PAW Based Analysis
and
Ntuple Converter

Edwin Soedarmadji

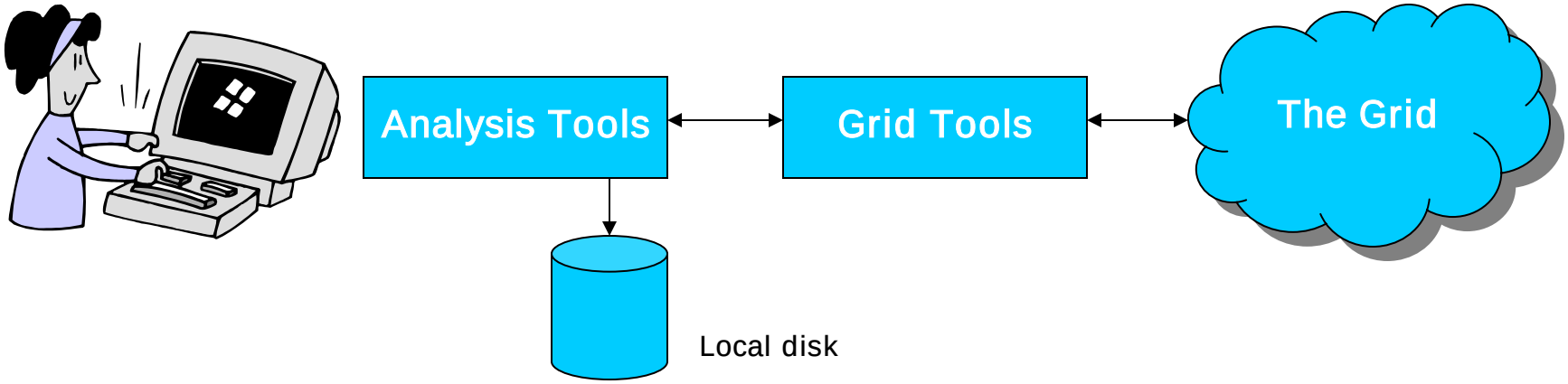
Caltech/CMS

May 22, 2002

Grid Analysis Environment

California Institute of Technology

Edwin Soedarmadji, May 22, 2002

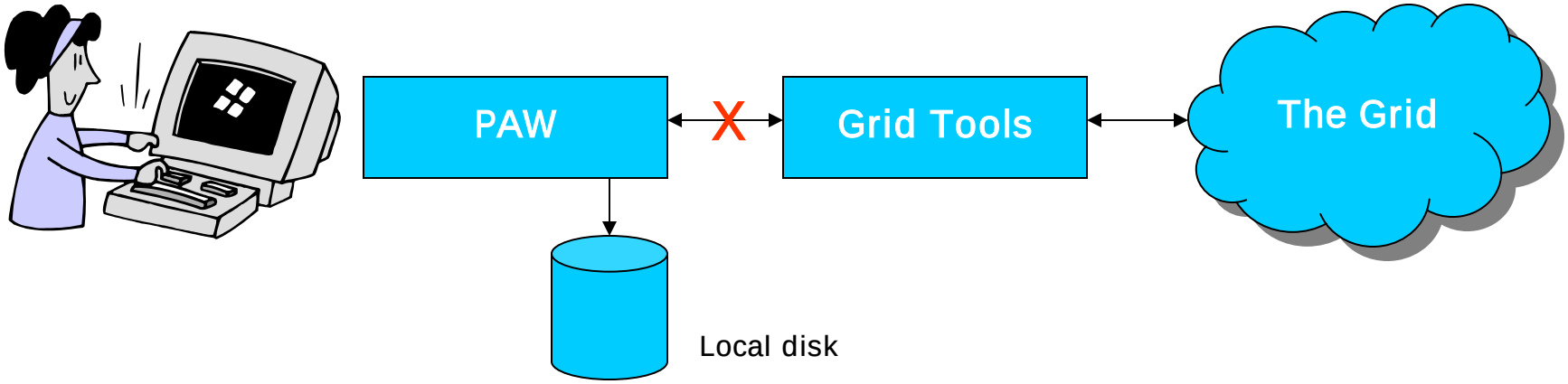


- Analysis tools such as PAW and ROOT will be a part of this environment
- In a Grid Analysis Environment, these tools will be enhanced to interact with other grid tools
- Enhancements to include grid features offer improvements in performance, manageability and scalability
- Need a simple analysis as a starting point and benchmark for performance and scalability measurements

