Karlsruhe Institute of Technology

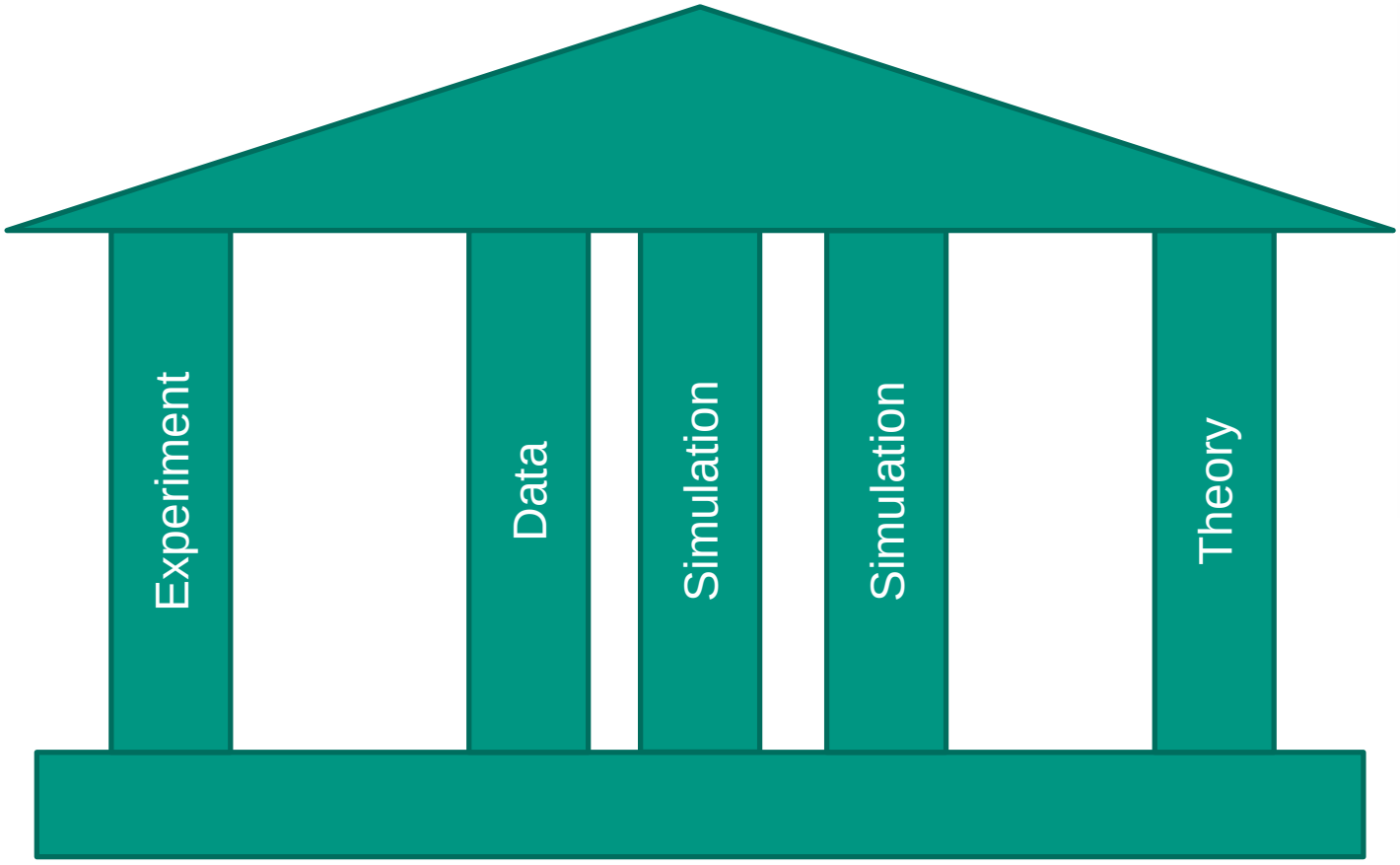
Enabling Data-Intensive Computing & the EOSC

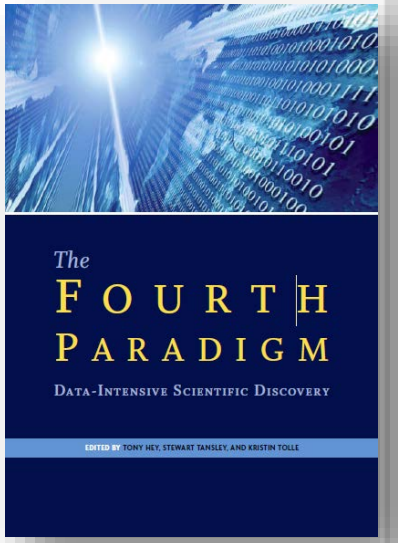
Achim Streit <achim.streit@kit.edu>

711. WE-Heraeus-Seminar „The Science Cloud“

Steinbuch Centre for Computing

Four pillars of Science

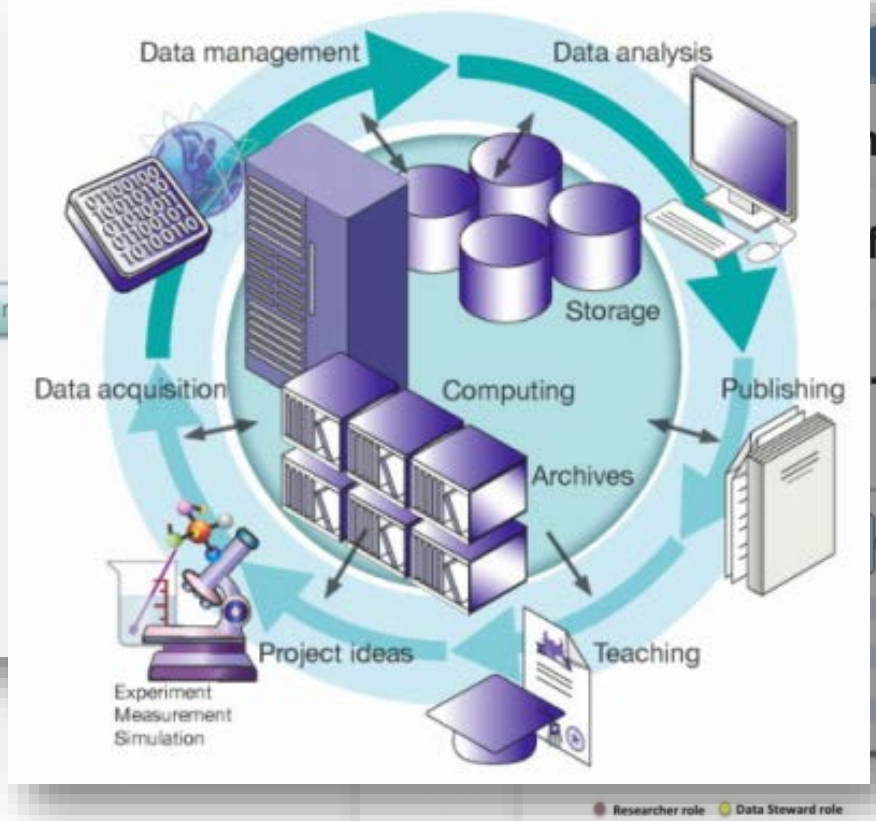




Tony Hey, Stewart Tansley, Kristin Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, ISBN 978-0982544204, http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Research Data Life Cycle

