# bwHPC:
# Hardware and Storage Architecture

**Peter Weisbrod, SCC, KIT**

Universität Heidelberg Zukunft seit 1386

Hochschule für Technik Stuttgart

Uni Freiburg

Universität Hohenheim 1818

Hochschule Esslingen University of Applied Sciences

Universität Konstanz

Universität Mannheim

Universität Stuttgart

Eberhard Karls Universität Tübingen

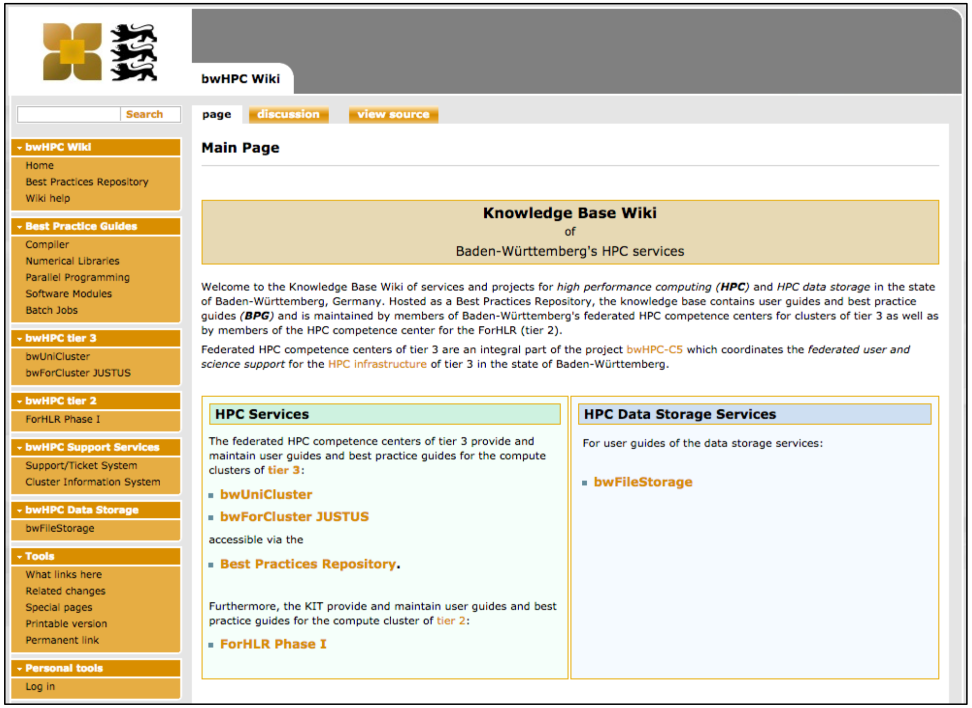
KIT Karlsruher Institut für Technologie

ulm university universität uulm

Funding: Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

# Reference: bwHPC-C5 Best Practices Repository

- Most information given by this talk can be found at http://bwhpc-c5.de/wiki:
    - Category:Hardware_and_Architecture
    - Or choose the cluster, then „Hardware and Architecture“ or „File Systems“



Main Page

**Knowledge Base Wiki**
of
Baden-Württemberg's HPC services

Welcome to the Knowledge Base Wiki of services and projects for *high performance computing* (***HPC***) and *HPC data storage* in the state of Baden-Württemberg, Germany. Hosted as a Best Practices Repository, the knowledge base contains user guides and best practice guides (***BPG***) and is maintained by members of Baden-Württemberg's federated HPC competence centers for clusters of tier 3 as well as by members of the HPC competence center for the ForHLR (tier 2).

Federated HPC competence centers of tier 3 are an integral part of the project bwHPC-C5 which coordinates the *federated user and science support* for the HPC infrastructure of tier 3 in the state of Baden-Württemberg.

