Digital Agenda:

1. **Accelerate Digital Transformation in Science**

To secure a broad, discipline and organization overarching access to and availability of digital information, the scientific information infrastructures are strengthened, extended, and better connected.
Natural science fundamental research is a central area of application of new digital methods and techniques. It is a significant driver for further developments. Increasing computational effort and complex data management is addressed by site overarching work techniques and the elimination of technological bottlenecks. Open access and long term data management must continue to take into account the requirements and specifics of the different research infrastructures. Young scientists acquire a unique expertise in data management. New services and holistic solutions can arise in future based on the know-how in fundamental research.
Challenges: Data Volumes/Rates

- HL-LHC resource estimates factors above flat budget scenario
- ALICE & LHCb: Triggerless readout in Run 3
- Belle II: 50x more data than Belle
- FAIR: 30 PB per year, 300,000 CPU cores
- CTA: several 10 PB per year
Challenges: Technological Evolution

And opportunities:

- Multicore machines
- GPUs
- SSDs
- Virtualization
- Machine Learning
- Artificial Intelligence
- Quantum computing
- ...
Approach:

» Find common solutions
IDT-UM Project

➢ Innovative Digital Technologies for Research on Universe and Matter

• Application of partners from
  ✓ Particle Physics (ATLAS, Belle II, CMS)
  ✓ Hadron and Nuclear Physics (ALICE, CBM, PANDA)
  ✓ Astroparticle Physics (Auger, CTA, IceCube)

➢ to develop experiment overarching solutions

• Evaluated by panel including computer scientists
  ➢ Got 3.6 M€ for 3 years, started October 2018
Project Partners
Scientific Computing Today

HEP User

12h job on x86, SL6
2 GB RAM
Dataset X

Grid Site
Batch system
x86 WN
SL6, 2 GB

Dataset X
Exp. Service
Scientific Computing Tomorrow

HEP User

Job
Dataset X

Cloud (Spot Market)

Grid Site
Batch system
x86 WN
SL6, 2 GB

Exp. Service

Dataset X

High Performance Center (HPC)

Uni-Cluster

GPU Cluster

...
Scientific Computing Vision

Astro User

Job Dataset X

HEP User

Bio User

GPU Cluster

Cloud (Spot Market)

Grid Site

VM oder Container

Exp. Service

High Performance Center (HPC)

Uni-Cluster

...

Astro User

HEP User

Bio User
Subject Area A

- Developments for the provision of technologies for the use of heterogeneous computing resources

<table>
<thead>
<tr>
<th>A1: Tools for integration</th>
<th>A2: Efficient Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Scheduling of cloud jobs</td>
<td>• Transient data caches</td>
</tr>
<tr>
<td>• Container technologies</td>
<td>• Transparent access to distributed data</td>
</tr>
<tr>
<td>• Database access</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A3: Workflow Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Optimization with data mining</td>
</tr>
</tbody>
</table>
Subject Area B

Application and test of virtualized software components in the environment of heterogeneous computing resources

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation and test on different platforms</td>
<td></td>
</tr>
<tr>
<td>• Storage and caching solutions</td>
<td></td>
</tr>
<tr>
<td>• Virtualized services (databases, monitoring, accounting)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B3: Virtualization of User Jobs</th>
<th>B4: Combined Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Requirement capture</td>
<td></td>
</tr>
<tr>
<td>• Determination and creation of runtime environment</td>
<td></td>
</tr>
<tr>
<td>• Creation of container and meta data</td>
<td></td>
</tr>
<tr>
<td>Test of complete system on different platforms regarding</td>
<td></td>
</tr>
<tr>
<td>• Installation and maintenance</td>
<td></td>
</tr>
<tr>
<td>• Performance</td>
<td></td>
</tr>
<tr>
<td>• Scalability</td>
<td></td>
</tr>
<tr>
<td>• Robustness</td>
<td></td>
</tr>
</tbody>
</table>
Example of a Common Solution

