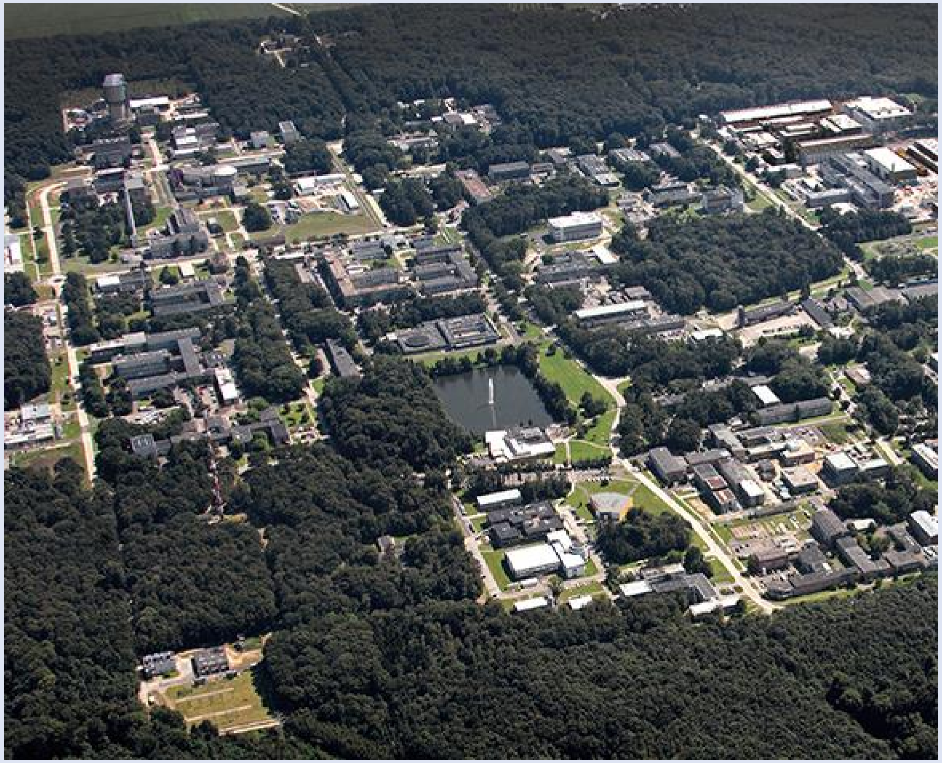
THE EVOLUTION OF THE HPC FACILITY AT JSC

2019-06-04 I D. KRAUSE (WITH VARIOUS CONTRIBUTIONS)

JÜLICH Forschungszentrum

RESEARCH AND DEVELOPMENT @ FZJ

on 2.2 Square Kilometres



FORSCHUNGSZENTRUM JÜLICH: AT A GLANCE

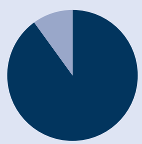
Facts and Figures

1956

FOUNDATION

on 12 December

Shareholders



90 %	Federal Republic of Germany
10 %	North Rhine-Westphalia

11

INSTITUTES

2 project management organizations

609.3

million euros **REVENUE total**

(40 % external funding)

5,914

EMPLOYEES

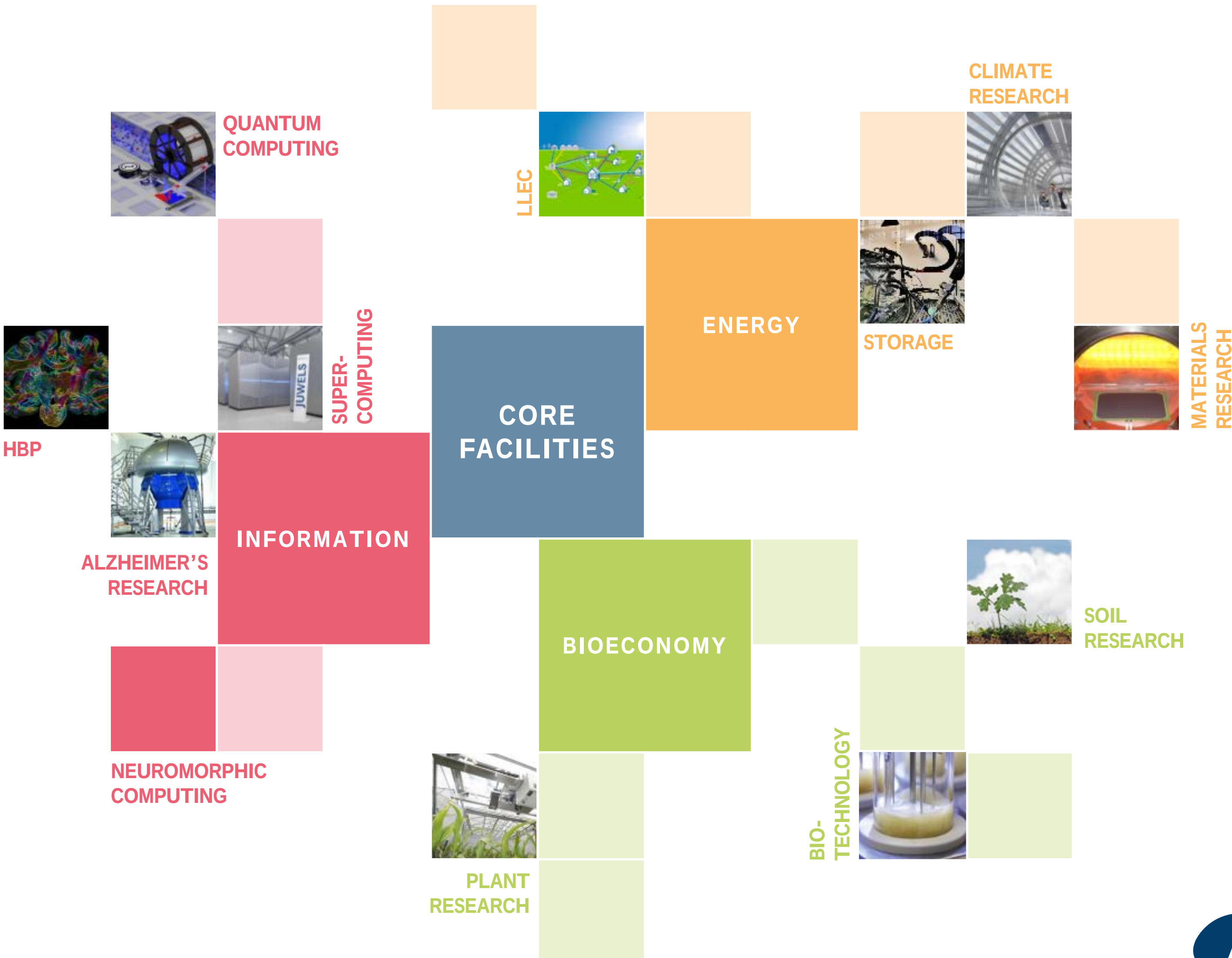
2,165	scientists
536	doctoral researchers
323	trainees and students on placement

