Infrastructure

Benedikt Riedel
UW-Madison

IceCube SCAP 2021
Jan 27 2021

ICECUBE
SOUTH POLE NEUTRINO OBSERVATORY

NSF

Infrastructure Changes since 2018

- OneNeck (commercial cohosting facility) replaced 222 West Washington in Dec 2019/Jan 2020
- Adding more external resources
  - Now getting more resources from outside WIPAC
  - Large-Scale Community Partnership Allocation at Frontera
- Exploring and transitioning to the cloud
  - E-CAS Phase 1 for Multi-Messenger Astrophysics
  - Cloud Burst experiments – See next talk for details
  - Transitioning more services to the cloud

Recommendations

- 2018-4: Single Sign On
- 2018-5: Unified data organization, management, and access
- 2018-9: Workflow and workload improvements
- 2018-10: Add resources
  - 10x increase resources
  - Adding more collaboration resources

Challenges

- IceCube is in transition – Discovery to Precision Science
- Diminishing returns of more hardware
  - Hardware/hosting we can afford will not solve resource shortfall
  - More hardware will be used inefficiently
- Significant technical debt to address
  - Evolving landscape for scientific computing, code, etc.
  - Data movement in simulation and data processing
  - CPU and GPU efficiency
  - Adjusting for resource pool – Larger resources for shorter time
- Machine Learning/Artificial Intelligence on the rise:
  - Specialized hardware required – Some only available in the cloud
  - More interactive computing – Iterative rather than batch computing