➢ COBaID (COBaID Opportunistic Balancing Daemon)
  • Overlay Batch System (OBS)
  • Pilot (resources) → Drone (resources and environment)
  • Adjustment of allocated resources to demands

➢ TARDIS (Transparent Adaptive Resource Dynamic Integration System)
  • Adapters for OpenStack, CloudStack, Moab, Slurm, HTCondor
Example of a Common Solution

Local Site:
- Users
- Job submission
- Overlay Batch System (OBS)
- usage monitoring

External Site:
- Access Point
- Resource Pool
  - VM
    - batch job
    - container
- schedule request and start resource

increase resources
utilization
decrease resources

integrate into OBS
jobflow
Subject Area C

- Deep Learning, Gain of knowledge by substantiated data-driven methods

<table>
<thead>
<tr>
<th>C1: Processing of Sensor Data</th>
<th>C2: Object Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Signal filter, noise suppression</td>
<td></td>
</tr>
<tr>
<td>- Processing of time dependent data</td>
<td></td>
</tr>
<tr>
<td>- Track and cluster reconstruction, jet forming, event reconstruction</td>
<td></td>
</tr>
<tr>
<td>- Questions of placement, order, assignment of data</td>
<td></td>
</tr>
<tr>
<td>- Extraction of small signals in case of large backgrounds</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C3: Network Accelerated Simulations</th>
<th>C4: Quality of Network Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Generative adverserial networks, adjustment of simulation to data</td>
<td></td>
</tr>
<tr>
<td>- Methods for the evaluation of the quality of network simulations</td>
<td></td>
</tr>
<tr>
<td>- Reduction of experimental systematic uncertainties</td>
<td></td>
</tr>
<tr>
<td>- Special learning strategies</td>
<td></td>
</tr>
<tr>
<td>- Prediction relevant information</td>
<td></td>
</tr>
<tr>
<td>- Uncertainty of predictions</td>
<td></td>
</tr>
</tbody>
</table>
Example of a Common Solution

- Selection of (background) events on generator level to save simulation time

- Collection of problems and solution being worked on
Subject Area D

- Event reconstruction: Cost- and energy-efficient use of computing resources

<table>
<thead>
<tr>
<th>D1: Track Finding</th>
<th>D2: Parameter Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Alternative algorithms, e.g. cellular automata</td>
<td>• Connection of GenFit2-ACTS</td>
</tr>
<tr>
<td>• Alternative architectures, e.g. GPUs</td>
<td></td>
</tr>
</tbody>
</table>
Example of a Common Solution

➢ A Common Tracking Software

- ACTS core, framework, simulation, data

![Diagram of Detailed geometry and ACTS Geometry with labels: Detailed geometry, DetectorElement, ACTS Geometry, Surface]
IDT-UM Further Information

➢ Web page: https://www.erum-data-idt.de/
➢ Mailing list: computing-verbund@lists.lrz.de
➢ erum-data-idt organization on github
➢ Next collaboration meeting on Thursday, April 2\textsuperscript{nd} at 15:00 in Bonn (during DPG conference)

Innovative Digital Technologies for Research on Universe and Matter

Progress in fundamental research on universe and matter (ErUM) is made by studying structures at smaller and smaller scales. The high resolution of modern instruments in particle, hadron and nuclear, and astroparticle physics results in huge amounts of research data, at the order of millions of terabytes. And the next generation of experiments will increase the dataset sizes even more, exceeding the growth expected from advances in storage technologies.
Computing Strategy Workshop

➢ Purpose: Agree on a common computing infrastructure strategy to address in particular the HL-LHC needs in the context of a general KAT, KET, KHuK strategy
  • Roles of research centers and universities?
  • Community or workflow specific resources or science cloud?
  • Resource projections?
  • Required technological developments?
  • Required political boundary conditions?
  • Relations among communities, to international partners and funding agencies?
  • Long term sustainability?

➢ Tentatively planned for May at GridKa
➢ Open to all who are interested in the topic
Wohin mit den gigantischen Datenmengen der Grundlagenforschung?