PAW Based Analysis

California Institute of Technology

Edwin Soedarmadji, May 22, 2002

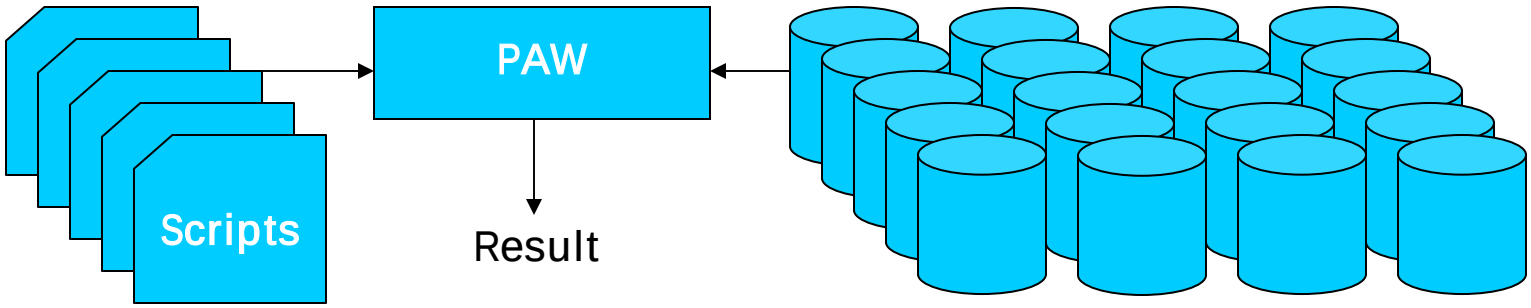


- PAW is a popular physics analysis tools
- Large user base → large base of user applications
- Not a grid-enabled analysis tools
- Used in many important analysis within CMS

JetMet PAW Analysis

California Institute of Technology

Edwin Soedarmadji, May 22, 2002

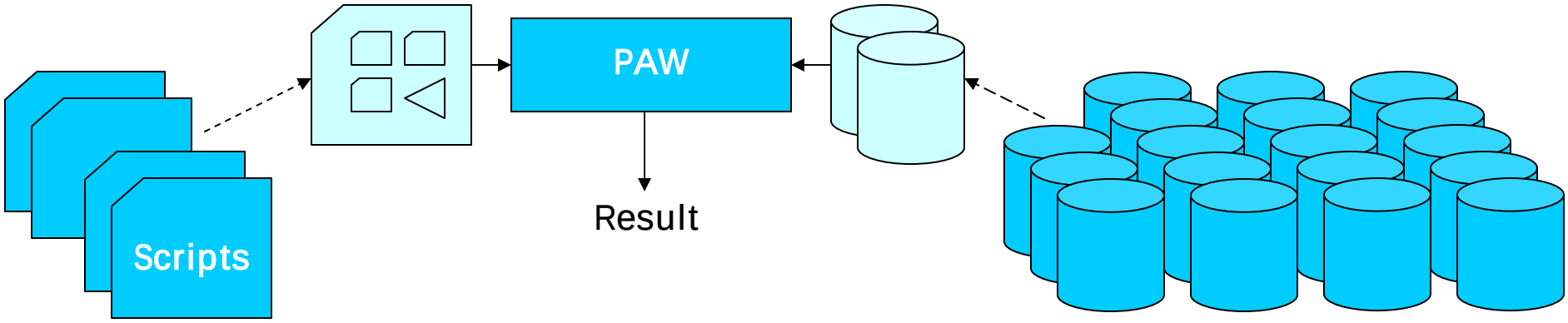


- One candidate is Pal Hidas' JetMet trigger analysis programs written in PAW kumac scripts
- Pal's analysis program is quite significant and imposes heavy requirements in terms of data and computation power
- This analysis is widely recognized and is representative of many other analysis programs in terms of requirements
- These scripts are chosen to be studied and enhanced with connectivity to RDBMS and other distributed computing facilities

