Scientific Computing

Stefan Kluth
MPP Project Review
15.12.2015

The third way

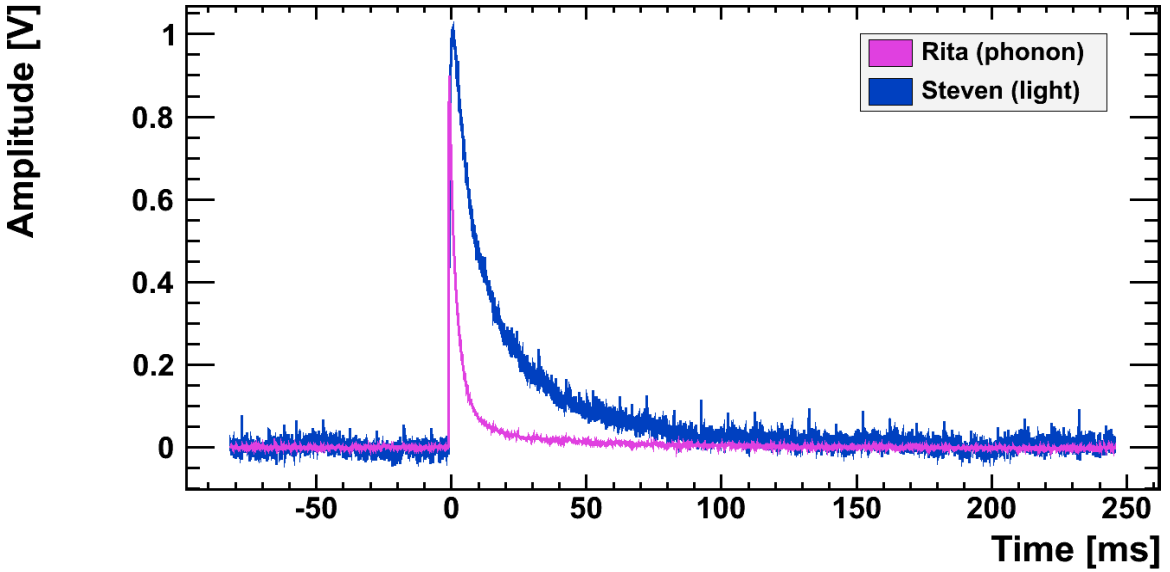
- The scientific method (simplified)
  - Experiment: design a setup and collect data, infer from data underlying principles; test theories
  - Theory: build up from fundamentals a mathematical framework to describe nature and make predictions; learn from experiment data
- With computers
  - Numerical simulation: translate abstract / unsolvable models into practical predictions, discover behavior
  - Find structures in (unstructured) data

Overview

- Example applications
  - CRESST
  - ATLAS
  - (Theory see Thomas Hahn talk)
- Data Preservation
- Software development example
  - DataBrix
- Resources
  - MPP, MPCDF, LRZ, Excellence Cluster (C2PAP)

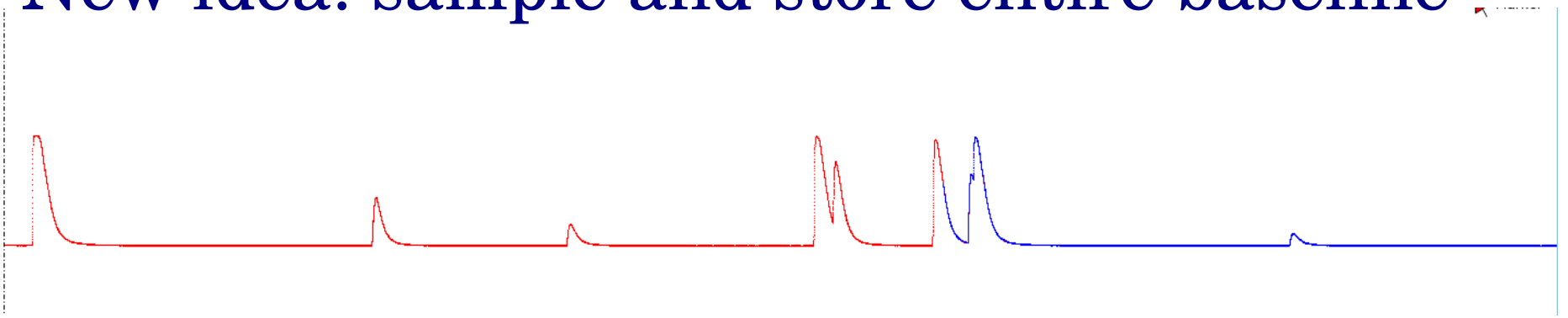
CRESST continuous data stream

- Current approach: hardware trigger activates the DAQ and the samples are stored



