Data Processing - Raw data to analysis and data sharing

Erik Blaufuss, University of Maryland

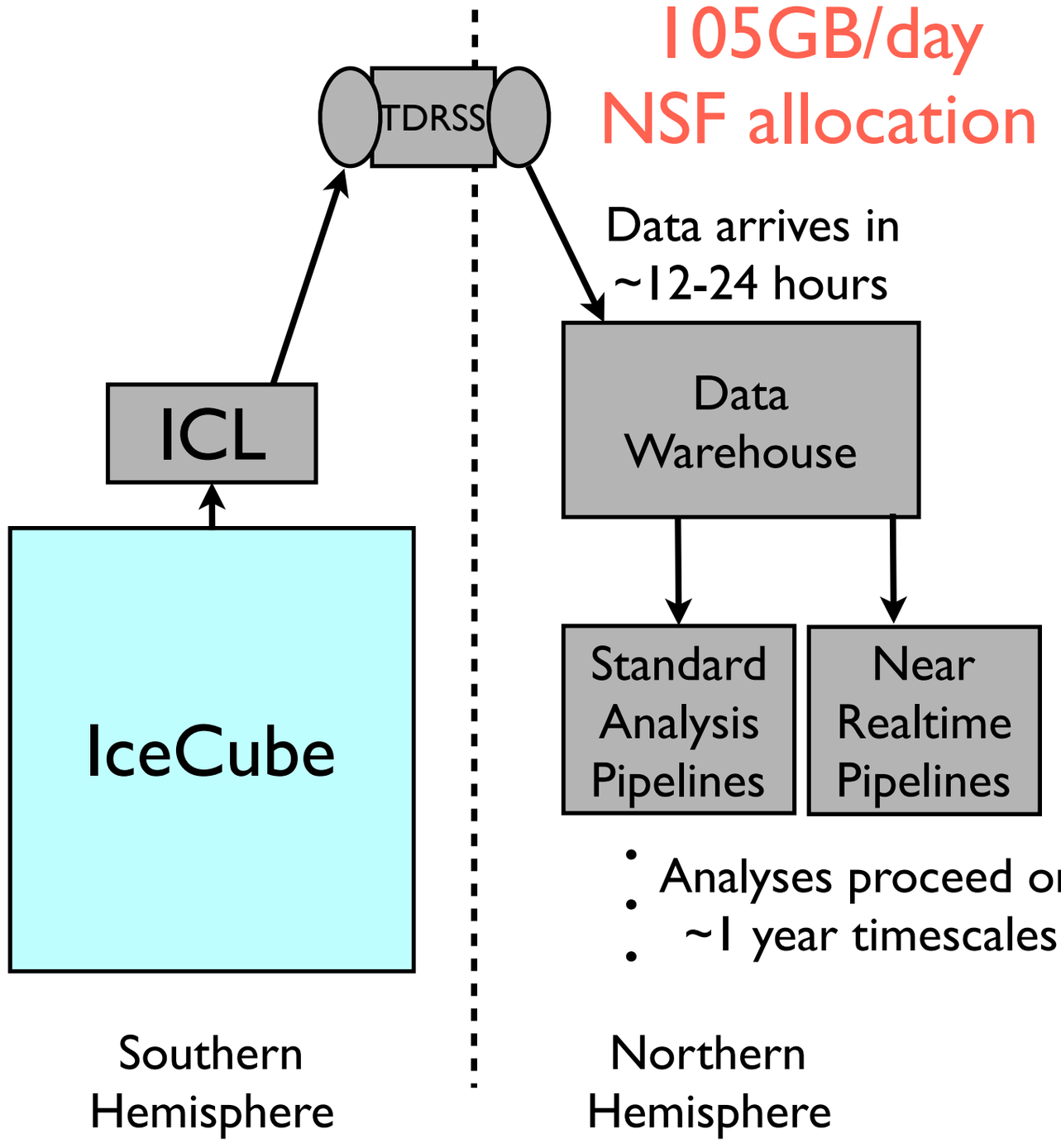
Science Advisory Committee, October 19-20, 2015