Subset of JetMet Analysis

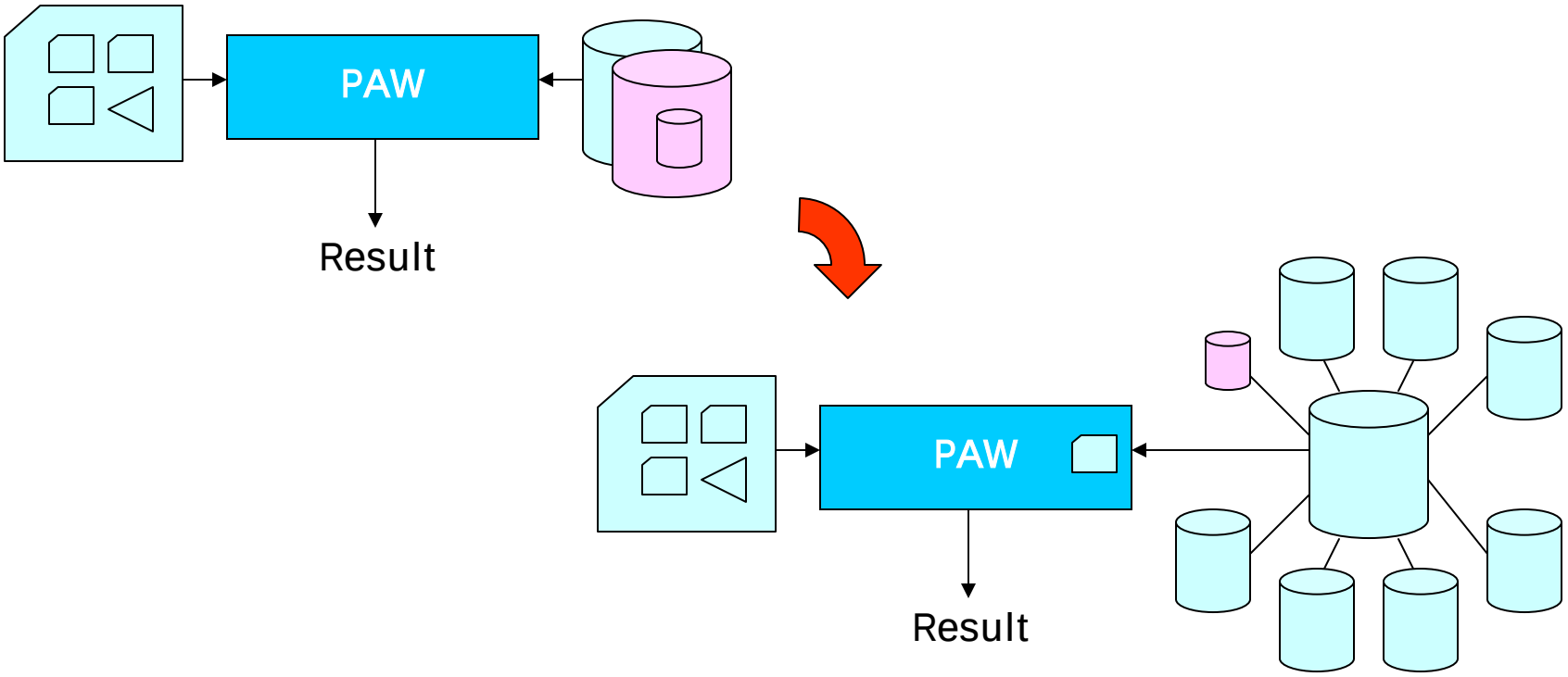
California Institute of Technology

Edwin Soedarmadji, May 22, 2002



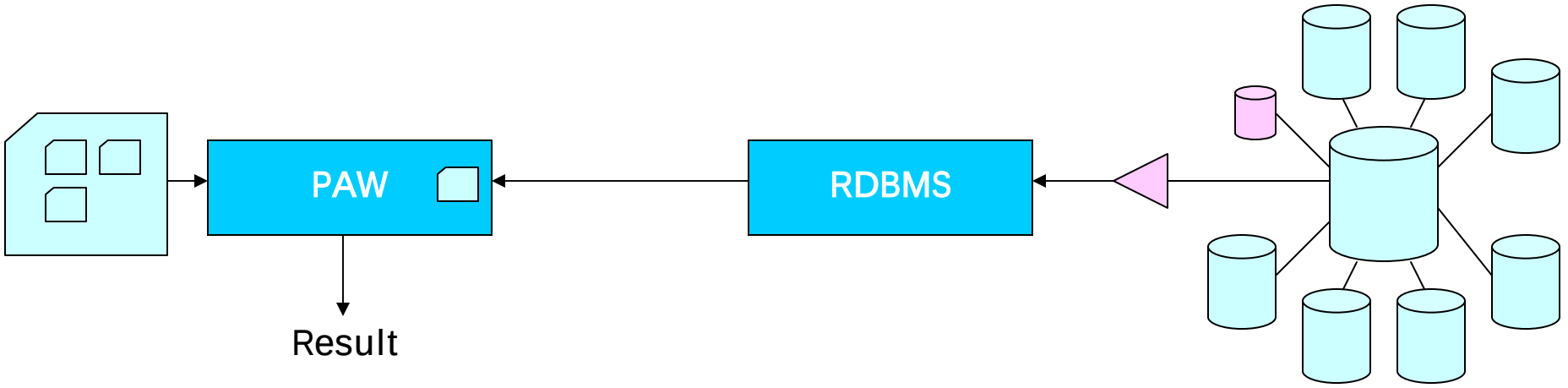
- We took only a small subset of the “kumacs3” scripts run by Pal Hidas at FNAL. This subset contains three main scripts other ancillary scripts.
- We copied only 1/10th of the data read by these scripts.
- The scripts are studied
  - Which data is read
  - What is done with the data
  - What is the script performance before enhancement

Enhancements (1 of 2)



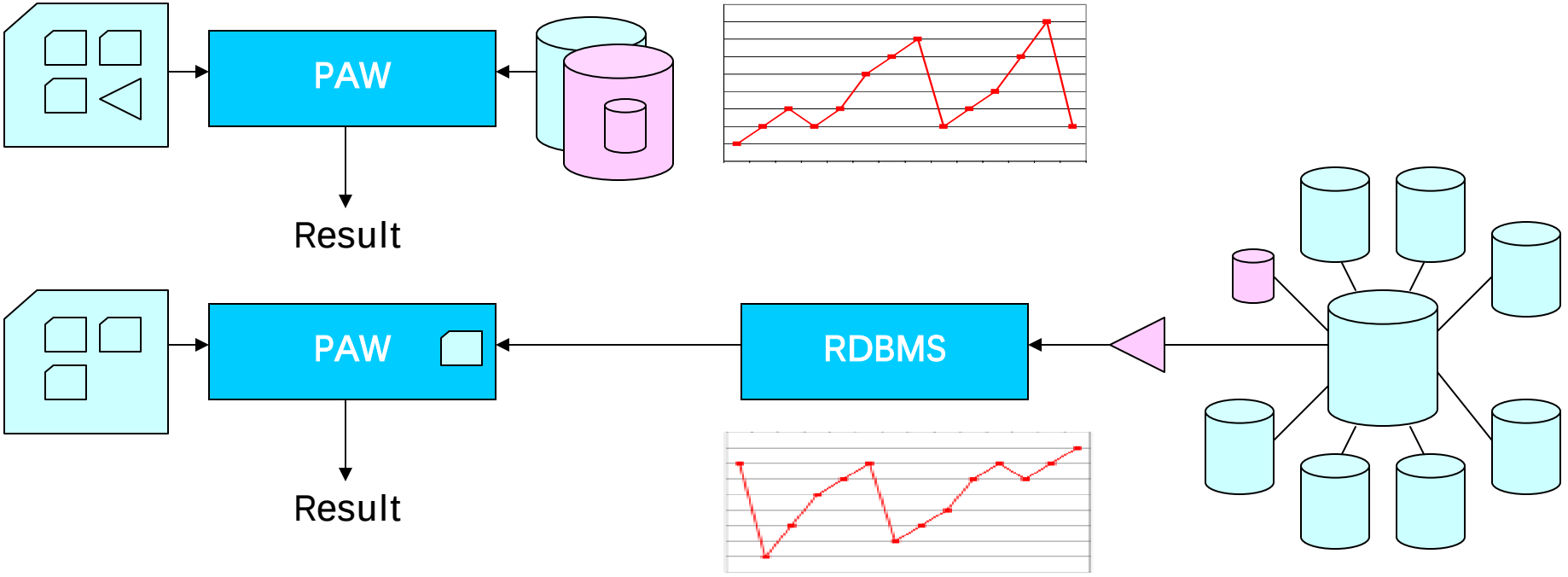
- Data will be moved out of ntuples into RDBMS
- Data will be split up into blocks. This allows the logical division of data into blocks in ntuples to manifest into data groups based on frequency of access

Enhancements (2 of 2)



- Data set filtering will be done on the server
- Data will be available as part of distributed system
- PAW analysis program will be modified to data from server

Measurements (1 of 2)



- Performance before and after enhancements will be compared. We hope for a significant speed-up

- User experience with enhanced analysis will be noted and used to fine-tune further enhancements. We hope for a much improved user analysis experience.