Nahuel Ferreiro

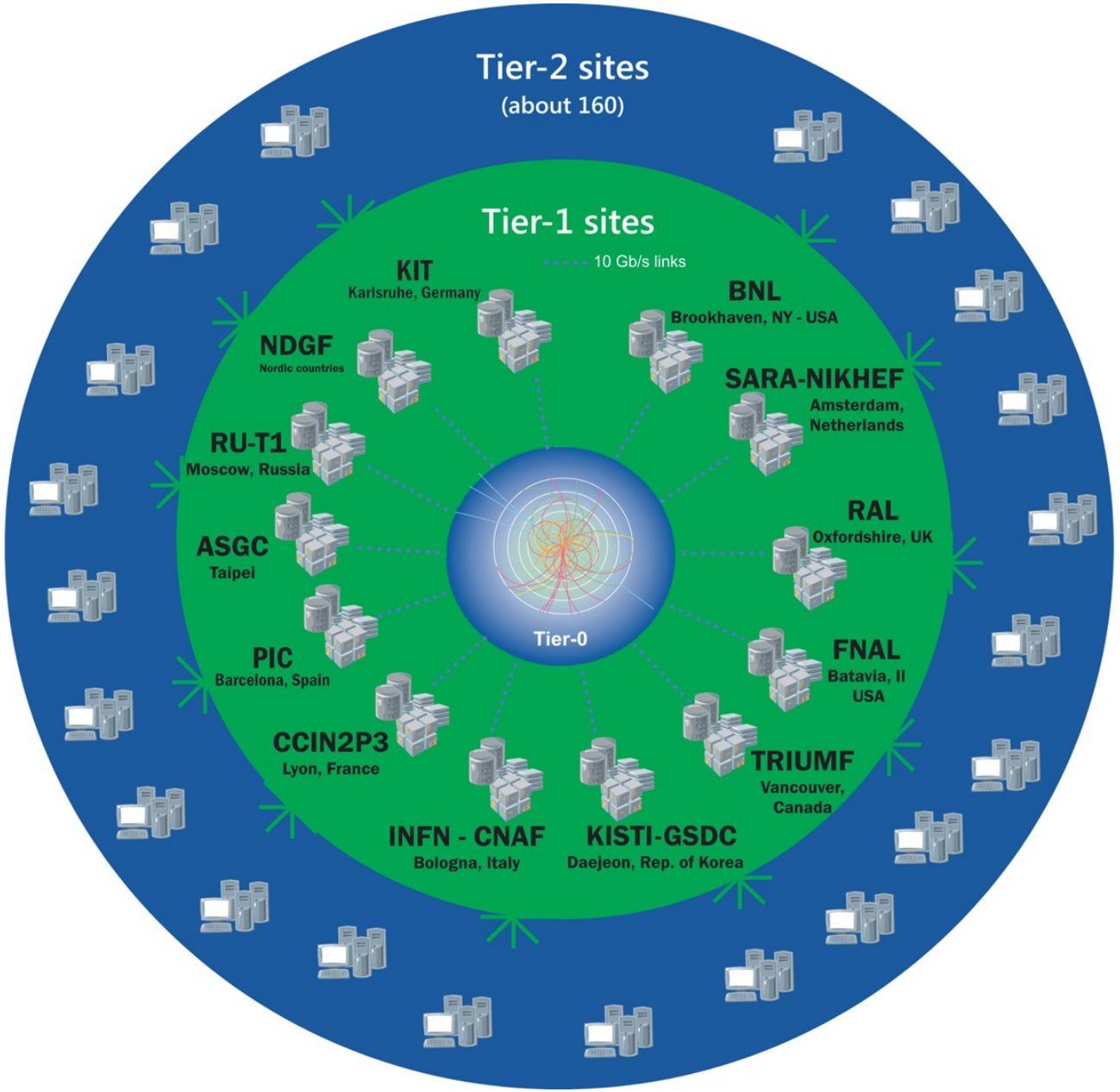
- New idea: sample and store entire baseline



CRESST continuous data stream

- Low energy threshold is key to low energy dark matter search
- Software trigger can perform better than hardware
- Allows to re-think trigger decisions
- 40 Channels read at 25 kHz, 16 bit: ~120 MB/min or ~50 TB/yr
- ~50 core-years CPU for trigger calculation
  - ~18 days on 1000 cores

