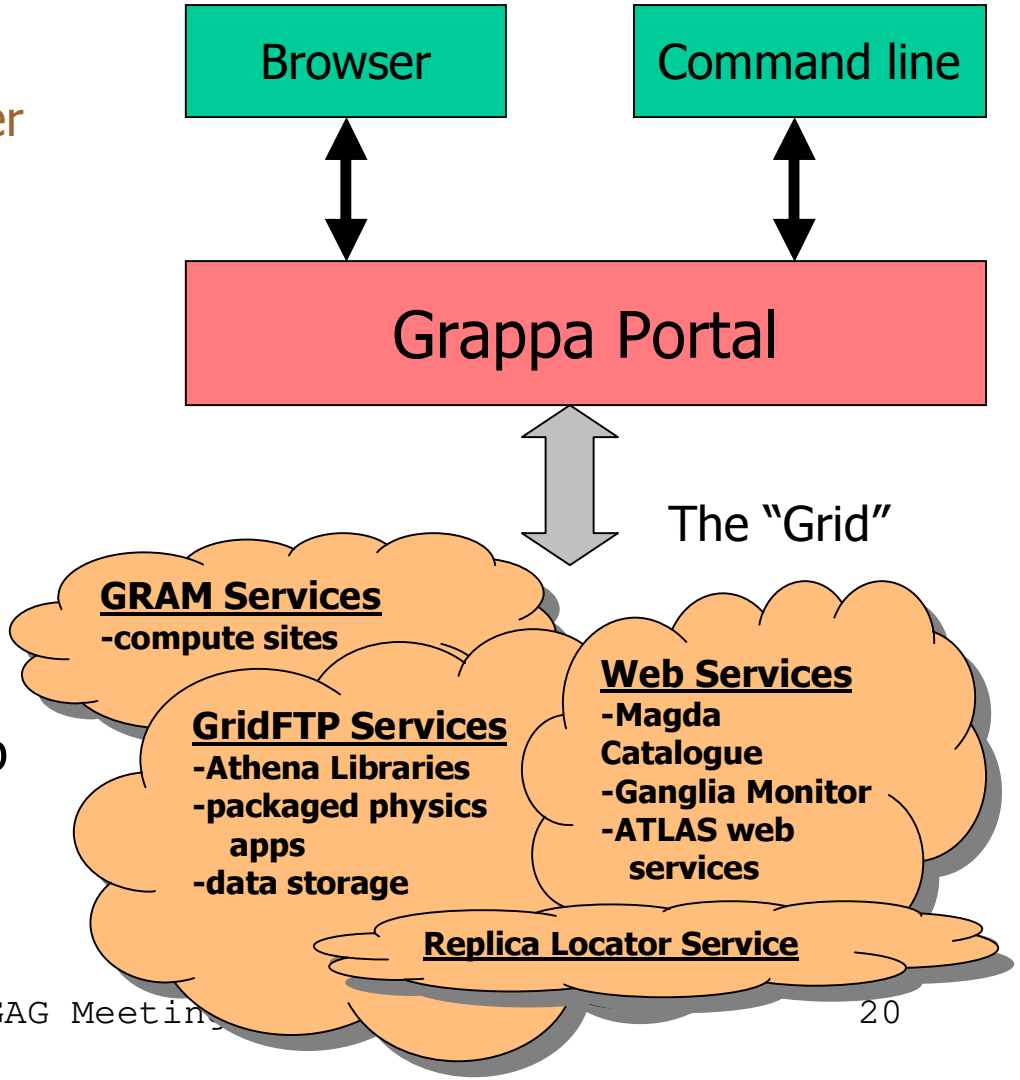
- Abstract data/workflows (Virtual Data DAG)
  - Resource locations unspecified
  - File names are logical
  - Data destinations unspecified
- Concrete data/workflows (DAGs for DAGMan)
  - Resource locations determined
  - Physical file names specified
  - Data delivered to and returned from physical locations



