# Data and Analysis Centres @ DESY & KIT

## Six Pillars in Particle Physics and beyond

Volker Guelzow (DESY) & Andreas Heiss (KIT)
Karlsruhe, November 2nd, 2017

DESY

HELMHOLTZ GEMEINSCHAFT

KIT Karlsruhe Institute of Technology

# Large Experiments supported today by DESY & KIT

- Atlas, CDF, CMS, LHCb, Alice, Belle & Belle 2, ILC, Babar, D0, Compass
- CTA, MAGIC, Icecube, HESS, Auger, FERMI, Taiga
- XFEL, Petra III, Flash

- DESY&KIT
- DESY only
- KIT only

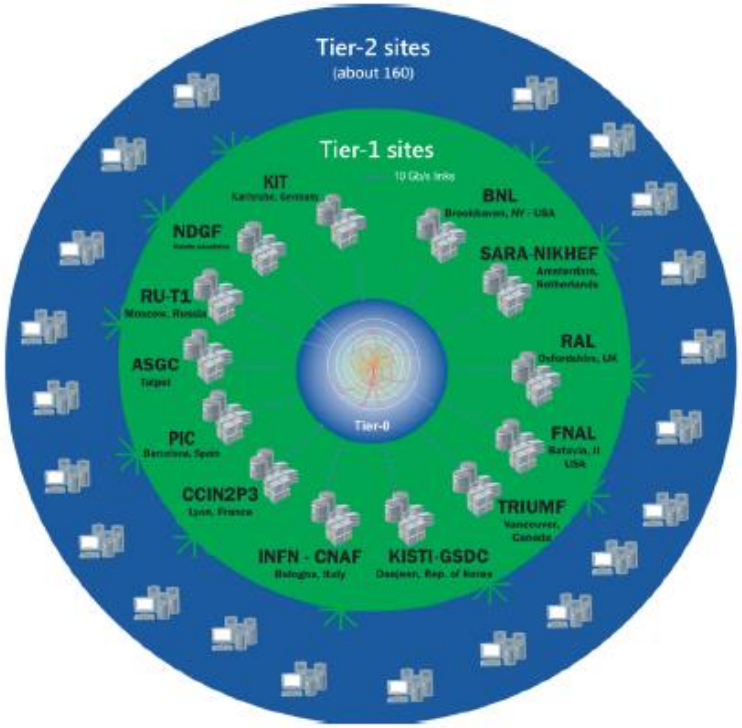
- This requires the highly professional operation of the very large computer centres @ DESY and KIT

# Worldwide LHC Computing Grid

**TIER-0 (CERN):**
data recording, reconstruction and distribution

**TIER-1:**
permanent storage, re-processing, analysis

**TIER-2:**
Simulation, end-user analysis




nearly 170 sites, 40 countries

> 500k cores

~ 1 EB of storage

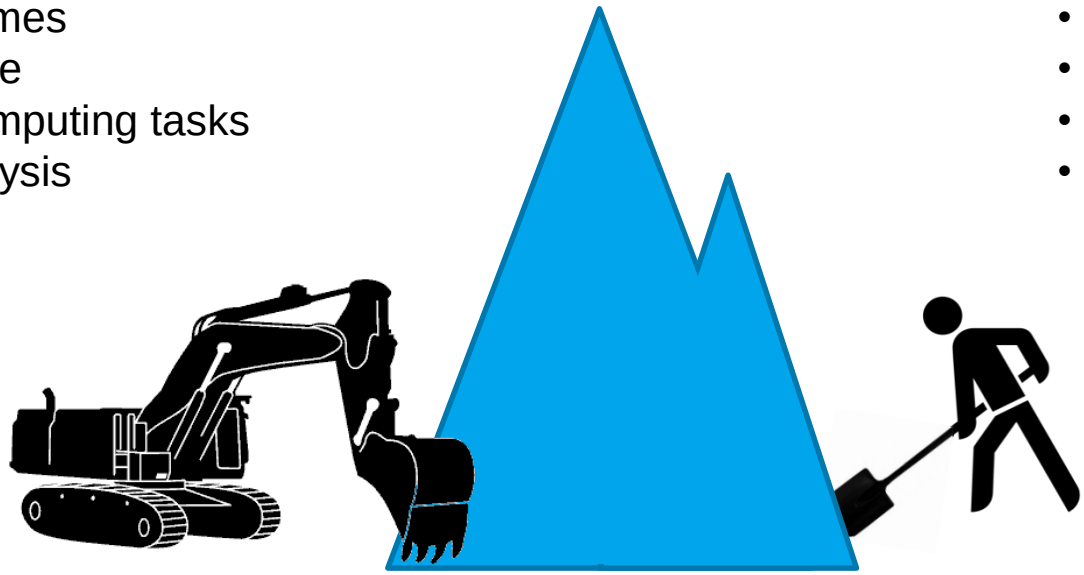
> 2 million jobs/day

10-100 Gb links

# National Analysis Farms (NAF) - Grid versus NAF

The Grid World
- batch
- centralized
- large volumes
- less flexible
- central computing tasks
- some analysis

The NAF World
- interactive
  - i.e. PROOF
- flexible
- fast turnaround
- final analysis
- user driven