867

VISITING SCIENTISTS

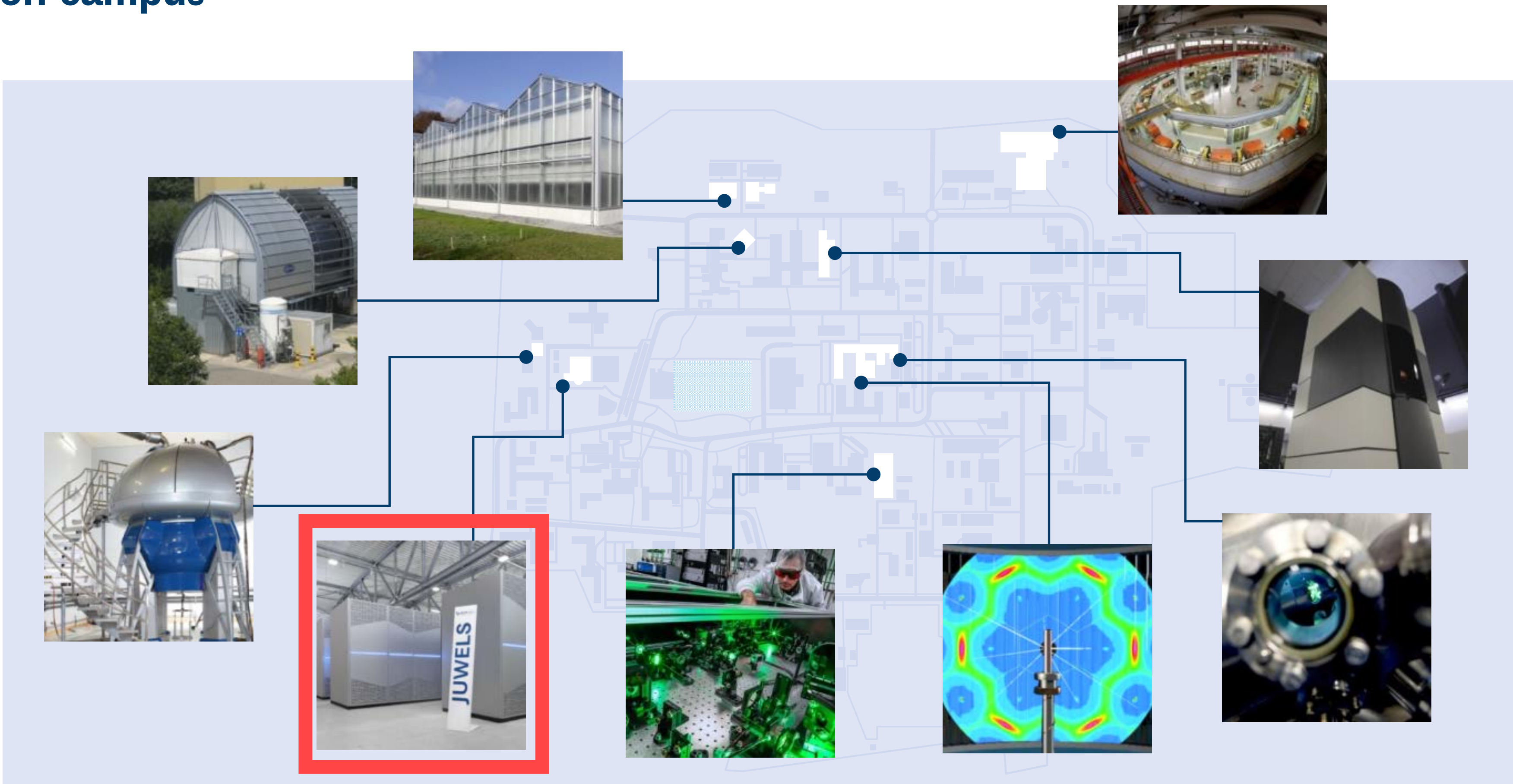
from **65** countries

STRATEGIC PRIORITIES



LARGE-SCALE INSTRUMENTS

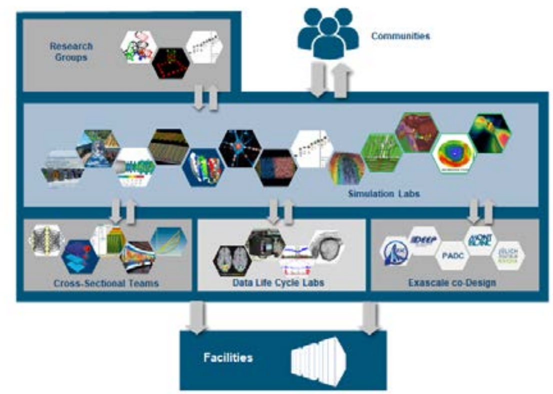
on campus



JÜLICH SUPERCOMPUTING CENTRE

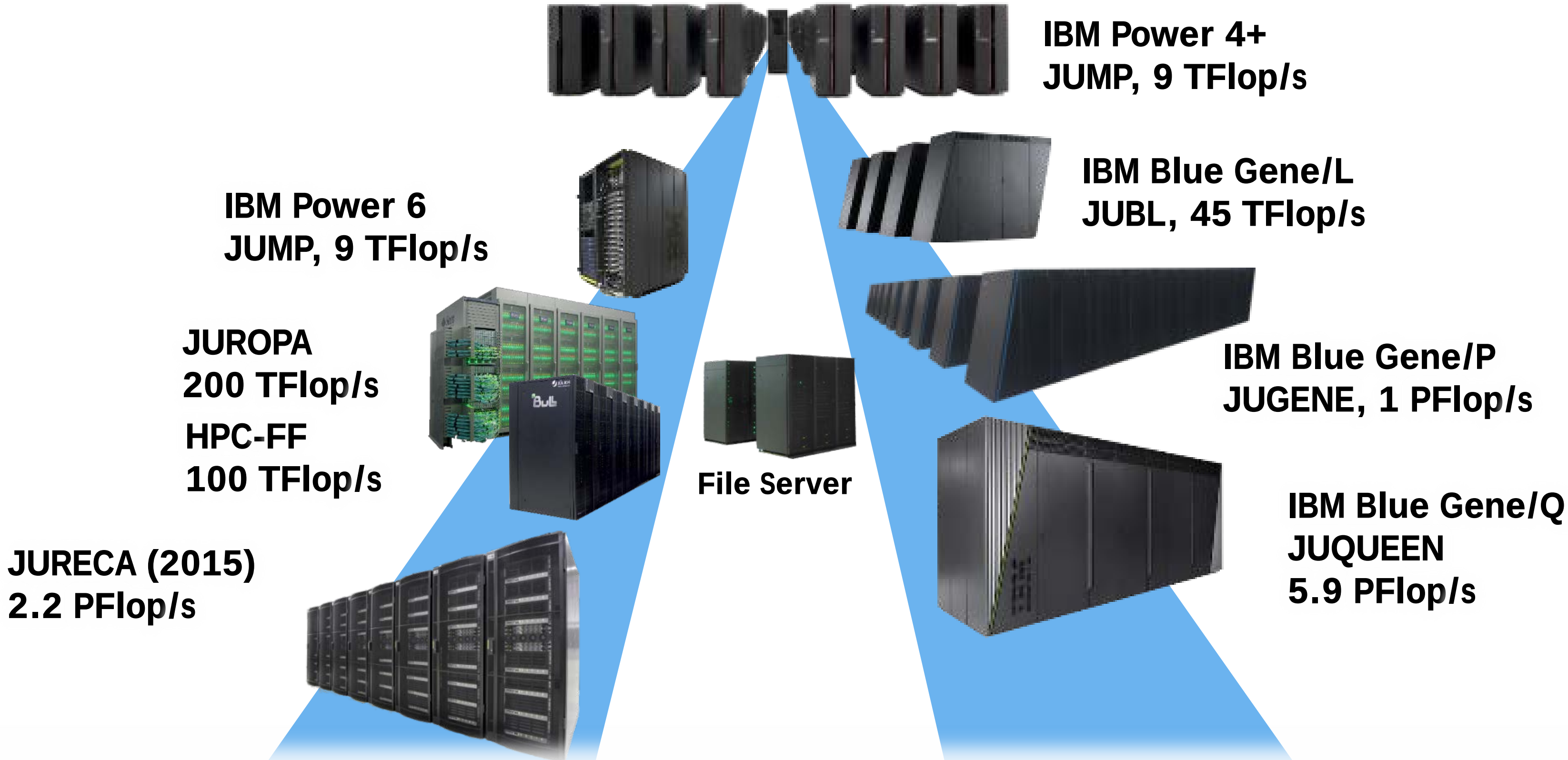
JÜLICH SUPERCOMPUTING CENTRE

- **Supercomputer operation for:**
  - Center – FZJ
  - Region – RWTH Aachen University
  - Germany – Gauss Centre for Supercomputing
    John von Neumann Institute for Computing
  - Europe – PRACE, EU projects
- **Application support**
  - Unique support & research environment at JSC
  - Peer review support and coordination
- **R-&-D work**
  - Methods and algorithms, computational science, performance analysis and tools
  - Scientific Big Data Analytics
  - Computer architectures, Co-Design
    Exascale Laboratories: EIC, ECL, NVIDIA
- **Education and Training**



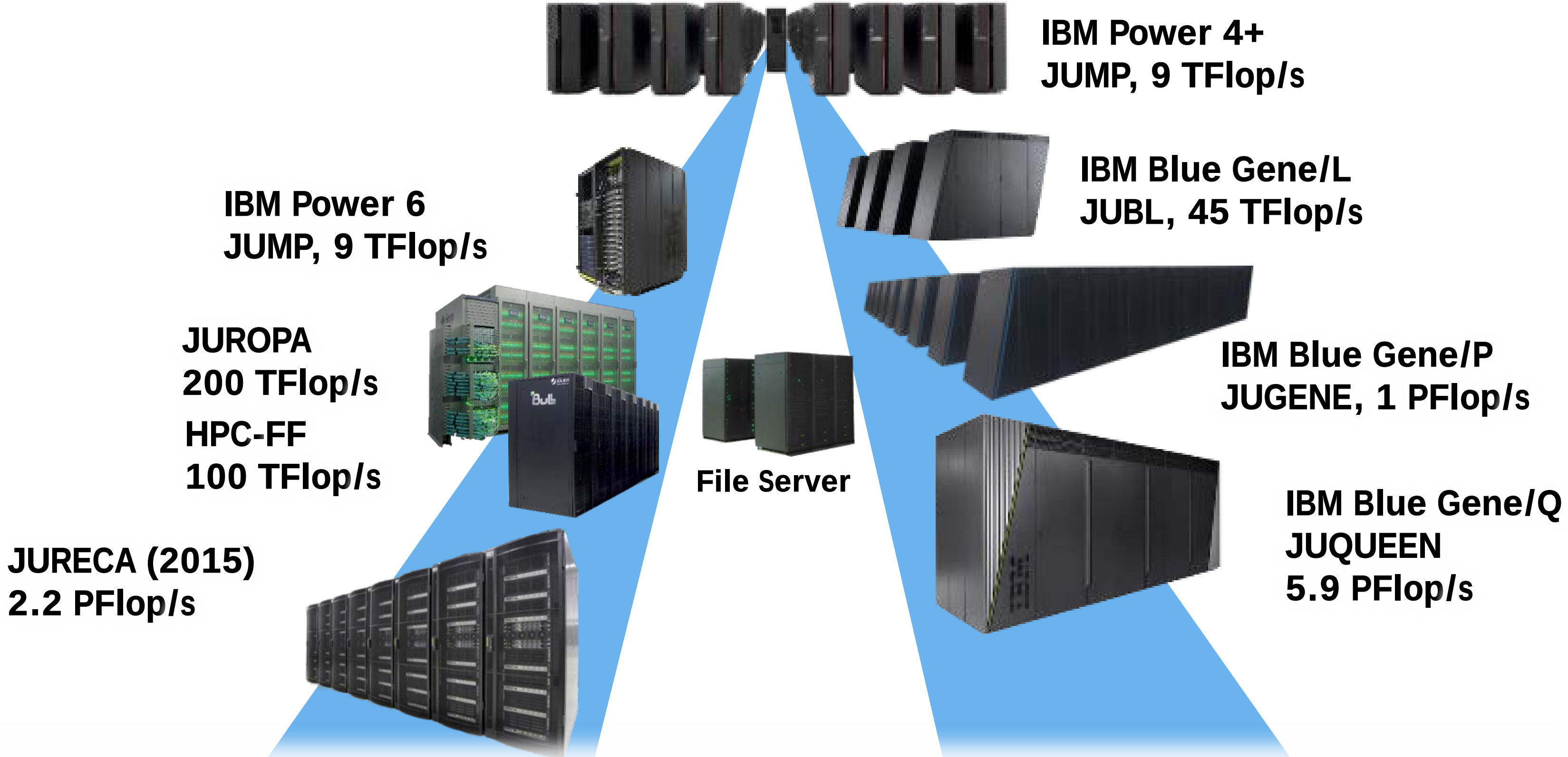
DEEP

JÜLICH Forschungszentrum



Dual architecture strategy: Addresses disparity of user requirements

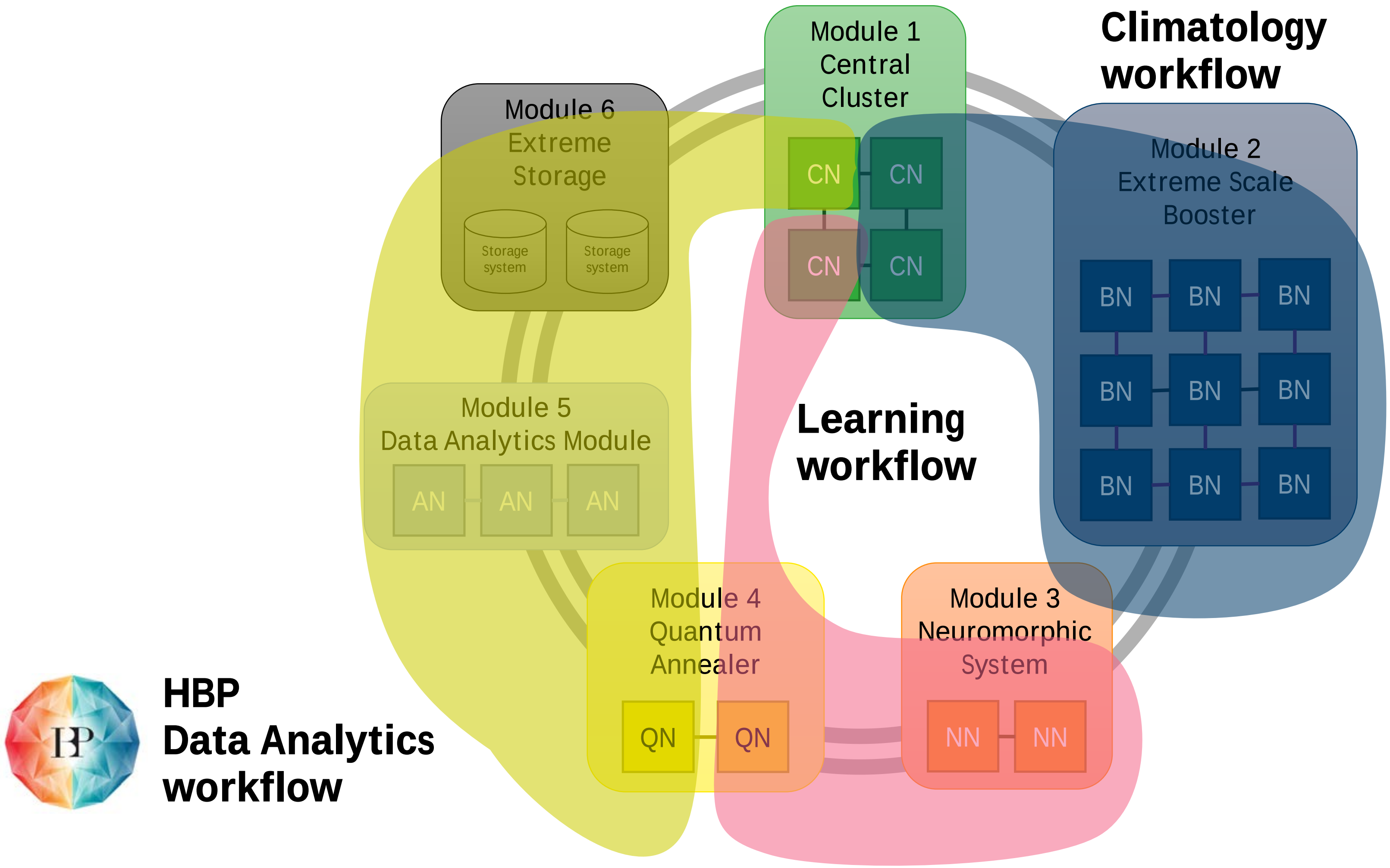
- Grand Challenge applications require extreme performance
- Not achievable with general purpose architectures (x86 clusters) due to cost & energy
- Highly scalable architectures not suitable for applications requiring high single node performance, large memory per core



Dual architecture strategy: Does not address dynamic requirements

- Parts of complex applications or workflows often have different requirements and scalability properties
- Traditional accelerated systems enforce static ratio of CPU / accelerator performance often wasting resources and energy