ATLAS WLCG



Tier-0: CERN

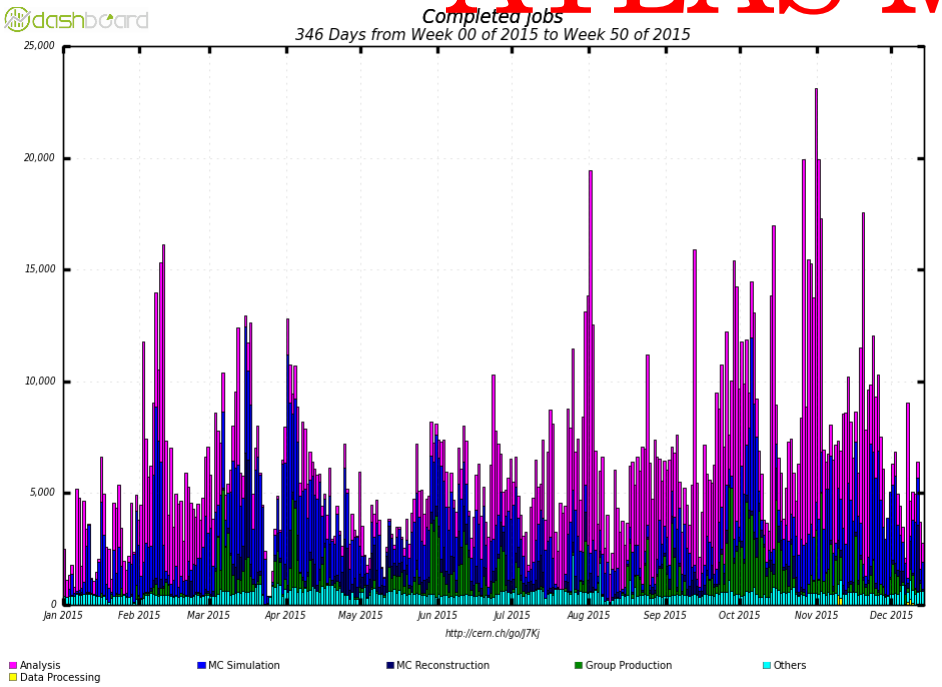
Tier-1: GridKa

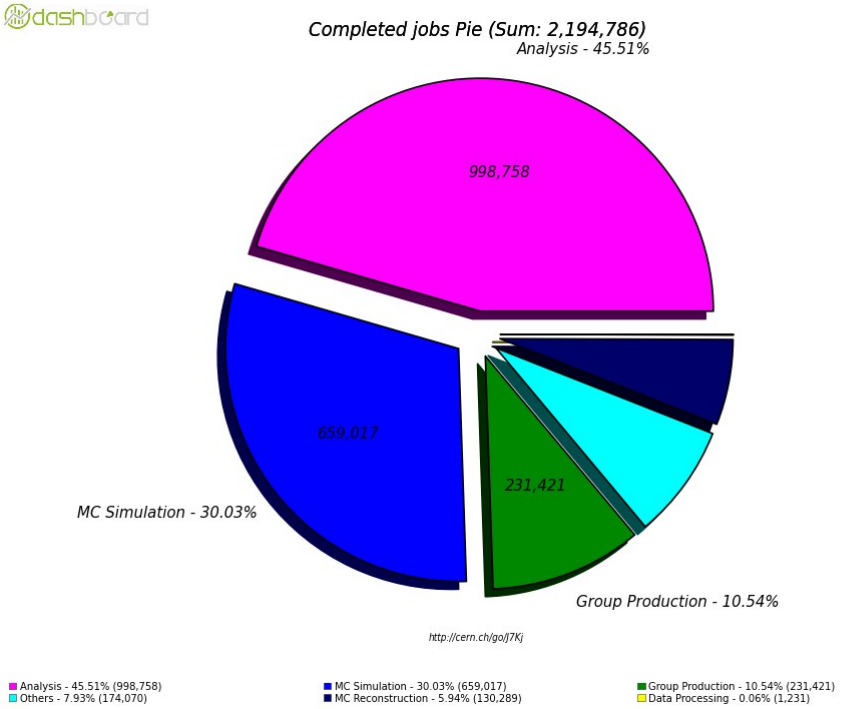
Tier-2: MPPMU

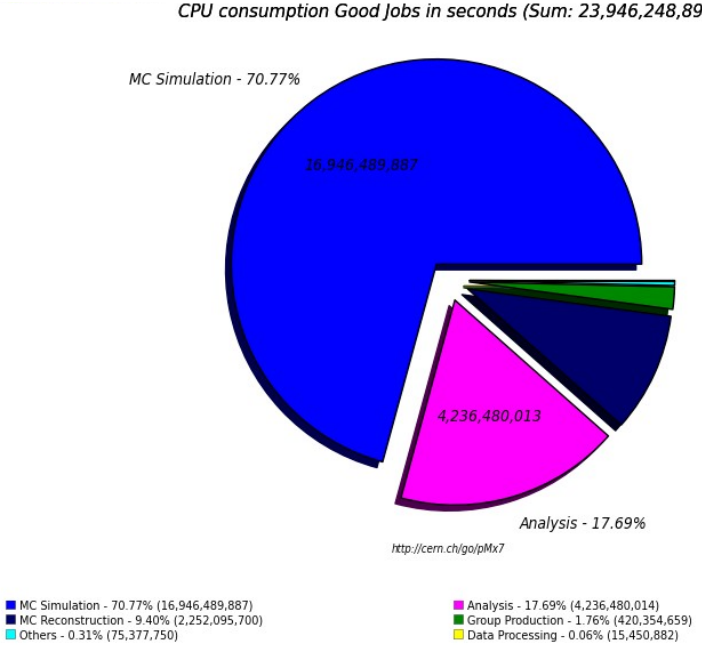
Originally hierarchical, moving to network of sites

