NSF Large Facilities Cyberinfrastructure Workshop

Panel on Integration, interoperability and reuse of CI solutions, practices
Sep 6 2017

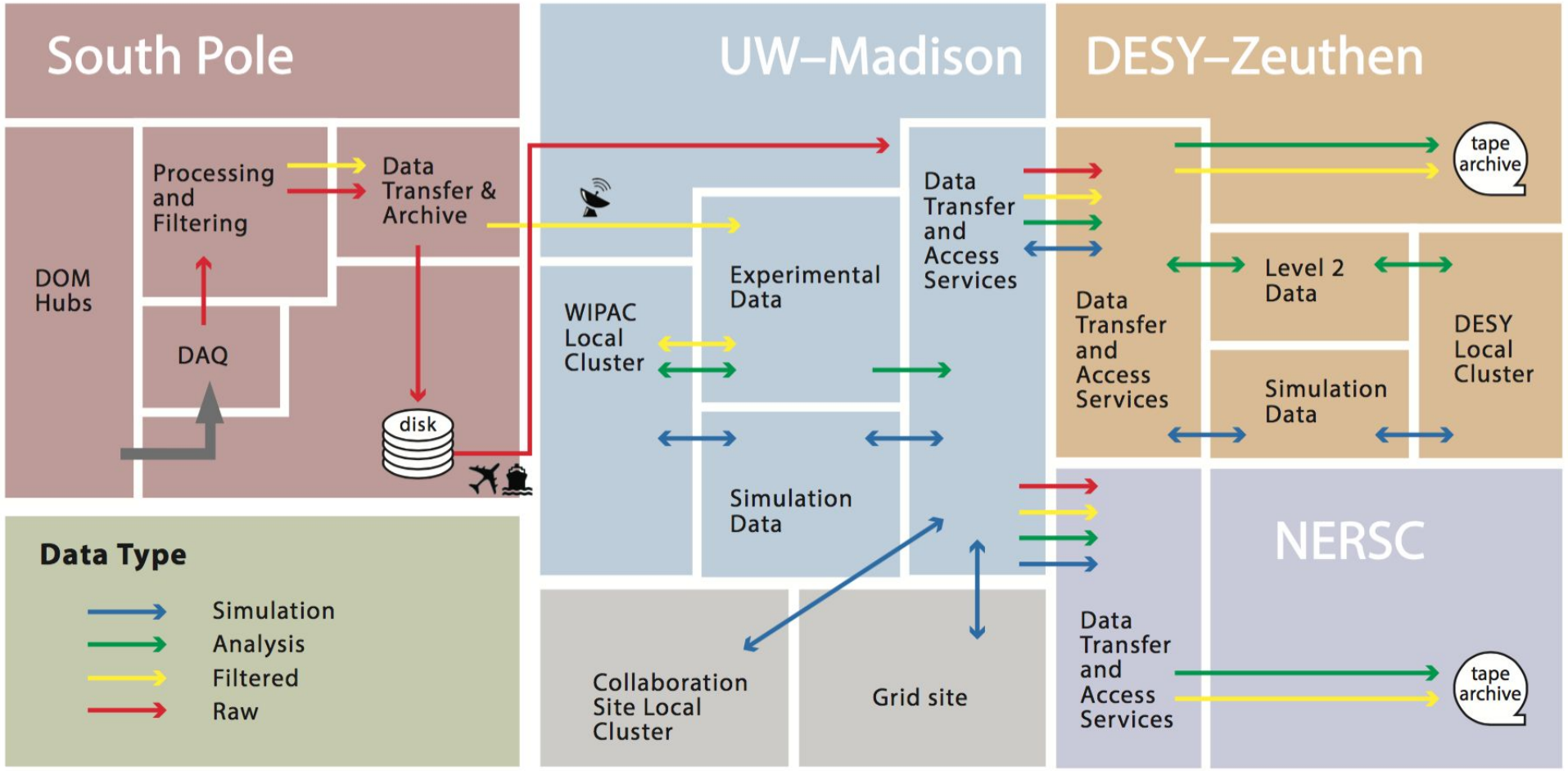
Cyberinfrastructure

Research environment that supports

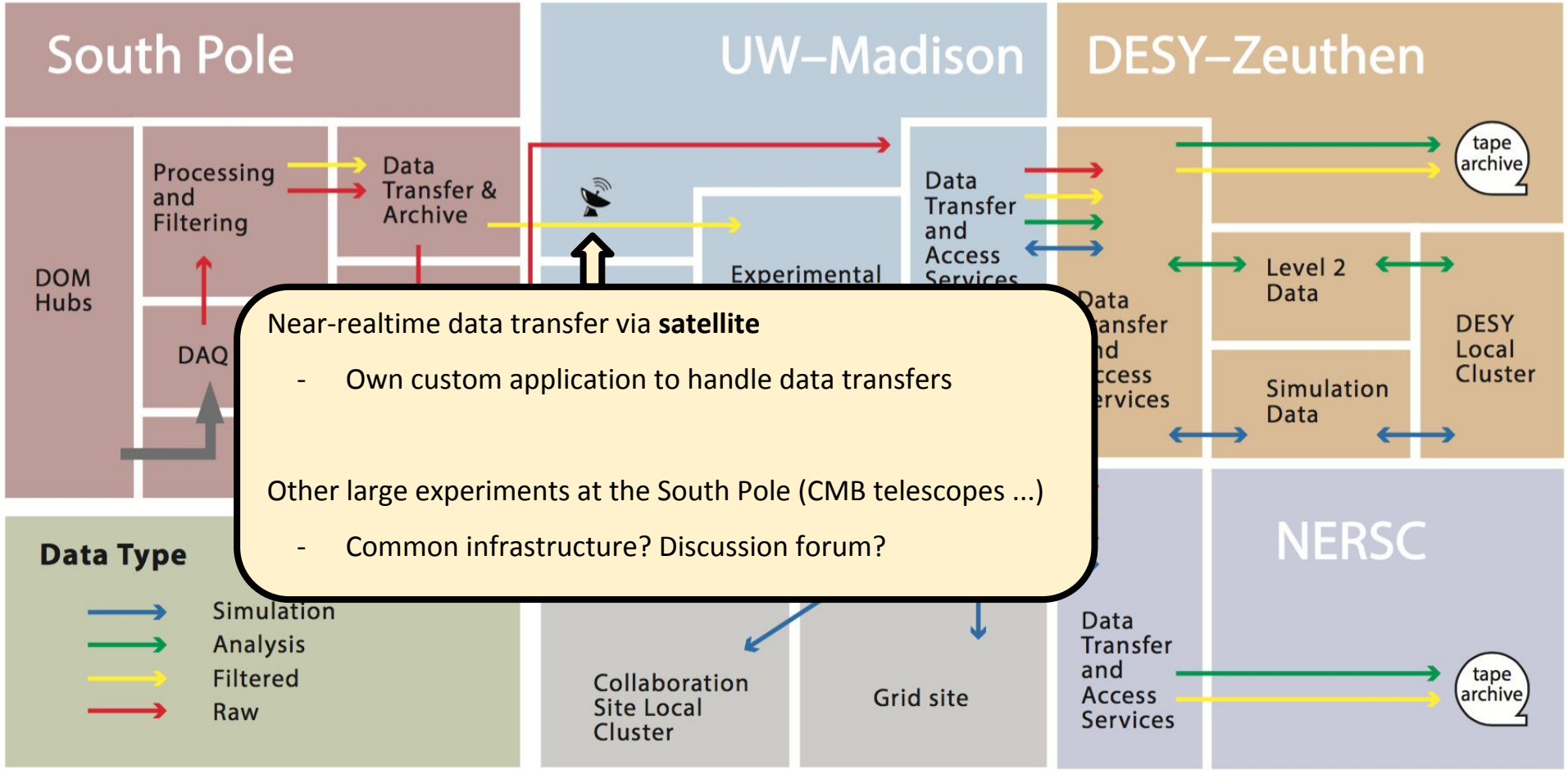
- **Data** acquisition
- **Data** storage
- **Data** management
- **Data** analysis
- **Data** integration

Technological and sociological solution to the problem of efficiently **connecting** laboratories, data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.