MODULAR SUPERCOMPUTING



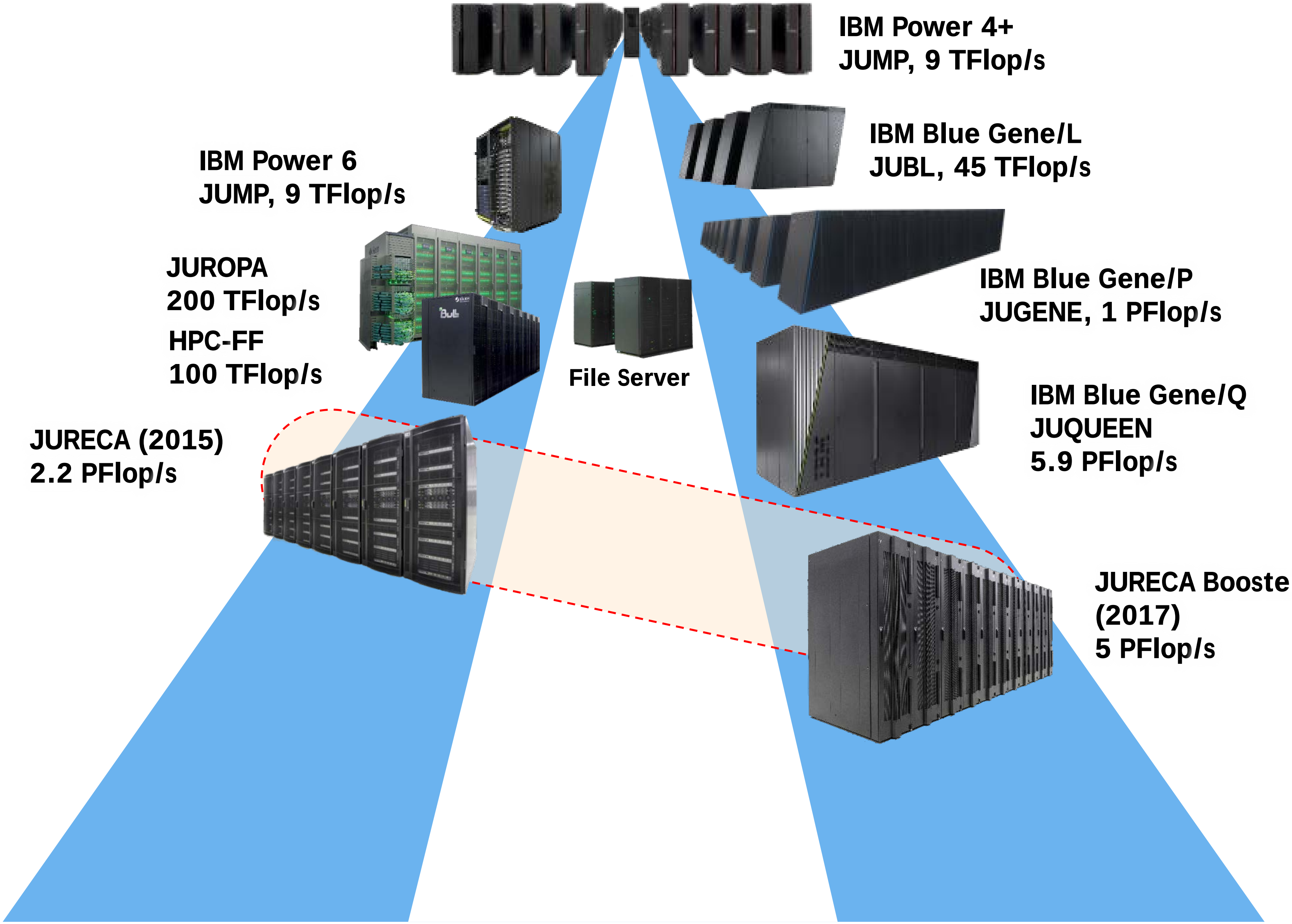
JÜLICH Forschungszentrum

DEEP PROJECT SERIES

DEEP, DEEP-ER, DEEP-EST: Exascale technology development

- 20+ partners
- 44 Mio € (30 M€ EU funded)

SEVENTH FRAMEWORK PROGRAMME



IBM Power 4+
JUMP, 9 TFlop/s
IBM Blue Gene/L
JUBL, 45 TFlop/s
IBM Power 6
JUMP, 9 TFlop/s
JUROPA
200 TFlop/s
HPC-FF
100 TFlop/s
IBM Blue Gene/P
JUGENE, 1 PFlop/s
File Server
IBM Blue Gene/Q
JUQUEEN
5.9 PFlop/s
JURECA (2015)
2.2 PFlop/s
JURECA Booster
(2017)
5 PFlop/s

JURECA CLUSTER+BOOSTER

JURECA

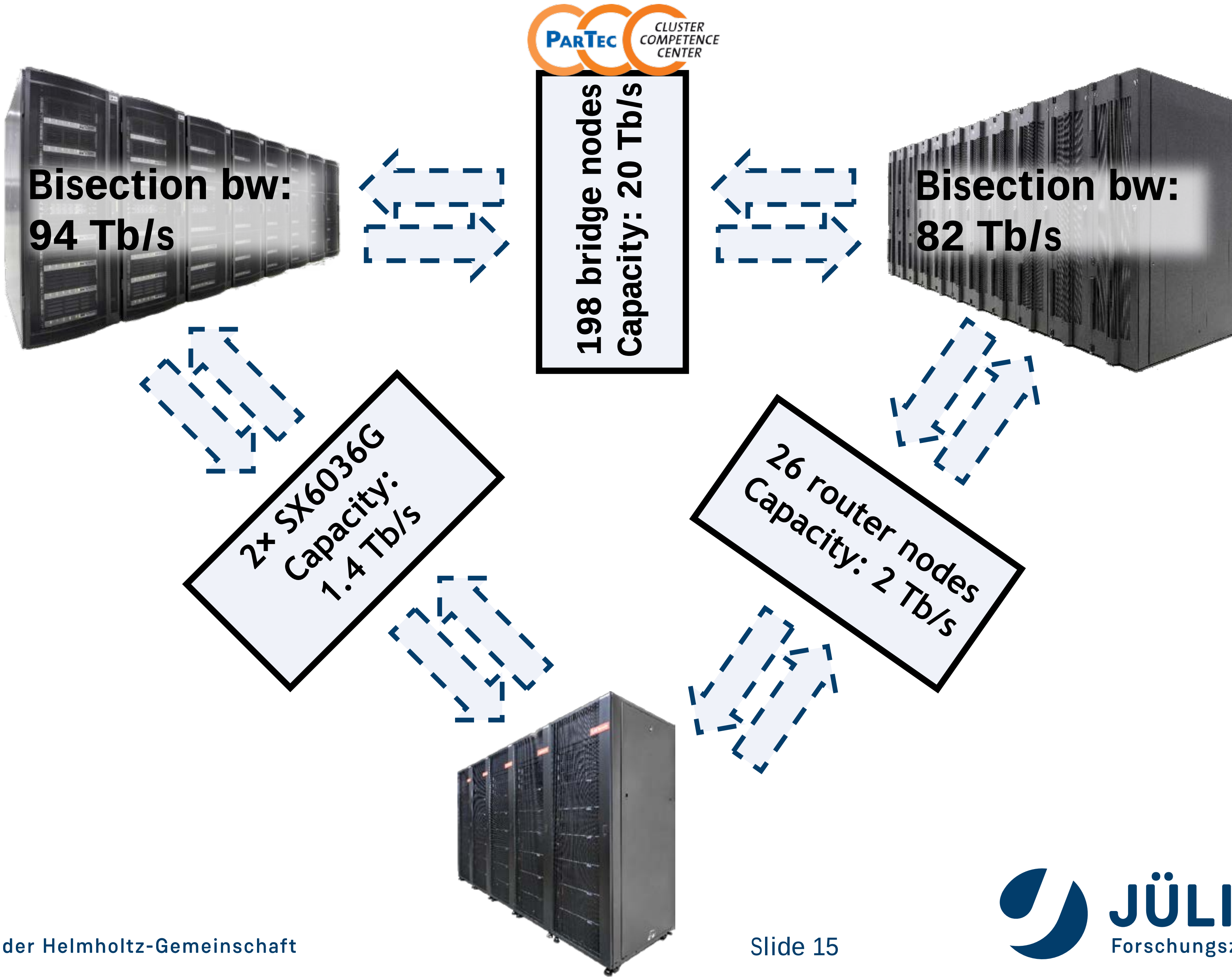
JURECA Cluster

- 1882 compute nodes based on dual-Socket Intel Xeon Haswell
- Mellanox InfiniBand EDR100 Gb/s network **!**
- Full fat-tree topology
- 2.2 PF/s

JURECA Booster

- 1640 compute nodes based on Intel Xeon Phi 7250-F
- Intel Omni-Path Architecture 100 Gb/s network **!**
- Full fat-tree topology
- 5 PF/s

JURECA CLUSTER+BOOSTER ARCHITECTURE



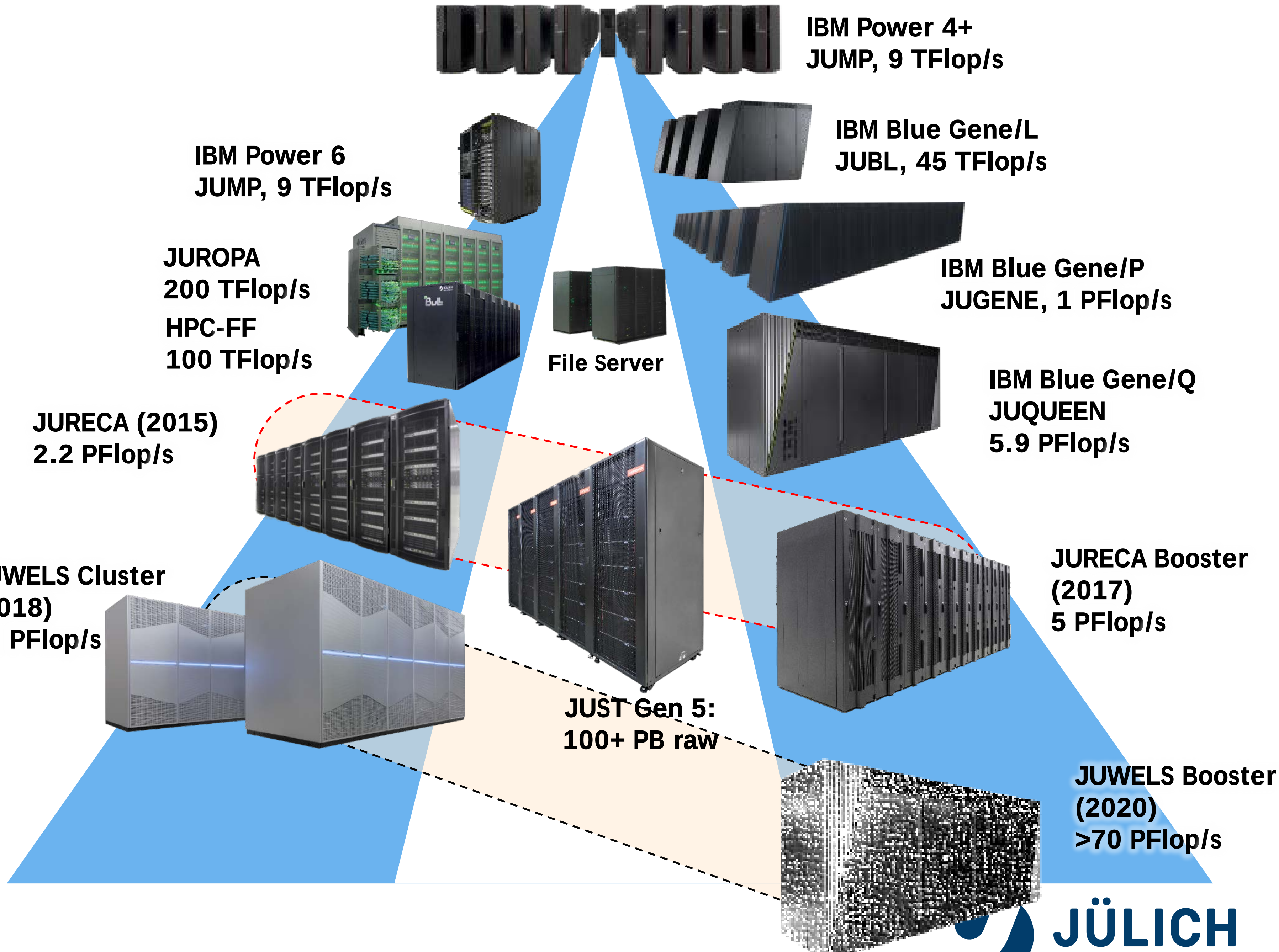
POC: FULL-SYSTEM LINPACK ON JURECA (NOV 2017)