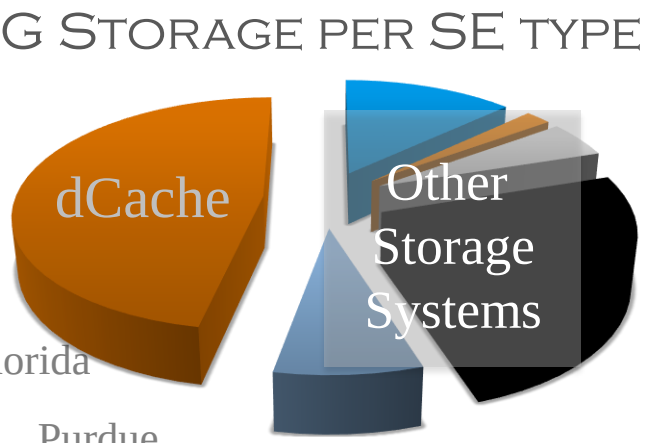
**LHC Data – the treasure**

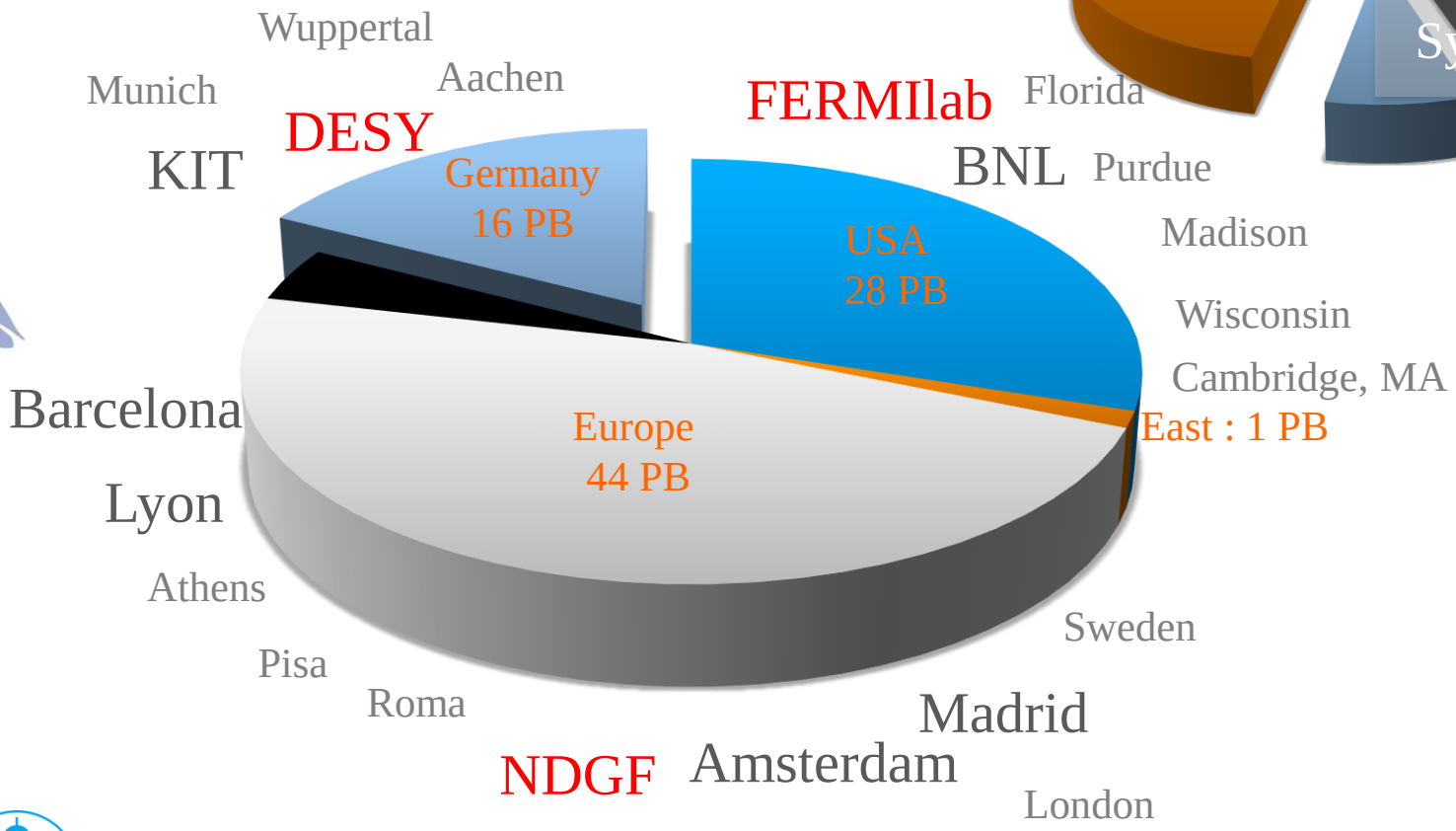
Worldwide User Community

German User Community

# dCache, a Data Management package

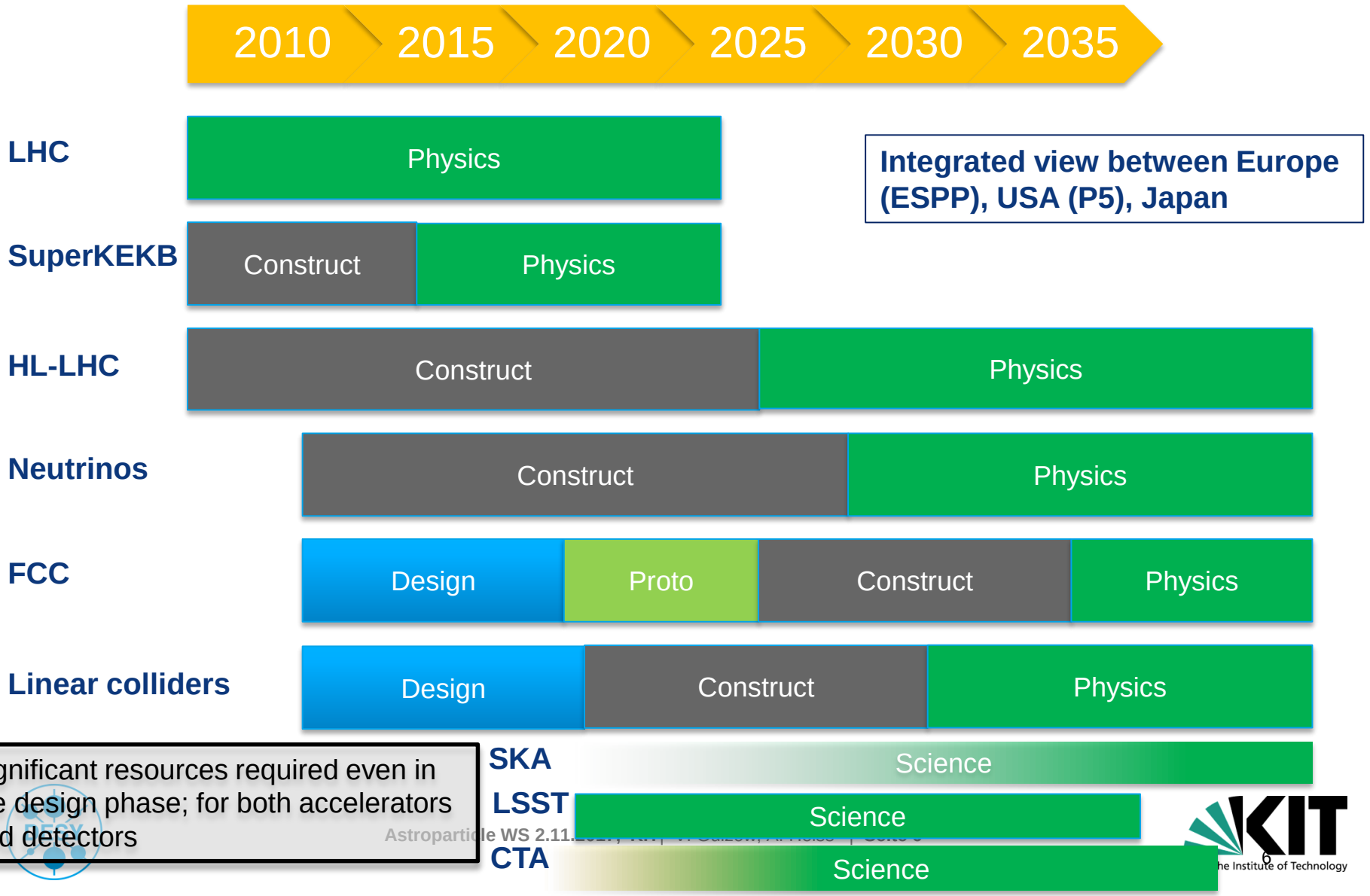
- 94 PB in total WLCG