**HPC Services**

The federated HPC competence centers of tier 3 provide and maintain user guides and best practice guides for the compute clusters of tier 3:

- **bwUniCluster**
- **bwForCluster JUSTUS**

accessible via the

- **Best Practices Repository.**

Furthermore, the KIT provide and maintain user guides and best practice guides for the compute cluster of tier 2:

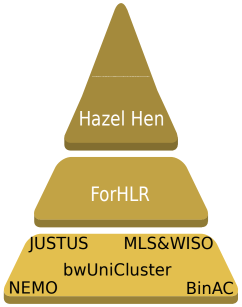
- **ForHLR Phase I**

**HPC Data Storage Services**

For user guides of the data storage services:

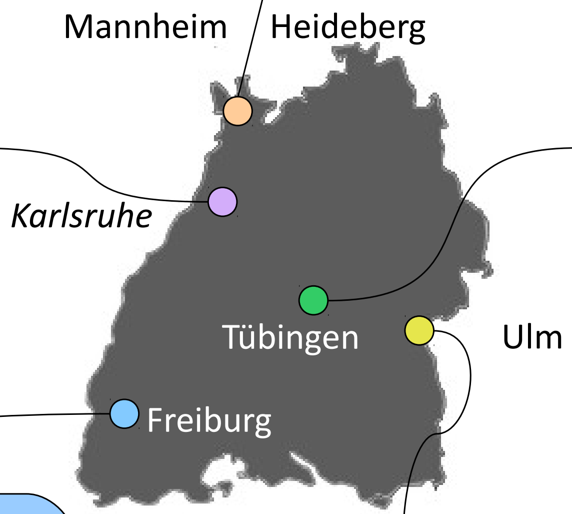
- **bwFileStorage**

# Clusters @ Tier 2+3



***bwForCluster MLS& WISO***
***(10/2015):***
Economics & Social Science,
Molecular Life Science

***bwUniCluster***
***(02/2014):***
General purpose,
Teaching & Education
***ForHLR I+II***
***(09/2014),(03/2016):***
Research, high scalability



***bwForCluster BinAC***
***(11/2016):***
Bioinformatis,
Astrophysics

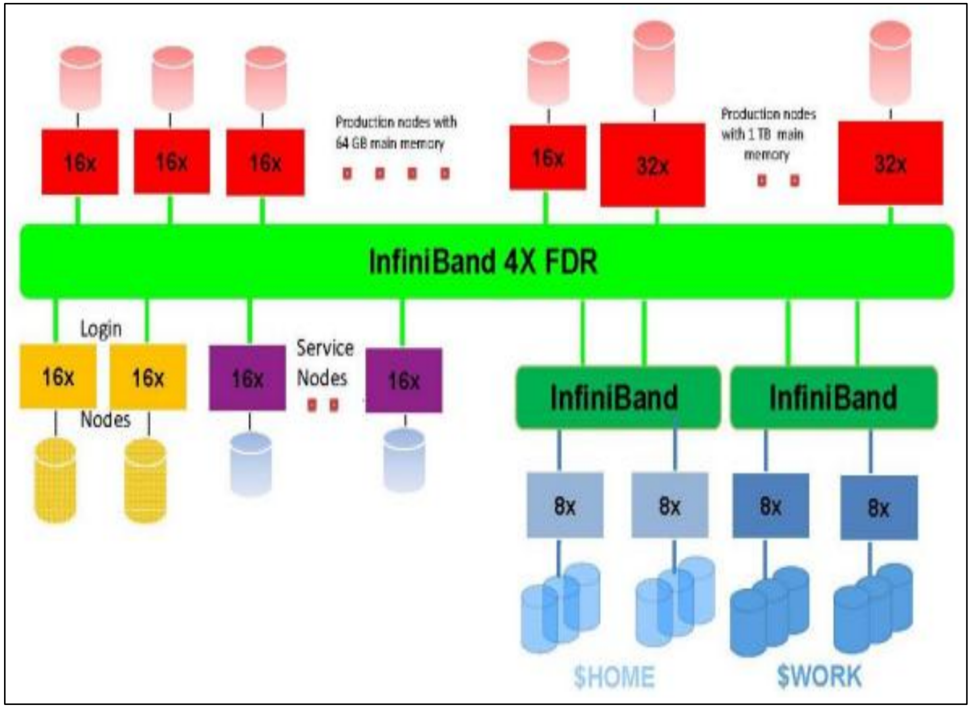
***bwForCluster NEMO***
***(09/2016):***
Neurosciences,
Micro Systems Engineering,
Elementary Particle Physics