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WHC00L2L4     5321904   336    40    84          26565.78            3.78257e+06
HPL_pdgesv() start time Sun Nov  5 00:23:35 2017

HPL_pdgesv() end time   Sun Nov  5 07:46:21 2017

        HPL Efficiency by CPU Cycle 5328300.353%
        HPL Efficiency by BUS Cycle 9446281.578%
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0030562 ...... PASSED
================================================================================
```

1760 Cluster nodes + 1600 Booster nodes + 120 bridge nodes

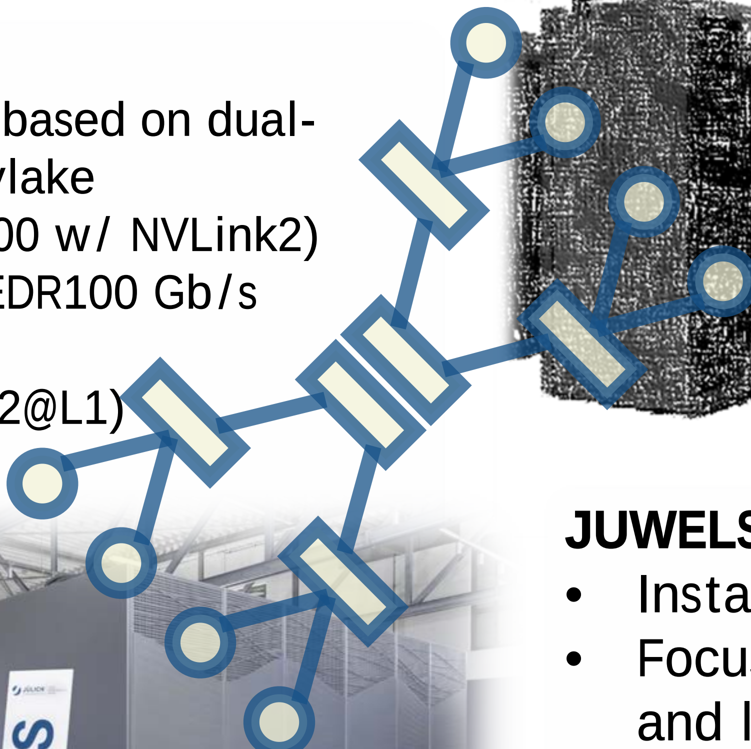

IBM Power 4+
JUMP, 9 TFlop/s
IBM Blue Gene/L
JUBL, 45 TFlop/s
IBM Power 6
JUMP, 9 TFlop/s
IBM Blue Gene/P
JUGENE, 1 PFlop/s
JUROPA
200 TFlop/s
HPC-FF
100 TFlop/s
File Server
IBM Blue Gene/Q
JUQUEEN
5.9 PFlop/s
JURECA (2015)
2.2 PFlop/s
JURECA Booster
(2017)
5 PFlop/s
JUWELS Cluster
(2018)
12 PFlop/s
JUST Gen 5:
100+ PB raw
JUWELS Booster
(2020)
>70 PFlop/s

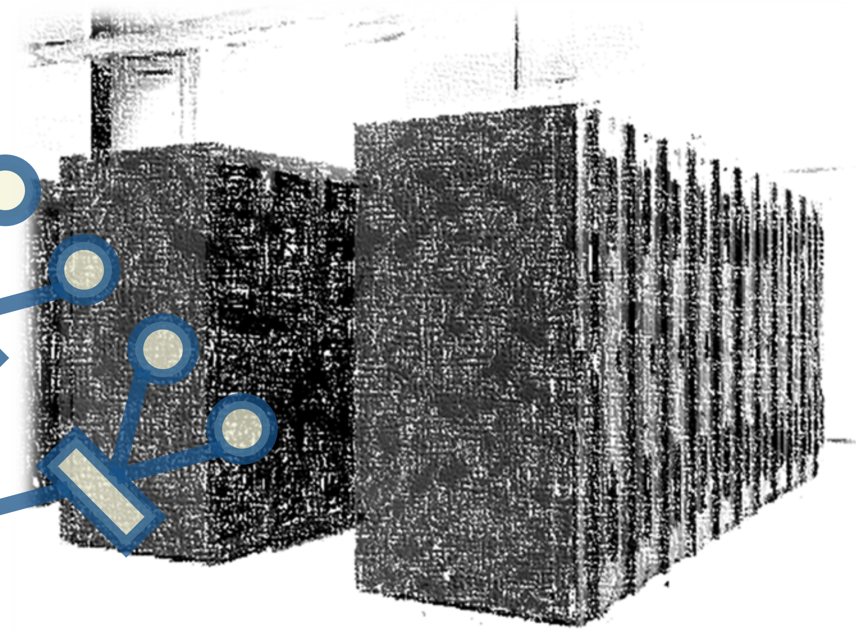
JUWELS CLUSTER (+ BOOSTER)

JUWELS

JUWELS Cluster

- 2511 compute nodes based on dual-Socket Intel Xeon Skylake
- 48 GPU nodes (4× V100 w/ NVLink2)
- Mellanox InfiniBand EDR100 Gb/s network
- Fat-tree topology (1:2@L1)
- 12 PF/s





JUWELS Booster

- Installation in 2020
- Focus on massively-parallel and learning applications
  - GPUs
  - Balanced network
- 50+ PF/s

DRIVING FACTORS FOR SYSTEM DESIGN

- Performance-per-€
- System balance
  - B:F ratio dropped from 1:7 (JURECA) to 1:16 (JUWELS)
  - Nvidia V100 (900 GB/s HBM2): 1:8
- Infrastructure constraints
  - Power envelope (F-per-W)
  - Cooling infrastructure
  - System density

EXASCALE PLANS WORLDWIDE

- **US:** Aurora @ ANL
  - Intel X86 + X^e GPU
  - Ca. 550 M€
- **US:** Frontier @ ORNL
  - AMD X86 + AMD GPUs
  - 1.5 EF, 40 MW, 500 M€
- **Japan:** Fugaku (Post-K) @ RIKEN
  - A64FX ARMv8 processor
  - ~ 1 EF (?), 40 MW, 810 M€
- **China**
  - Three prototypes:
    Sugon (accelerated)
    Tianhe (accelerated)
    Hygon (many-core based)
  - 30+ MW

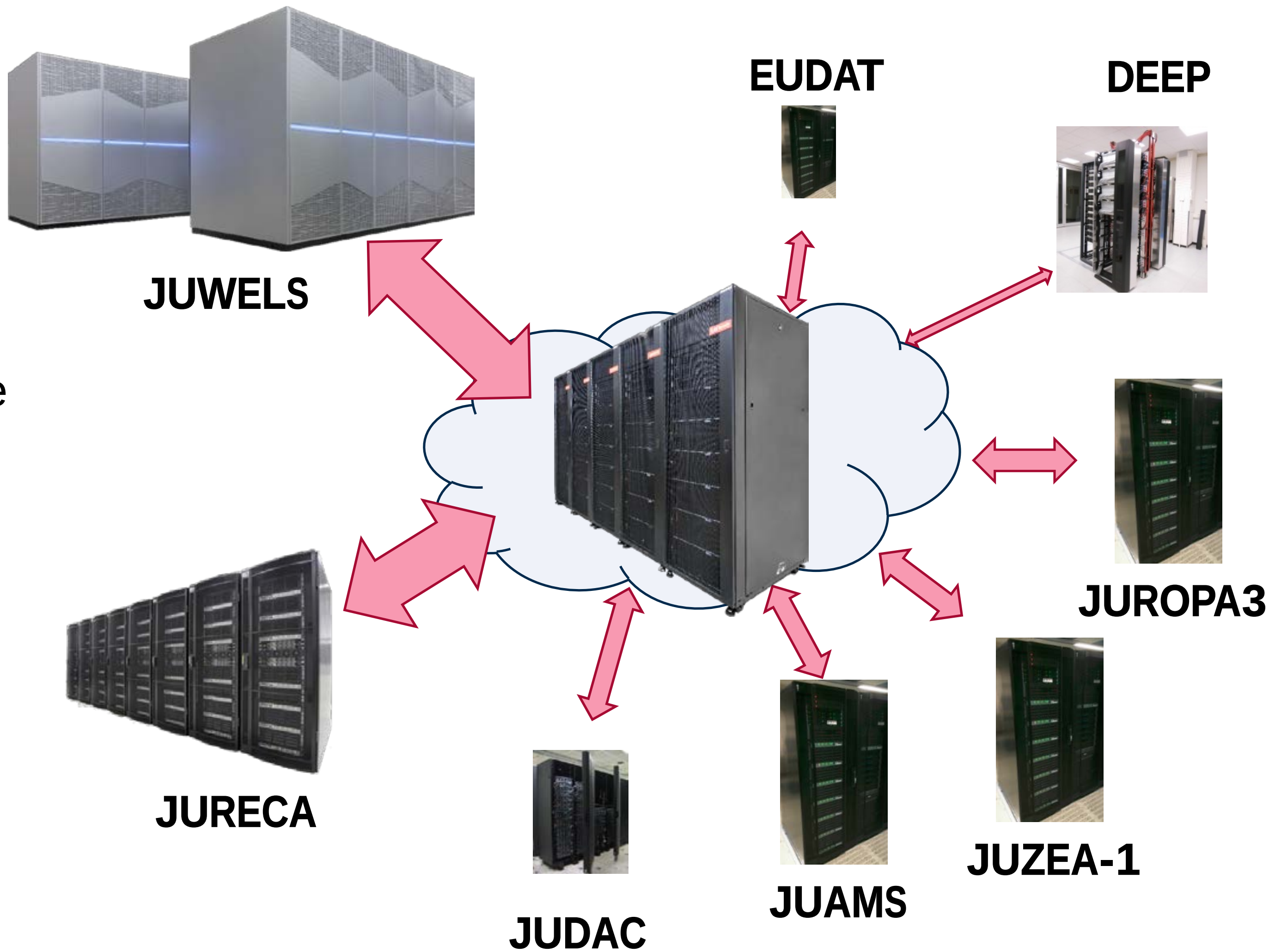
EXASCALE IN EUROPE

- Goal: Regain position among Top-3 global players