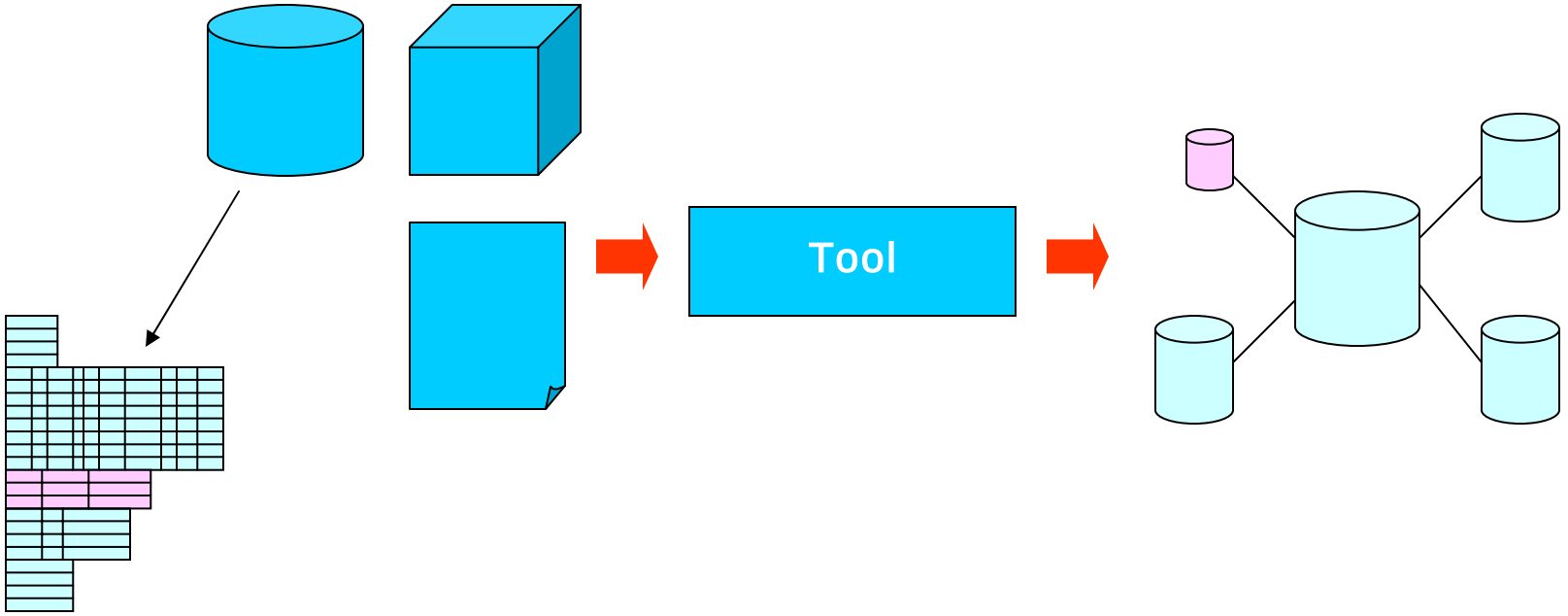
Measurements (2 of 2)

- Performance measurement before enhancements has completed. Conversion of data and filter is in progress.
- Once work is finished on the smaller data set, the full 50GB worth of data will be transferred.