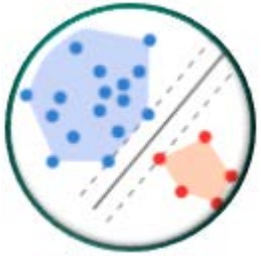
KIT SCC/LSDMA DLC



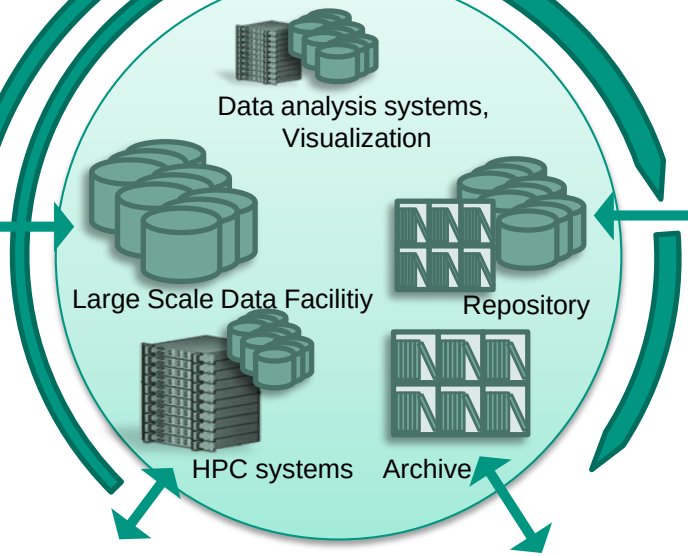
The KIT RDM DLC

Source: https://www.slideshare.net/EUDAT/the-data-lifecycle-eudat-summer-school-yann-le-franc

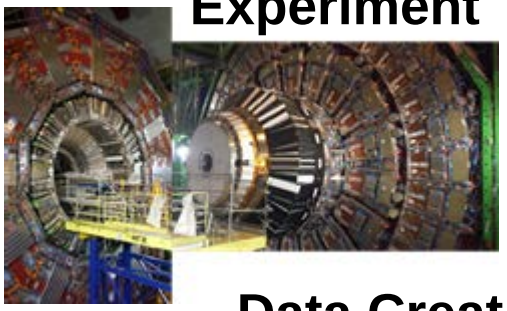
Enabling Data-Intensive Computing



Data Analysis Visualization



Experiment



Data Creation

Publications

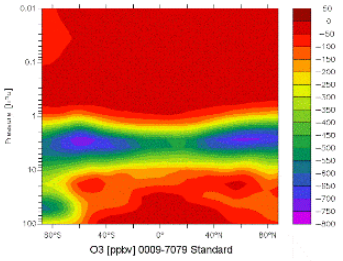
Archive

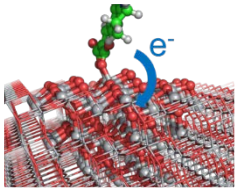
Simulation

Enabling Data-intensive Computing

Supercomputing

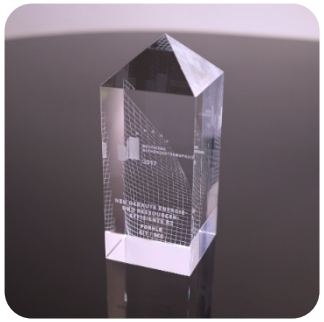
- Operation of **HPC systems**
  - **ForHLR:** Tier-2 system in Germany, 34,800 cores with > 1.4 PetaFlop/s peak, peer-review access for users in Germany
  - **bwUniCluster:** Tier-3 system in the state BaWü, HPC capacity system with 18,300 cores, shareholder ownership with all 9 state universities
- **Joint R&D** with scientific communities & KIT institutes
  - Application optimisation, scaling, model enhancements
  - Simulation Labs in Helmholtz Programme
- **HYIGs** MBS + FiNE
- **Innovation drivers** for SMEs
- **Architect** for HPC environment in BaWü





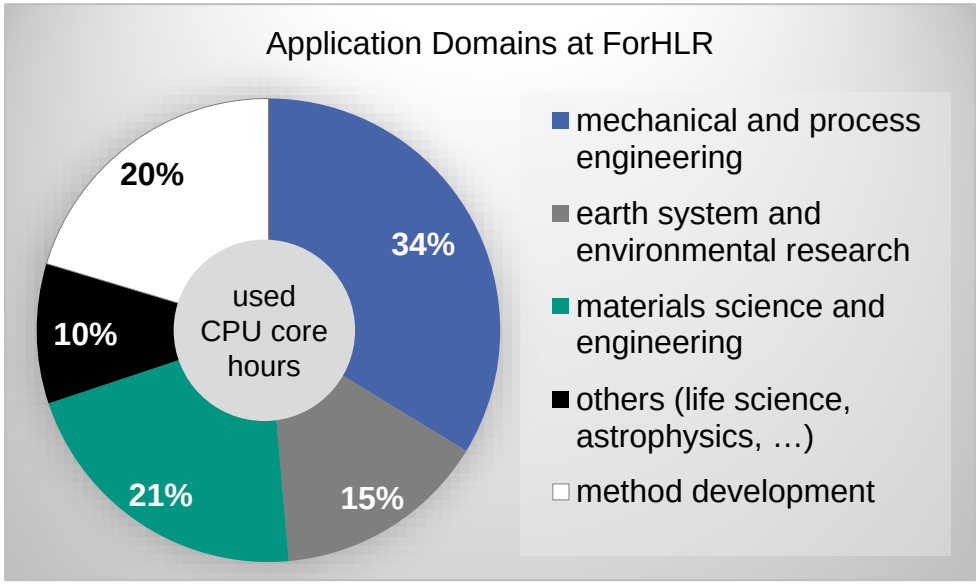
ForHLR: Forschungshochleistungsrechner Karlsruhe

- Third-party funded mid-range national (Tier-2) supercomputer
- 34,800 compute cores
- 1.4 PetaFlop/s peak
- Peer-review access
- Self-designed cooling concept