- Plan for two Exascale systems in 2022-2023 (one with European technology)
- EuroHPC Joint Undertaking
  - Petascale systems
  - Pre-exascale systems (2-3), 500 M€ total
  - Hosting entities to be announced soon
- European Processor Initiative

STORAGE INFRASTRUCTURE

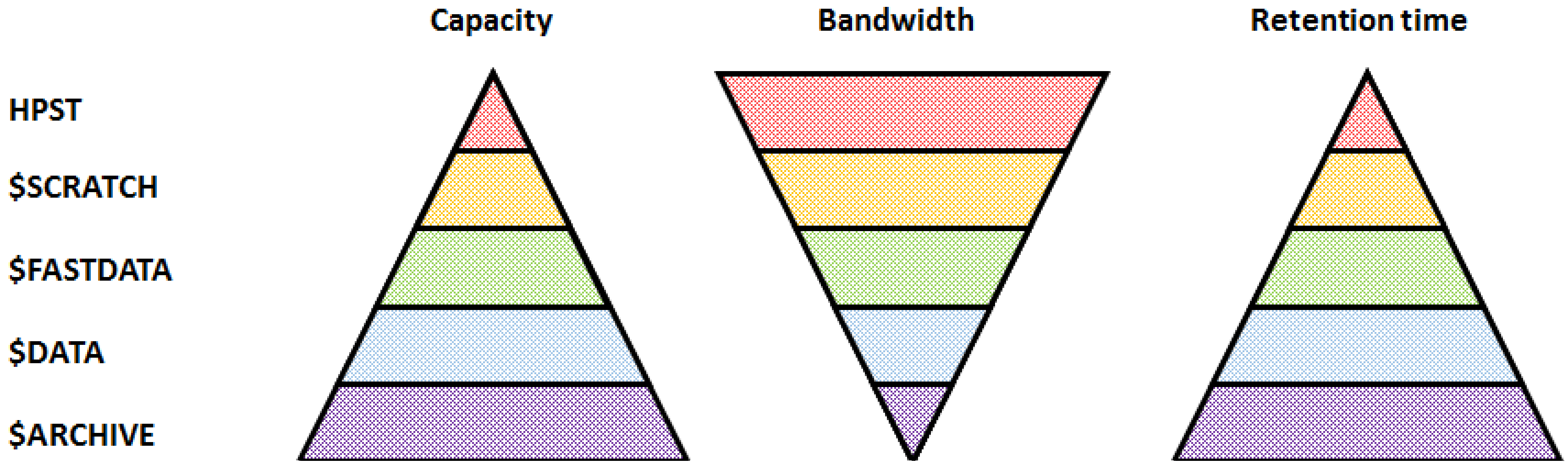
CENTRALIZED STORAGE INFRASTRUCTURE

- Spectrum Scale (GPFS)
- GPFS Native RAID (End-to-End data integrity) for some file systems
- Cross mounted on HPC systems
- Based on facility Ethernet fabric

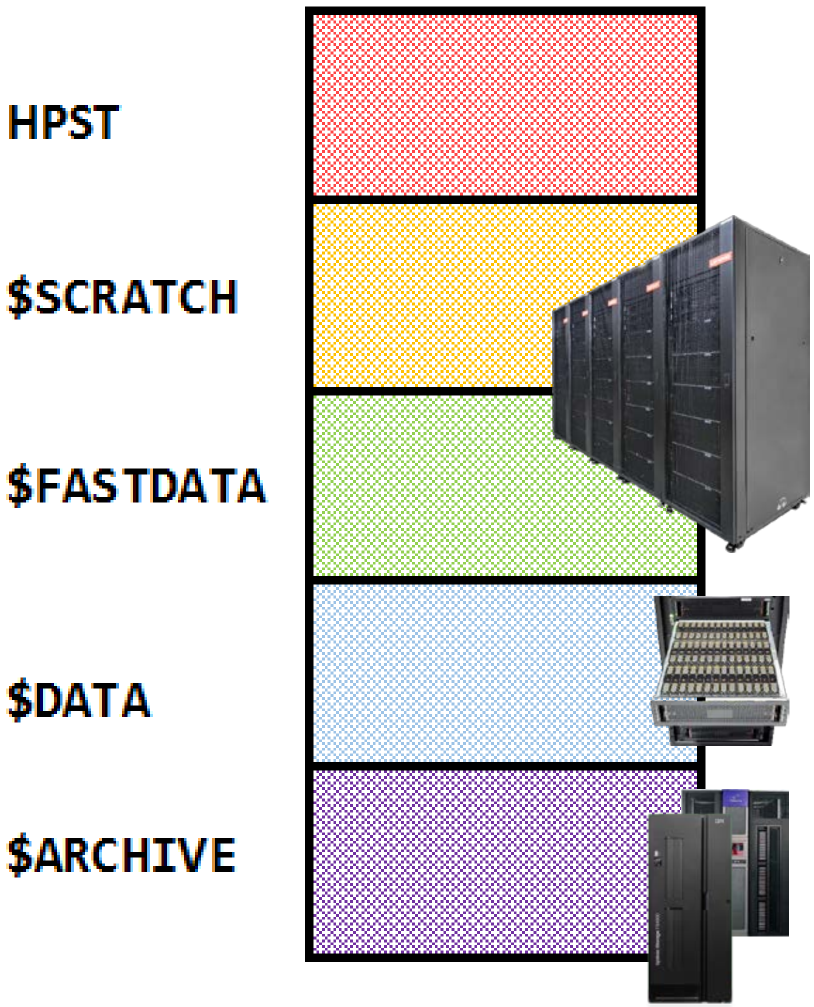


JUST: TIERED STORAGE INFRASTRUCTURE

- Different storage tiers (STs) with different optimization targets
  - Utilize most economic technology for data type and usage scenario
  - **H**igh-**P**erformance **ST**, **L**arge **C**apacity **ST**, e**X**tended **C**apacity **ST**, archival ST



JUST: MULTI-TIER STORAGE SYSTEM



Bandwidth optimized, capacity limited storage (NVM based)

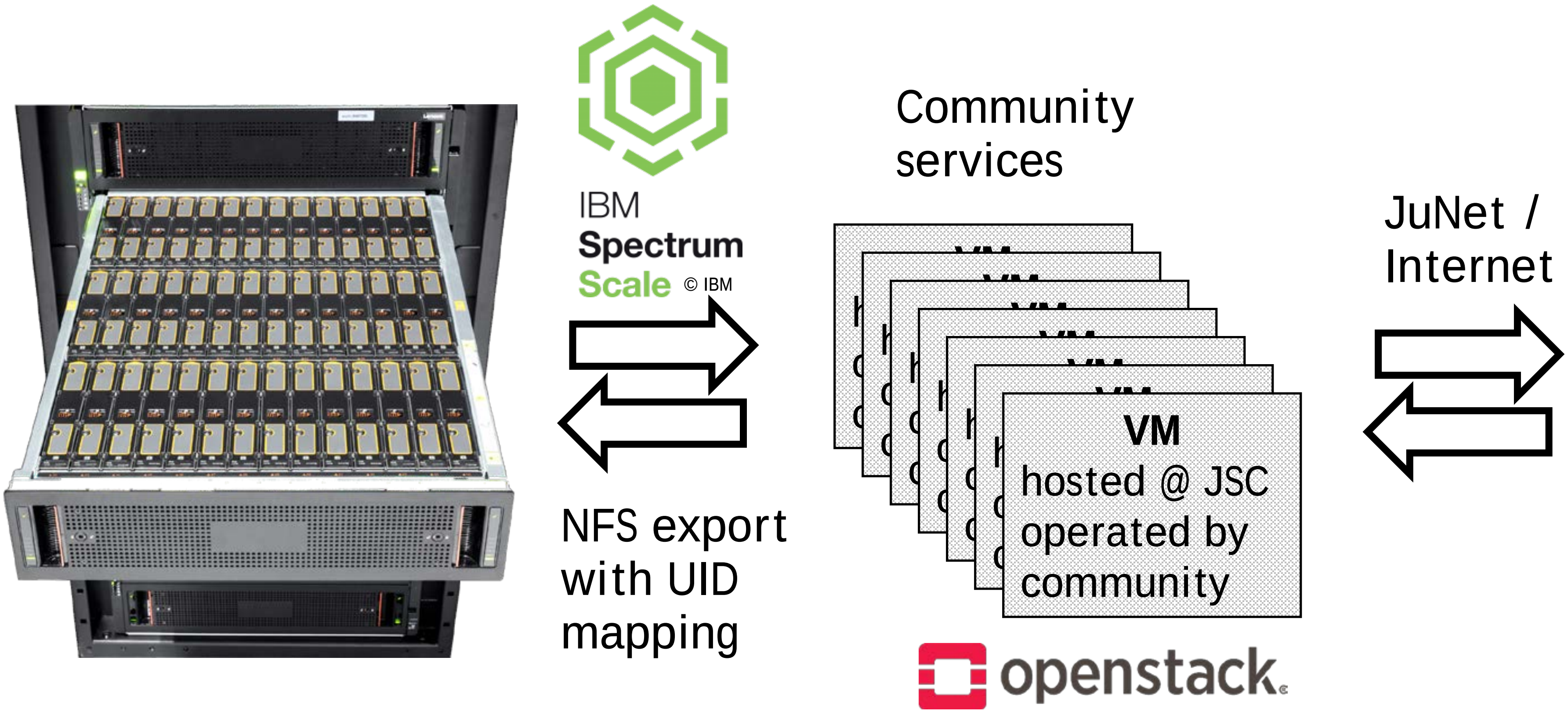
Capacity and bandwidth balancing HPC storage (temp.)

Capacity and bandwidth balancing HPC storage (pers.)

High capacity, low bandwidth storage (campaign use cases)

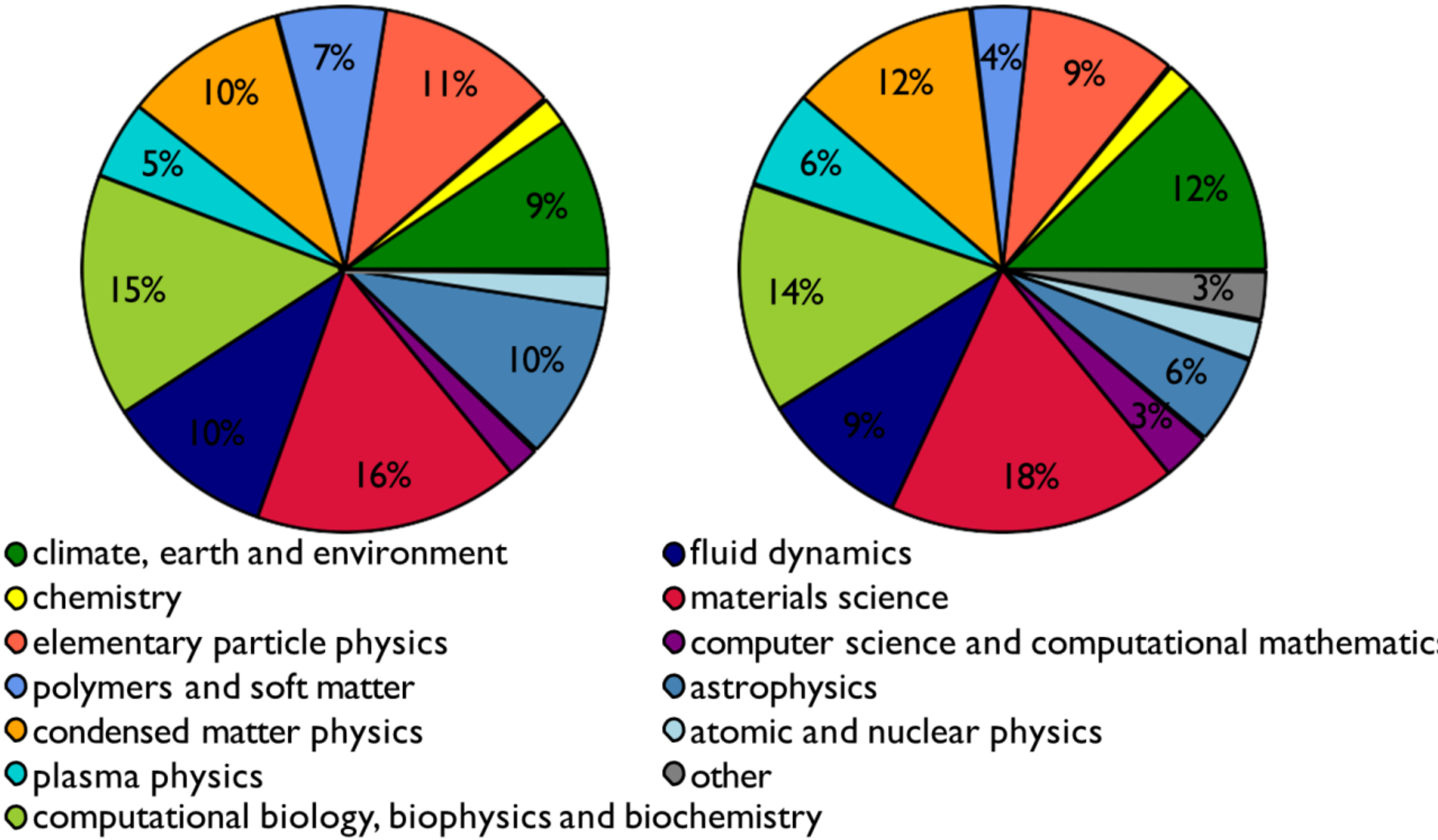
Archival storage based on high latency media

JUST: $DATA AND THE CLOUD



Limitations regarding performance and access control apply (single UID for data)

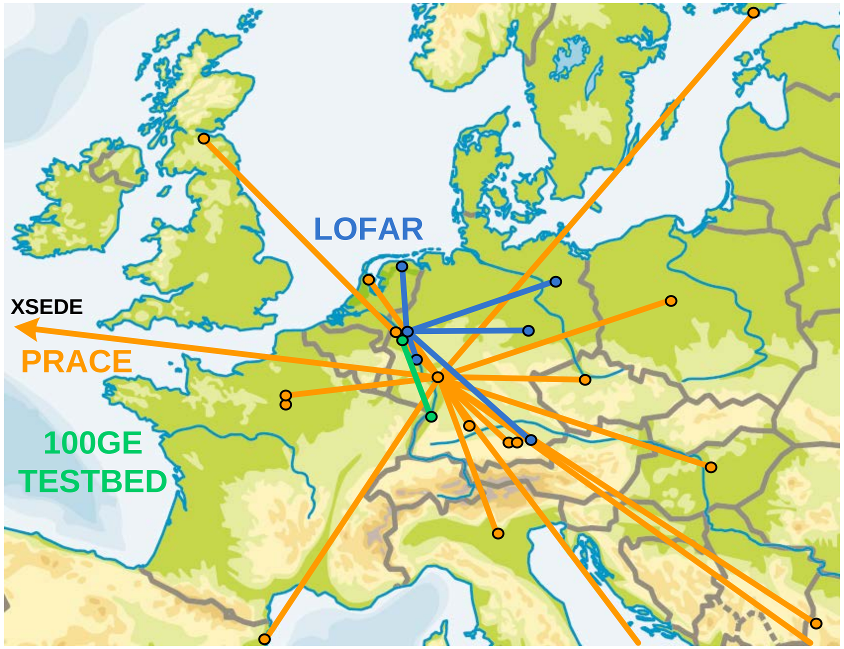
MULTI-USER SUPERCOMPUTING INFRASTRUCTURE



Allocated compute time (left) and number of projects (right) on JURECA by scientific field (Nov. 2015 - Apr. 2016).

COMMUNITY-SPECIFIC SERVICES

- Examples of tailored services for communities
  - **Radioastronomy:** 7× 10 Gb/s networking for German LOFAR antenna
    housing of correlation cluster
    long-term archive
  - **Lattice QCD:** QPACE-3 housing
  - **AMS:** Data analysis system
  - **ESM:** JUWELS partition
  - **Neuroscience:** Brain atlas,
    HBP PCPs,
    Fenix research infrastructure



JÜLICH
Forschungszentrum

AMS DATA ANALYSIS

Raw Data → DB CERN

Raw Data

ROOT Files

ROOT Files

ACQT Files

Raw Data
40 TB p.a.

ROOT Files
ca. 160 TB p.a.

ACQT Files
ca. 20 TB p.a.