MAGIC, CTA, Belle 2 following this model, our Tier-2 supports this

ATLAS MPP Tier-2





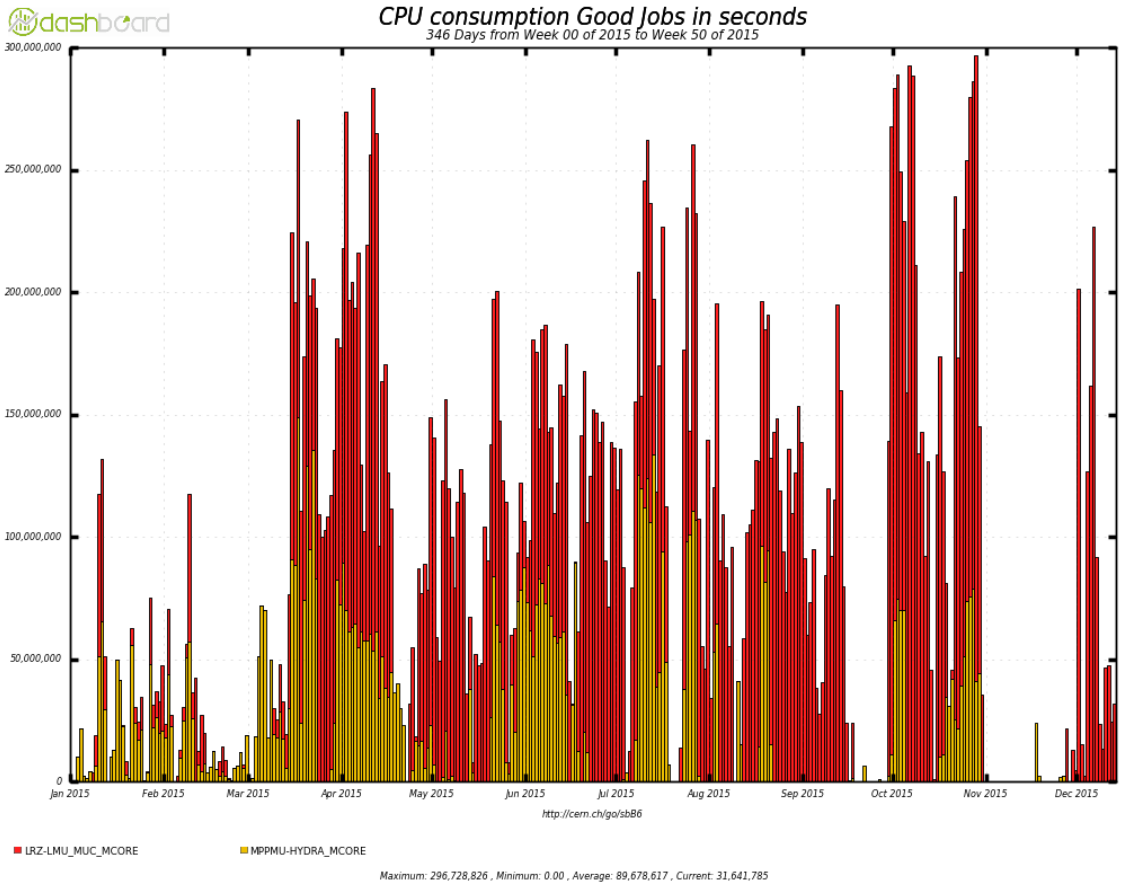


50% nominal Tier-2

1/60 of total ATLAS Tier-2

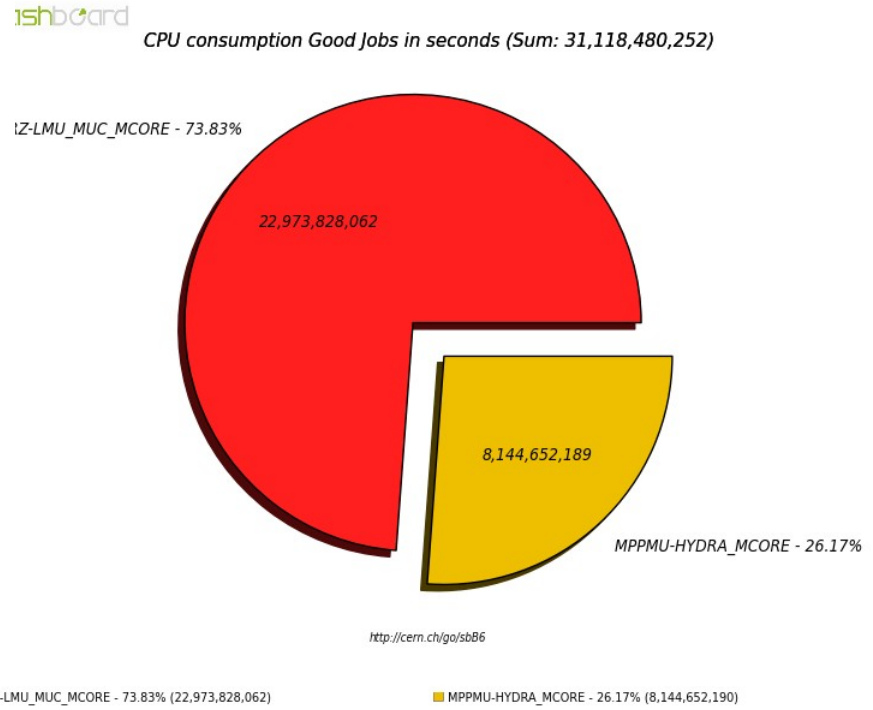
Incl. above pledge contributions

ATLAS @ Hydra



Hydra (MPCDF)
SuperMUC (LRZ)