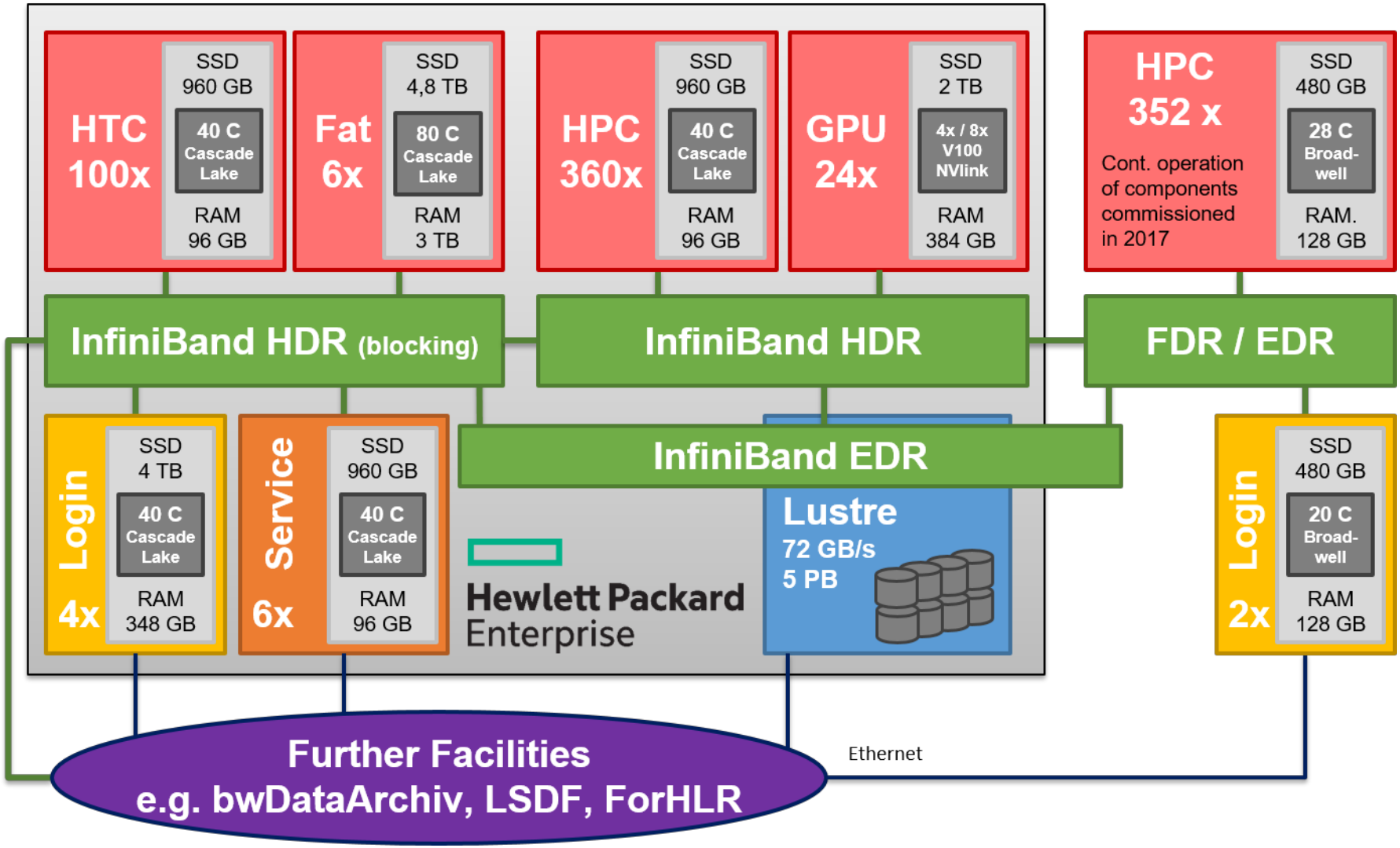


1st prize German Data Center Award 2017

Newly built energy and resource efficient data centers

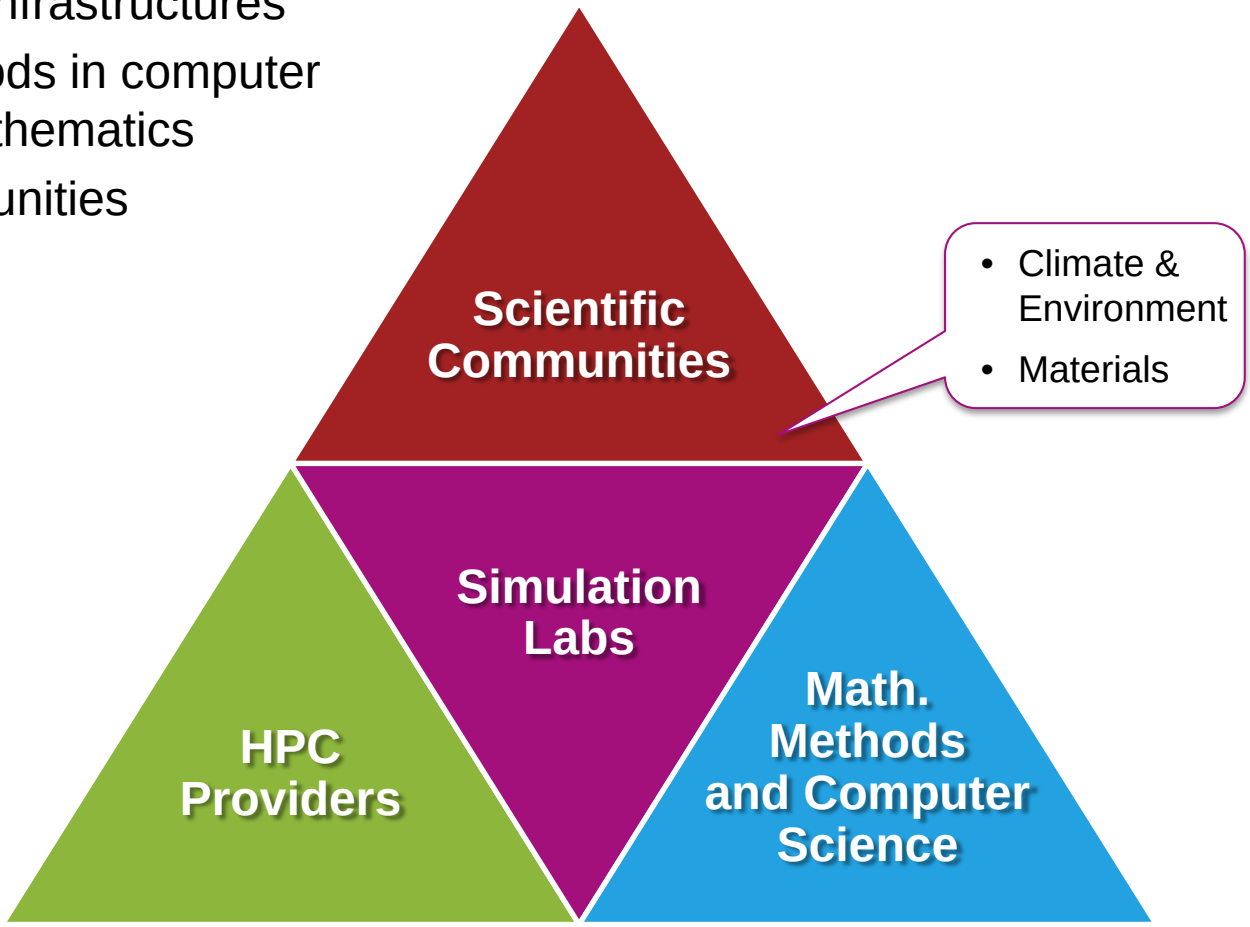


New HPC Tier-3 System with ML / AI Support



Simulation Labs (SimLabs)

- Bridging between
  - Providing HPC infrastructures
  - Research methods in computer science and mathematics
  - Scientific communities
- By performing **interdisciplinary joint R&D**



Enabling Data-intensive Computing

Big Data

- Operation of **GridKa**
  - German Tier-1 in WLCG for an international community
- Operation of the **Large-Scale Data Facility**
  - Multi-disciplinary data centre for climate research, systems biology, energy research, etc. in BaWü
- **Joint R&D** with scientific communities
  - Generic data management research
  - Data Life Cycle Labs in Helmholtz Programme
- **Innovation driver** for SMEs
- **Active role** in large projects & initiatives

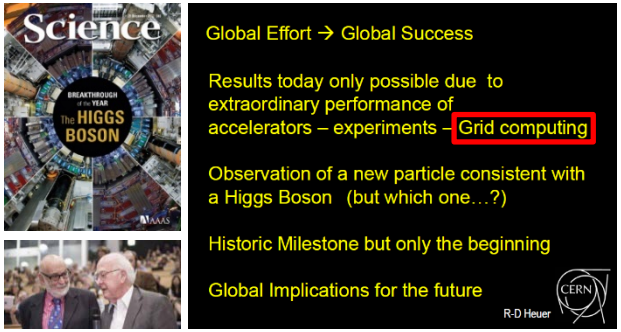
GridKa

- **Data and analysis center for particle and astroparticle physics**

PIERRE AUGER OBSERVATORY · BABAR · Belle II · ALICE · ATLAS EXPERIMENT · CMS · LHCb

~10 000 physicists

WLCG Worldwide LHC Computing Grid



Conclusion slide of R.D. Heuer, July 4, 2012

- A **cornerstone** of the Worldwide LHC Computing Grid (WLCG)
- **Integral part** of the LHC data processing chain