Overview

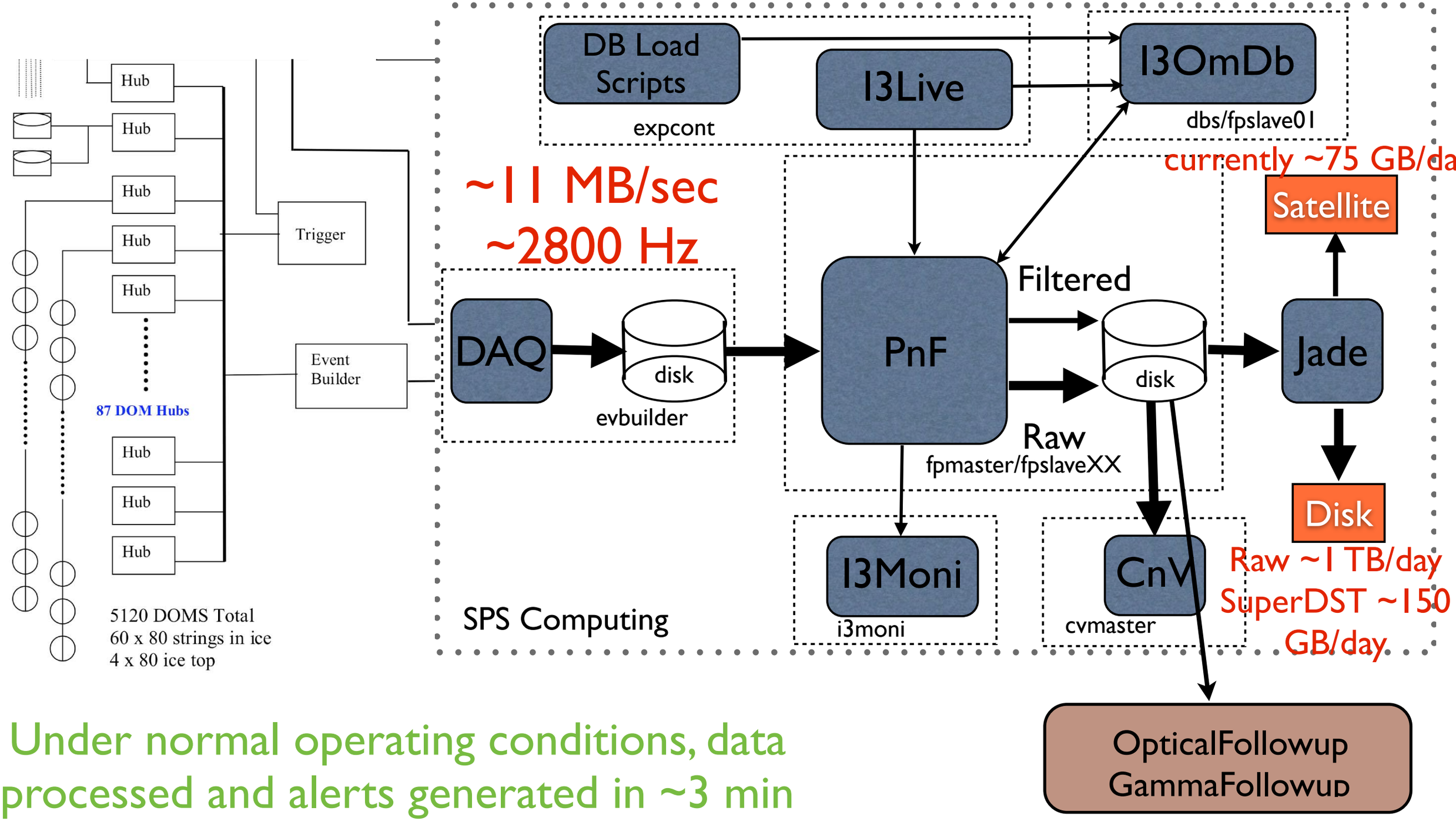
- Overview of data processing in IceCube
  - Robust online calibration, reconstructions and data formats
  - Stability of event selections
  - Multi-year analyses simplified for data and simulation in many channels
- Realtime efforts
- Data releases

Data Flow overview

- Data processing divided into two regimes, online at South Pole and offline at IceCube data warehouse
- Online
  - Apply robust calibrations and waveform feature extraction
  - Filter content set by working groups via Trigger, Filter, and Transmission Board
  - Host realtime processing & alerts
- Offline
  - Performing best reconstructions and data selections (higher CPU)
  - L2 processing now done as part of TFT process with production tools
  - Near-realtime pipeline system
  - High level data samples for neutrino analyses (Level3+).



Online systems at South Pole



Under normal operating conditions, data processed and alerts generated in ~3 min

Online stability - SuperDST

- Strategy: Apply BEST calibration and feature extraction at South Pole in online filter
  - Check pulses are a correct representation of the waveform
  - SuperDST format: Save compact form of all pulses found online for transmission north (time, charge: <10% of waveform size)
    - Save waveforms ONLY for those DOMs/events (~few %):
      - Reconstructed waveform poorly agrees with original
      - High charge in a single DOM observed (<<1% of hits)
      - High total event charge (<1% of all events)
      - IceTop waveforms - contain detailed information of particles in shower front
- SuperDST events saved to disk at pole -> long term data archive
  - Processing older tapes from IC86 to obtain full SuperDST archive