Conversion Tool

California Institute of Technology

Edwin Soedarmadji, May 22, 2002



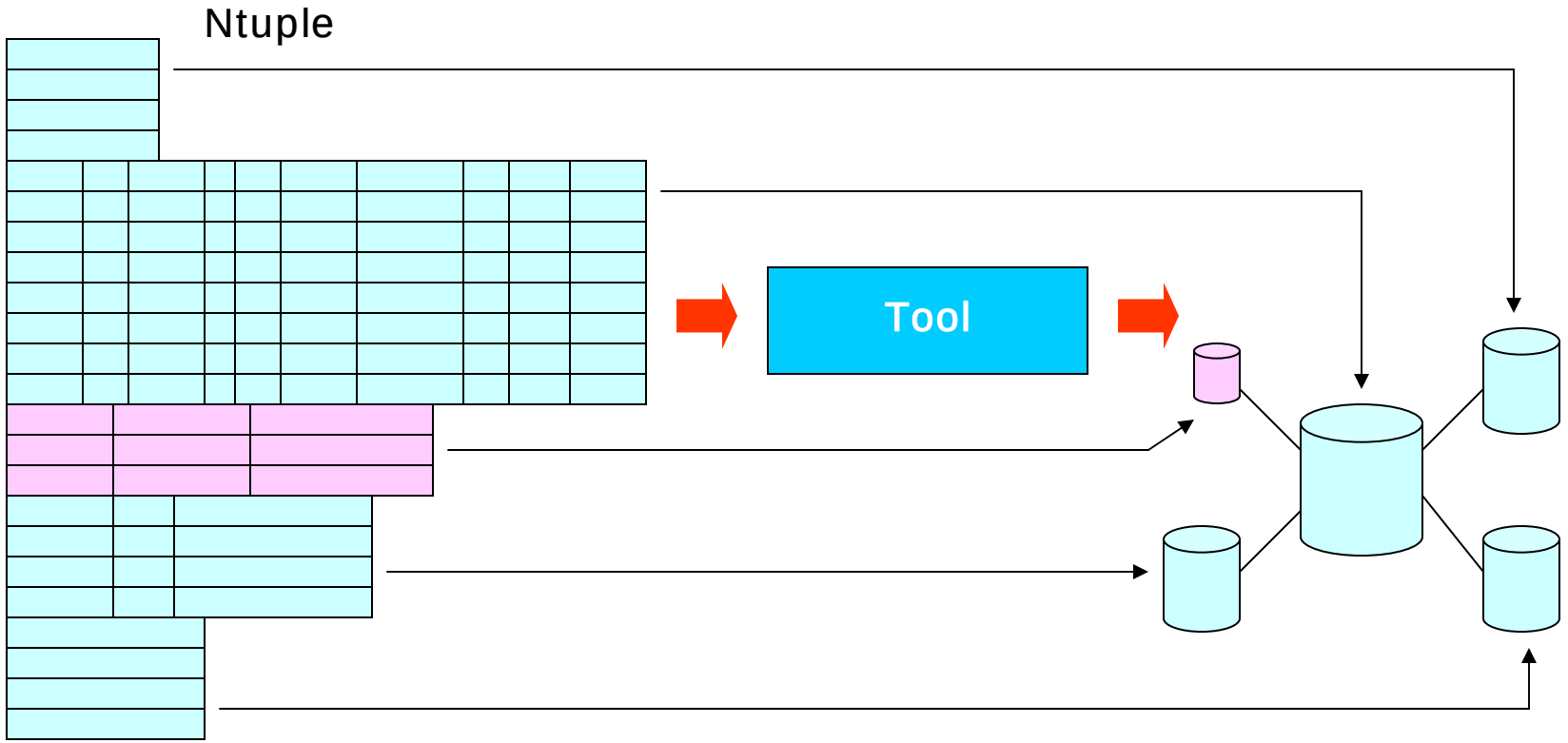
- The dominant data formats in CMS are ntuples, Objectivity databases, ROOT files, etc.

- A tool to read these data formats into an RDBMS will be a very useful tools in a grid analysis environment based on analysis tools using those data formats

Conversion Tool

California Institute of Technology

Edwin Soedarmadji, May 22, 2002



- For PAW Analysis work, a simple ntuple conversion tool is being developed

- Ntuple blocks are split into separate tables linked to a general table

Summary and Conclusion

- Grid Analysis Environment includes standard analysis tools that need to be converted or enhanced to interact with grid tools (at least until those tools natively support such interaction)
- A sample PAW analysis code is analyzed to develop a prototypical interface between a popular analysis tool (PAW) and currently available grid tools
- Both analysis data and code needs to be converted
- A conversion tool to migrate PAW ntuple data into RDBMS is currently in progress