- 8 Tier 1's, 60 Tier 2's




Contact : Patrick Fuhrmann @ DESY

# HEP Facility timescale



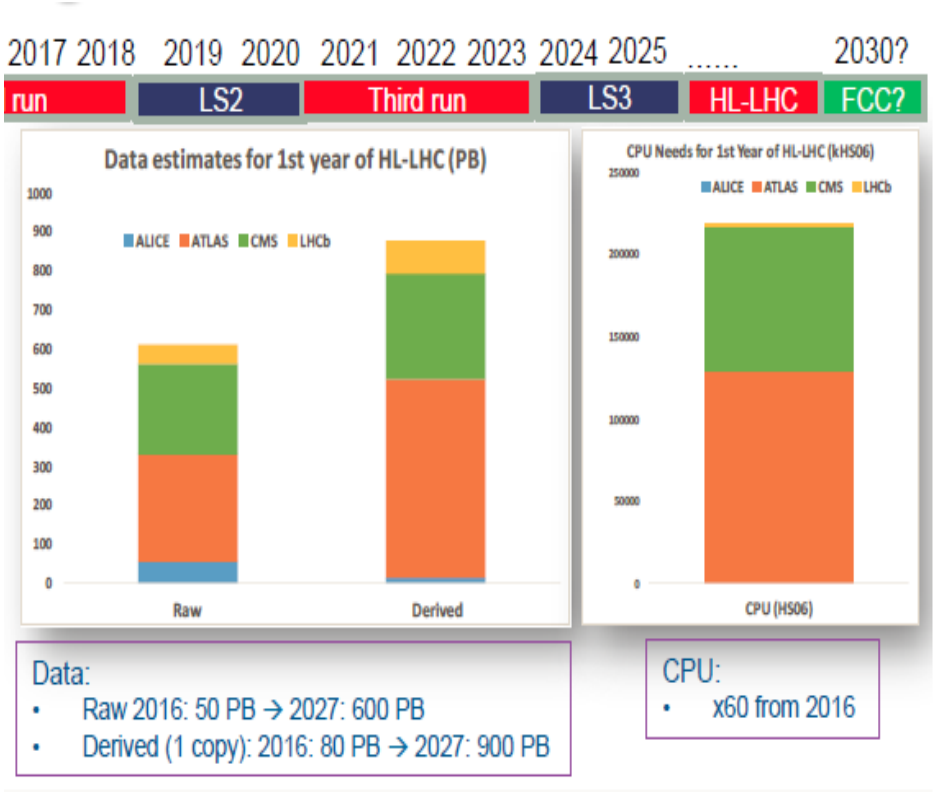

**Integrated view between Europe (ESPP), USA (P5), Japan**

Significant resources required even in the design phase; for both accelerators and detectors