Online stability

- IceCube collaboration and working groups determine contents of online filter via TFT
  - Determines allocations of ALL pole resources (Bandwidth, CPU, filters)
  - Controls contents of Offline L2.
    - Improved L2 processing lag from ~1 year to ~1-2 wk
- Annual (~April) request for changes to filters and triggers
  - Several filters used in core analyses: stable (mostly unchanged) since IC86-2012 season. Review is still treated as a filter checkpoint.
  - Opportunity for new filters/triggers to be added
    - Examples: FixedRateTriggers, Hit spool data collection
  - SuperDST allowed some "just in case" event selections to be retired
- Stability has allowed for analyses and simulation to treat these seasons as one continuous data sample.
  - Several working groups will be coupling L3 processing to L2
  - Large background simulation samples remain a big challenge

Online stability - Robust neutrino selections

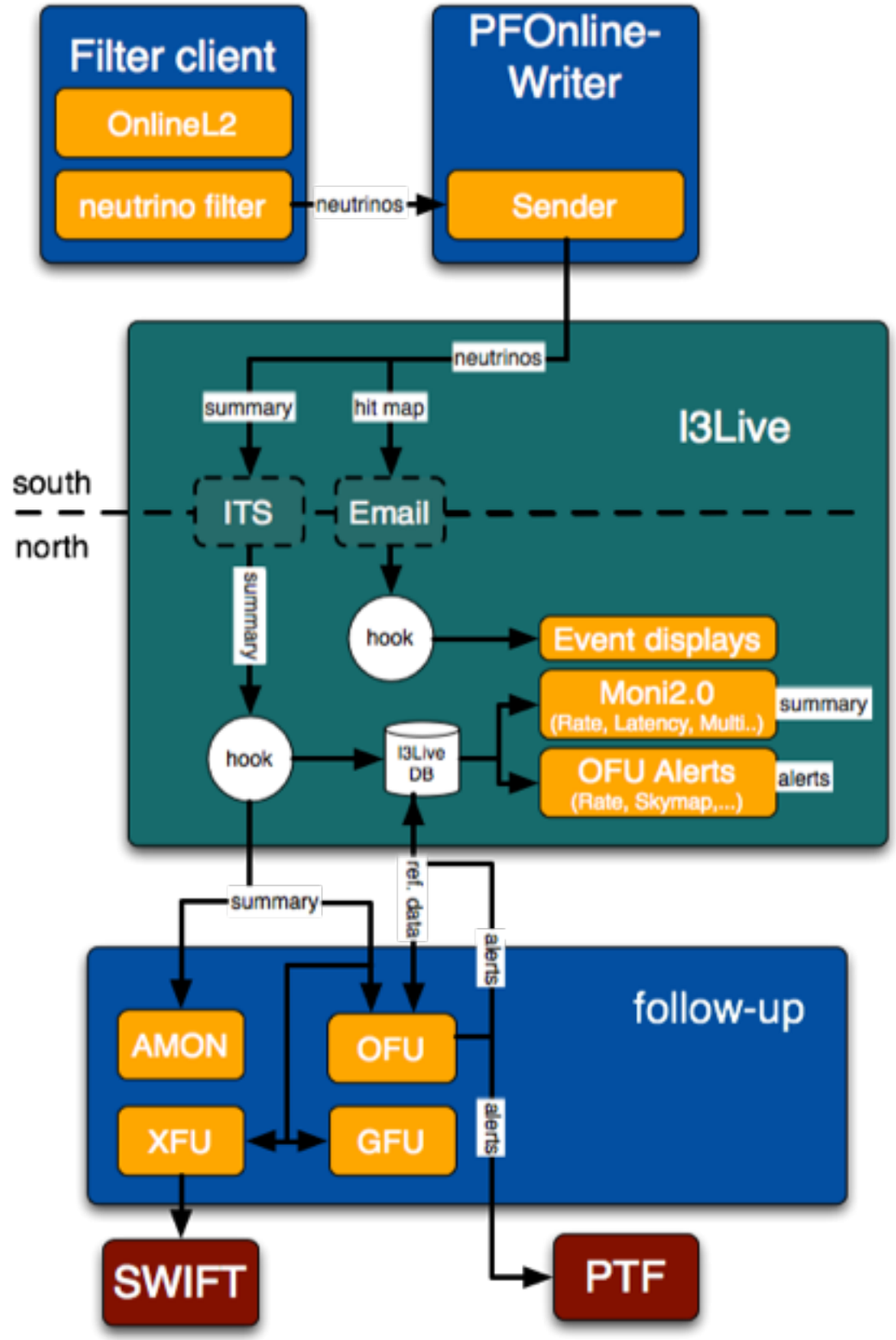
- For several years, IceCube has had a robust Online L2 selection, where:
  - Are selected from the entire sky by selecting high quality tracks (~5-6 Hz)
  - Apply more advanced and cpu-intensive reconstructions to these events (MPE fit, split topology fits, energy reconstructions)
  - Single neutrino candidates are selected from here and used to search for evidence of flaring sources
    - Neutrino doublets (<3 degrees in <100 seconds) will alert ROTSE, PTF and/or Swift XRT
    - Significant excesses of neutrino events from a catalog of potential TeV gamma ray sources will trigger followup observations by Magic or Veritas
- OnlineL2 selection and reconstructions are also used as a basis for several analyses
  - GRB track searches
    - OnlineL2 + event quality preselection + offlineBDT cut => neutrinos
  - Fast-response analysis (ATEL followups, Crab flare, etc)