RAW Data

Disk

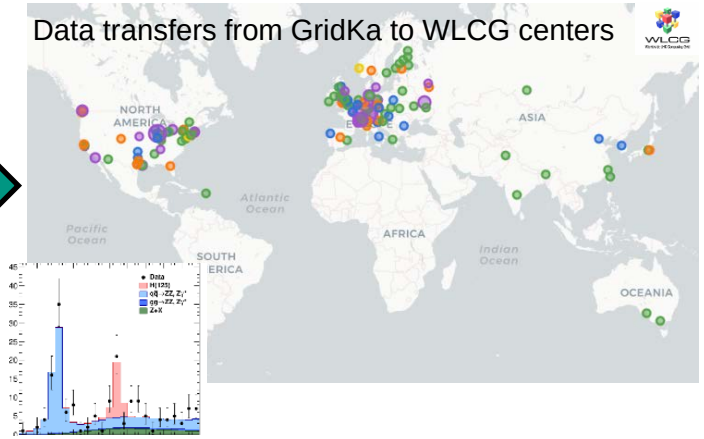
GridKa

- RAW data processing
  e.g. event (re-)reconstruction
- Event selection
- Physics and detector simulation
- Analysis

Tape Archive

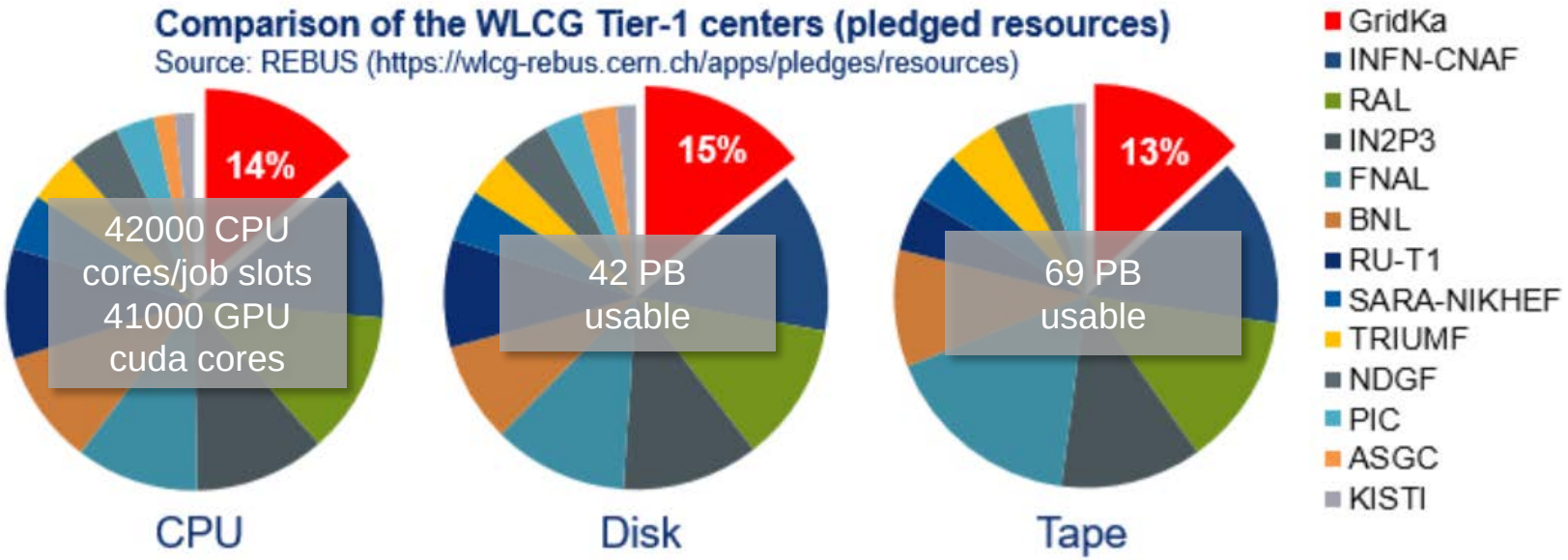
Data

Analysis and simu-lation jobs

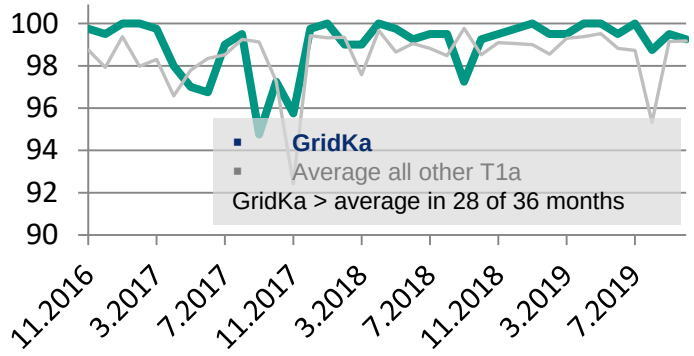


GridKa Contribution to WLCG

1/7 of total WLCG Tier-1

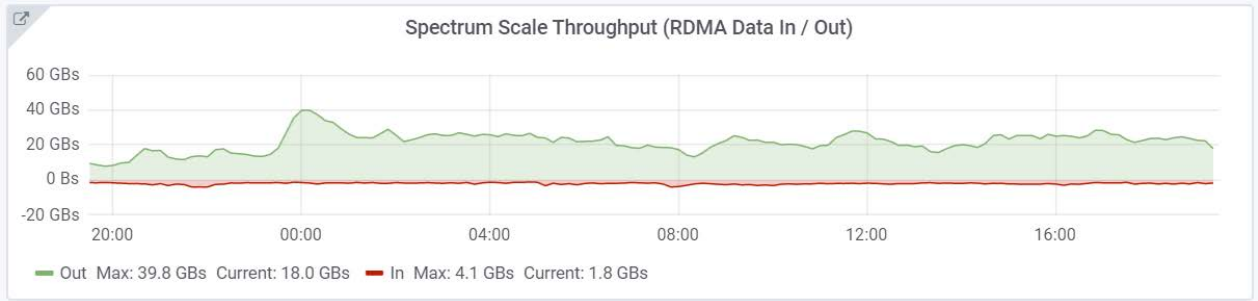


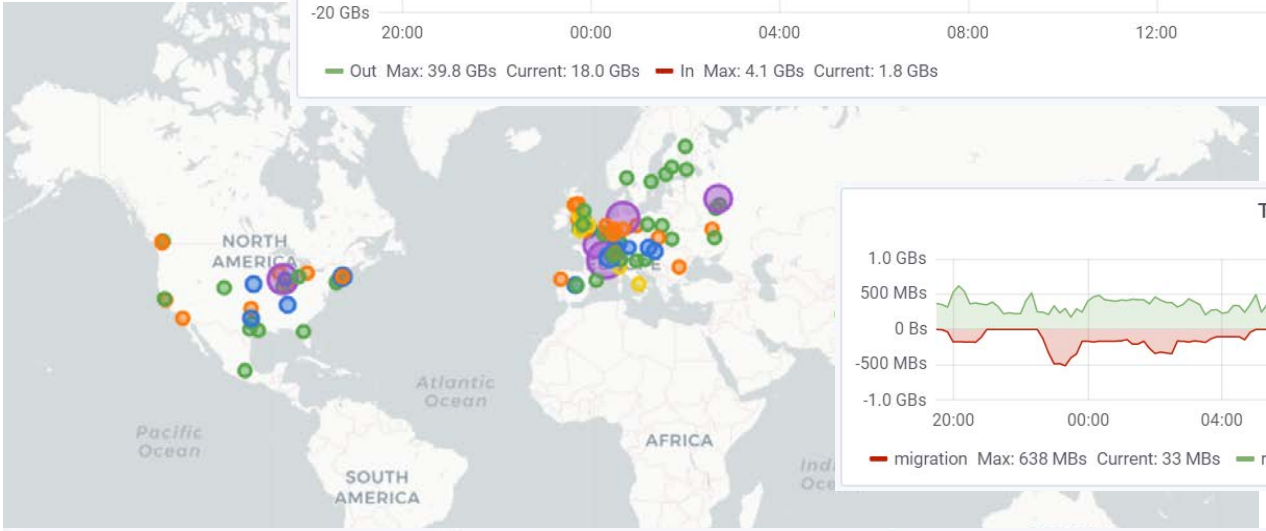
Reliability measured by LHC experiments

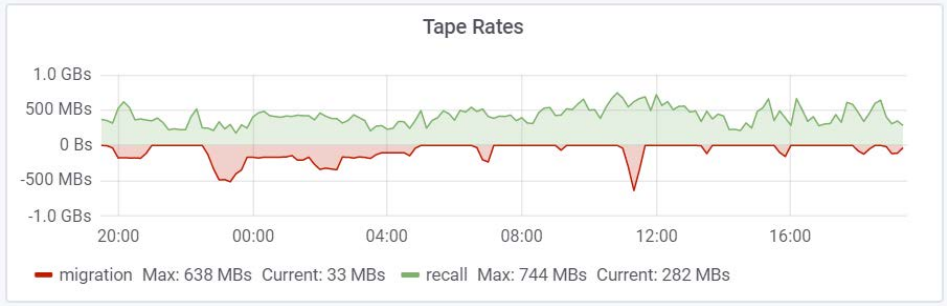


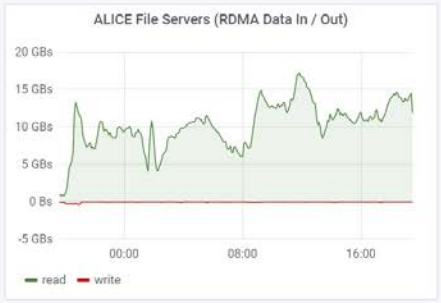
232 M core-hr
20 M jobs
57 PB in