***bwForCluster JUSTUS***
***(12/2014):***
Computational Chemistry

# System Architecture

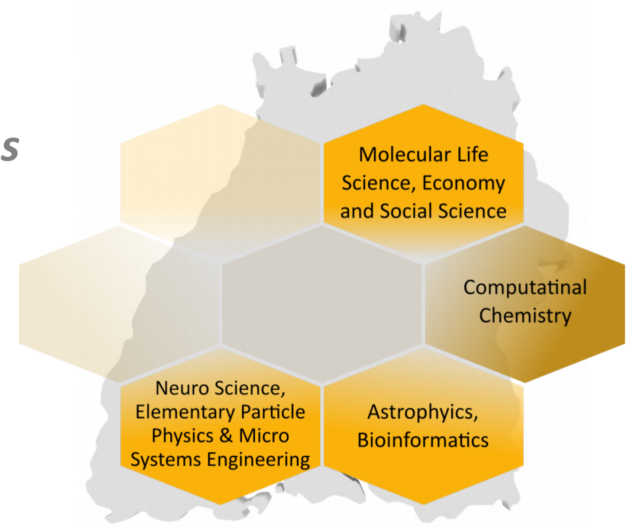
# System and Storage Architecture (bwUniCluster)

- each (compute/login) node has sixteen Intel Xeon processors, local memory, disks and network adapters, connected by fast InfiniBand 4X FDR interconnect
- Roles:
  - Login Nodes
  - Compute Nodes
  - File Server Nodes
  - Administrative Server Nodes



# bwUniCluster

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

- General purpose HPC entry level incl. education
- Universities are Shareholders
- Federated operations, multilevel fairsharing

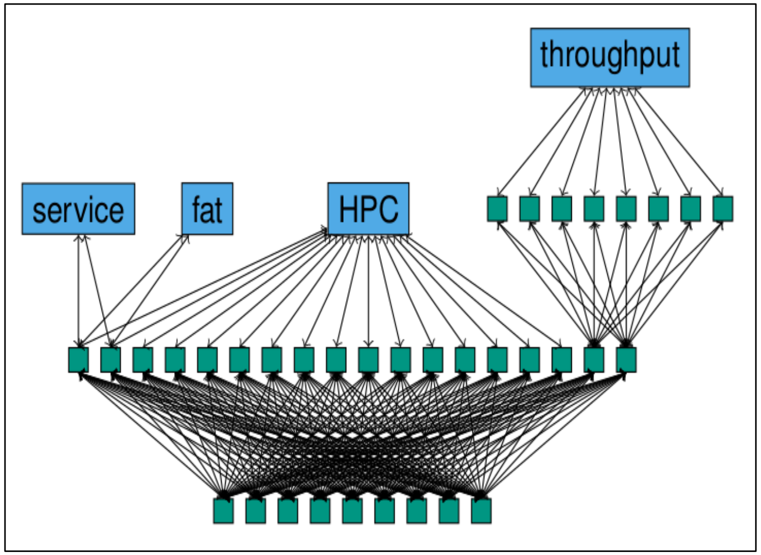
| | Thin | Fat | *In Preparation* |
|---|---|---|---|
| # nodes | 512 | 8 | *352* |
| Core/node | 16 | 32 | *28* |
| Processor | 2.6 GHz (Sandy Br.) | 2.4 GHz (Sandy Br.) | *2.0 GHz (Broadwell)* |
| Main Mem | 64 GiB | 1024 GiB | *128 GiB* |
| Local Storage | 2 TB HDD | 7 TB HDD | *480 GB SSD* |
| Interconnect | InfiniBand 4x FDR | | *InfiniBand FDR/EDR* |
| Blocking | 1:1 (50%), 1:8 (50%) | | *1:1* |
| PFS – HOME | 427 TB Lustre | | |
| PFS – Workspaces | 853 TB Lustre | | |