About 1 nominal Tier-2
Almost 600k jobs



Hydra full with other (big) jobs, opportunistic use only

SuperMUC grant for ATLAS

DPHEP

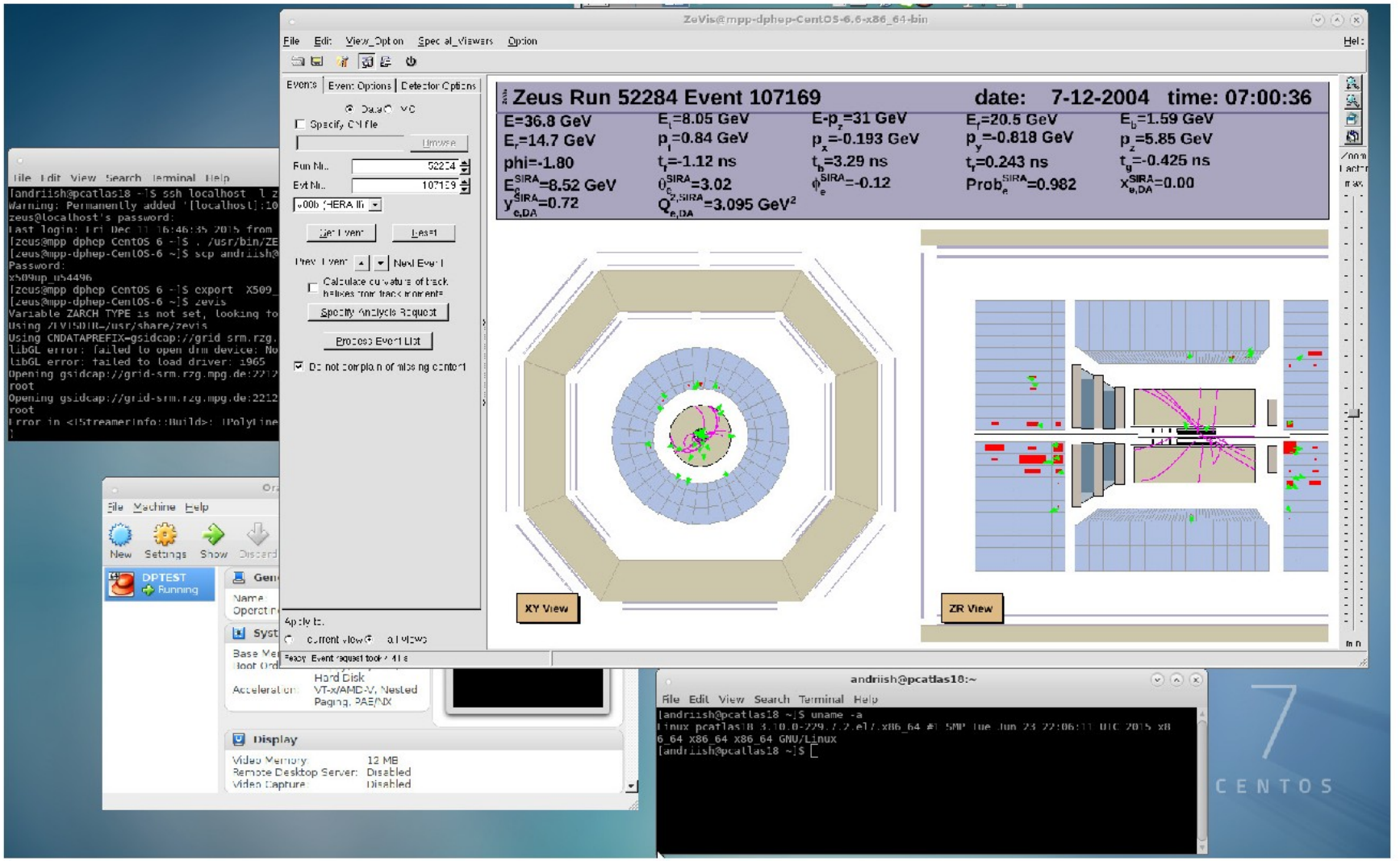
Andrii Verbytskyi

- MPP has several experiments with valuable data and ongoing analysis activity
- H1 and ZEUS @ HERA
- OPAL @ LEP and JADE @ PETRA
- See Andrii Verbytskyi talk
  - and previous project reviews since 2000

DPHEP

- Save bits: copy data to MPCDF
  - Provide access via open protocols (http, dcap)
  - Use grid authentication (X509)
  - About 1 PB (H1, ZEUS, OPAL, JADE), goes to tape library
- Save software: installation in virtual machine
  - Provide validated environment (SL5, SL6, ...)
- Save documentation: labs, inspire, …
  - Older experiments: scan paper-based documents

ZEUS event display@MPP connected to MPCDF



HW independant: VirtualBox on 64-bit CentOS7 runs 64-bit CentOS6.

Outside of DESY: ZEVIS in MPP reads via dCap ZEUS data from MPCDF.

Current status of H1&ZEUS DP

Data/MC	ZEUS	H1
DESY archive	Processed data/MC ntuples	Raw data/MC, processed data/MC
DESY available online	Everything	up to 80%
DESY access	NFS, from 2 machines in DESY+BIRD	NFS, from 2 machines in DESY+BIRD
