

Distributed Interactive Data Analysis

ATLAS SW workshop
Grid session

David Adams

BNL

May 29, 2002

Contents

- Definition
- Request processing
- Production *vs.* analysis
- Applications
- Plans and progress
- Datasets

Definition

Distributed

- User and data at different sites
 - In general data may distributed over multiple sites

Interactive analysis

- Interactive configuration and response
- View histograms, correlation plots, events
- Partial results for long requests (1 min?)

Data

- EDO's and later summary ntuples

Request processing

Configuration

- Specify input data
- Define transformation
 - Output includes histogram and tuples

Execution

- Locate data at sites
- Divide and assign jobs to sites
- Monitor progress
- Collect output (including partial results)

Production vs. analysis

Commonalities

- Distributed production and analysis share a great deal of (GRID & ATLAS) infrastructure
- May have common applications
 - *E.g.* athena

Resource allocation

- Either dedicated analysis resources or
- Give priority to analysis activities
 - Suspend production when an analysis request is received

Applications

Athena

- Allows access to arbitrary EDO's through StoreGate

ROOT

- Popular HEP tool for data analysis
- Keep the user interface
- Extend for distributed processing over the grid

Plans and progress

Short term goals

- Interactive analysis of MC generator data distributed over multiple nodes (10+) on a single cluster.
- Same for data through GRID portal
- Same for multiple (2+) GRID portals

Plans and progress (cont)

Identified components

- Data view defined by dataset
 - See the following discussion
- Application: distributed ROOT
 - We have begun developing a RootSlave that will start up on a remote node and process commands
- GRID tools
 - Expect to use existing tools

Datasets

See talk in DB session

Dataset interface

- Event range: Collection of event ID's
- Content: Collection of content ID's
 - In ATLAS, content ID is type and string key
- Event data: For each event ID-content ID pair:
 - A means to access the corresponding EDO or
 - A flag indicating the EDO is not included

Datasets (cont)