IceCube Data Flow



IceCube Data Flow



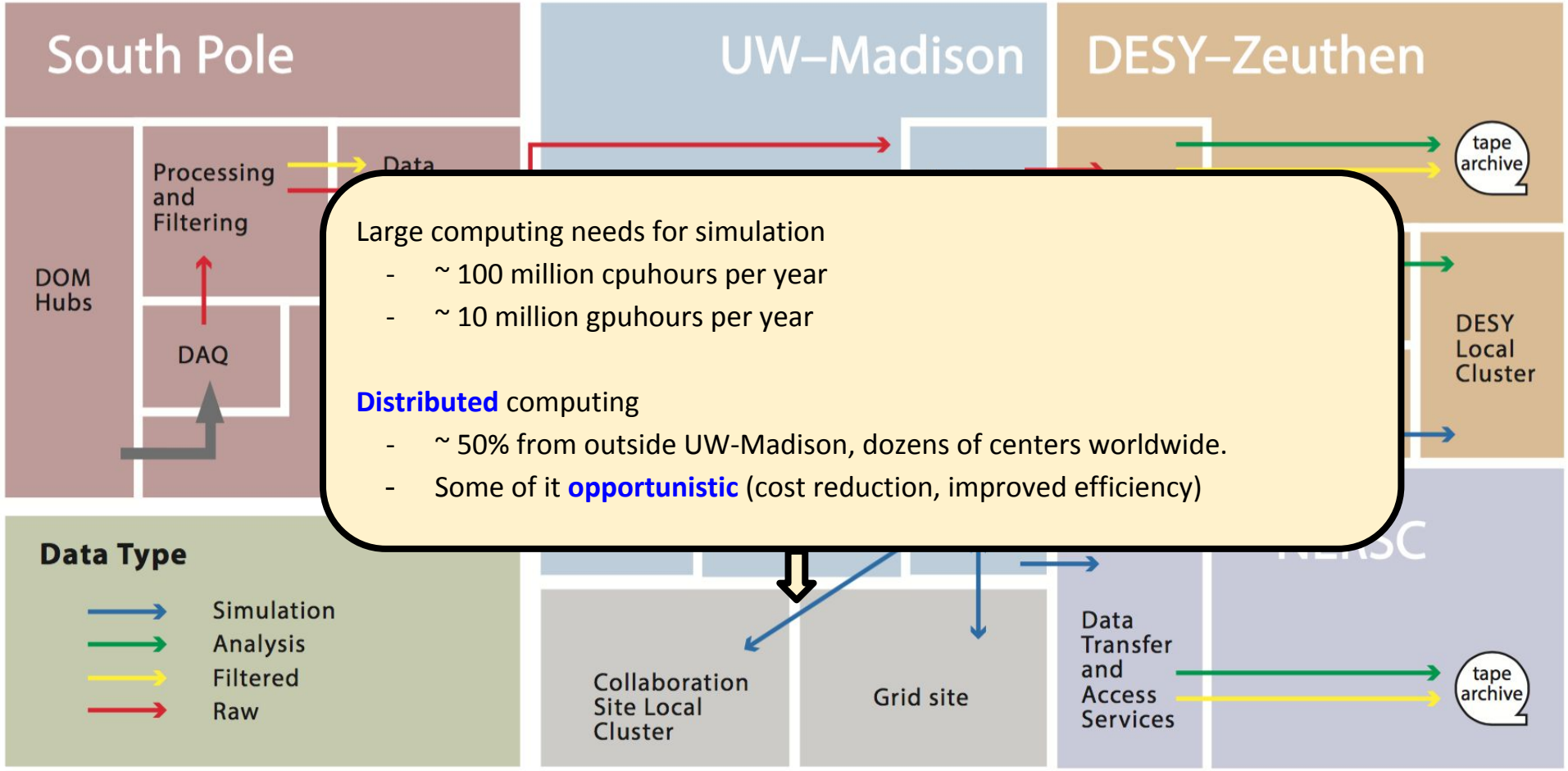
Near-realtime data transfer via **satellite**