MPCDF/MPP online+archive	As in DESY+raw data	As in DESY (online)
MPCDF/MPP access	Multiprotocol, worldwide with ZEUS VO cert.	Multiprotocol, worldwide with H1 VO cert.
Software		
DESY reconstruction	No	Yes
DESY MC generation	No	Yes
DESY analysis	Yes	Yes(up to 5y)
DESY user storage	Yes, limited, on 2 machines in DESY	Yes, limited, on 2 machines in DESY
DESY environment	2 machines in DESY+BIRD(up to 5y)	2 machines in DESY+BIRD(up to 5y)
MPCDF/MPP reconstruction	Yes	Planned
MPCDF/MPP MC generation	Yes	Planned
MPCDF/MPP analysis capability	Yes	Planned
MPCDF/MPP user storage	Yes, unlimited, MPCDF SE+CephFS	Yes, unlimited, MPCDF SE+CephFS
MPCDF/MPP environment	CentOS7 virtual machine available	CentOS7 virtual machine planned
Documentation		
DESY analysis primer/manual	Archived web-server	Archived web-server
DESY legacy notes, drafts etc.	InSpire+DESY library	InSpire+DESY library
DESY preservation paper/note	No	No
MPCDF/MPP analysis primer/manual	Relies on DESY	Relies on DESY
MPCDF/MPP legacy notes, drafts etc.	InSpire+DESY library	InSpire+DESY library
MPCDF/MPP preservation paper/note	First draft is available (+A.G.)	Planned