Realtime astrophysical events

- Since the discovery of an astrophysical neutrino signal (HESE, tracks, cascades), we've had many requests for more prompt notifications
  - Several MOUs in place with optical, radio, gamma-ray, gravitational telescopes
    - If neutrinos arise from a transient source, prompt followup is required.
- Now deploying and testing neutrino selections to generate alerts in realtime as these events happen
  - HESE event selection
  - VHE track selection
- Expect this to expand to more astrophysical events

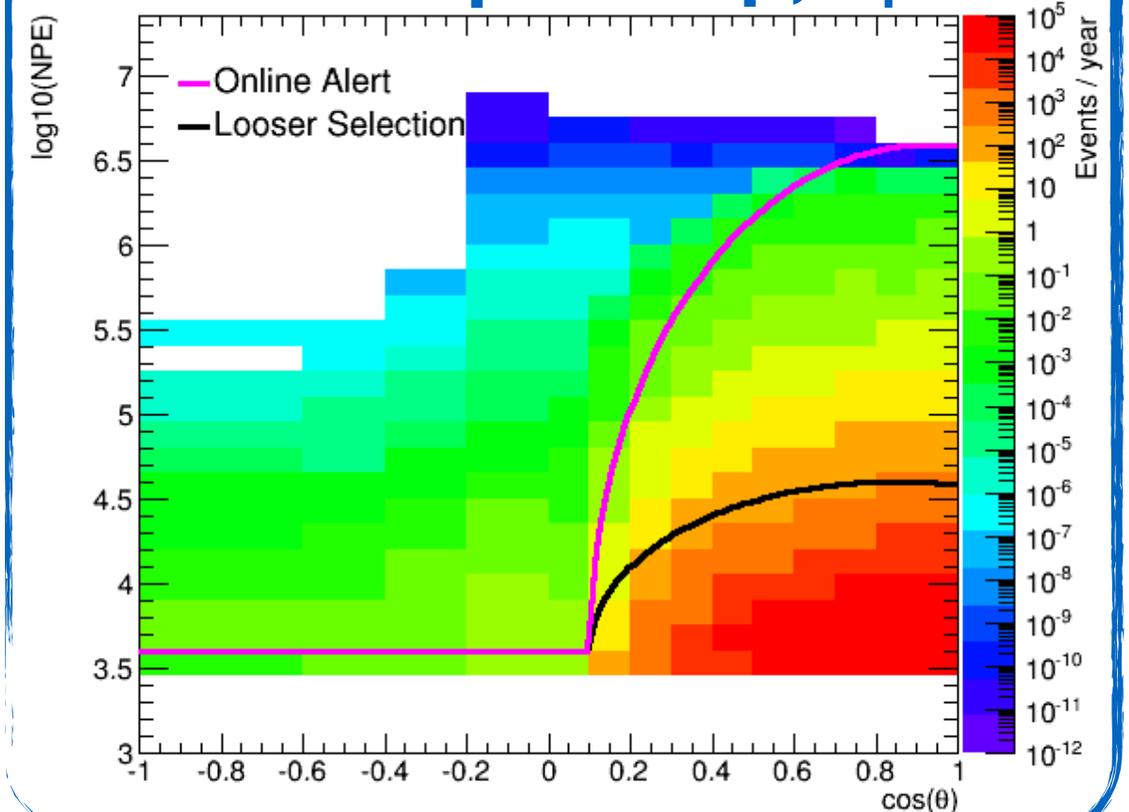
Tools

- A broad infrastructure to support these analyses us coming together
  - “Realtime” - event selection, reconstruction and alert generation are done as quickly as possible and alerts go out immediately.
    - General latencies are ~2-3 minutes
- “Near realtime” - additional event reconstructions and analysis are done as soon as the Filter selected data arrives at UW
  - General latencies are a few to ~24 hrs
- Making broad use of I3Live infrastructure to support alert/event rapid transmission north and detector monitoring.