AMS DATA ANALYSIS

- Dedicated cluster (70 nodes, 2.8K cores)
  - Initially as partition/share of JUROPA
- Community requirements
  - High I/O bandwidth
  - CVMFS on compute nodes & external connectivity
- Why dedicated resources?
  - Easier customization
  - **Plus:** Customizable scheduling & internal job prioritization capabilities
  - **Minus:** Burst out to large systems more complicated

HTC & SUPERCOMPUTER: CHALLENGES

- **Mentioned yesterday**
  - Network connectivity
  - (Lack of) local disk
  - FUSE
- **Additionally**
  - Workload mix & scheduling
  - Allocation policies?

INTERACTIVE LARGE-SCALE ANALYSIS WORKFLOWS FOR ESM

Use-case driven infrastructure optimization for large-scale data analysis

Supercomputing facility

- Large multi-user infrastructure
- Batch processing & utilization-focused operation mode
- Increasingly heterogeneous

ESM data analysis community

- Increasing performance requirements
- Complicated community SW stacks
- Need for interactivity & capable of adaption to resource changes

© Pangeo Team

© Dask core developer

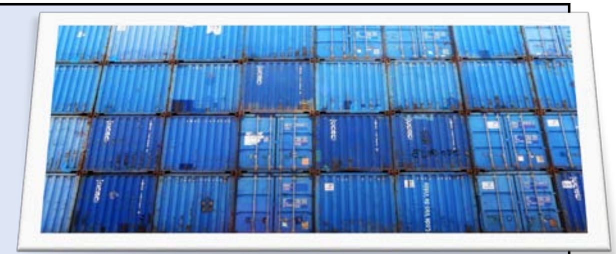
OPTIMIZATION FOR ESM DATA ANALYSIS

Accessibility

- Distributed interactive (e.g., web-based) access to supercomputing resources for analysis
  ⇨ Evaluation of Jupyter Hub service

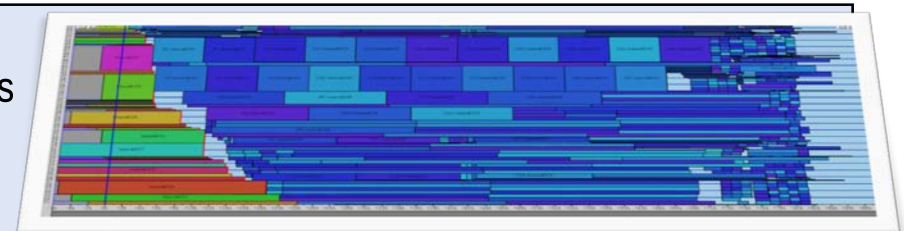
Usability

- Support for complex community analysis software stacks
  ⇨ Portable & performing implementation via containerized execution
  ⇨ Evaluation of Joint community & facility support/operations approach



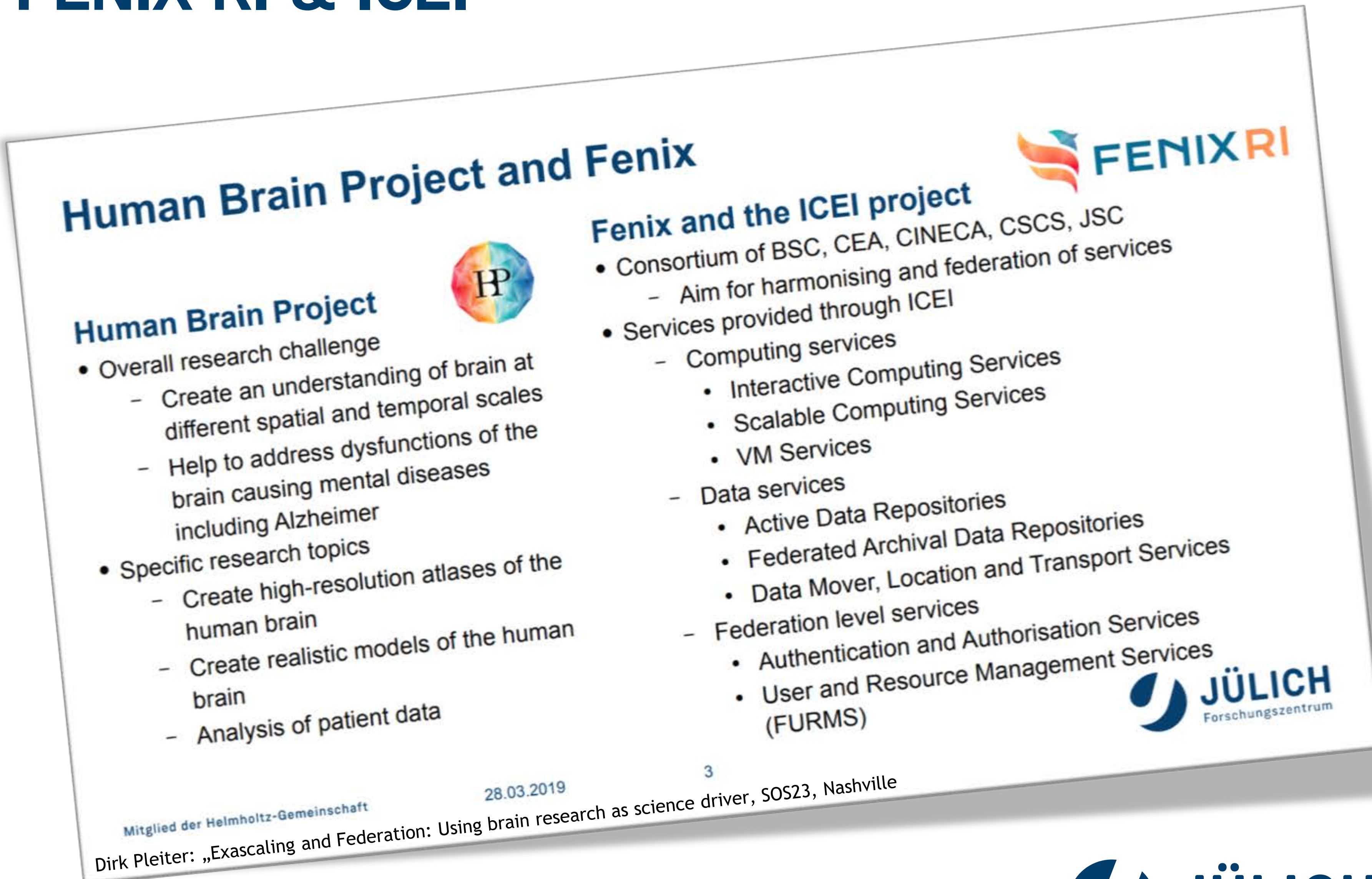
Interactivity

- Improve support for interactive ESM analysis workloads on supercomputers
  ⇨ Evaluation of scheduling & job management options (e.g., preemption)
  ⇨ Leverage resilience features of ESM data analysis software



JÜLICH Forschungszentrum

FENIX RI & ICEI



FENIX RI & ICEI

Data Sharing and Federated Data Stores: Requirements

Integration in AAI + consistent access control

- Exceeding local control domains → challenge of agreeing on common policies
- Different storage technologies do not provide compatible access control mechanisms