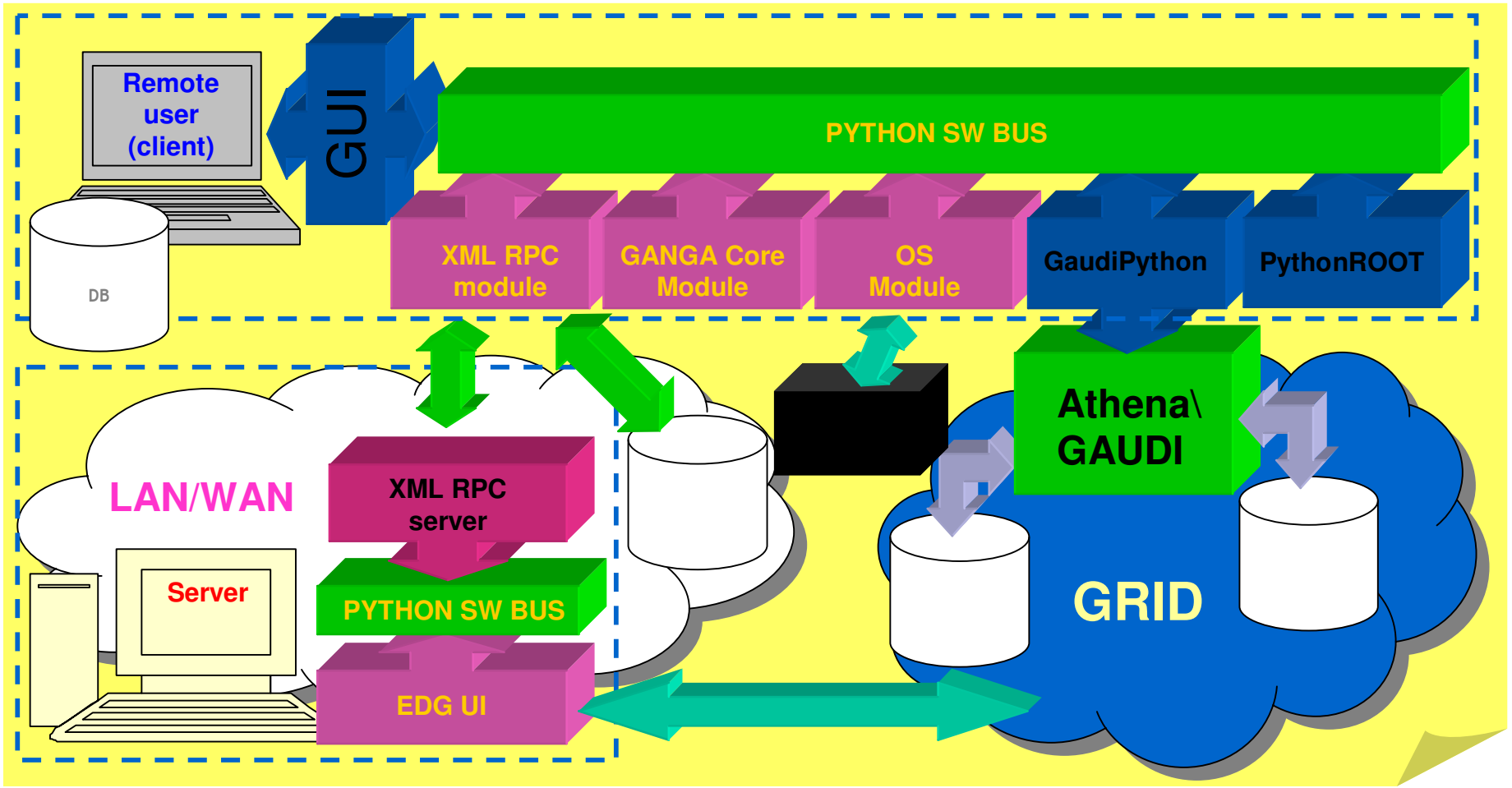
- History of a Data Analysis (like CVS)
- "Check-point" a Data Analysis
- Analysis Development Environment
- Audit a Data Analysis

Grappa: A Web Portal

- User Xbook (notebook scripts)
  - scripts customized for the user (as simple or complex as desired)
  - Athena xbook customized for ATLAS job submission
  - Many scripting languages supported (Jython, Python, Perl…)
- Xbooks server
  - A Jetspeed portlet
  - manages requests from user xbook
- Jetspeed: Apache-Jakarta web portlet framework
- Tomcat server: provides the web server