110 PB out
0 downtime

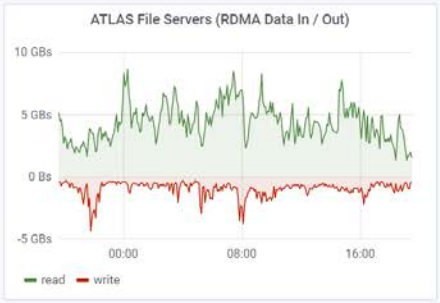
GridKa – some Grafana plots…

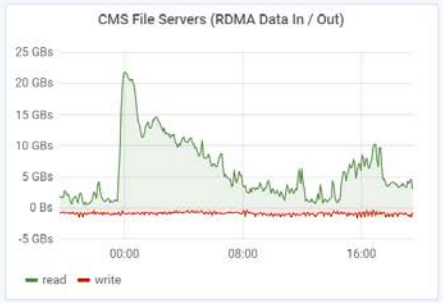


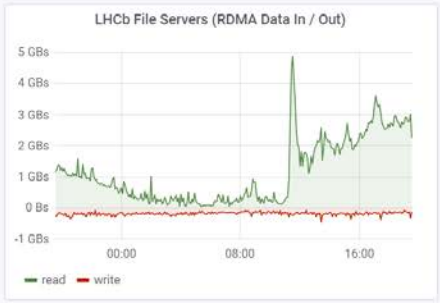








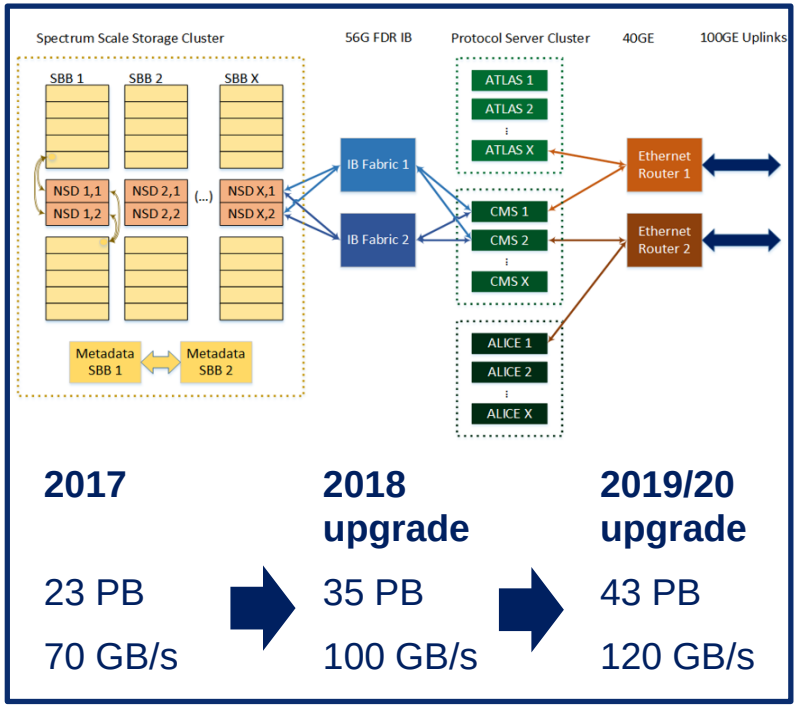




Addressing Changing Computing Models

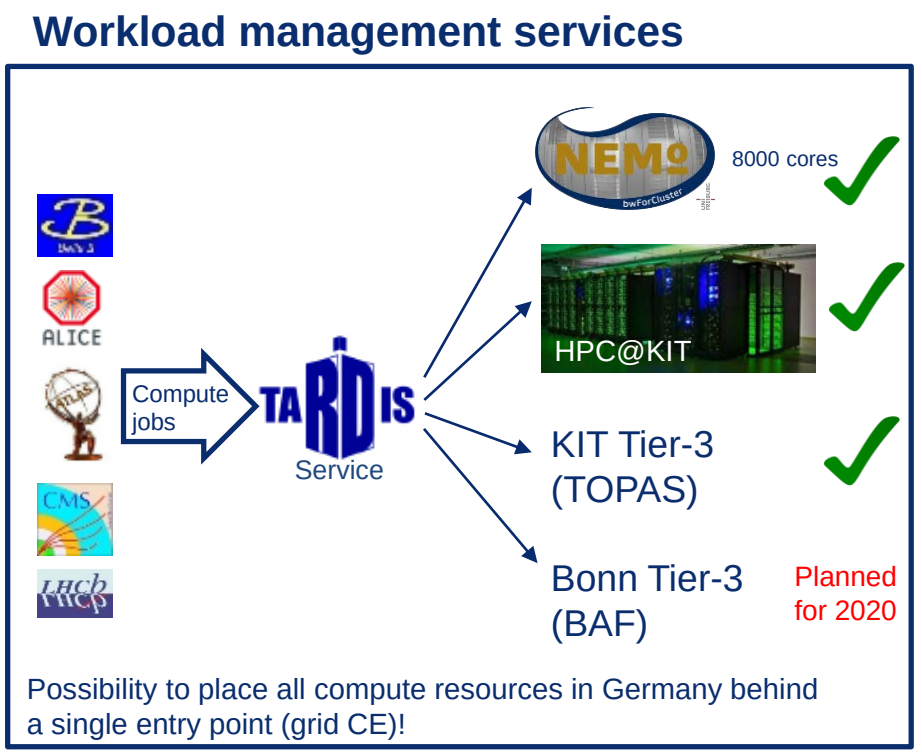
- **Software Defined Online Storage**
  - Data access becomes less predictable
  - Increasing data access from remote compute sites
    - Dedicated sites (WLCG)
    - Opportunistically used sites (HPC, cloud)
- **Powerful Networks**
  - Redundant links to CERN (100 + 2x10 Gbit/s) and to DFN (2x100 Gbit/s)