DESY	Finished	Finished, but not optimal	Significant advance	Moderate advance	Will not be done
MPCDF/MPP	Finished	Finished, but not optimal	Significant advance	Moderate advance	Will not be done

Current status of OPAL&JADE DP

Data/MC	OPAL, Host=CERN	JADE, Host=DESY
Host data	Raw/processed, data/MC on CASTOR/EOS	Probably
Host access	Multiprotocol, CERN	No
MPCDF/MPP available online	Raw/processed, data/MC	Raw/Processed Data/MC
MPCDF/MPP archive	Raw/processed, data/MC	Raw/Processed Data/MC
MPCDF/MPP access	Multiprotocol, worldwide with OPAL VO cert.	Multiprotocol, worldwide with ZEUS VO cert.
Software		
Host reconstruction	Yes	No
Host MC generation	Yes	No
Host analysis	Yes	No
Host user storage	CERN users only	No
Host environment	Default CERN	No
MPCDF/MPP reconstruction	Yes (M.S.)	Planned
MPCDF/MPP MC generation	Yes (M.S.)	Update/Planned
MPCDF/MPP analysis	Yes	Planned
MPCDF/MPP user storage	Yes, unlimited, MPCDF SE+CephFS	Not needed
MPCDF/MPP environment	CentOS7 VM available	AIX → Fedora17 PPC or CentOS7 x86_64 VM planned
Documentation		
Host analysis primer/manual	CERN web-server	No
Host legacy notes, drafts etc.	InSpire+CERN library	InSpire+DESY+J.O.
Host preservation paper/note	No	No
MPCDF/MPP analysis primer/manual	Relies on CERN	Yes (Update!)
MPCDF/MPP legacy notes, drafts etc.	InSpire+CERN library	InSpire+DESY library+
MPCDF/MPP preservation paper/note	Yes, early stage update	Yes (Update!)

Host	Finished	Finished, but not optimal	Significant advance	Moderate advance	Will not be done
MPCDF/MPP	Finished	Finished, but not optimal	Significant advance	Moderate advance	Will not be done

Software: DatABriCxx / dbrx

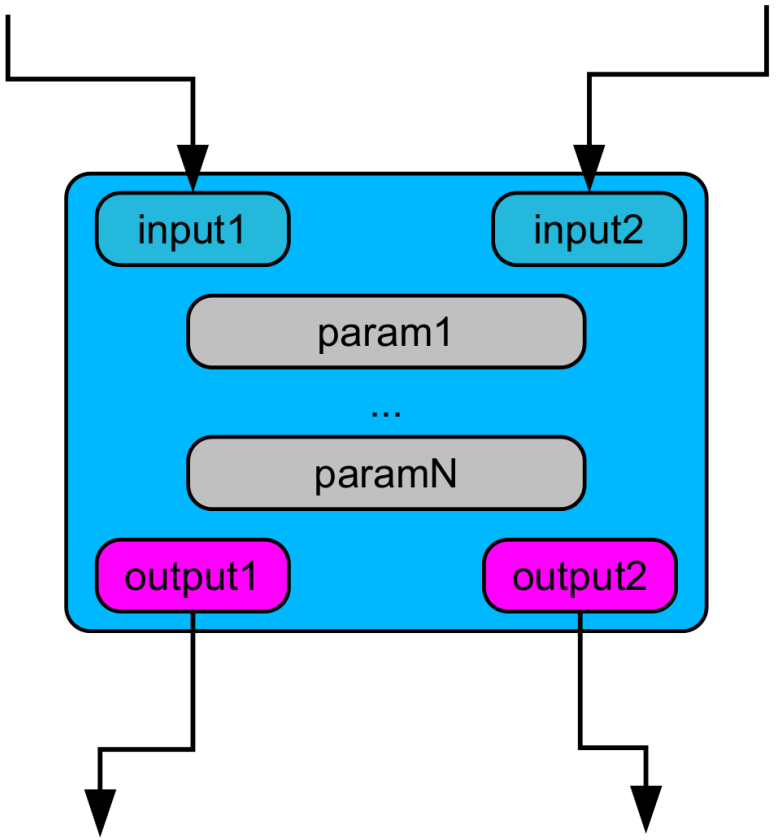
- Many small experiments have organically grown computing