Storage accessible from outside the data be centre

- Need to move away from silo approach

Web-based clients

- No proprietary clients, easy to deploy by any user

Persistent references

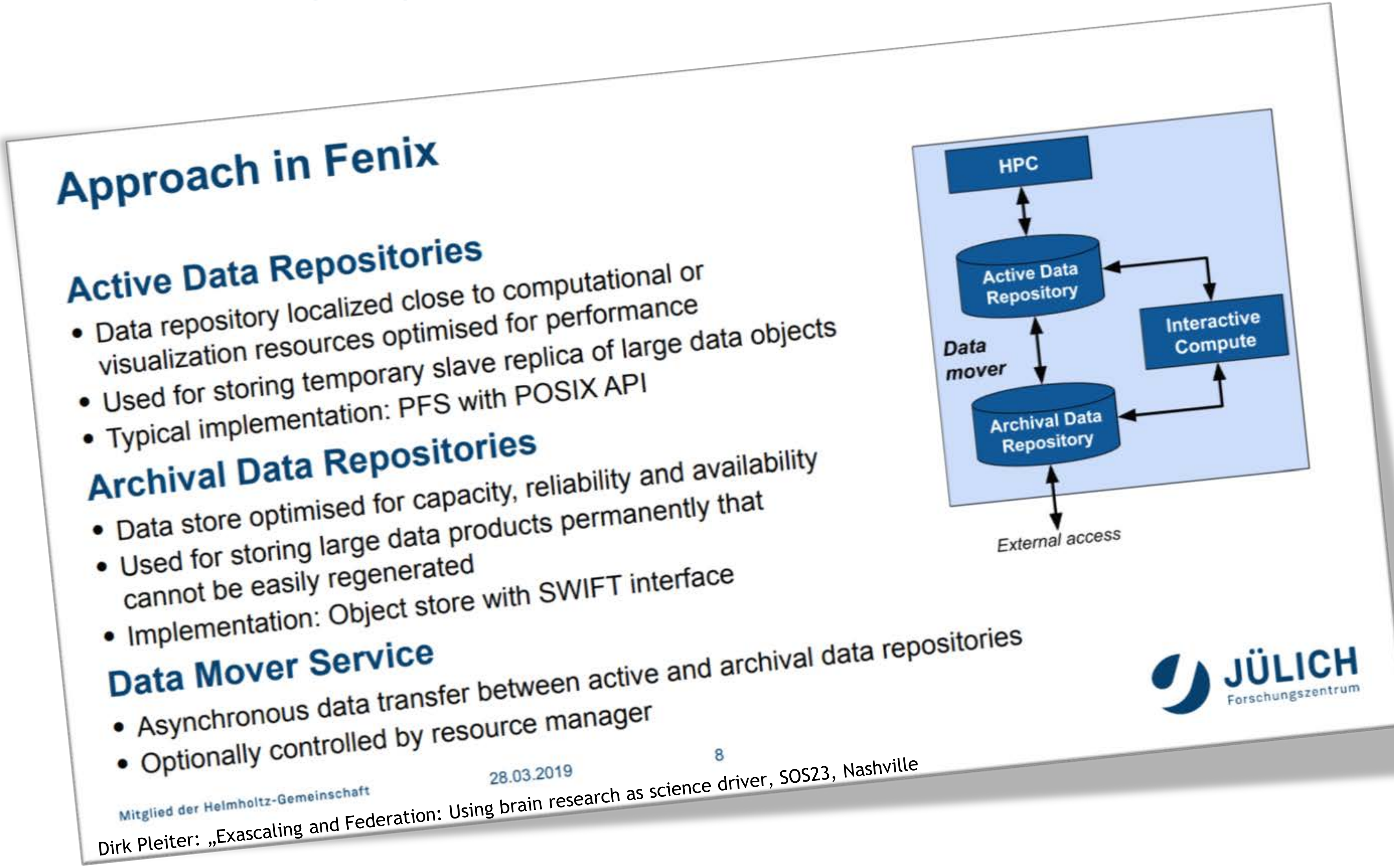
- Keep data findable

JÜLICH Forschungszentrum

Mitglied der Helmholtz-Gemeinschaft 28.03.2019 7

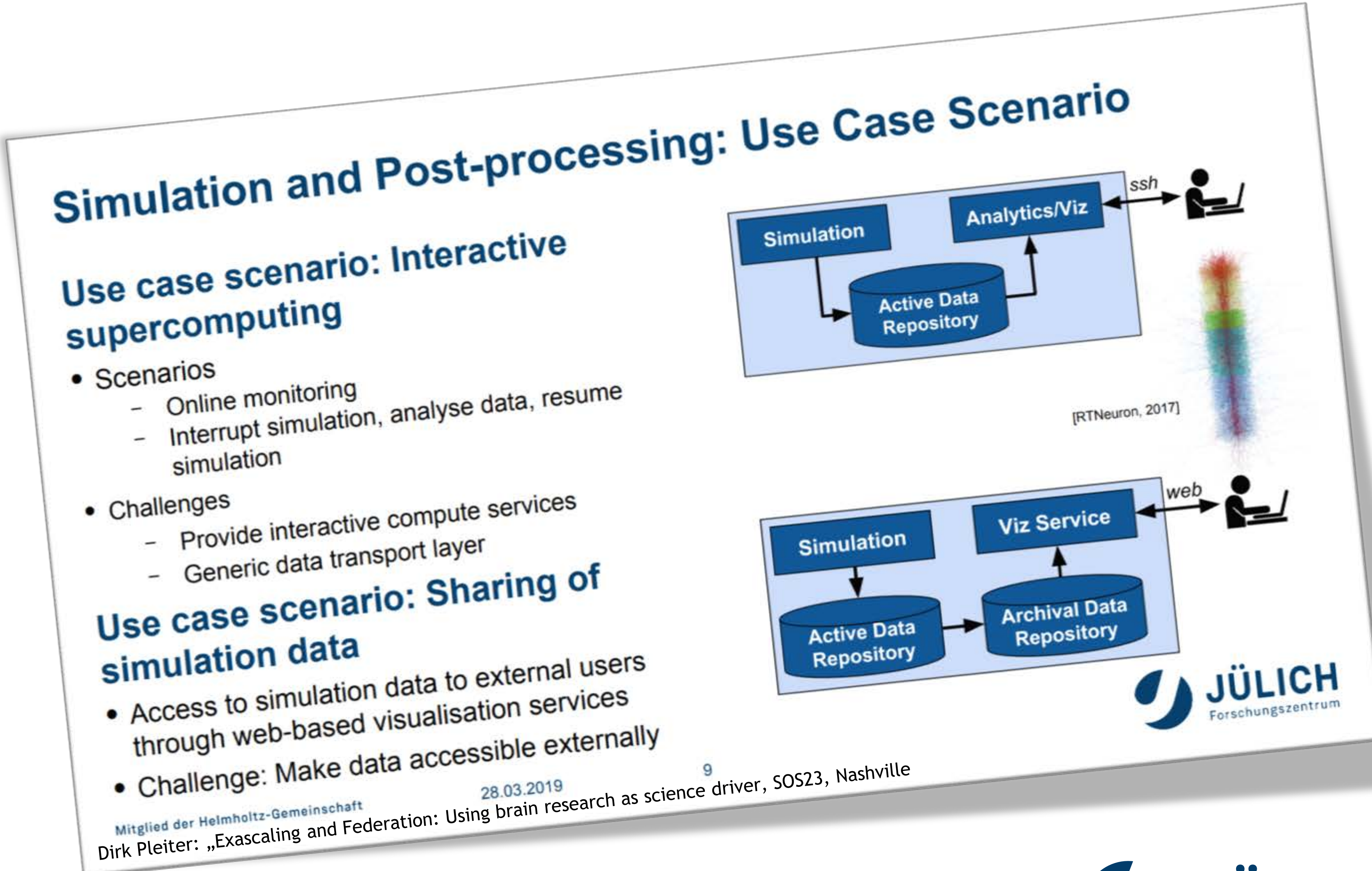
Dirk Pleiter: „Exascaling and Federation: Using brain research as science driver, SOS23, Nashville

FENIX RI & ICEI



Dirk Pleiter: „Exascaling and Federation: Using brain research as science driver, SOS23, Nashville

FENIX RI & ICEI



COMMONALITIES OF COMMUNITY REQ.

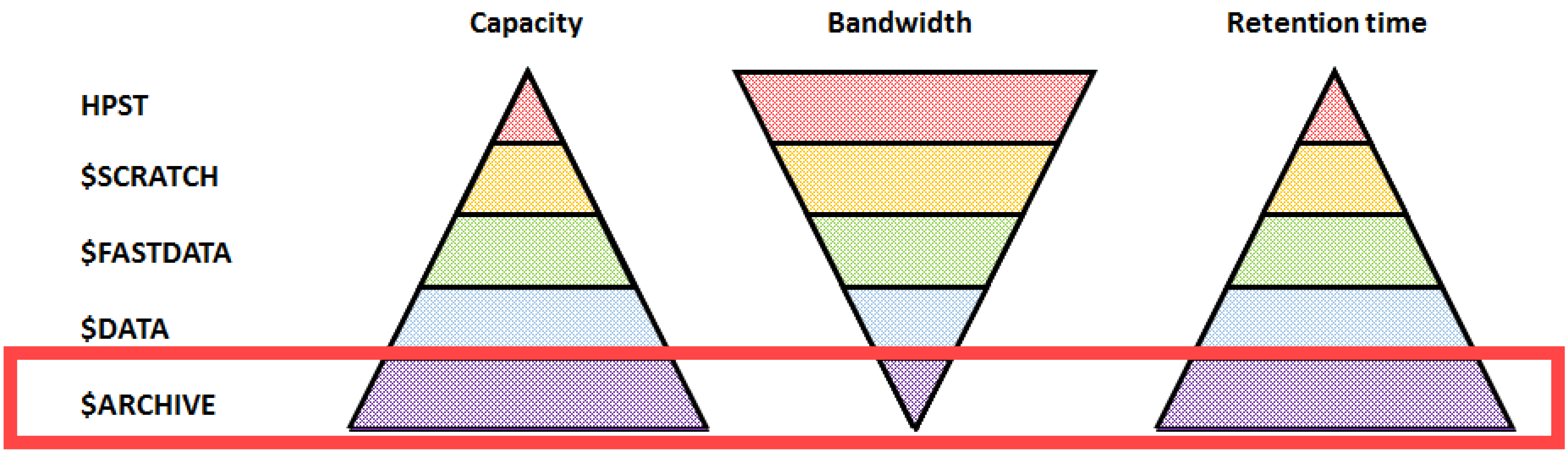
- Supercomputing as part of a web of distributed infrastructure components
  - External instruments, community data repositories, use of multiple data centers
  - Data sharing requirements, data-based community services
- Interest in support of interactive workloads to augment batch-processing
  - May lead to policy and scheduling changes, but: different requirements
- Response ⇨ new APIs & AAI mechanism (web & cloud technologies)

THANK YOU

BACKUP SLIDES

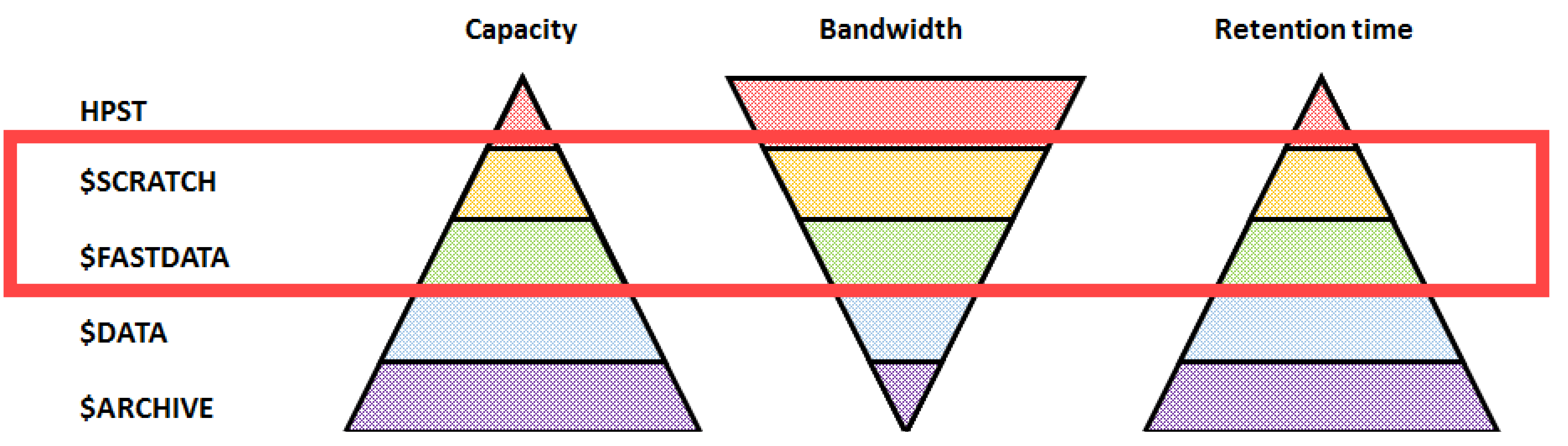
JUST: ARCHIVAL STORAGE

- **Archival storage** for cold data
  - 200+ PB of capacity on tape
  - POSIX file system (**`$ARCHIVE`**) on HPC frontend systems and data access nodes



JUST: LCS TIER

- **HPC-focused** data storage
  - ~40 PB of capacity on disk, up to 500 GB/s bandwidth
  - GPFS file systems (**$SCRATCH**, **$FASTDATA**, **$PROJECT**, **$HOME**) accessible on HPC frontend and compute nodes, data access nodes

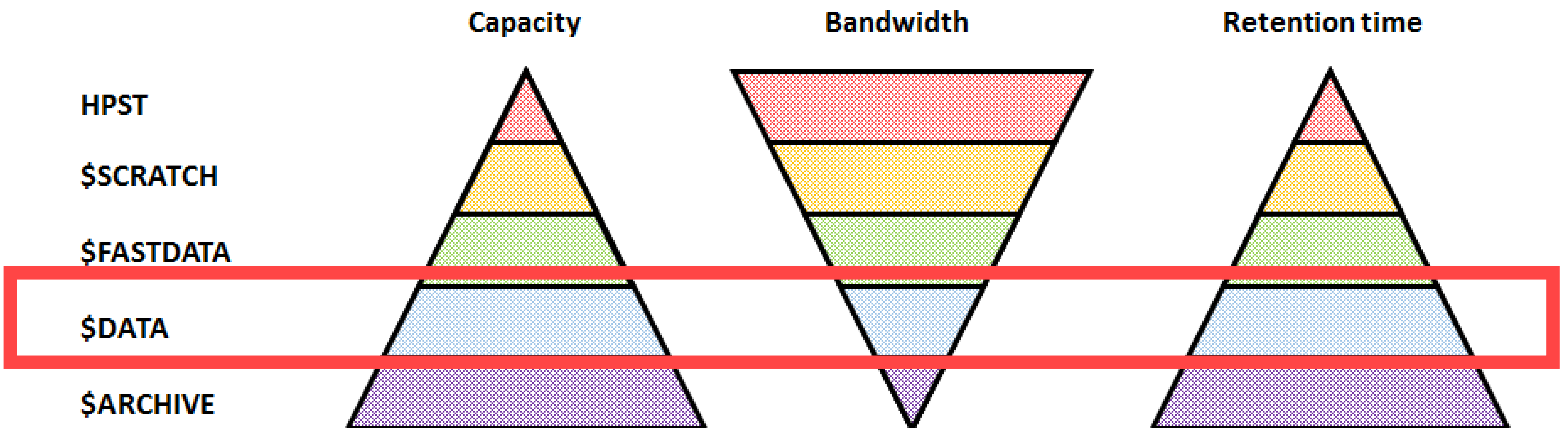


JUST: LCS TIER

- **$HOME**, **$PROJECT**: Data storage for user and compute projects
  - Low bandwidth, low capacity, with tape backup
  - Accessible on CNs, FENs, DANs
- **$SCRATCH**: Temporary storage for SC workloads
  - High bandwidth, adequate capacity, not reliable (90 d data retention time)
  - Accessible on CNs, FENs, DANs
- **$FASTDATA**: Reliable storage for HPC-processing of valuable data
  - Good bandwidth, limited capacity
  - Reliable: snapshots, but: no regular backup