Scalable online storage technology: throughput, IOPs, capacity



Addressing Changing Computing Models

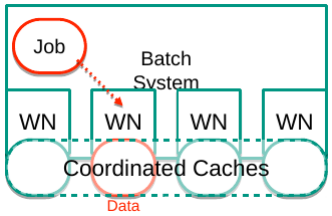
- **Software Defined Online Storage**
- **Powerful Networks**
- **Leverage Additional Opportunistic Resources**
  - More heterogeneous computing resources (CPUs, GPUs, …)
  - Long-term and opportunistic access to HPC, cloud, ….
  - Resources that the experiments even do not know about!
  - Hide additional resources behind a single entry point visible to the experiments‘ central workload management

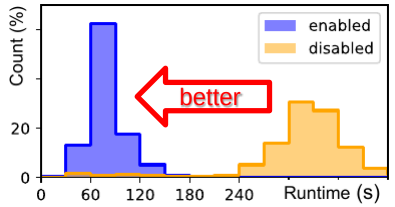


Addressing Changing Computing Models

- **Software Defined Online Storage**
- **Powerful Networks**
- **Leverage Additional Opportunistic Resources**
- **Optimized Resources and Increased Computing Efficiency**
  - Specialized resources
    - Innovative ideas and improvements to speed-up analysis tasks
    - Optimized configurations of hard- and software
  - Sophisticated data and workload management

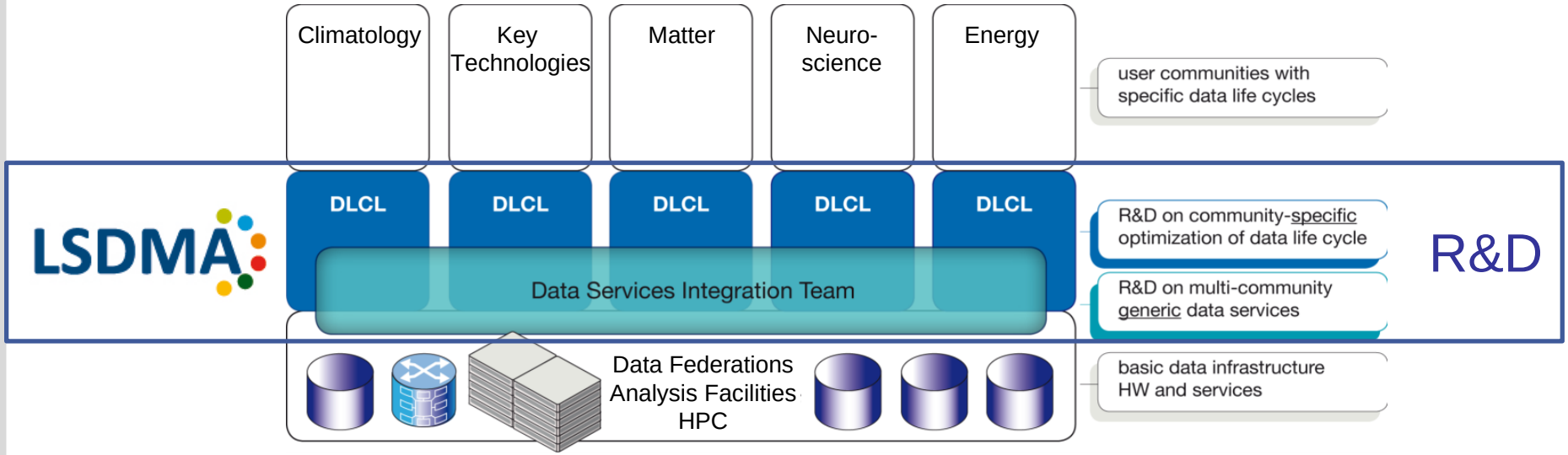
Highly optimized analysis cluster





- Performance increase through data locality
  - Coordinated data placement on local caches in compute nodes
- Performance boost for certain types of computing tasks
  - Example (top right corner): CMS calibration jobs
- Prototype cluster with 862 CPU cores and GPUs
  - Currently promising tests by CMS and Belle-II
  - Other experiments from 2020

Concept of Data Life Cycle Labs (DLCL)
from 2011



Data Life Cycle Labs

Joint R&D with communities

- Optimizing the data life cycle
- Specific data analysis tools and services

Data Services Integration Team

Generic, multi-community R&D

- Interface between federated data infra-structures and DLCLs resp. Communities
- Integration of data services in scientific working process

Highlight from SDL Earth System Science: Compression Methods for Floating-Point Data

- New climate models produce **several TBs** of data at each simulation run
- Nowadays the **bottleneck** is not about solving the differential equations, but the **storage of the output**
- The **goal** of compression methods is to identify and reduce the redundant information in the data

Prediction-based compression

223.48	221.71	221.54	222.87	?	What will be the next value?
				222.40	Average?
				222.87	Last value?
				224.20	Last difference?
				221.47	Seasonal information?

Uğur Çayoğlu et al.
IEEE e-Science 2019
DOI: 10.1109/eScience.2019.00032

- Methods developed at SCC are on average **10% better** than previously developed compression methods for floating point data

Source: https://code.mpimet.mpg.de/attachments/download/16625/r2b02_europe.png

Helmholtz Analytics Framework (HAF)
co-coordinated by KIT

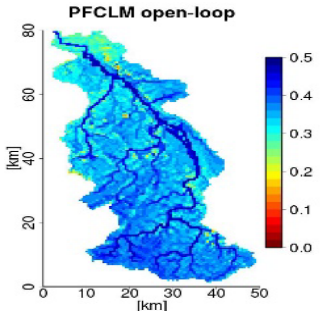
- Create **data analytics techniques** in a systematic manner
  - Domain-specific as well as generalizable and standardized
  - Use case driven co-design between domain scientists, data experts and infrastructure professionals

Karlsruher Institut für Technologie
DESY
JÜLICH FORSCHUNGSZENTRUM
dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT
DLR
HelmholtzZentrum münchen
German Research Center for Environmental Health

www.helmholtz-analytics.de

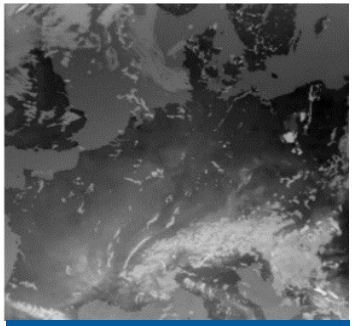
- Facts & Figures
  - 3.5 years started 10/2017
  - 6 Mio. €
  - 23 FTE