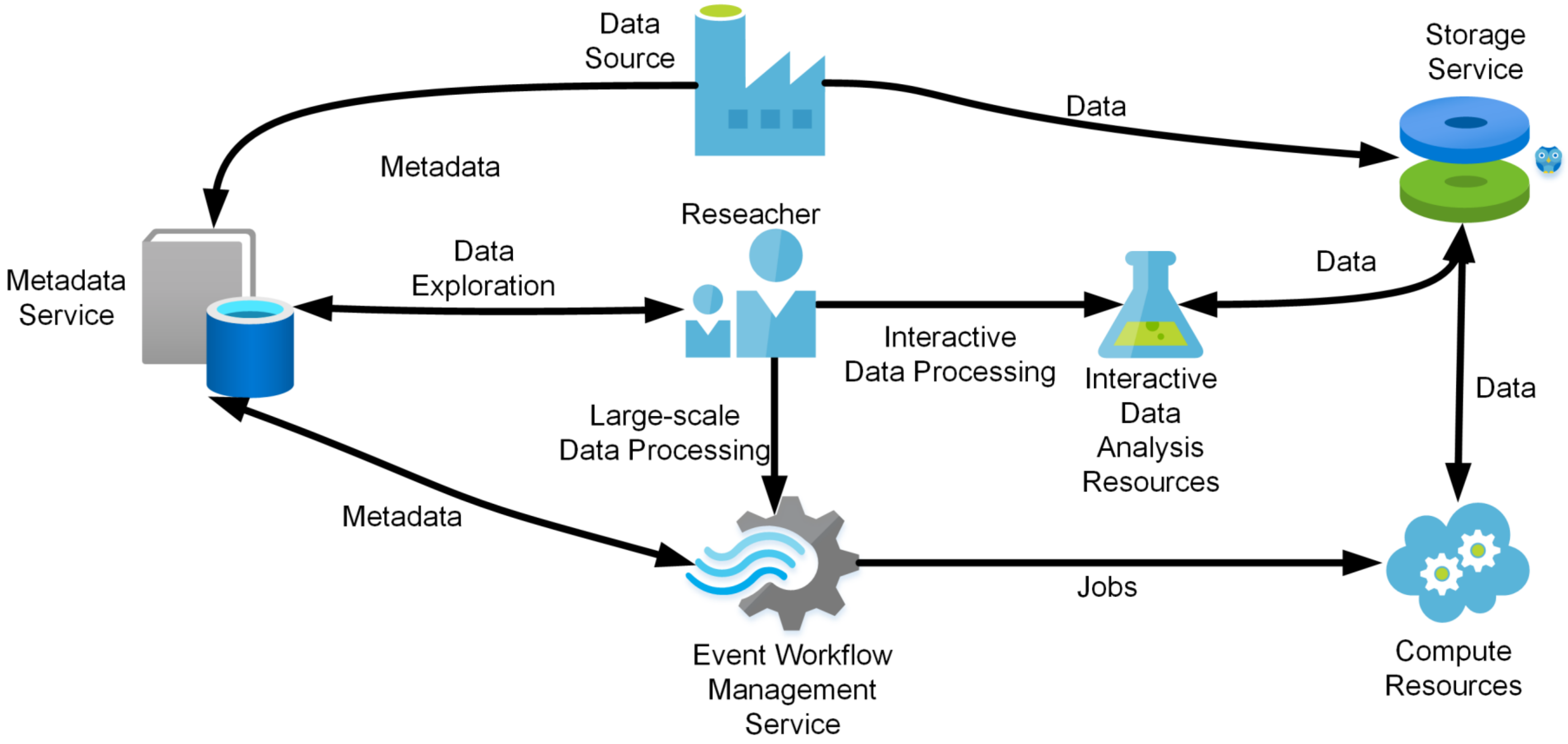
Single-Sign On/User Management

- Picked **Keycloak** for user management and SSO provider through SAML/OIDC/OAuth2
  - Considered cloud vendors (Okta, Auth0, AWS, GCP) and self-hosted version (COManage, KeyCloak)
  - Cloud vendors too expensive and syncing for SSH access not always a supported feature
  - COManage had no clear upgrade path – IceCube is growing, unclear how we want to differentiate between different parts
- Currently deploying and creating automated user sign-up system
- Biggest hurdle is LDAP <-> Keycloak for SSH connections

Data Management

- First steps towards a data management platform
  - File Catalog
    - What experimental data and simulations is where
    - File integrity check
  - Long-Term Archive – Automated data check and backup of raw data coming from pole to NERSC
- Storage infrastructure needs re-design
  - Alternative filesystems
    - Ceph has matured as an alternative for Lustre
    - Currently in the R&D phase – Smallish Ceph cluster to test experience and learn before putting all the data on it
    - POSIX filesystems no longer preferred interface for distributed filesystem
  - Data size is becoming unwieldy for individual users, analysis groups are moving to common samples – Metadata can help people

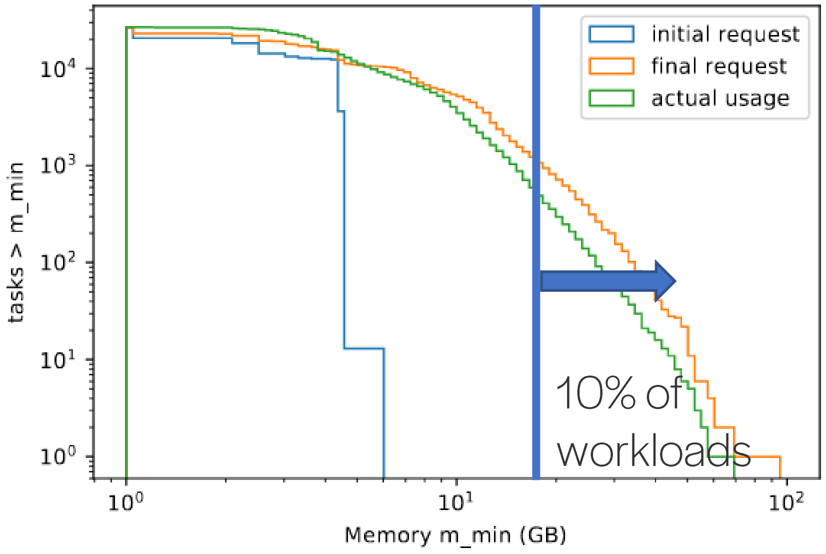
Data Management


Data Source
Data
Storage Service
Metadata
Reseacher
Data Exploration
Data
Metadata Service
Interactive Data Processing
Interactive Data Analysis Resources
Data
Large-scale Data Processing
Metadata
Jobs
Event Workflow Management Service
Compute Resources