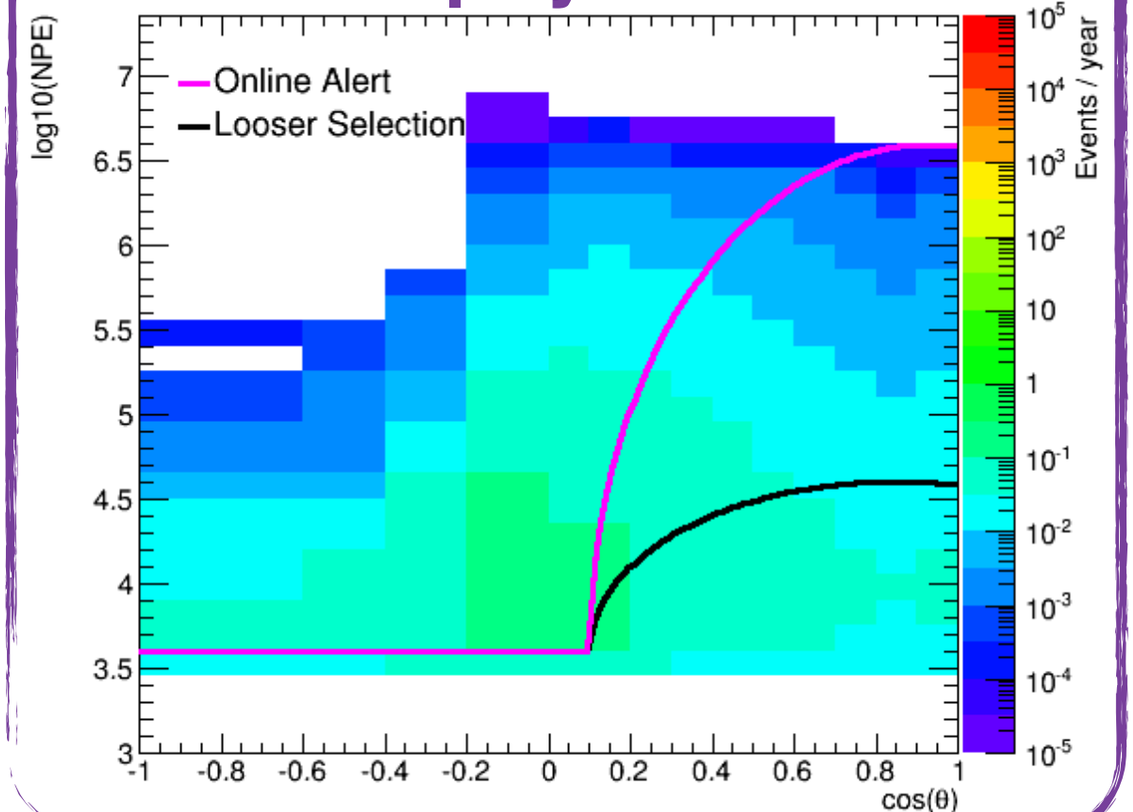


New OFU/GFU/XFU realtime system

Atmospheric μ, ν_μ

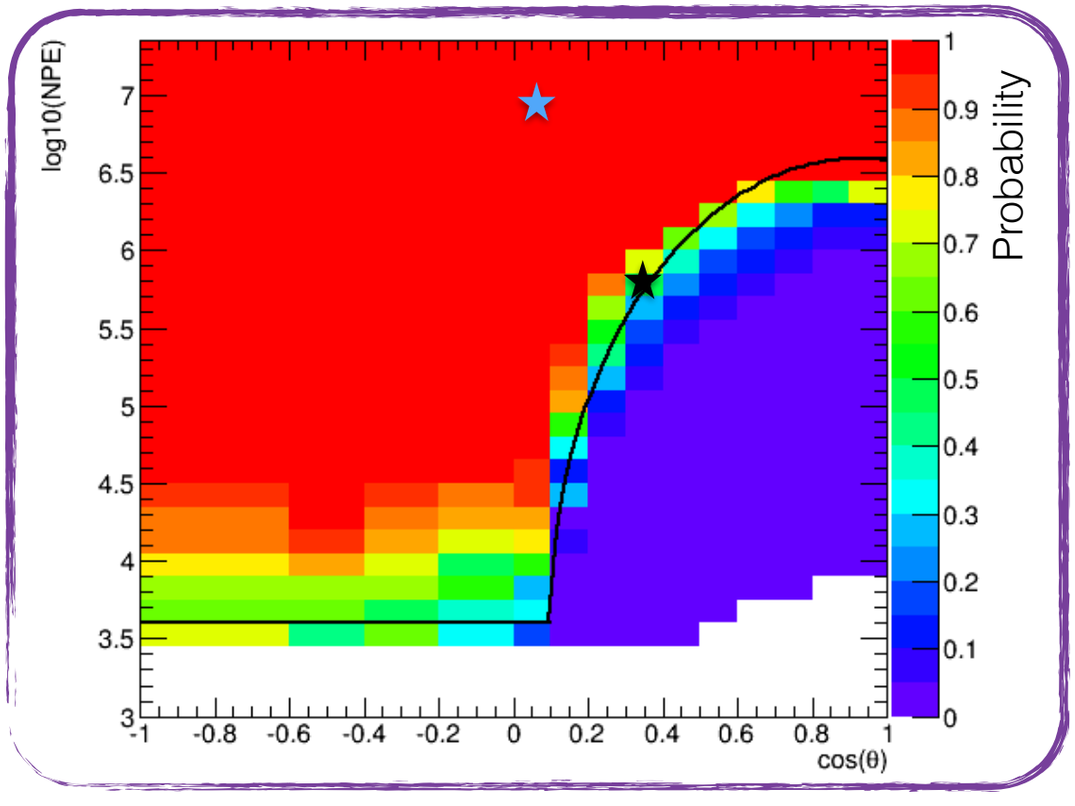


Astrophysical E^{-2}



Sample	Yield / year
Atmos. μ	0.52
Conv. Atmos. ν_μ	1.20
Prompt Atmos. ν_μ	0.19
Total Background	1.91
Astro. ν_μ	4.09

VHE track selection



Realtime concerns

- As these realtime systems are maturing, several concerns have come up that we are working through
  - Ensure realtime processes are robust
    - Missing strings can compromise vetos
    - Many Moni2.0 reports available in realtime
  - What to report
    - Generally: time, direction w/ error and "signalness"
    - Some online reconstructions not accurate
      - HESE cascades -> initial report with no direction, followup with results from fast followup
- Need to ensure any overlapping alerts have consistent content
- How to report: AMON GCN alerts (public alerts/MOU partners?)
  - Public alerts wider distribution, but concerns about being "scooped"?
  - MOU partner agreements have clear control over publication

External responses

- We need a uniform method for replying to externally generated alerts
  - E.g: VHE gamma-ray flares, ANTARES ATEL, LIGO
- In process of assembling several pieces to respond to these
  - Stable of well understood (sig+bkg response) analyses that have neutrino selections ***already*** run
- Machinery in place to easy run these analyses based on input time, direction, spatial extent
- Someone to check data quality, run machinery, check results
- Someone to authorize a quick reply