HAF Use Cases



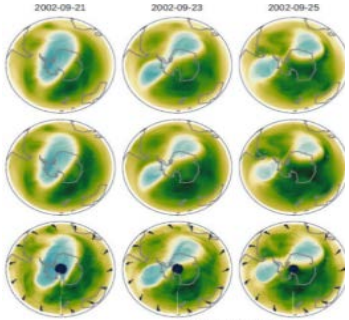
UC 1: Terrestrial Monitoring and Forecasting

source: FZJ-IBG-3



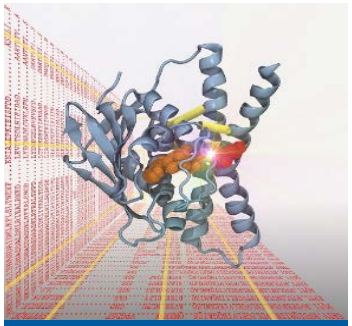
UC 2: Cloud and Solar Power Prediction

source: FZJ-IEK-8



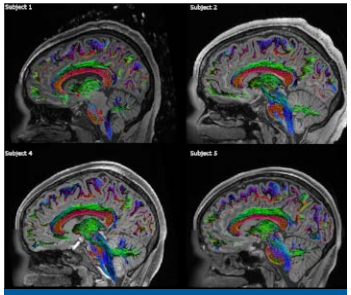
UC 3: Stratospheric Impact on Surface Climate

source: KIT-IMK



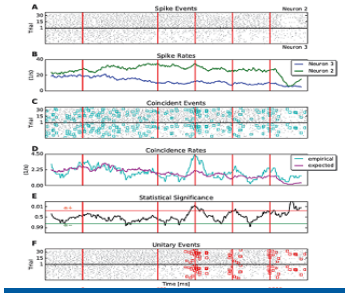
UC 4: Hybrid Data Analysis and Integration for Structural Biology

source: KIT-SCC



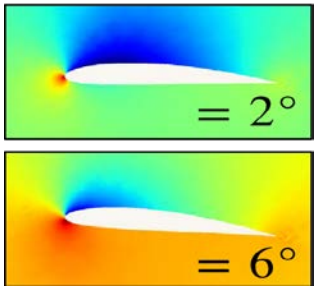
UC5: Image-based High-Throughput Cohort Phenotyping

source: FZJ-INM-1



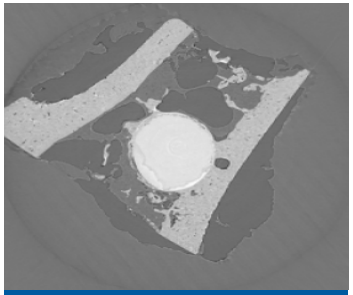
UC 6: Multi-scale, multi-area Interaction in Cortical Networks

source: FZJ-INM-6



UC 7: Virtual Aircraft

source: DLR

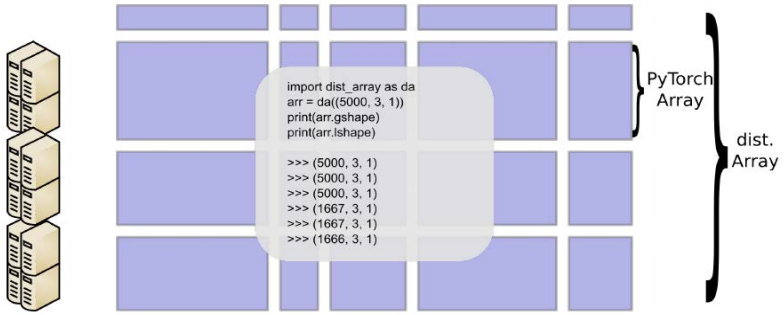


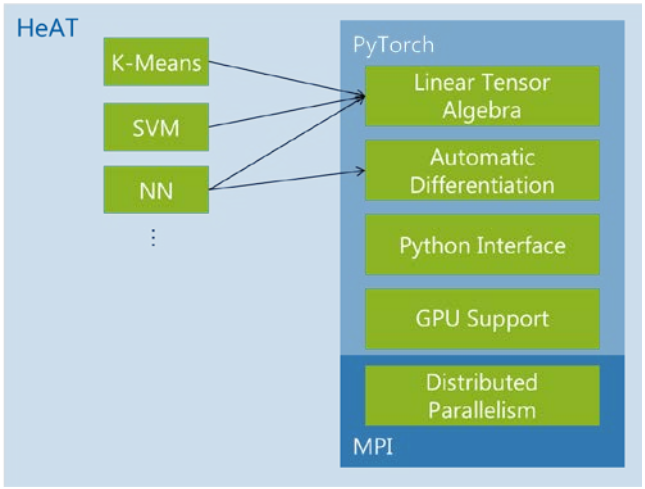
UC 8: Automatic Volumetric Interpretation

source: DESY-IT

HAF – Helmholtz Analytics Toolkit (HeAT)

- **Aim:** develop a generic method for AI on modern, parallel and distributed systems and computing architectures (GPUs)
- Open-source Python data analysis library
  - Parallel, distributed and GPU-accelerated tensor and algorithm implementations
  - Bleeding edge distributed auto-gradient computation for large-scale data-parallel and model-parallel neural networks
  - GitHub Repository https://github.com/helmholtz-analytics/heat





SCC projects landscape – issuing the European federated data infrastructure

Governance
Policies
Skills/Training

Data / Security
Policies
Architecture

Services
Software
Integration

IT services
infrastructure
and Support

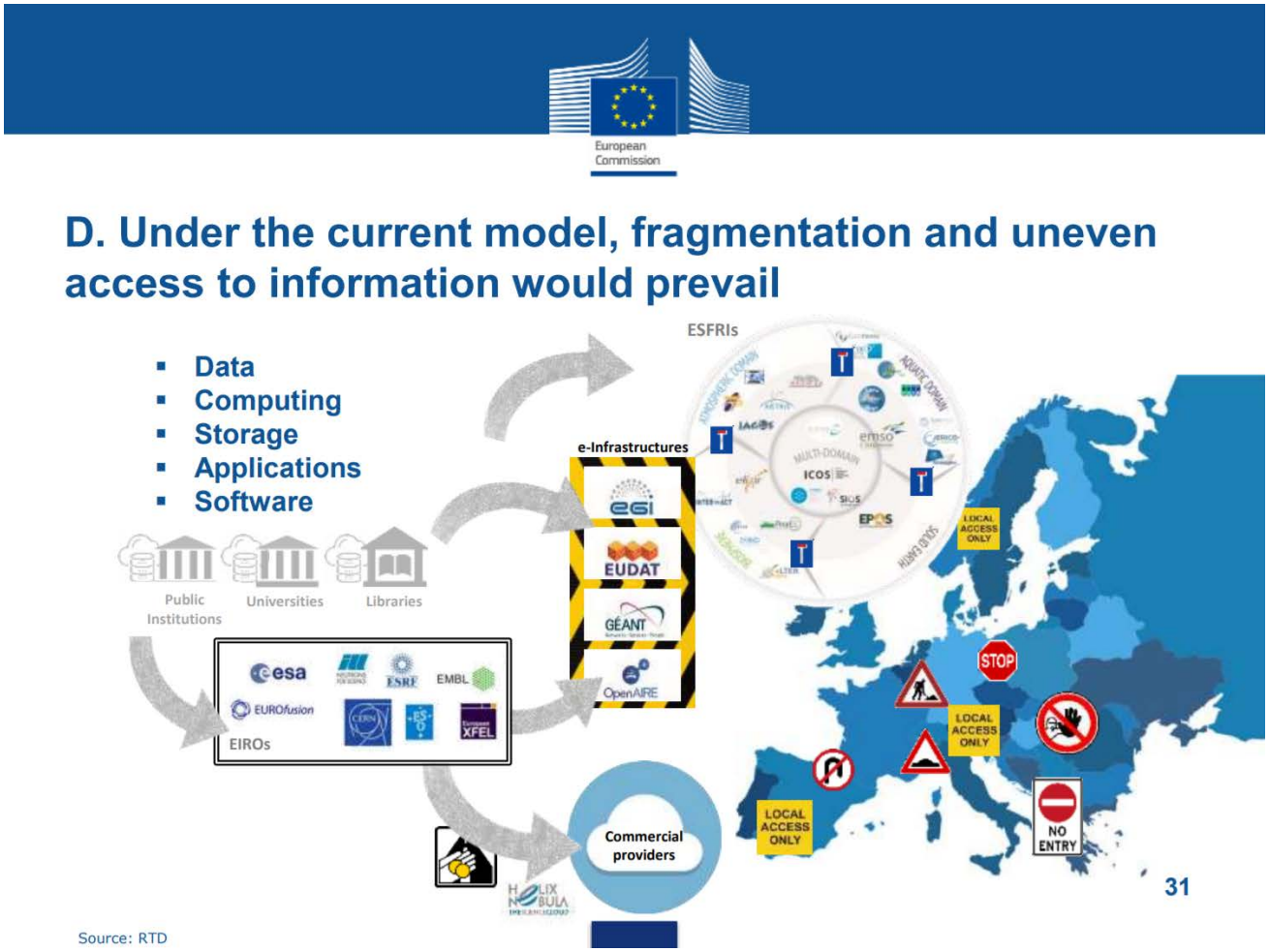
inception | development | piloting | production

