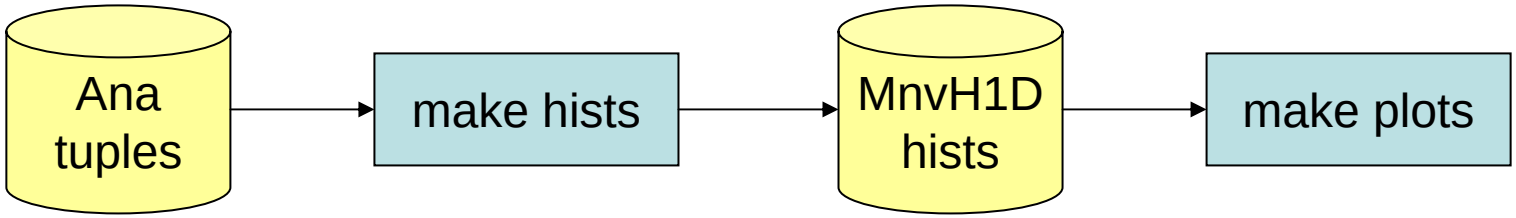
End-User Analysis Processing

Jaewon Park

April 8 2014

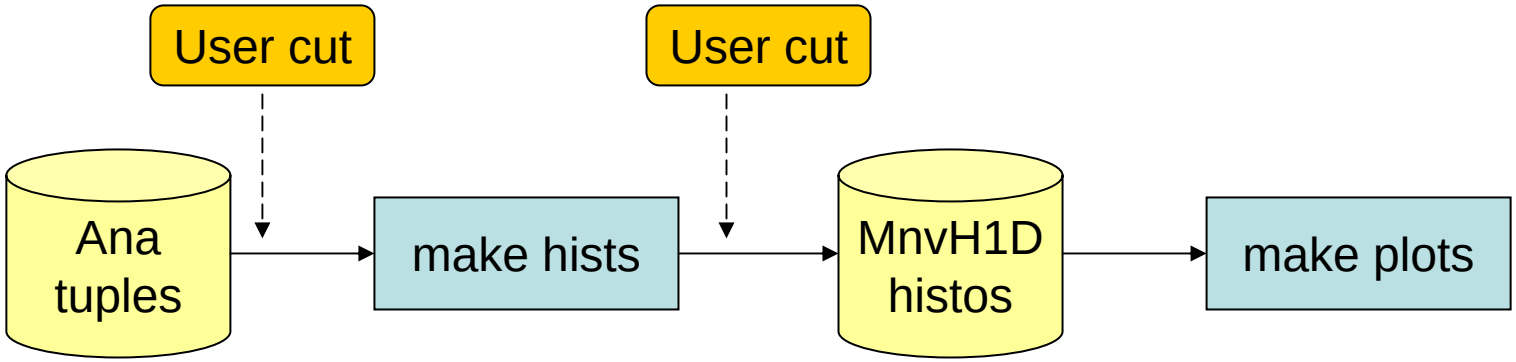
Minerva data management review

End-User Analysis



- Use common package PlotUtils to make intermediate histo and final plots
- Main functionalities in the PlotUtils
  - Error band calculation using multi-universes
  - Adding other source of systematic uncertainties
  - Common plotting
    - Data-MC comparison, ratio, error band, etc
- Use also use common Unfolding utils for the unfolding
- User applies event selection on the sample
  - Earlier stage analysis tries various cuts to optimize the cut
- User's own code is necessary to do the sideband tuning, background subtraction or efficiency correction
- User runs make_hists and make_plots interactively
- make_hists takes most of time