  - Currently this is only the analysis coordinator and spokesperson.

External responses

- Flaring shifters
  - Members of the collaboration would be responsible for running these analyses, checking results and data quality
  - Given rare nature of alerts: considering inserting fake alerts routinely
- Response panel
  - ~5 senior experts able to quickly meet, review flare search results, issue appropriate response
- Working to organize these responses over the next couple of months

Data releases

- IceCube astrophysical neutrinos are a valuable resources for the community. We want to support this
- Data releases in conjunction with publications
  - Careful evaluation of models by others possible with more detailed data releases.
    - Includes enough information (effective areas, resolutions, etc) to repeat analysis with alternative models.
    - Examples: HESE samples, thru going muon samples and, GRB model evaluation tools
      - http://www.icecube.wisc.edu/science/data
- Alerts are releases of single neutrino information

Data releases

- Considering wider data releases
  - Publication linked to a larger "IceCube neutrino catalogs" of some sort
    - Similar to the Fermi catalogs
    - Focus on astrophysical only?
      - All neutrino candidates (mostly atmospheric neutrinos)
        - Need careful consideration of systematic errors for lower energy neutrinos/oscillations
      - Again concerns about being "scooped" by others
        - Several analyses take time due to complicated systematics
- NSF mandated primary data release
  - Not particularly useful for others,