Experimente in der Grundlagenforschung sind Speicherfresser: 10 Millionen DVDs bräuchte man für die Daten, die jährlich am CERN anfallen. Mit innovativen Verarbeitungsmethoden wollen Forschende eines Computing-Verbundes dieser Datenflut Herr werden.
ErUM Data

- IDT-UM is a pilot project of ErUM Data
- ErUM Data action plan of BMBF expected this year
- Input from ErUM communities collected last year
  - Federated infrastructure
  - Big data analytics
  - Data management

- Communities:
  astro particle physics (KAT),
  particle physics (KET), astronomy (RDS),
  hadron and nuclear physics (KhuK),
  accelerator physics (KfB),
  research with neutrons (KFN) / synchrotron radiation (KFS) / ions (KFSI)

<table>
<thead>
<tr>
<th>Scientists with doctoral degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFS</td>
</tr>
<tr>
<td>RDS</td>
</tr>
<tr>
<td>KHuK</td>
</tr>
<tr>
<td>KET</td>
</tr>
<tr>
<td>KFSI</td>
</tr>
<tr>
<td>KAT</td>
</tr>
<tr>
<td>KfB</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>1,500</td>
</tr>
<tr>
<td>1,500</td>
</tr>
<tr>
<td>1,300</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>2,300</td>
</tr>
<tr>
<td>1,500</td>
</tr>
<tr>
<td>1,500</td>
</tr>
<tr>
<td>1,300</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>8,400</td>
</tr>
</tbody>
</table>
ErUM Data Community Input

➢ Workshop at BMBF 4./5.10.2018

5 Recommended measures and cost estimates

5.1 Federated infrastructures
5.2 Integration of workflows to exploit infrastructures
5.3 Comprehensive management of research data
5.4 Modern Big Data Analytics in fundamental research
5.5 Scientists’ integrated web working environment
5.6 Tenure-Track programme: knowledge in digitization
5.7 Partnership for Innovative Digitization
5.8 Cost estimates


➢ Strategy document of all ErUM communities given to BMBF on 2.5.2019
ErUM Data Cost Estimates

Full Time Equivalents

<table>
<thead>
<tr>
<th>Position Description</th>
<th>MEuro/y</th>
<th>MEuro/position in 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflows to exploit infrastructures</td>
<td>100</td>
<td>0.072</td>
</tr>
<tr>
<td>Management of research data</td>
<td>100</td>
<td>0.072</td>
</tr>
<tr>
<td>Big Data Analytics in physics research</td>
<td>200</td>
<td>0.072</td>
</tr>
<tr>
<td>Scientist’s web working environment</td>
<td>100</td>
<td>0.072</td>
</tr>
<tr>
<td>Tenure track ErUM programme + 1 RA*</td>
<td>100</td>
<td>0.158</td>
</tr>
<tr>
<td>Total FTE</td>
<td>600</td>
<td></td>
</tr>
</tbody>
</table>

*RA=Research Associate

Due to the long-term nature of the responsibilities, these positions should ideally have long-term perspectives.

Cost estimate of recommended measures /MEuro

<table>
<thead>
<tr>
<th></th>
<th>2020</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Time Equivalents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>17.8</td>
<td>32.1</td>
<td>47.3</td>
<td>52.6</td>
<td>58.3</td>
<td>60.0</td>
<td>61.8</td>
<td>63.7</td>
<td>65.6</td>
<td>67.6</td>
</tr>
<tr>
<td>2</td>
<td>5.0</td>
<td>10.0</td>
<td>15.0</td>
<td>20.0</td>
<td>25.0</td>
<td>30.0</td>
<td>30.0</td>
<td>30.0</td>
<td>30.0</td>
<td>30.0</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Total</td>
<td>25.8</td>
<td>45.1</td>
<td>65.3</td>
<td>75.6</td>
<td>86.3</td>
<td>93.0</td>
<td>94.8</td>
<td>96.7</td>
<td>98.6</td>
<td>100.6</td>
</tr>
</tbody>
</table>