- Own custom application to handle data transfers

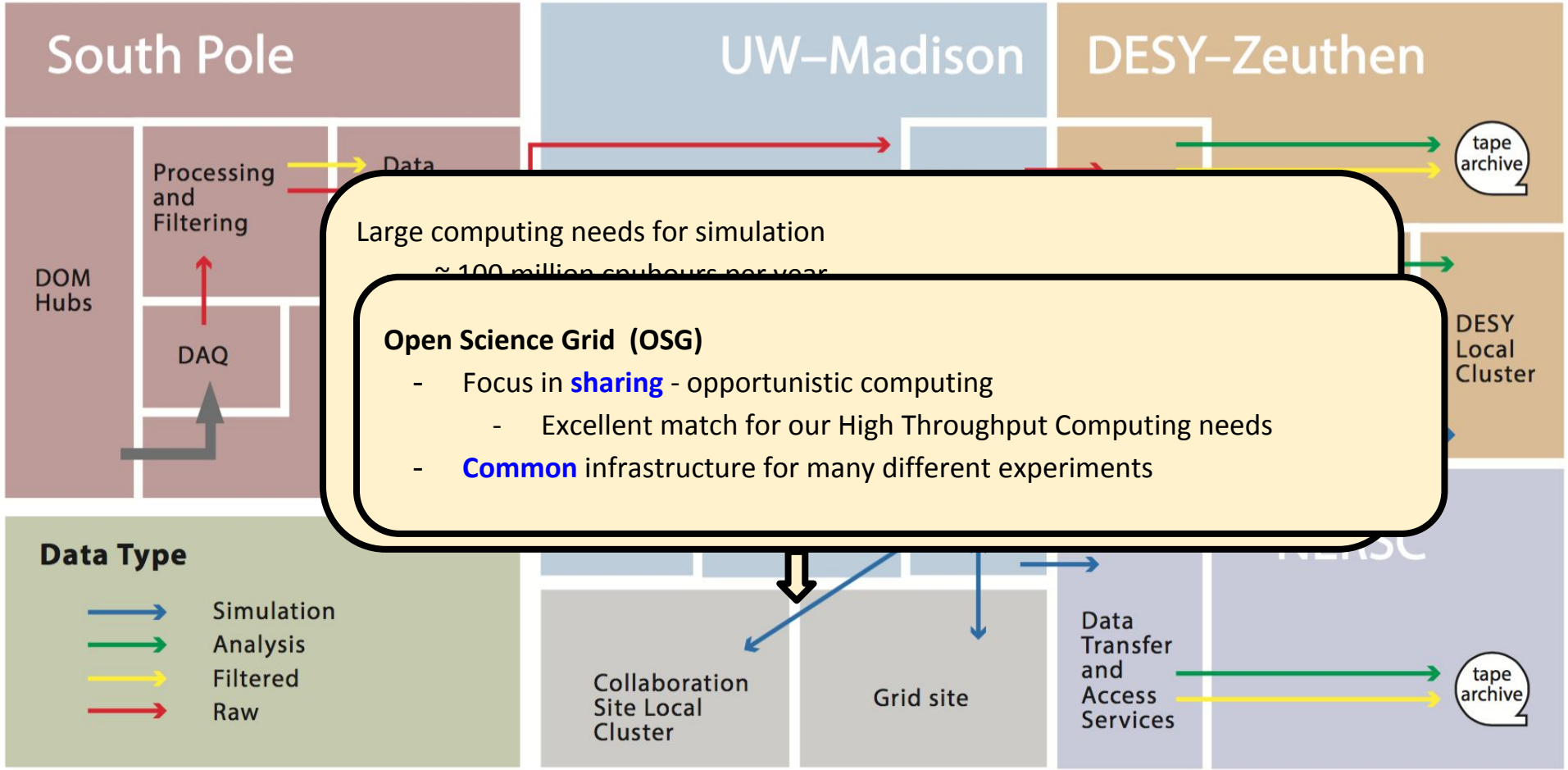
Other large experiments at the South Pole (CMB telescopes ...)

- Common infrastructure? Discussion forum?