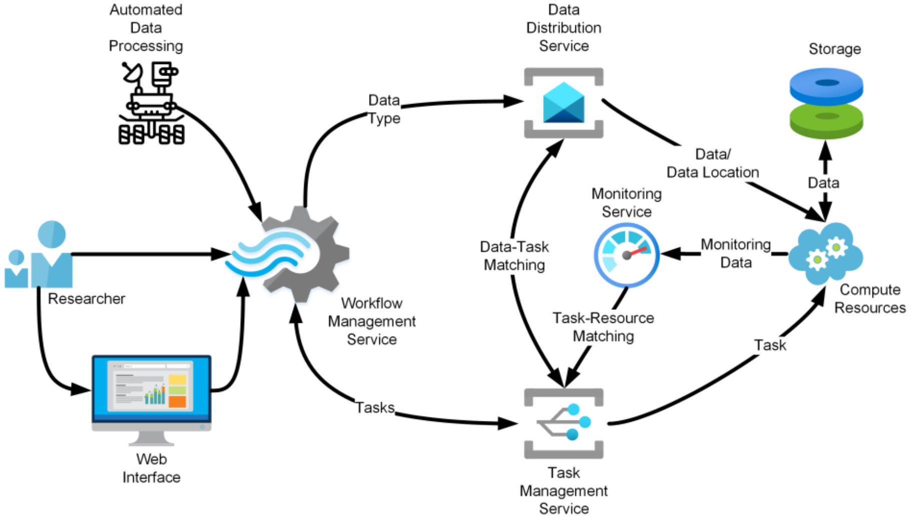
- Observation Management Service
- Data management and workflow pieces
- Researchers are meant to operate on data using metadata
- Storage details are not important
- Submitted proposal to fund part of development