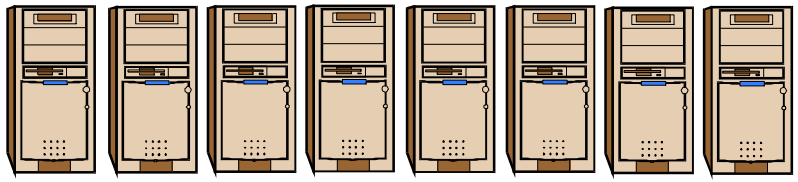
Ganga: Python Bus Design



DIAL:
Distributed Interactive Analysis of Large datasets

- DIAL provides a connection between
  - Interactive analysis framework
    - Fitting, presentation graphics, …
    - E.g. ROOT
  - and Data processing application
    - E.g. athena for ATLAS
    - Natural for the data of interest
- DIAL distributes processing
  - Among sites, farms, nodes
  - To provide user with desired response time

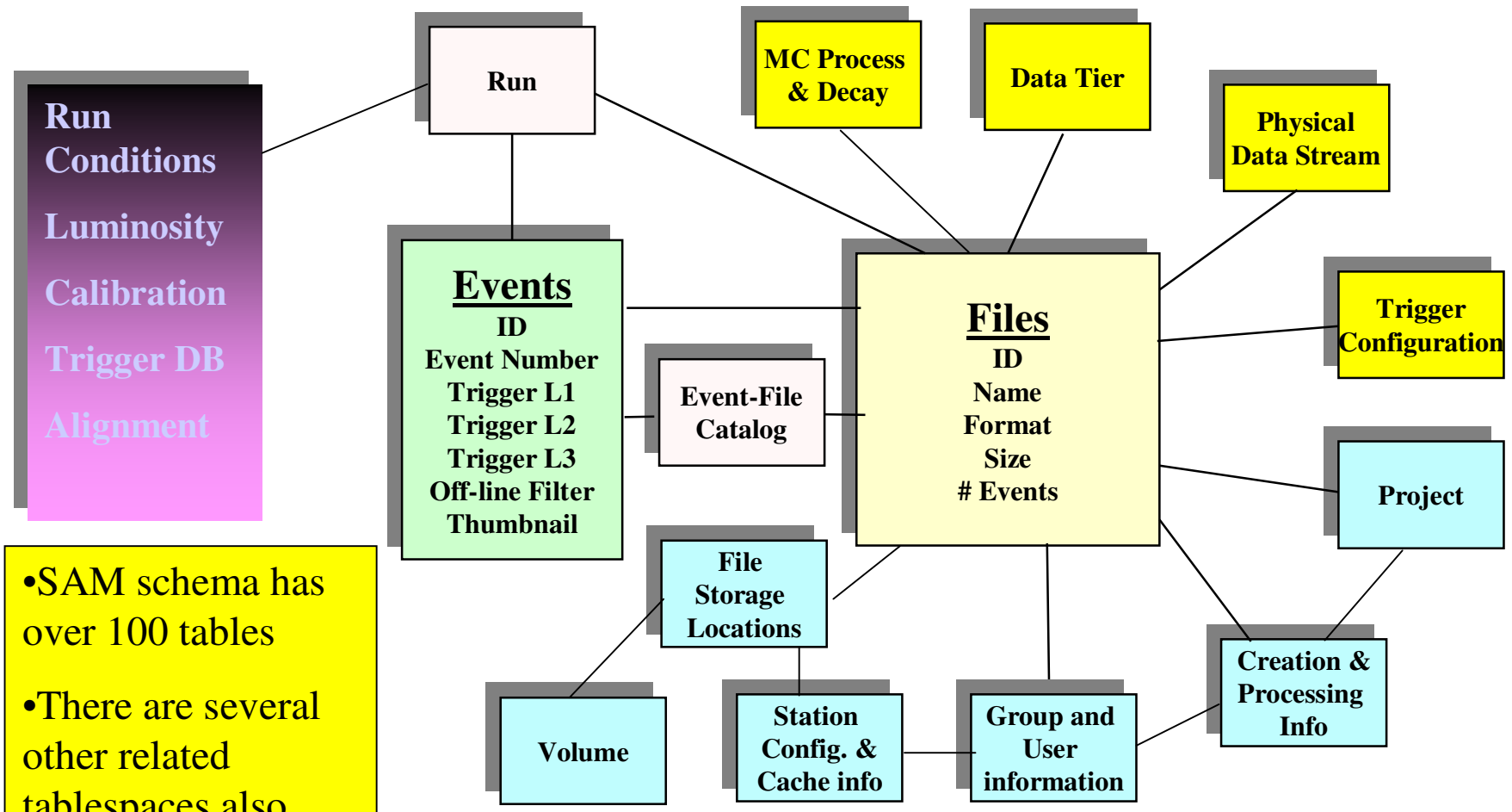
Interactive analysis
e.g. ROOT, JAS, ...