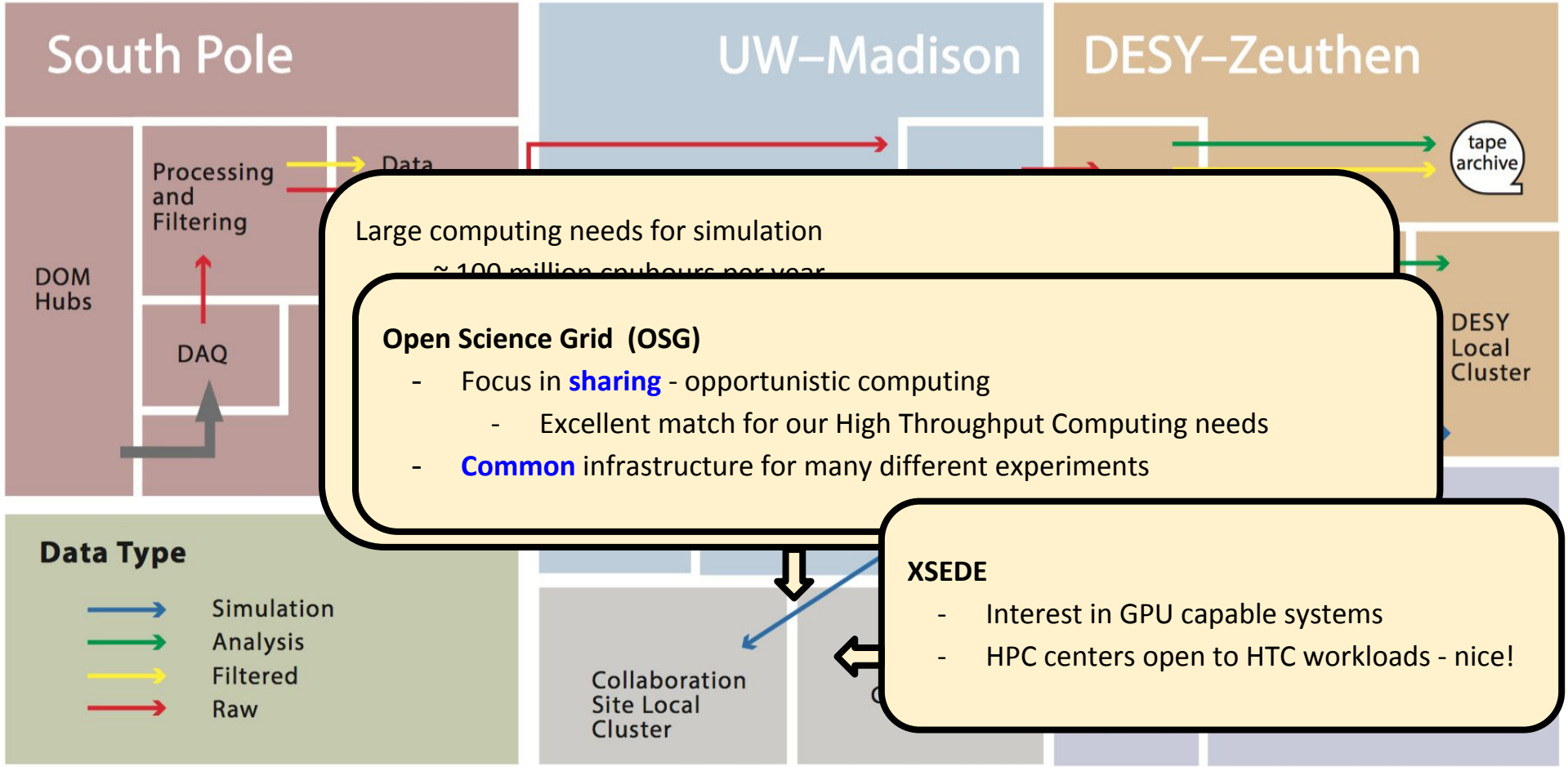
IceCube Data Flow



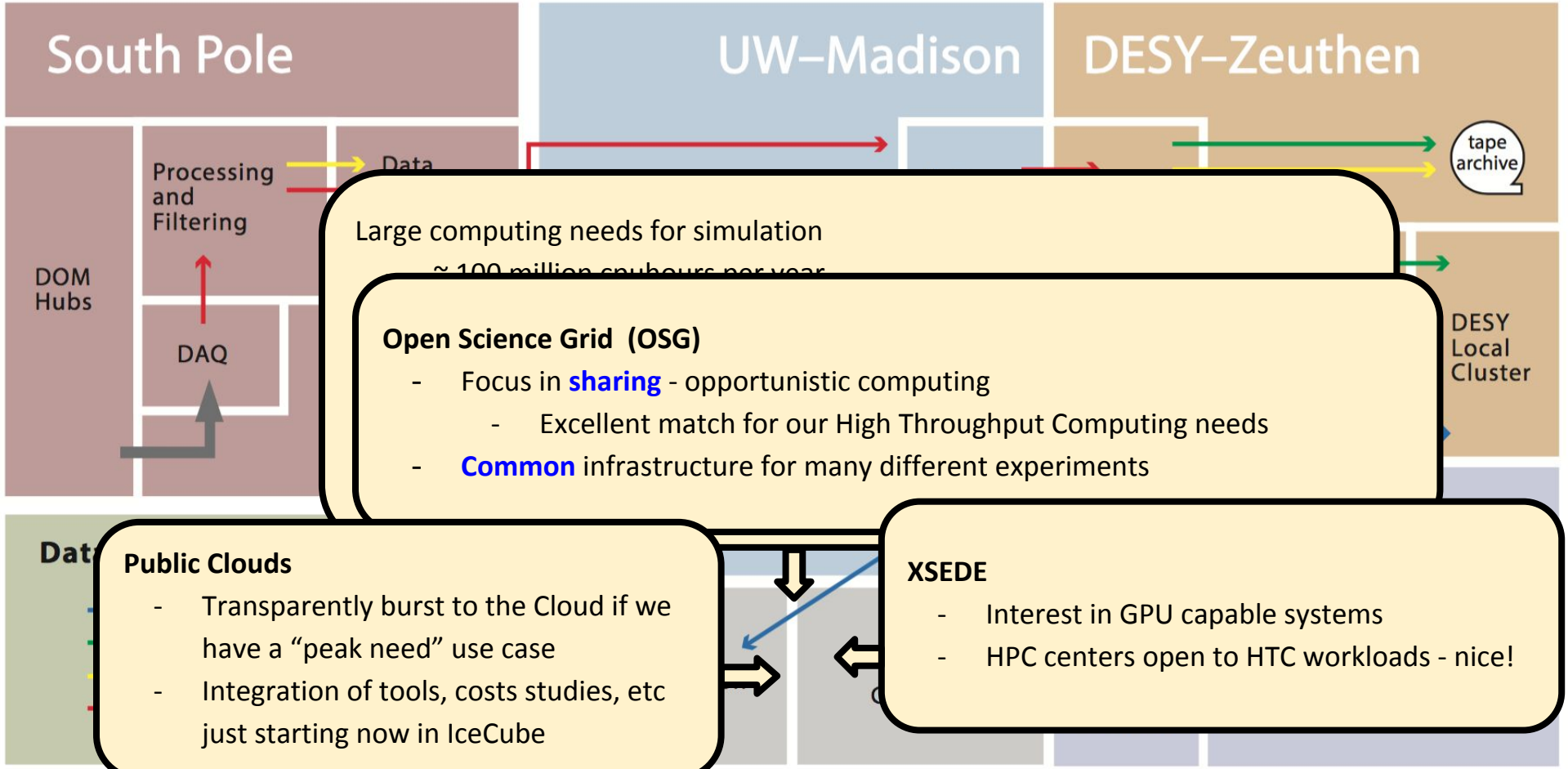
IceCube Data Flow



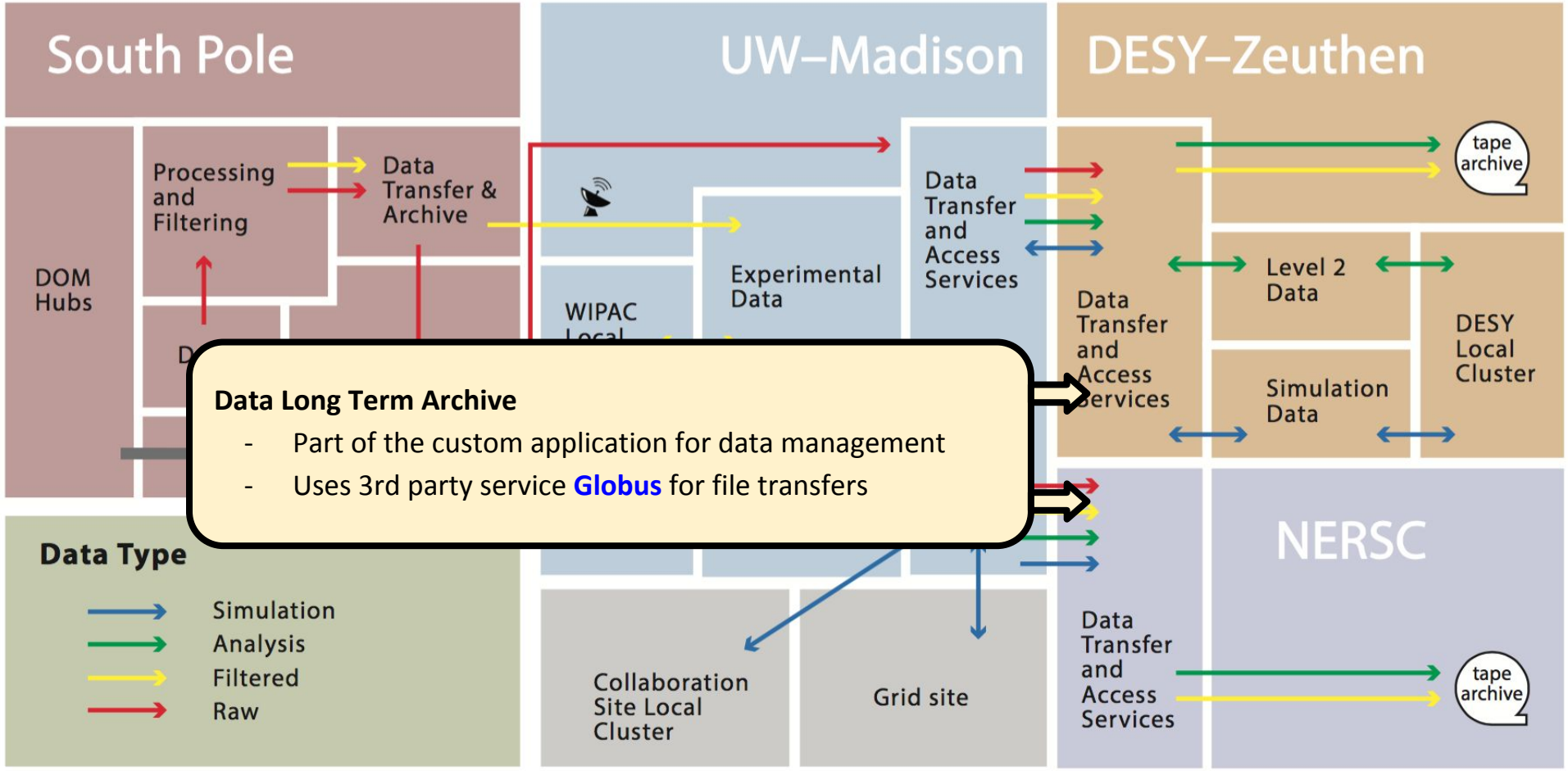
IceCube Data Flow



IceCube Data Flow



IceCube Data Flow



How do your projects discover and evaluate available solutions?

Acknowledge that we are “not alone”, try to be active part of larger community:

- Experimental Particle Physics, High Throughput Computing, Distributed Computing, Polar Sciences …

… in practice (email forums & regular workshops)

- Open Science Grid
- Computing in High Energy Physics
- HTCondor, Globus
- National Data Service
- Polar CI workshop, NSF Polar CI Program

How do your projects deal with changing availability of CI?

Try to keep abreast of new CI by staying connected to the community (workshops ...)

- Currently rely a lot on permanent professional IT staff to “become experts” of available CI and to integrate it with our scientific workloads
    - Often a steep learning curve
- Some good experiences with external support - XSEDE ECSS