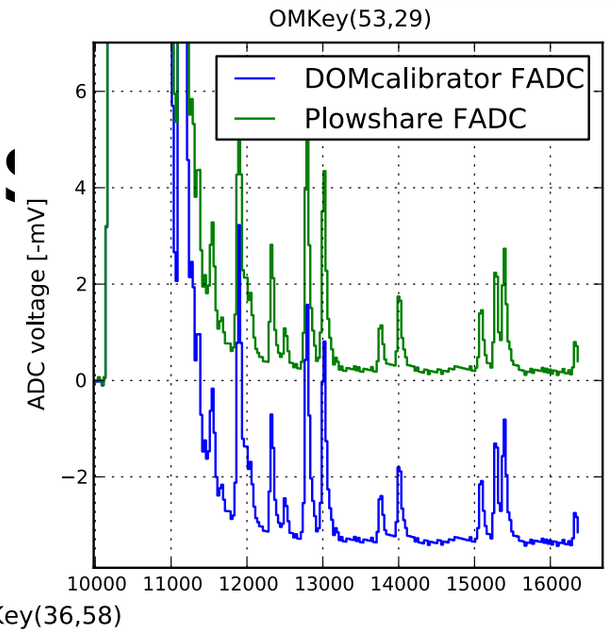
Summary

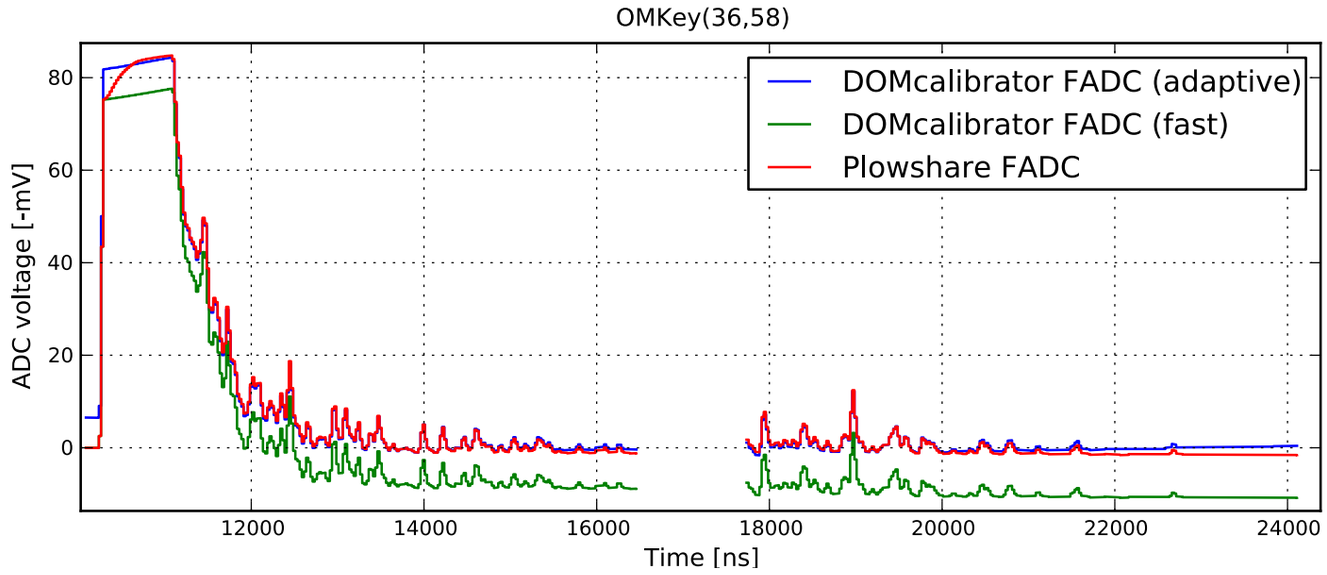
- Data processing has been streamlined over the last several seasons
  - *Time to analysis* of a data sample is decreasing for many key analyses
    - Robust and well characterized neutrino selections will be key in the move to more realtime analyses.
- Realtime searches and external alert response plans advancing
- Data releases to support scientific community