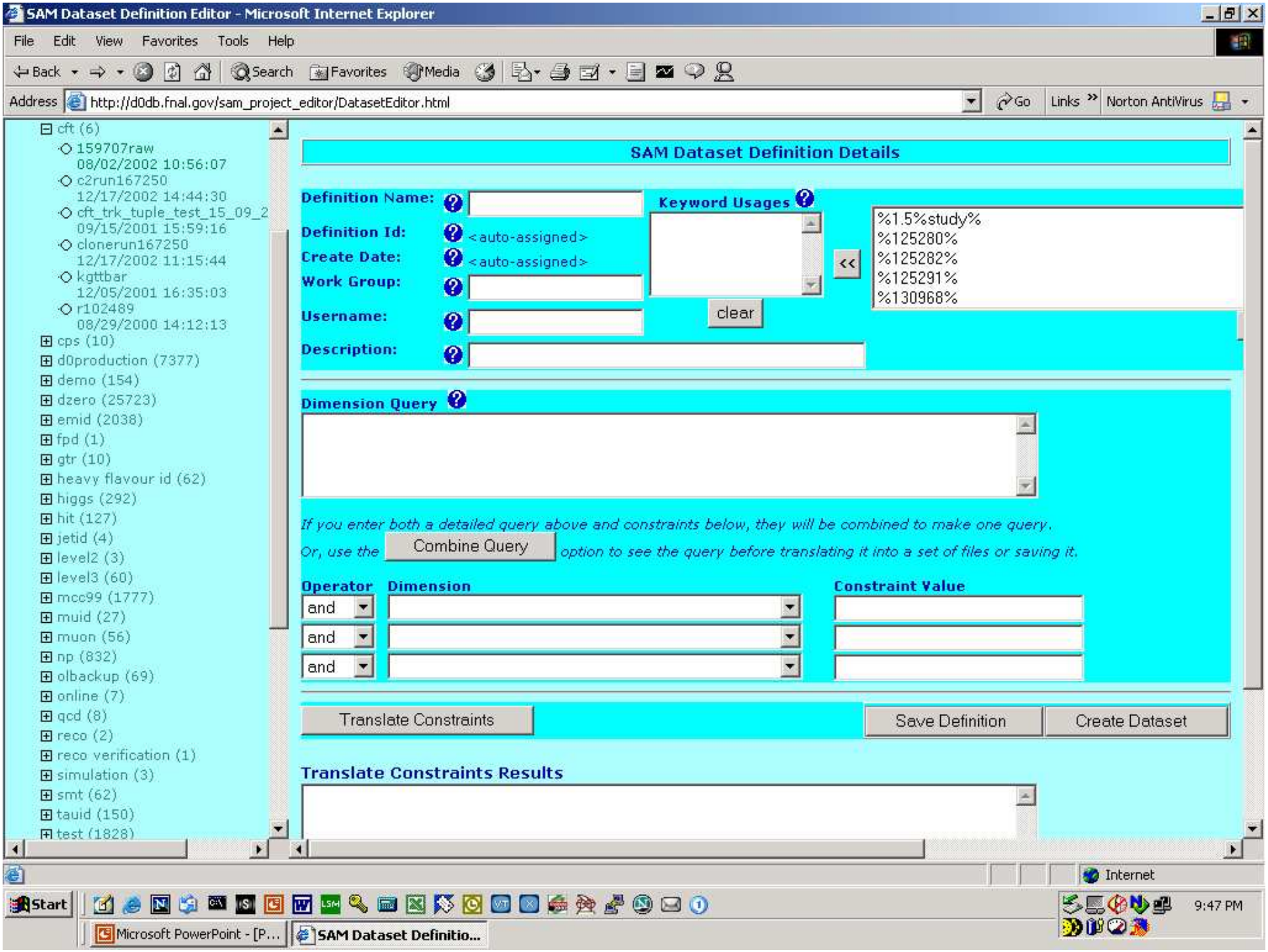
Distributed processing running data-specific application