Can increased awareness/reuse of CI solutions increase interoperability across facilities?

YES!

For us, examples of success

- OSG/XSEDE/HTCondor for distributed workload management
- Globus for data transfer services

… of course the question becomes then “sustainability” …

Can community efforts in integration, interoperability and sustainability lead to well defined interfaces that facilitate access to and incorporation of new technologies?

YES, BUT … there is no lack of challenges

Example: current end of support announced by Globus on all x509 auth libs and tools.

- Globus-url-copy transition to Globus CLI/Python SDK and the Globus cloud transfer service.

Focus:

- **integrating** existing infrastructure rather than top-down development
- **open** interfaces, provide more than one alternative - sustainability

What are the most critical CI gaps that you would like to be addressed?

Curation, long-term preservation and sharing of data, reproducibility of scientific results.

Operational costs: tools exist to perform data analysis at large scale - but are still too complicated to operate

Identify and adopt new CI (open-source, widely used in industry ...) that can be applied to solving scientific data analysis problems

- Connecting academia to the rapidly growing “data science” or “bigdata” industry
- Funding for CI experts - competing with industry for “devops” or “data science” experts is a challenge

What are the most critical CI gaps that you would like to be addressed?

Domain gap (CS-ES) : difficulty for computer scientists to explain algorithms/methods to domain scientists, and vice versa

- Disconnect between the “fancy” techniques developed by CS researchers and the real questions that experimental scientists try to answer
  - Training non-CS researchers to be able to understand the output of CS research and its applicability to specific science problems
- Interdisciplinary partnerships “CS - Domain Science” are key for success

Inter-domain communication: scientists from different fields often do not interact much. There is potential for common CI, but it is hard.