Workflow and Workload

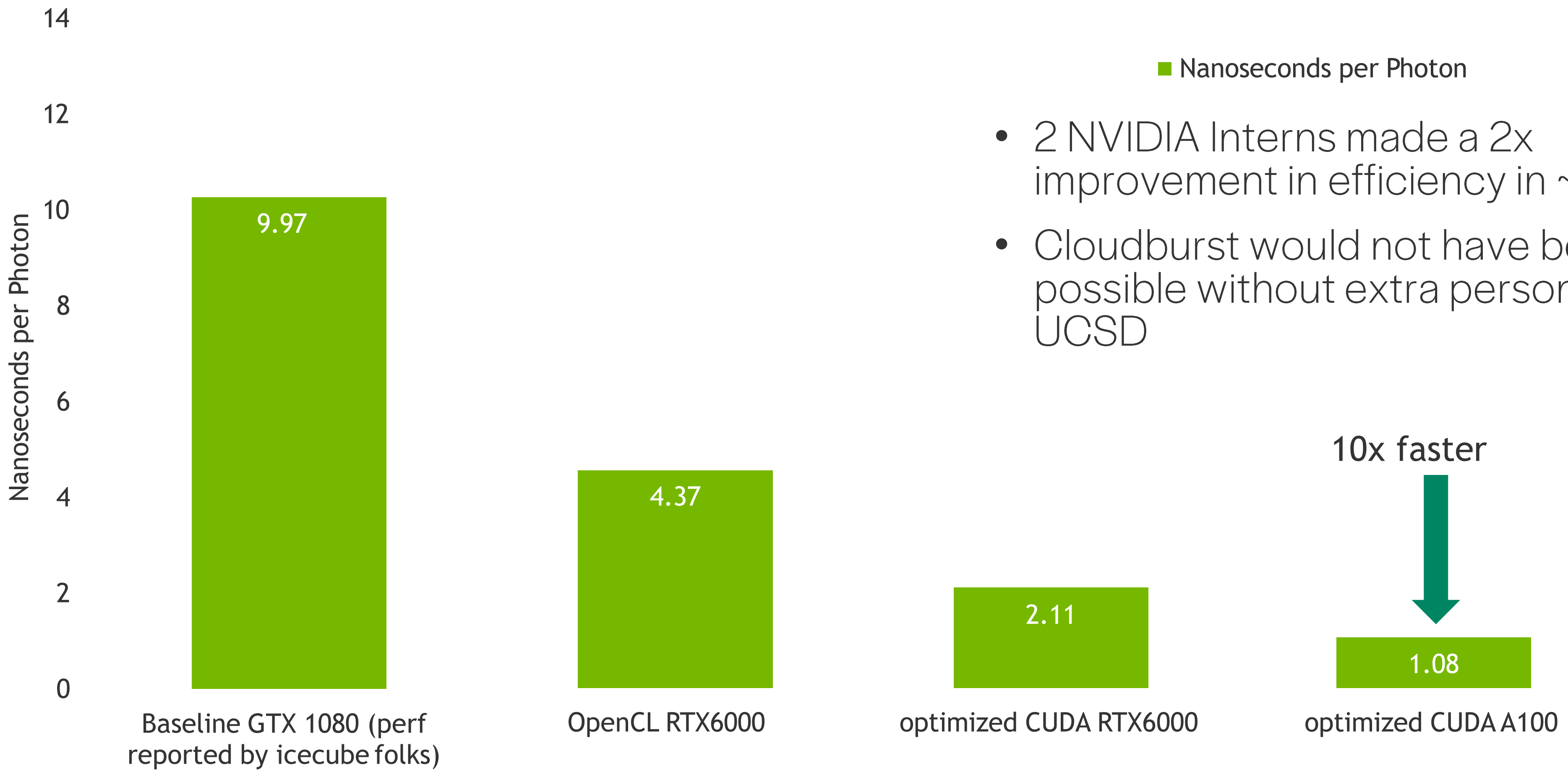
- IceProd2 has now replaced IceProd1 for all production needs – More in D. Schultz's talk
- Optimizing workflow and workloads lacks person-power
  - Many places for optimization and simplification – Person power is missing
  - Memory usage of reconstructions (light yield parametrization biggest consumer)
  - Mixing high and low memory events
  - CORSIKA simulation
    - Consumes 34-98% of workflow run time
    - 1M showers ≅ 2 seconds detector lifetime
  - Data transfer to/from job – Time wasted waiting for data movement



10% of workloads



Workflow and Workload – Why person-power?



- 2 NVIDIA Interns made a 2x improvement in efficiency in ~6 months
- Cloudburst would not have been possible without extra person from UCSD

Non-shuffled case using 1.46e8 photons

Non M&O Funding

- Awarded
  - CESER Award – Funds Long-Term Archive development
  - Exploring Cloud for the Acceleration of Science Phase 1 – Cloud funds
  - EAGER for Exa-Scale Demo – Awarded through UCSD, only cloud credits
- Not awarded
  - Convergence Accelerator – Data portal combining neutrino events and gamma-ray catalog
  - Humans Advancing Research in the Cloud
  - Mid-Scale R1 – GPU cluster hosted at UCSD
  - 4 AI Institutes
  - CC* with UW HEP group
  - Requested resources from PRACE – No definite response
- Submitted
  - CSSI for Workflow Management with Events
  - CSSI for Machine learning work
  - MRI with UW researchers
  - 4 HDR Institute proposals
- Planned
  - Keeping our eyes open for DOE opportunities