European Open Science Cloud

Disclaimer

- I’m using material from other presentations and webpages, all sources are specified, links are provided
- Very good source of information is the EOSC portal https://www.eosc-portal.eu/
- Objects with a red border contains a hyperlink for more information
- KIT is member in the EOSC-related projects EOSCpilot (already finished), EOSC-hub, EOSCsecretariat.eu, EOSC-synergy, EOSC-Pillar

Until recently…



Source: https://ec.europa.eu/research/openscience/pdf/eosc_strategic_implementation_roadmap_large.pdf

Vision…

European Commission

“The EOSC will offer 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States.”
(from https://www.eosc-portal.eu/about/eosc)

Source: https://ec.europa.eu/research/openscience/pdf/eosc_strategic_implementation_roadmap_large.pdf

Evolution

- **04-2016** EOSC is proposed by the EC as part of the European Cloud Initiative to establish a competitive data and knowledge economy in Europe
- **10-2016** First report of the EOSC High Level Expert Group (HLEG) contains initial recommendation to realise the EOSC
  https://op.europa.eu/en/publication-detail/-/publication/2ec2eced-9ac5-11e6-868c-01aa75ed71a1
- **…** intensive consultations with member states and stakeholders
- **06-2017** First EOSC Summit with the ratification of the EOSC Declaration by more than 70 institutions
  https://eosc-portal.eu/sites/default/files/eosc_declaration.pdf
- **03-2018** EC presents the implementation roadmap for the EOSC
  https://ec.europa.eu/transparency/regdoc/rep/10102/2018/EN/SWD-2018-83-F1-EN-MAIN-PART-1.PDF
- **11-2018** EOSC HLEG publishes 2nd and final report "Prompting an EOSC in practice“

  https://op.europa.eu/en/publication-detail/-/publication/5253a1af-ee10-11e8-b690-01aa75ed71a1

  FAIR data HLEG publish the report "Turning FAIR into reality"

  https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1
- **11-2018** Official launch of the EOSC & https://www.eosc-portal.eu/

Landscaping of current EOSC projects

- **General overview:** https://www.eosc-portal.eu/about/eosc-projects
- Call INFRAEOSC-05-2018-2019, part a), **to support the EOSC governance**, see https://cordis.europa.eu/programme/rcn/703191/en

EOSCsecretariat.eu
Setup and management of the EOSC Secretariat supporting the EOSC Governance

- Call INFRAEOSC-05-2018-2019, part b), **to coordinate national and thematic initiatives**, see https://cordis.europa.eu/programme/rcn/703191/en

EOSC-Pillar

EOSC SYNERGY

+ NI4OS, ExPaNDS, EOSC-Nordic

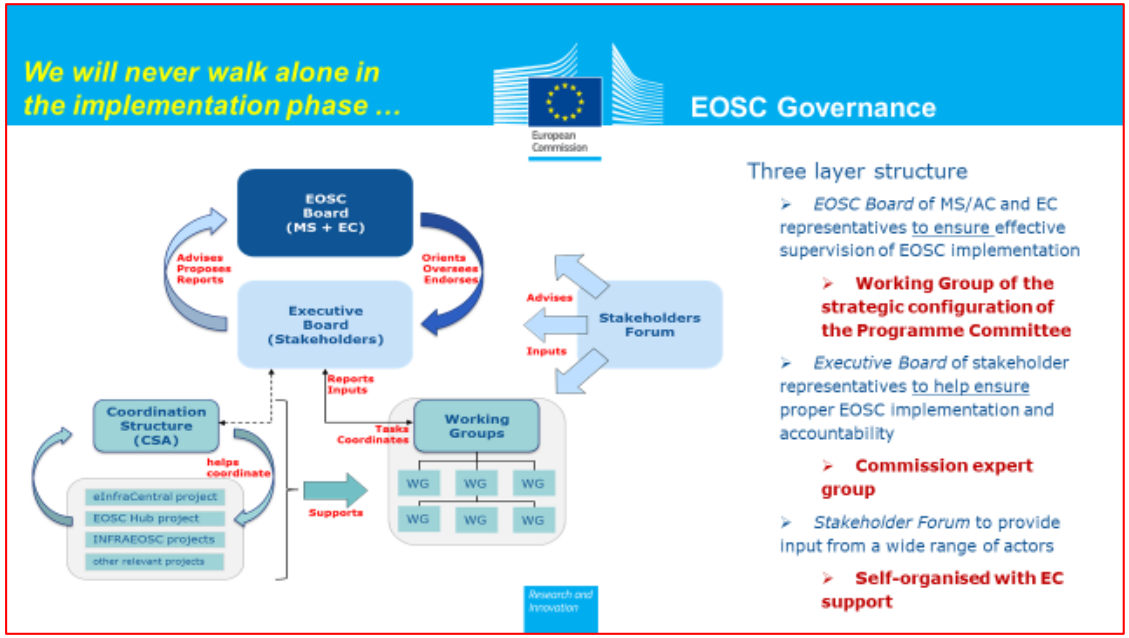
- Call INFRAEOSC-04-2018 **to connects ESFRIs with EOSC**, see https://cordis.europa.eu/programme/rcn/703194/en

PaNOSC, SSHOC, ESCAPE, EOSC-Life, ENVRI-FAIR

- Call Call INFRAEOSC-06-2019-2020 **to optimize the EOSC-portal and connect thematic clouds**, see https://cordis.europa.eu/programme/rcn/703192/en

EOSC Enhance

EOSC Governance Framework



EOSC Executive Board:

List of appointed members

- **Chair** Karel LUYBEN representative of CESAER
- **Vice Chair** Cathrin STÖVER representative of GEANT

Organisations and their representatives

1. CESAER represented by Karel LUYBEN
2. CESSDA ERIC represented by Ronald DEKKER
3. EMBL represented by Rupert LÜCK
4. European Spallation Source ERIC represented by John WOMERSLEY
5. GÉANT represented by Cathrin STÖVER
6. OPENAIRE represented by Natalia MANOLA
7. Research Data Alliance (RDA) represented by Juan BICARREGUI
8. Science Europe represented by Stephan KUSTER

Individual experts

1. Sarah JONES
2. Jean-Francois ABRAMATIC
3. Jan HRUSAK

Sources: slide 21 of https://www.eoscpilot.eu/sites/default/files/burgelman-2018_eosc_stakeholderforum.pdf, https://www.eosc-portal.eu/governance, https://ec.europa.eu/info/news/results-call-applications-selection-members-expert-group-members-executive-board-eosc-2018-nov-23_en

Thank you, Questions?