  - Accessible on CNs, FENs, DANs

JUST: XCS TIER

- **Multi-purpose capacity-focused** data storage
  - **Multiple goals:** Fills gap between HPC and archival data storage (campaign use case); facilitates data sharing and federation; new interfaces (object storage)
  - Introduction in Q4 2018 (**$DATA**), service and capacity expansions planned in steps (2019+)

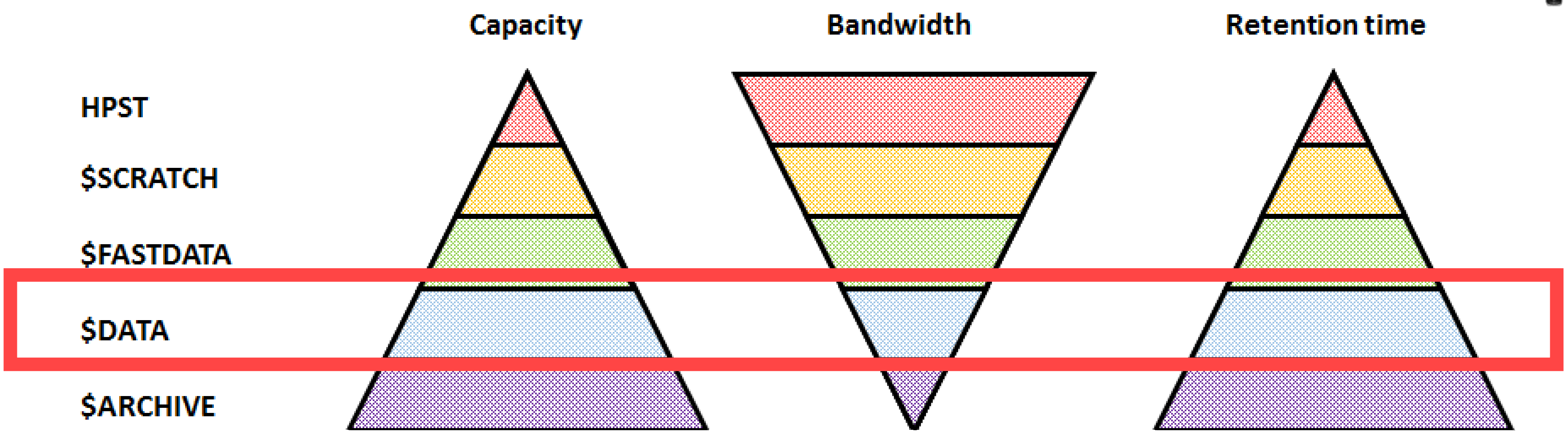


JUST: XCS TIER

- Multi-purpose storage tier
  - POSIX access via Spectrum Scale for HPC users (campaign storage use cases)
  - Allow POSIX access from selected sources outside of the SC facility perimeter
  - Object-storage access → long-term strategy
- Procurement in 2017, phased installation 2018-2021
  - Raw (**!!**) capacities of
    - Q2 2018: 40 PB (10 TB drives), Q3/Q4 2018: 12 PB
    - 2019: 12-14 PB, 2020, 2021: 14-28 PB
    - Σ = **92 - 132 PB** capacity

JUST: XCS TIER

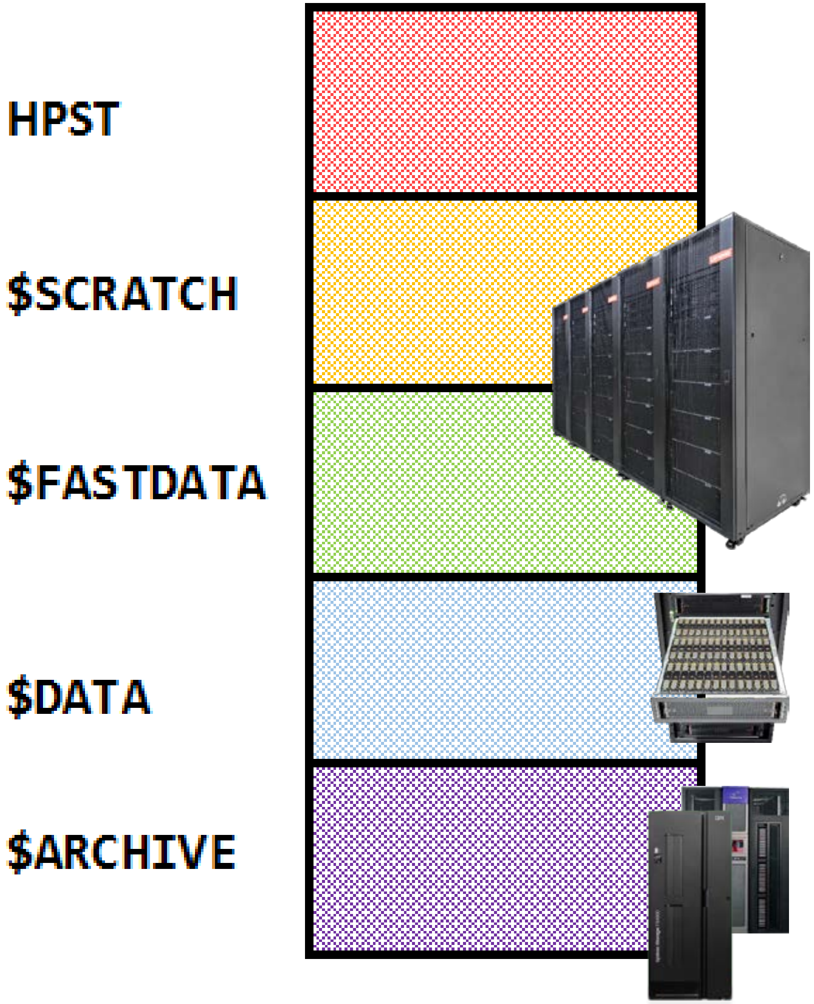
- **Multi-purpose capacity-focused** data storage
  - Initial usable capacity: 15 PB, extensions in 2018+ planned
  - **2018:** GPFS file system **$DATA** opened for data projects
  - **2019:** Access from cloud-hosted VMs to community data
  - **End of 2019:** Introduce object-storage space on XCST hardware



JUST: XCS TIER

- **$DATA:** Campaign storage file system
  - Low bandwidth (10+ GB/s), high capacity
    - High-capacity through incremental growth implies performance variability
  - Reliable: Protection against accidental data deletion via snapshots
    - Service does not include tape backup
  - Accessible on FENs, DANs
    - Currently no access on CNs offered

JUST: WHERE SHOULD MY DATA GO?



Data used now (± hour) on the SC infrastructure

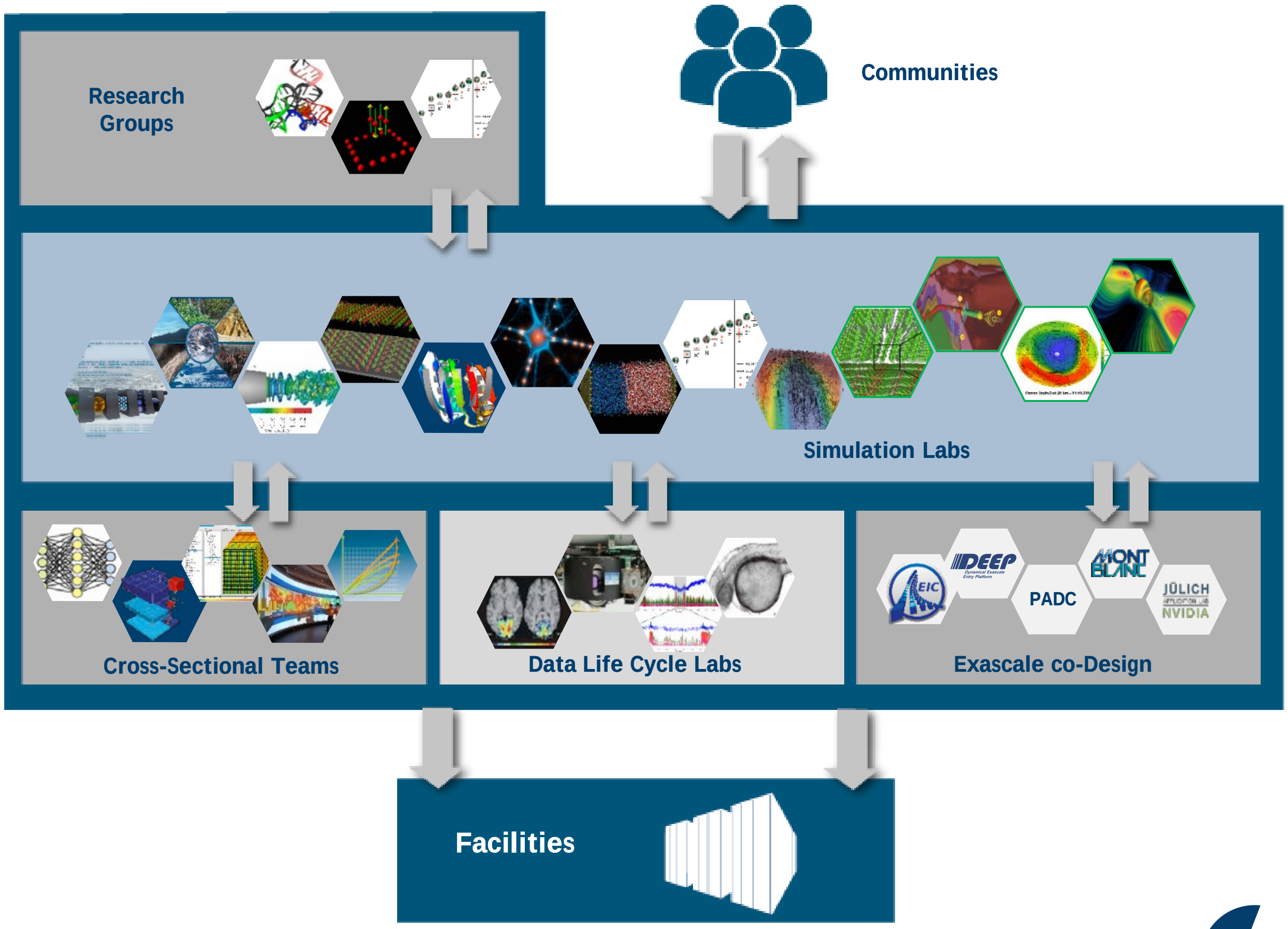
Data used now (± week) on the SC infrastructure

High-value data used now and soon again by HPC workloads

Data used (again) next month or w/o related HPC workload

Data used not within the next months

SIMLABS



Research Groups
Communities
Simulation Labs
Cross-Sectional Teams
Data Life Cycle Labs
Exascale co-Design
DEEP
PADC
MONT BLANC
JÜLICH NVIDIA
EIC
Facilities