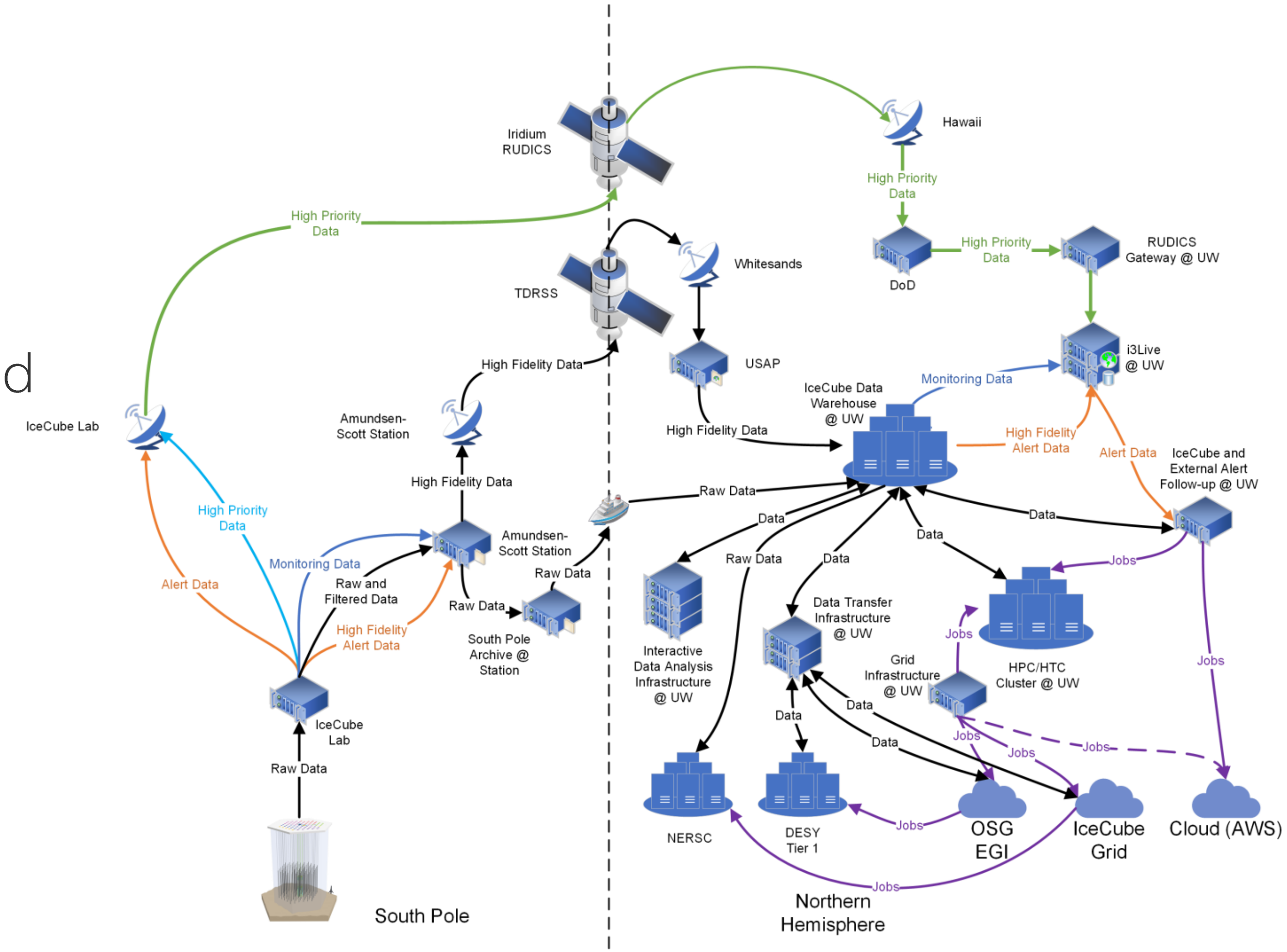
Thank you!