# Big Data, are we still leading?

- Flagship projects SKA and HL-LHC with Exabytes/Year in about 2025
- Various other Projects like
- European XFEL ~ 100 PB/Year before 2020
- CTA, …..

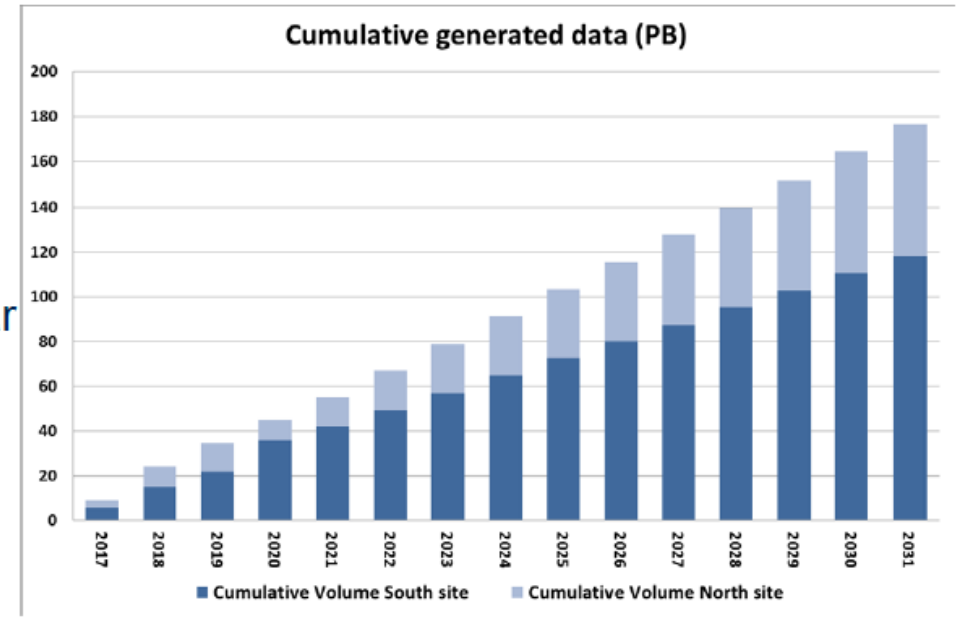
- Big Data not only a matter of
- size
- Big Data means open data



# CTA data volume

- Data to be archived for 30 years of operation + 10 years
- One reprocessing per year (2 versions kept)
- Resulting new data per year
  - Raw data: 4 PB/y
  - Processed data: 4 PB/y
  - Monte-Carlo data: 20 PB