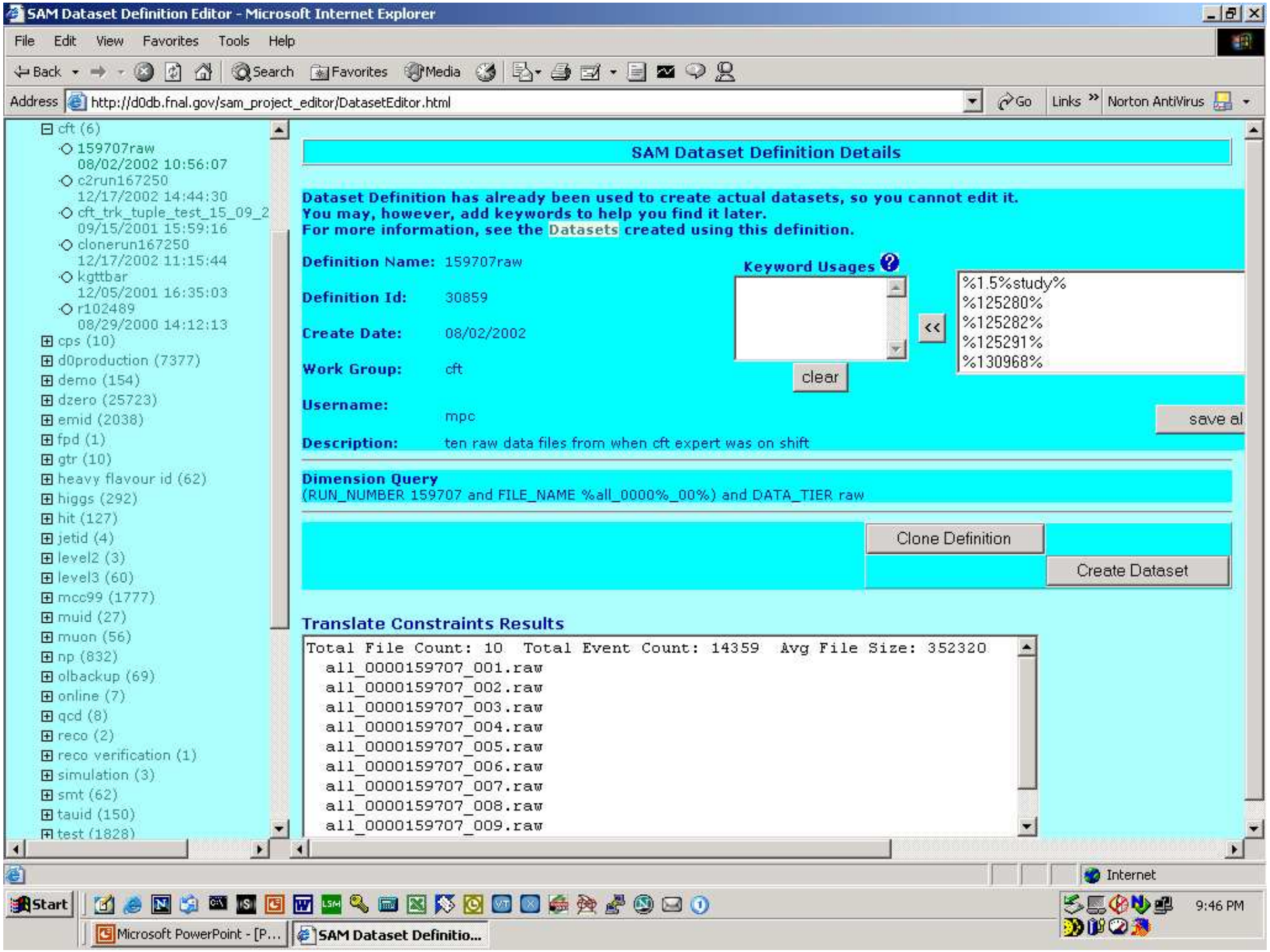
SAM Simplified Database Schema

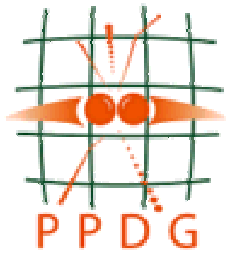
(quite complicated in reality)



- SAM schema has over 100 tables
- There are several other related tablespaces also available