Ana Tuple Output



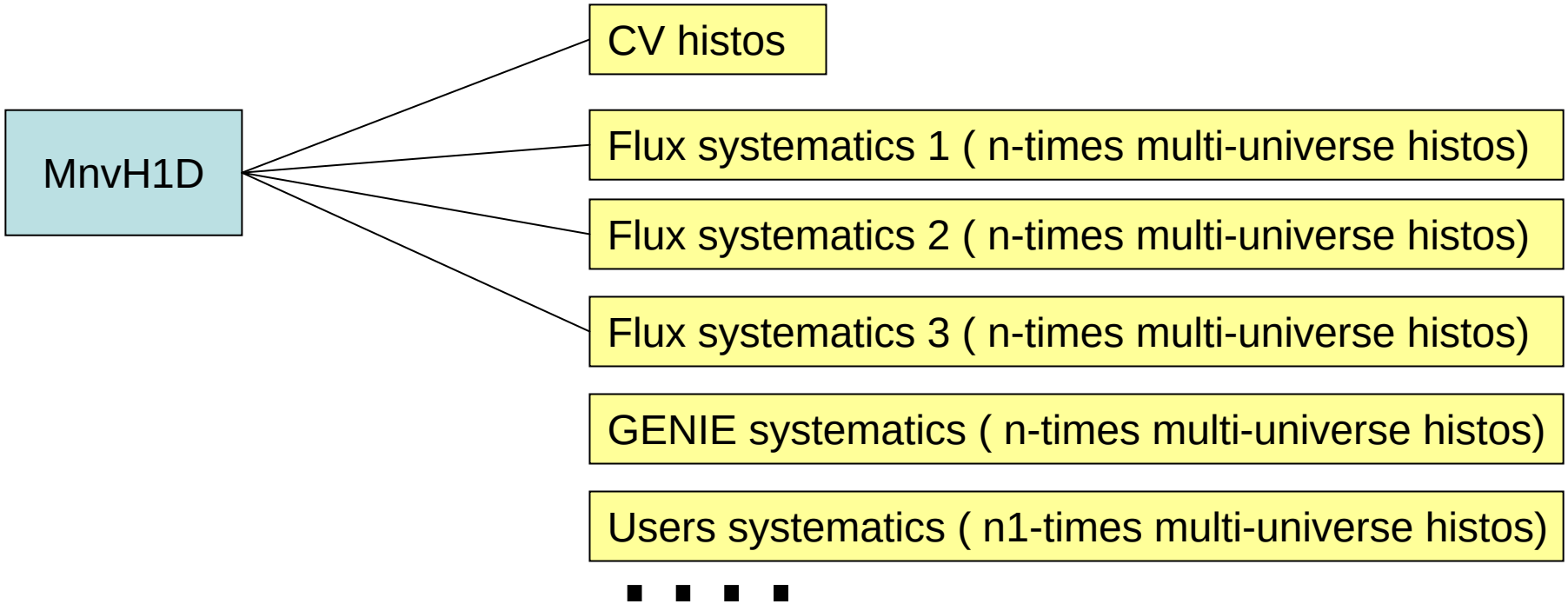
- Separate truth tree is written for every gate
  - It's used for efficiency calculation
- (reconstructed) ana tuples is selectively written per gate
- Successfully reconstructed event candidates will be written on ana tuple
  - It often writes events with preliminary event selection
- Ana tuple output per file varies strongly depending on analysis
- Ana tuple output and processing time also depends on number of multi-universe
  - Currently maximum number of multi-universe is 1000
  - User often uses smaller number of multi-universe (100 to 300)
  - MC processing takes long than data

Most User Task is IO-Intenstive

- Ana tuple
  - Ana output is generally much smaller than reco input
  - 1 ana processing uses 4-5 subruns reco input files

- Ana tuple use
  - Typical make_hists is IO-intensive
    - It has many input files and writes out small files
  - It uses most of time to read entry and skip it
  - Sideband tuning or unfolding can be a little bit of CPU-intensive

MnvH1D and Multi-universes



- Multi-universes are given by weight factors
- Each histogram is filled for each multi-universe using the weight
- Three categories of beam systematic have each multi-universes
- One typically has a combined GENIE systematics multi-universe
- In case we want to know individual physics parameters systematics, we need to write out more parameter multi-universes

POT Counting

- POT calculation script is used
  - Check if subrun is complete
  - Check if number of event is consistent with GENIE
  - Ignore too tiny POT data file
- Later analysis uses calculated total POT for POT normalization

CC Inclusive

- Minerva1 playlist: ~1600 MC ana tuples (~5.7E20 POT)
- Ana tuple
  - Data: 1MB per file
  - MC: 240 MB per file
- Make hists produces 71 MB file (data & MC)
  - It takes ~8 hours (minerva 1 playlist)

CC Coherent Pion

- Ana tuple
  - Data: 250 kB
  - MC: 30 MB per file (5 subruns)
- Make hists produces 7 MB (100 universe)
- 5 minutes for minerva1
- Signal is rare but clean
  - Preliminary event selection already makes ana tuple output quite small

Nu-E Scattering

- Ana tuple files (all low energy files)
  - 3800 MC ana tuple files (~390MB per file)
  - 3700 data ana tuple files (~18 MB per file)
- Use skim on files before do make_hists
  - Parallel skip jobs are run interactively (~8 hours)
- make_hists: ~1hours
- sideband tuning: 1-2 hours
- make_plots: ~10minutes

Summary

- User end analysis is mostly not too bad except
  - Limited common production due to disk issue
- Recent update on ana processing allows to process ana using files in dcache
  - Previously ana tuple creation suffered due to occasional grid bottle neck
- Interactive running for make_histo is OK for most case
- Current way would be a problem in ME data analysis
- In principle, make_hists can be done on grid
  - Need common organized analysis code and script for easy grid running