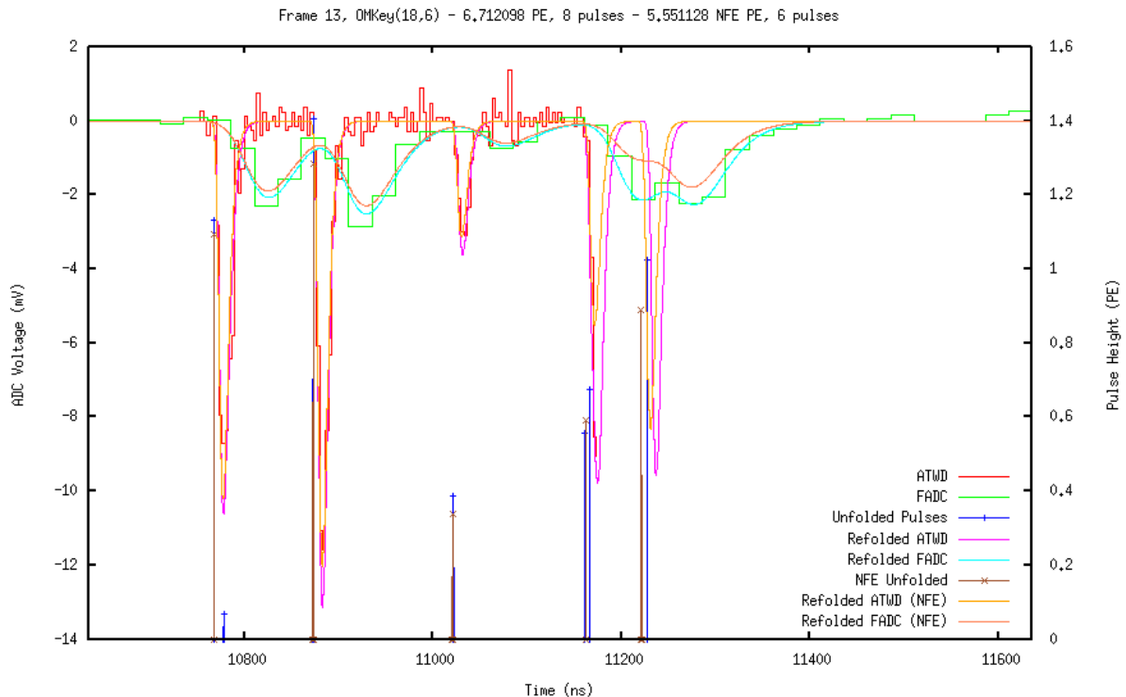
- Backup

2013-Improved calibrations

- Improved calibrations
  - Better baseline measurement and correction
  - Better droop correction
  - Seamless transition from ATWDs to FADC
- Improved feature extractions
  - Improved deconvolution of waveform into individual pulses
  - Naturally handles ATWD/ FADC transition
- Enabled:
  - Reconstructions that use ALL pulses
  - Better energy reconstruction tools with improved resolution

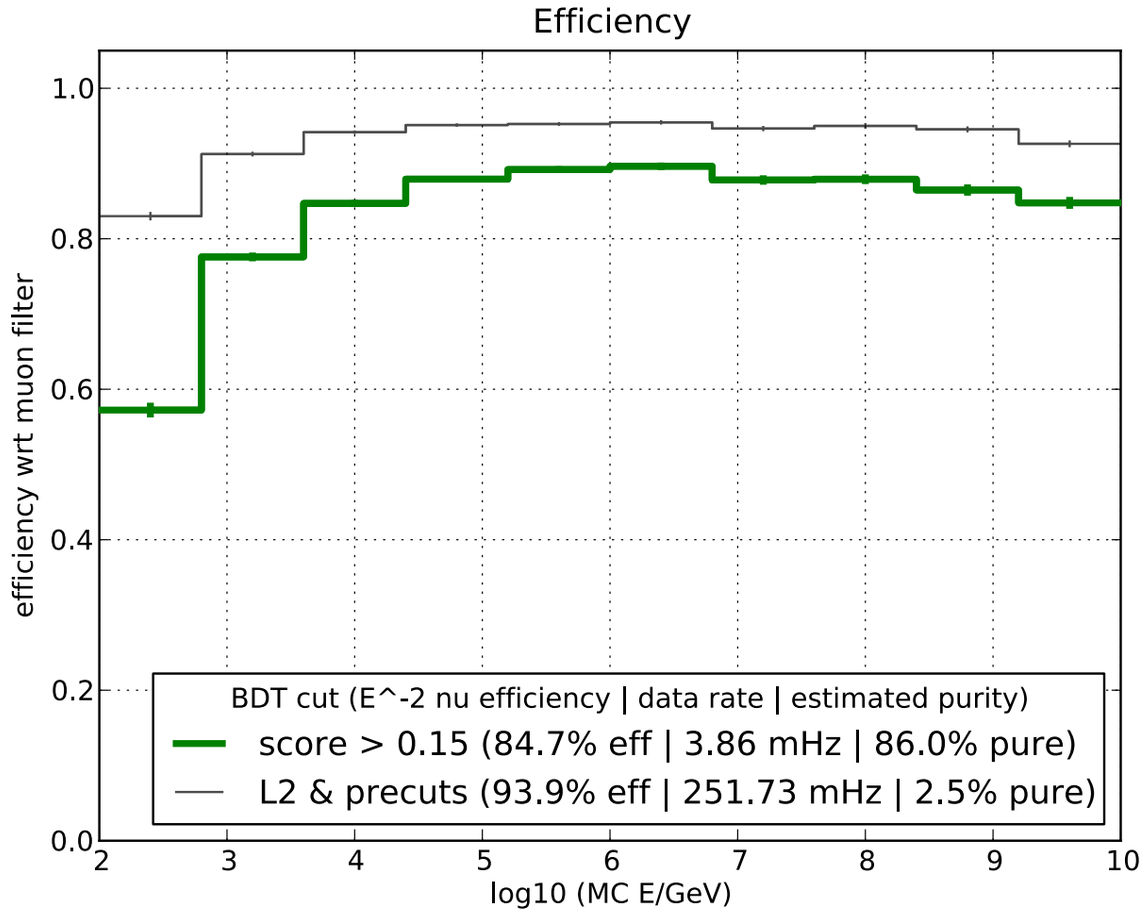






Maturation of Online analyses(2)

- This effort paying dividends: IC86-2011 GRB analysis
  - Based completely on online neutrino selection
  - OnlineL2 + event quality preselection + offlineBDT cut => neutrinos



Highest efficiency GRB search to date, obtained with values we calculate in realtime ONLINE.

Online analysis latency