  => 12 PB/year



- Data flow

**SKA1-LOW**

**Global internet traffic ~360 Tb/s**

**(Cisco: 2016)**

~2 Pb/ s

SKA SQUARE KILOMETRE ARRAY — LOW-FREQUENCY APERTURE ARRAY

7.2 Tb/ s

8.8 Tb/ s

SKA SQUARE KILOMETRE ARRAY — SIGNAL AND DATA TRANSPORT

SKA SQUARE KILOMETRE ARRAY — CSP — CENTRAL SIGNAL PROCESSOR

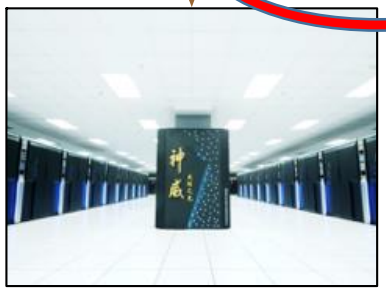
~50 PFlops

~5 Tb/ s

SKA SQUARE KILOMETRE ARRAY — SDP — SCIENCE DATA PROCESSOR

~250 PFlops

~300 PB/ yr

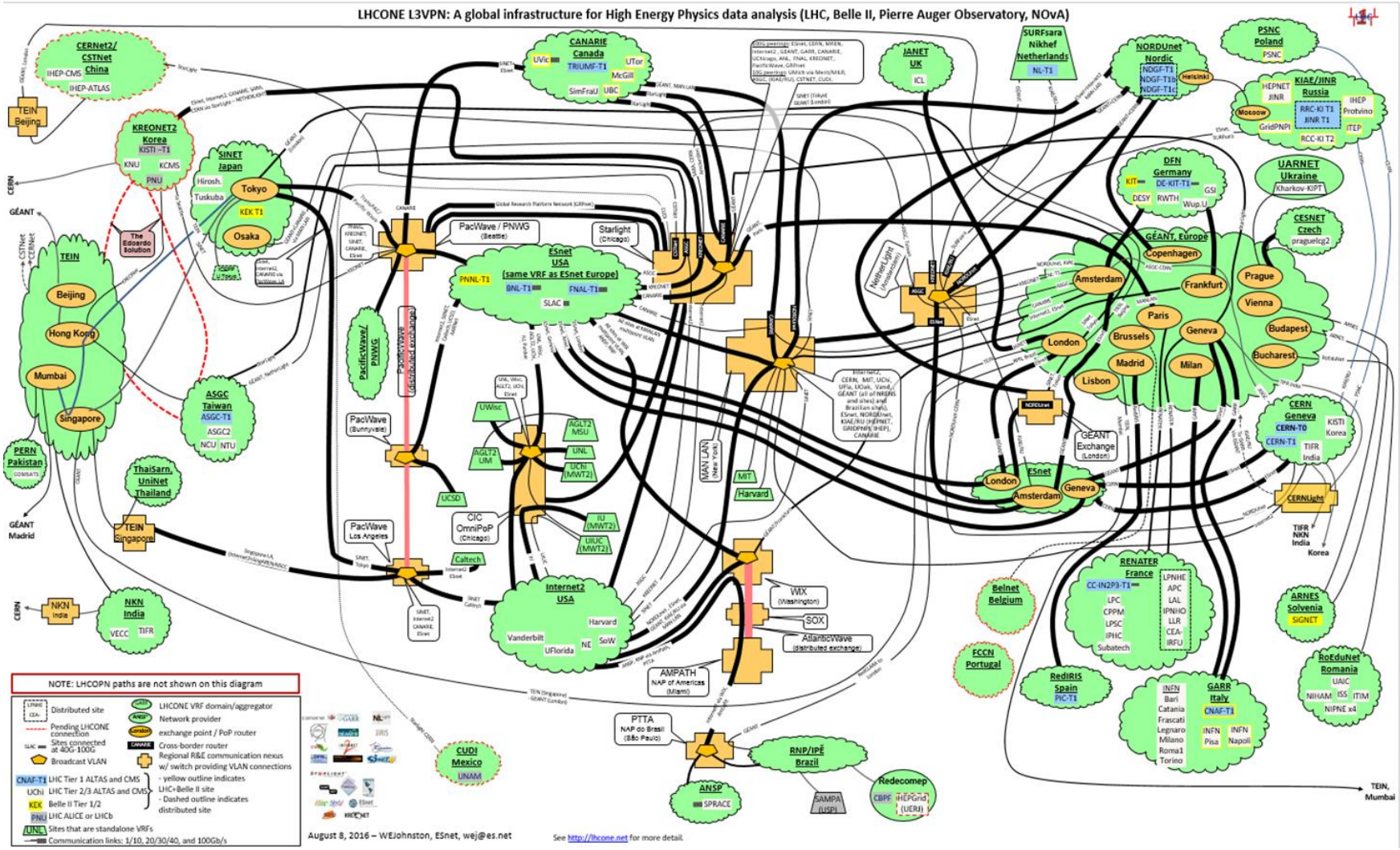


**SKA1-MID**

# Hardware technology perspective

- 20% increase of compute power/year per $
- 15% increase of disk capacity/year per $
- Tape will still improve very much but the role will change
- Only a few vendors, this is risky
- Evolution, no disruptive changes
- Application development for multicore/GPU‘s needed
- A rapid network development to Tbit/s

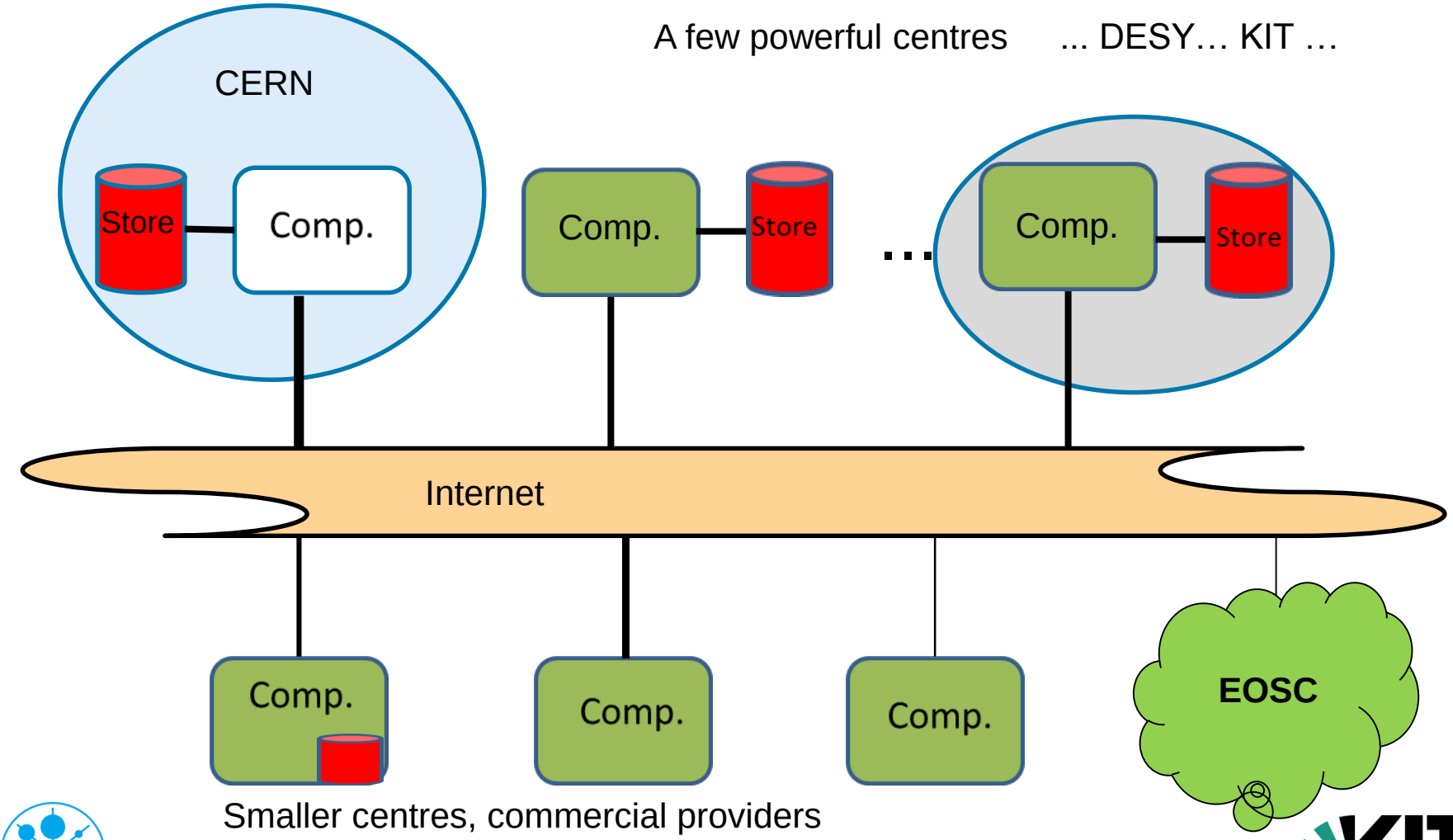
# Connectivity, a key to science: example LHCONE