Due to the long-term nature of the responsibilities, these positions should ideally have long-term perspectives.
ErUM Data Organization

- **Guidelines for organization across communities developed by digitization board**

<table>
<thead>
<tr>
<th>Coordinator</th>
<th>Overview Board (OB)</th>
<th>Speaker / Co-Speaker</th>
<th>Digitization Board (DB)</th>
<th>Resource Provider Board (RB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 Committee Chairs, 1 Resource Provider, 1 Representative of the BMBF</td>
<td>Speaker, Co-speaker, 8 Experts from committees, 1 Resource Provider, 5 Topic Coordinators</td>
<td>10 Resource Providers, 8 Experts from committees</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Administrative Office (AO)</th>
<th>Annual Conference of the ErUM-Data Working Groups</th>
<th>International Advisory Board (IAB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone coordination, includes 1 Administration coordinator &amp; Team</td>
<td></td>
<td>ca. 5 from Science, Industry</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic Boards</th>
<th>Topic Federated Infrastructures Board: Coordinator, Experts Compute power Utilization Workflows</th>
<th>Topic Big Data Analytics Board: Coordinator, Experts Algorithms Autonomization Control &amp; preservation</th>
<th>Topic Research Data Board: Coordinator, Experts Data models Management Curation</th>
<th>Topic User Interface Board: Coordinator, Experts Scientists questions Developers work User support</th>
<th>Topic Knowledge distribution Board: Coordinator, Experts Tenure track programme Workshop, schools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
The aim of the national research data infrastructure (NFDI) is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally. The NFDI will bring multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.
PAHN-PaN

The PAHN-PaN Consortium
Particle, Astroparticle, Hadron & Nuclear Physics accelerate the NFDI

Task area 1: Developing workflows and tools for data management
Task area 2: FAIR data lifecycle concepts and open data
Task area 3: Data analysis procedures and services
Task area 4: Real-time data analysis and selection

Cross-cutting topic A: Synergies
Cross-cutting topic B: Services
Cross-cutting topic C: Professional training, education, outreach
Further National Context

- KAT Digital Committee
- KET Computing and Software Panel
- KHuK Computing Committee
- DPG: AKPIK
- DPG: Physics and Information is one of four topics of 175\textsuperscript{th} anniversary celebration

Arbeitskreis Physik, moderne Informationstechnologie und Künstliche Intelligenz

...
International Context

➢ International collaborations
➢ HEP Software Foundation: Community White Paper
  ➢ Improvements in software efficiency, scalability and performance
  ➢ Enable new approaches that can radically extend physics reach
  ➢ Ensure the long-term sustainability of the software
➢ IRIS-HEP
➢ SIDIS: Software Institute for Data Intensive Science
➢ New journal: Computing and Software for Big Science
➢ EOSC: European Open Science Cloud
➢ …
ESCAPE Project

Horizon 2020 funded project

Goals:

- Prototype an infrastructure adapted to the Exabyte-scale needs of the large science projects.
- Ensure the sciences drive the development of the EOSC
- Address FAIR data management

Data centres: CERN, INFN, DESY, GSI, Nikhef, SURFSara, RUG, CCIN2P3, PIC, LAPP, INAF

Science Projects
- HL-LHC
- FAIR
- KM3Net
- ELT
- EURO-VO
- (LSST)
- SKA
- CTA
- JIVE-ERIC
- EST
- EGO-VIRGO (CERN, ESO)
Summary

➢ Digitization offers opportunities to address the challenges of increasing data rates and volumes in fundamental research on universe and matter

● Development of common solutions encouraged by funding agencies

➢ Pilot project with partners from particle, hadron and nuclear, and astro particle physics

➢ ErUM Data can have high impact on our field of research

➢ A lot is currently happening in the field of digitization

➢ You are part of this and can shape the future of science and society
Backup
Technological Evolution

42 Years of Microprocessor Trend Data

Year