SAM Dataset Definition Editor - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://d0db.fnal.gov/sam_project_editor/DatasetEditor.html
Go Links Norton AntiVirus
cft (6)
159707raw 08/02/2002 10:56:07
c2run167250 12/17/2002 14:44:30
cft_trk_tuple_test_15_09_2 09/15/2001 15:59:16
clonerun167250 12/17/2002 11:15:44
kgttbar 12/05/2001 16:35:03
r102489 08/29/2000 14:12:13
cps (10)
d0production (7377)
demo (154)
dzero (25723)
emid (2038)
fpd (1)
gtr (10)
heavy flavour id (62)
higgs (292)
hit (127)
jetid (4)
level2 (3)
level3 (60)
mcc99 (1777)
muid (27)
muon (56)
np (832)
olbackup (69)
online (7)
qcd (8)
reco (2)
reco verification (1)
simulation (3)
smt (62)
tauid (150)
test (1828)
SAM Dataset Definition Details
Definition Name:
Definition Id: <auto-assigned>
Create Date: <auto-assigned>
Work Group:
Username:
Description:
Keyword Usages
clear
<<
%1.5%study%
%125280%
%125282%
%125291%
%130968%
Dimension Query
If you enter both a detailed query above and constraints below, they will be combined to make one query.
Or, use the Combine Query option to see the query before translating it into a set of files or saving it.
Operator
Dimension
Constraint Value
and
and
and
Translate Constraints
Save Definition
Create Dataset
Translate Constraints Results
Internet
Start
Microsoft PowerPoint - [P...
SAM Dataset Definitio...
9:47 PM


SAM Dataset Definition Editor - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://d0db.fnal.gov/sam_project_editor/DatasetEditor.html
Go Links Norton AntiVirus
cft (6)
159707raw
08/02/2002 10:56:07
c2run167250
12/17/2002 14:44:30
cft_trk_tuple_test_15_09_2
09/15/2001 15:59:16
clonerun167250
12/17/2002 11:15:44
kgttbar
12/05/2001 16:35:03
r102489
08/29/2000 14:12:13
cps (10)
d0production (7377)
demo (154)
dzero (25723)
emid (2038)
fpd (1)
gtr (10)
heavy flavour id (62)
higgs (292)
hit (127)
jetid (4)
level2 (3)
level3 (60)
mcc99 (1777)
muid (27)
muon (56)
np (832)
olbackup (69)
online (7)
qcd (8)
reco (2)
reco verification (1)
simulation (3)
smt (62)
tauid (150)
test (1828)
SAM Dataset Definition Details
Dataset Definition has already been used to create actual datasets, so you cannot edit it.
You may, however, add keywords to help you find it later.
For more information, see the Datasets created using this definition.
Definition Name: 159707raw
Definition Id: 30859
Create Date: 08/02/2002
Work Group: cft
Username: mpc
Description: ten raw data files from when cft expert was on shift
Keyword Usages
<<
clear
%1.5%study%
%125280%
%125282%
%125291%
%130968%
save al
Dimension Query
(RUN_NUMBER 159707 and FILE_NAME %all_0000%_00%) and DATA_TIER raw
Clone Definition
Create Dataset
Translate Constraints Results
Total File Count: 10 Total Event Count: 14359 Avg File Size: 352320
all_0000159707_001.raw
all_0000159707_002.raw
all_0000159707_003.raw
all_0000159707_004.raw
all_0000159707_005.raw
all_0000159707_006.raw
all_0000159707_007.raw
all_0000159707_008.raw
all_0000159707_009.raw
Internet
Start
Microsoft PowerPoint - [P...
SAM Dataset Definitio...
9:46 PM



Technologies being studied by CS-11

- Portals, Portlets
  - XCAT, Jetspeed, etc
- Web Services
  - SOAP, WSDL, UDDI, etc
- Grid Services
  - Resource Broker, RLS, Virtual Data System, etc
  - OGSA/I

Conclusion: Interactive Data Analysis Tools

- CS-11 is making good progress in defining
  - Use-cases and requirements (middle stages)
  - APIs for different Grid/Web Services (early stages)
- Several participating groups in CS-11 are also developing prototypes/tools
  - PROOF, JAS w/ CoG, Clarens, Grappa/GANGA, etc
  - Represents a good sampling of different ideas
- Work is underway to understand better what OGSA/I brings to the picture