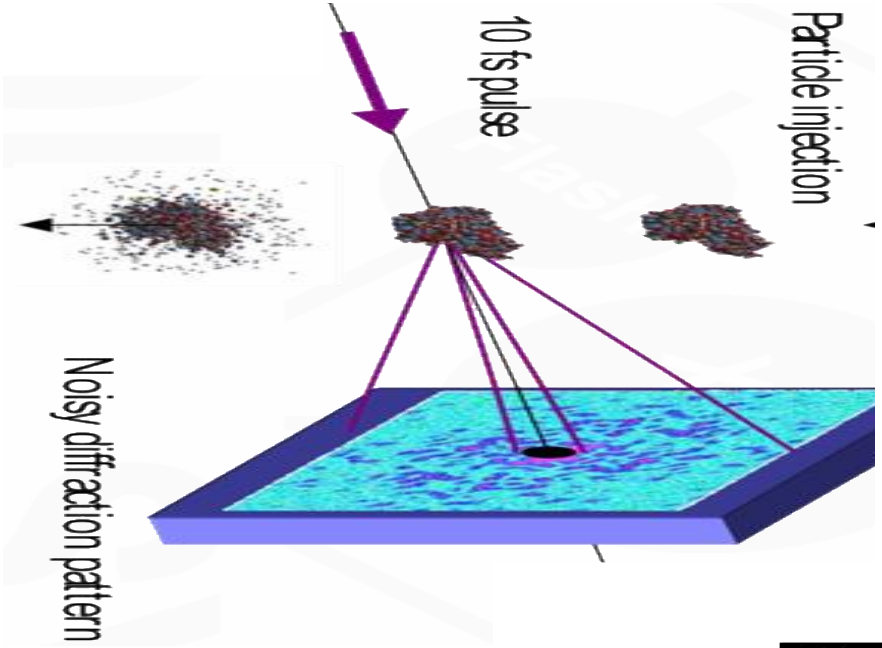

Karlsruhe Institute of Technology

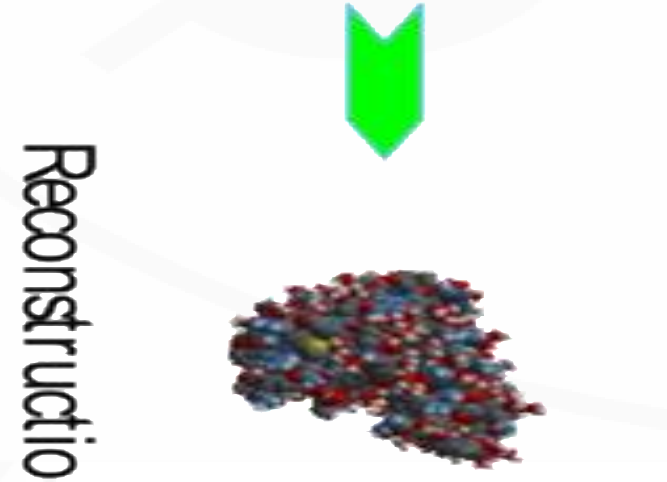
# A possible HL-LHC computing model



# Similar methods across disciplines:

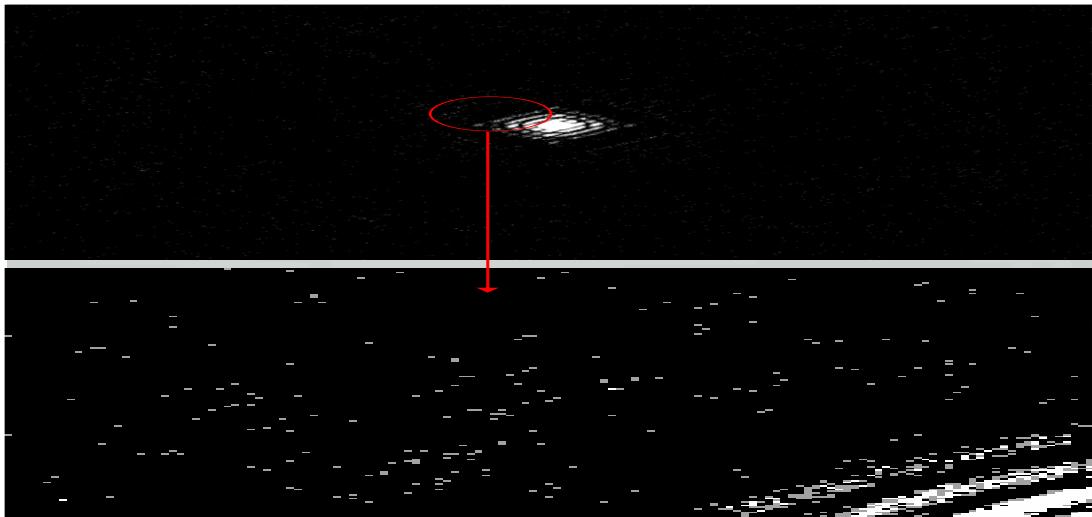
# European XFEL: Imaging of interference data



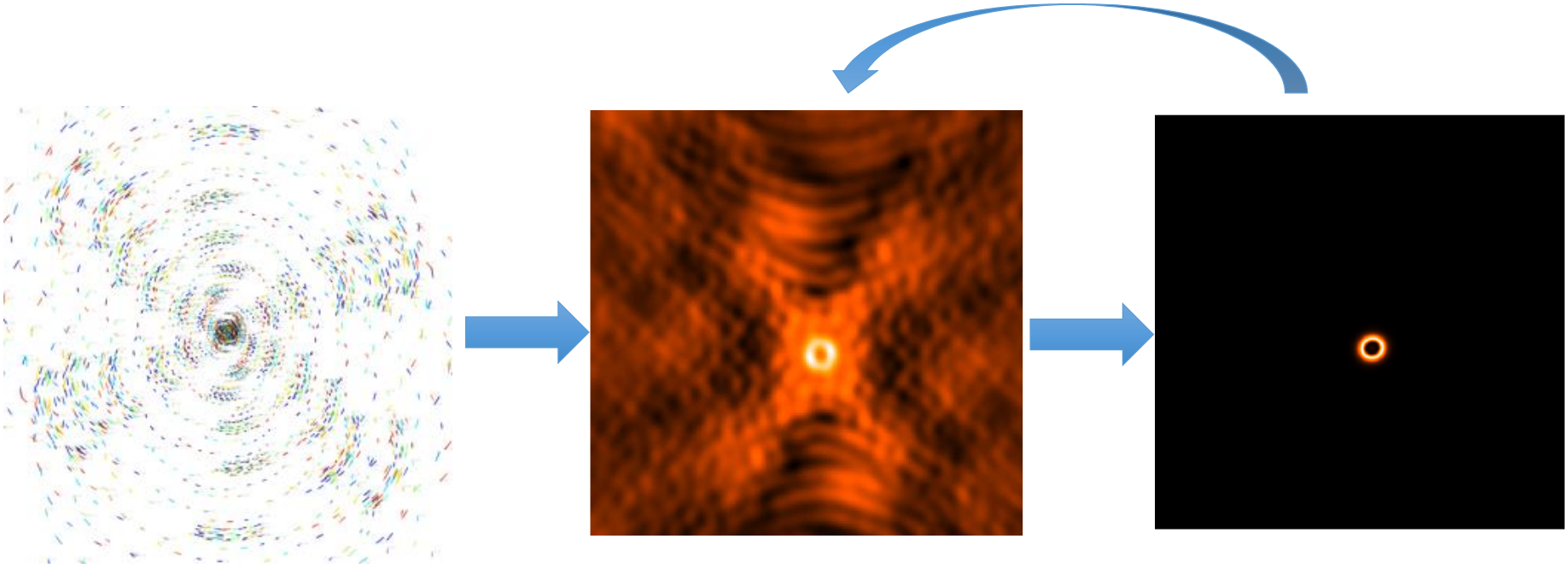




~ **Fourier-Trafo**[diffraction data]

Problem:
Noise in diffraction data
(due to photons from
transportation fluid)



