

# DH IO Modules Project Status Report.

F.Ratnikov, RUTGERS

## Introduction

- DH IO Modules project was started in the very beginning of Y2K
- Immediate problems:
  - need direct access to events to keep runsections compact on the output
  - need chain like structure of input and output stream - data branch lifetime is spread over many sequential files - branch driven design
- Framework:
  - was designed for essentially sequential access
  - IO was designed as file driven - SeqRootDiskFile
- To fulfill CDF data convention and DH requirements necessary functionality was included into DHMods

# Design of the DH IO Modules Project

- Project has essentially modular structure
  - well defined and well separated interfaces to
    - Edm
    - DFC
    - DH
    - ROOT
- Project has all the necessary hooks to handle multi-branch event structure properly
- Due to lack of the necessary infrastructure functionality modules perform many tasks not specific to DH itself

# DHInput Functionality

- Select data by any combination of dataset, fileset, or file names
- Full “include” and “exclude” support
- Extra restriction on required run# and runsection# can be applied
- Access both DH data and local private files
  - Accept wildcards for local files
- 100% compatibility with FileInputModule
- Communication with DFC to obtain full list of requested data
- Communication with DH to deliver requested data in the most effective order

# DHInput Functionality (cont)

- Process events in natural order, build catalog of events in the file (necessary for correct output file production)
  - Fast operation using EventInfo branch (any data except RAW)
  - Slow operation using LRIH information (RAW data)
- Navigation in the file
  - Skip events forward and backwards
  - Direct access to the event by run#/event#
  - Insert necessary BOR records when run# is changed
- Read/Write events by ROOT buffers without expanding to separate objects
  - Speed up IO bandwidth by a factor of 5
  - Is necessary for concatenating FARM output files
- Input events can be filtered by run# and event#

# Output Module Functionality

- Specify output by file name or by dataset name
- Assign data file name according to the CDF convention
- Collect statistics for the output file
- Collect output files in given directory
- Can put FILE record into DFC
- Split output data into files of given size
- Keep runsections compact in the file
- Save the intermediate status of the file to minimize reprocessing in case of job crash
- Creates new ERS records when EmptyRunsection condition is detected
- Can create many data branches synchronized with primary data branch

# Crash Recovery

- The goal is to continue data processing after the job has crashed due to any reason
  - with minimal reprocessing overlap
  - while keeping DFC consistent at any time
- Sophisticated procedure is developed
- It is semiautomatic
  - Algorithm is well defined but requires manual work
  - The close coordination between Input and Output is required to make procedure mostly automatic
  - Automatic procedure can be implemented on the Framework level where coordination of Input and Output is possible

# Documentation

CDF/5336

## Input and Output Modules user guide

[Fedor D. Ratnikov](#) ([Rutgers](#))

*This is a changing document; this version has date 9/21/01 11:22 PM.*

Look for the latest version: <http://rutpc7.fnal.gov/ratnikov/Docs/DHIOModuleReference.htm>

This is a hypertext document: cross-references will be lost on hardcopy.

### Contents

- [Introduction](#)
- [Input](#)
  - [Include InputModule into executable](#)
  - [Specification of the input data](#)
  - [Examples](#)
  - [Input options](#)
  - [Selection of events to be processed](#)
  - [Navigating in the input file](#)
  - [Hints to access events of the Commissioning Run](#)
- [Output](#)
  - [Include OutputModule into executable](#)
  - [Options for output streams](#)
  - [Examples](#)
- [Command line specification of the input and output files](#)
- [File Content Catalog](#)
- [Terms and definitions](#)

# Still Missed

- Clean up of output from obsolete BOR records
  - several BOR for the same run
  - BOR for those runs when all events were filtered out
- Minor issues of current user requests

# Conclusions

- **DH IO Modules project has successfully reached all the original goals, DHInput and DHOutput provide all the necessary functionality**
- It fulfills mutually contradictory requirements having high performance for the FARM operation and being flexible for user convenience
- Modules are in use for a long time. No operation problems were detected
- DH IO Modules include hooks necessary to support multi-branch event structure