Oliver Schulz

  - Does not scale to bigger collaborations
  - Does not scale to much bigger data sets
  - Does not scale to much longer lifetimes
- Provide a framework
  - Selection and organisation of algorithms
  - Configuration and parameter handling
  - Services (IO, calibrations, …) provided

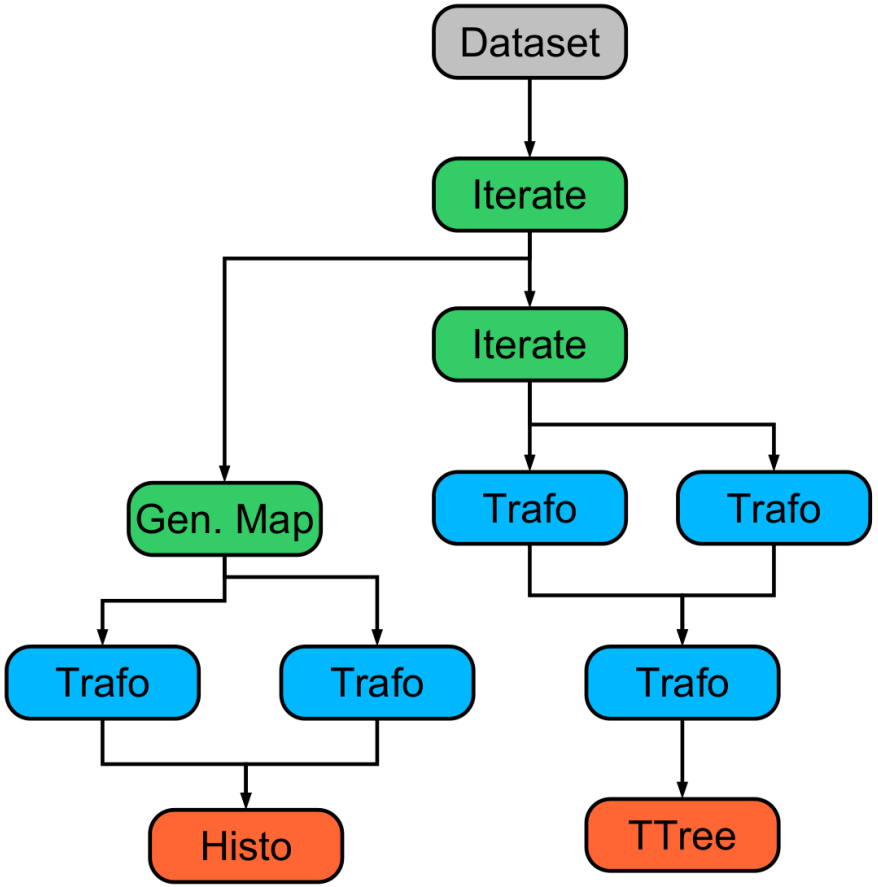
dbrx

Algorithm: receive input from uptream, hold parameters (configuration), provide output to downstream (transformation)



1 to n: Mapper (event loop),
1 to 1: transform (hits to tracks, fits, …)
n to 1: reducer (histgrammer, Output, final calculations, ...)

dbrx



Brics are analysis building blocks

Brics form a directed acyclic graph
a.k.a. tree

Framework schedules brics according to data flow dependencies, naturally parallel

GUI possible via http

GeDet, CRESST, GERDA (partially), COBRA, …

LGPL, get it from github

Resources

- MPCDF
  - MPP Linux cluster: 84 nodes, ~1.000 cores, > 2.5 PB storage
  - Hydra: 4.110 nodes, ~83.000 cores, fast inter-connect, 338 nodes with dual Nvida Tesla K20X
- LRZ
  - SuperMUC: >12.000 nodes, 241.000 cores, fast interconnect
- Excellence Cluster Universe
  - C2PAP: 128 nodes, >2000 cores, fast interconnect, SuperMUC integration

Resources

- MPP
  - > 200 desktop PCs via condor batch system (need to sort out kerberos login)
  - Brand new PC with Intel Xeon Phi 5100
    - 60 core, 8 GB, 1 GHz
    - Theory group

Summary

- Scientific computing essential for our success
- Many activities at MPP
  - From software development to data preservation
- Resources: MPP, MPCDF, LRZ, C2PAP
- All centers provide application support
  - Porting to parallel platforms, performance tuning, …
- Transition to HPC in many of our research areas