# SKA: Imaging of interference data



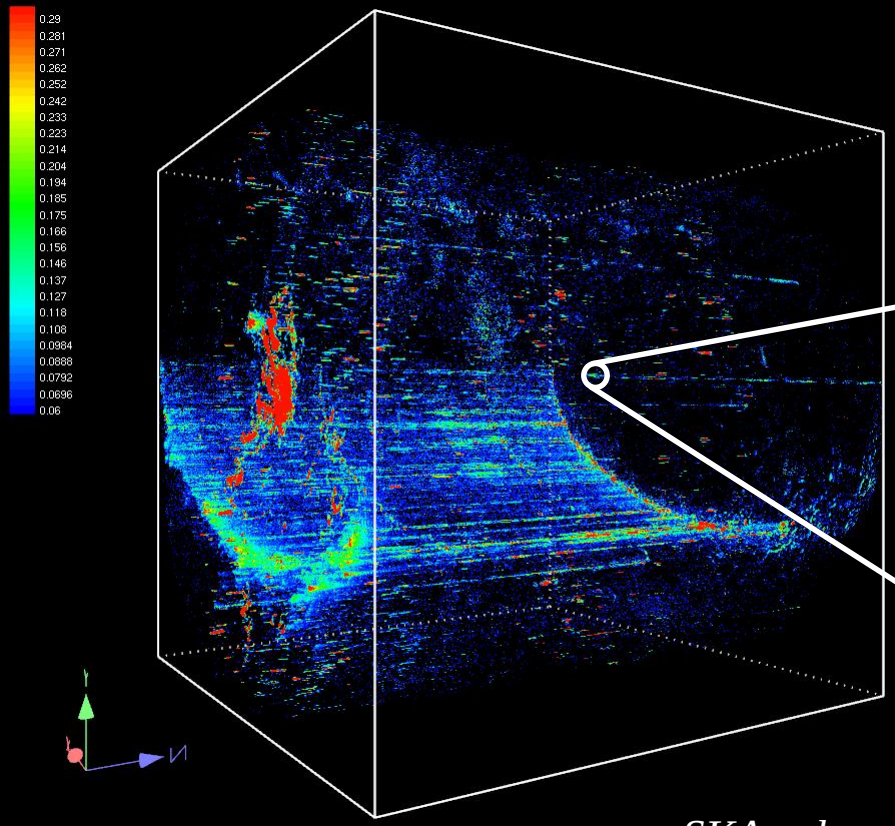
Interference Data:
Visibility V(u,v)

Dirty Image
= **Fourier-Trafo**[V]

Image
= Clean[Dirty Image]

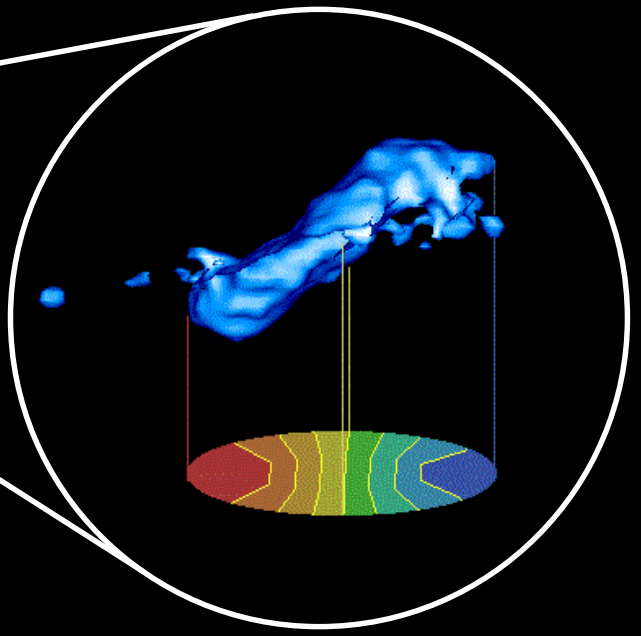
Problem: Low signals, strong noise
- Gridded uv-data
- Sky, calibration, ...

# Interference, Visualisation, Classification




*ASKAP HI Cube (Jurek et al. 2010)*

*Automated detection and calculation of galaxy rotation curves in HI surveys*



*SKA cube ~ 0.9 PB*

# … and similar challenges …

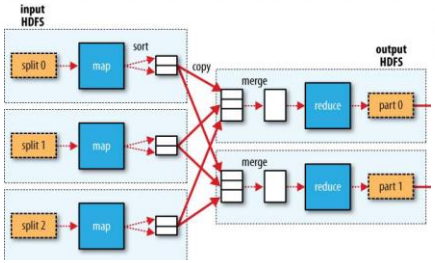
How do we get performance and manage data volume?

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

Approach: Build on BigData Concepts

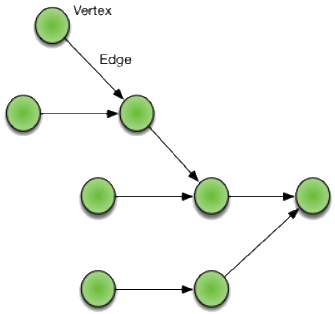
"data driven" graph-based processing approach receiving a lot of attention

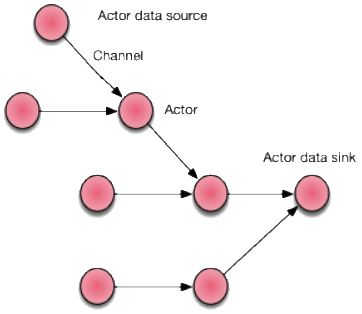
Inspired by Hadoop but for our complex data flow




Hadoop

Graph-based approach




UNIVERSITY OF CAMBRIDGE

Paul Alexander

# The 6 pillars (quick summary):

- Data availability: accessable for member of experiments via standards
- Analysis: Grid&Cloud based solutions, NAF‘s, Analysis Frameworks
- Simulations & Methods developement: Large activities in experiments, HEP Software Foundation
- Open Access: All data are open (after embargo periode)
- Education: GridKa School, schools of the Terascale Alliance, CERN schools, Graduate schools, Curricula
- Archive: Part of Experiments, Data Preservation Initiative