Questions?

BACKUP

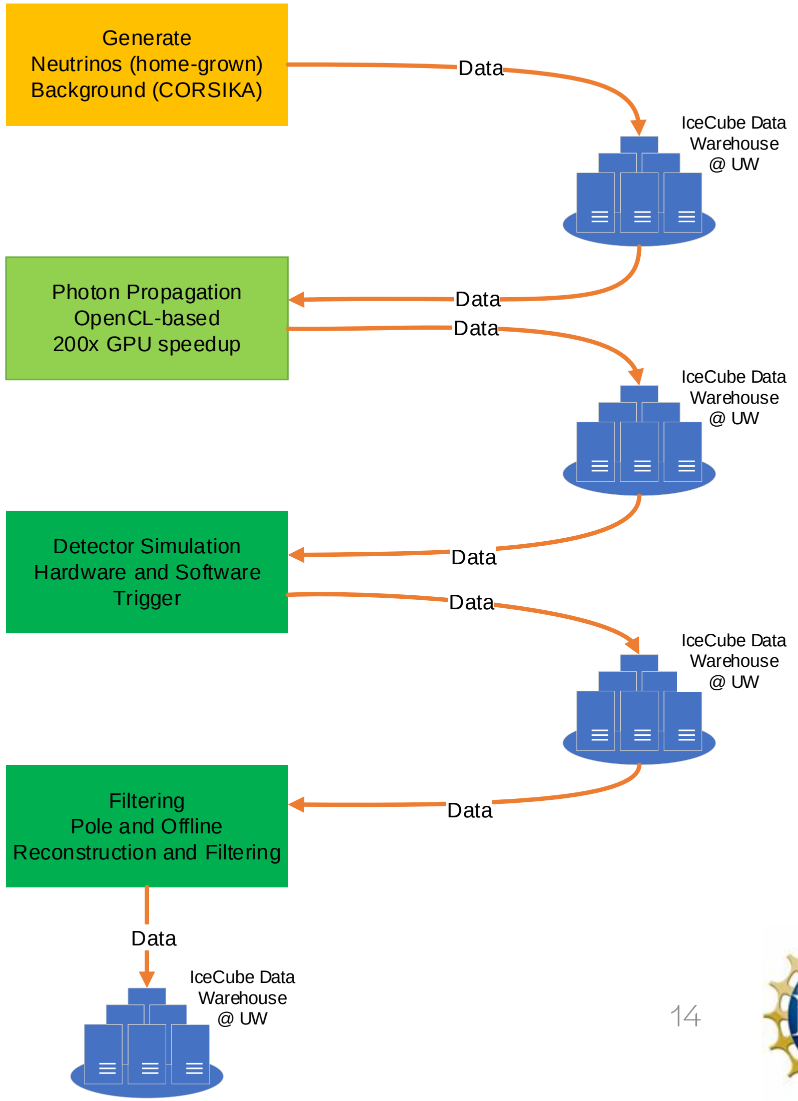
Data Flow

- Pole Filtered Data arrives via satellite - Arrives at UW-Madison and is processed further
- Raw data is written to archival disk at pole, retrieved once a year
- Raw data is archived at NERSC
- Filtered data is archived at DESY



Simulation Chain

- Fairly straightforward particle physics-like workflow
- Big constraint is lack of dedicated resources
    - No data aware scheduling
    - Lots of data movement – Lots of time wasted to move data
- Different steps can have drastically different requirements



Generate
Neutrinos (home-grown)
Background (CORSIKA)
Data
IceCube Data Warehouse @ UW
Data
Photon Propagation
OpenCL-based
200x GPU speedup
Data
IceCube Data Warehouse @ UW
Data
Detector Simulation
Hardware and Software
Trigger
Data
IceCube Data Warehouse @ UW
Data
Filtering
Pole and Offline
Reconstruction and Filtering
Data
IceCube Data Warehouse @ UW