# System Properties (1)

- Compute node types:
  - Thin: for applications using high number of processors, distributed memory, communication over InfiniBand (MPI)
  - Fat: for shared memory applications (OpenMP or explicit multithreading)
  - Other types exist on some clusters
- Processor types:
  - (older ← → newer)
    … – Sandy Bridge – Ivy Bridge – Haswell – Broadwell – ...
- Main memory:
  - Useful to know when requesting resources (pmem, mem) during batch job submission

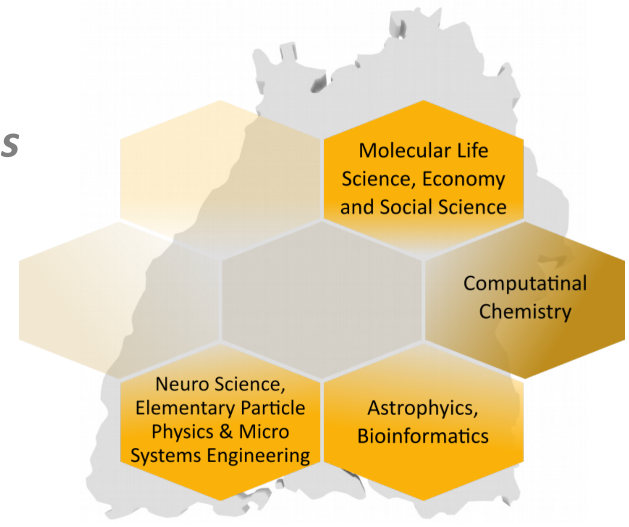
# System Properties (2)

- Local Storage:
  - Size and read/write performance interesting when using local file system ($TMP / $TMPDIR)
- InfiniBand:
  - (older ← → newer, higher speed, lower latency)
    … – QDR – FDR – EDR – …
  - Or Omni-Path instead
- Blocking:
  - Ratio of uplink and downlink bandwidth
  - Non-blocking if equal
  - Example bwUnicluster:
    both blocking and „fat tree“ area



# bwUniCluster

*Federated HPC tier 3 resources*



Molecular Life Science, Economy and Social Science

Computatinal Chemistry

Neuro Science, Elementary Particle Physics & Micro Systems Engineering

Astrophyics, Bioinformatics

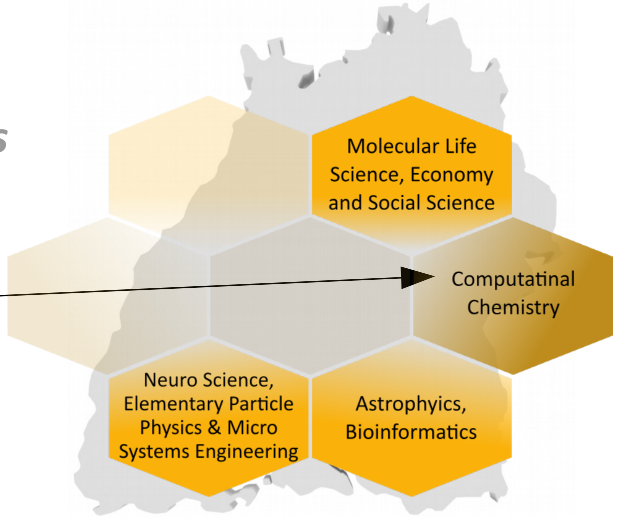
Selected characteristics:

- General purpose HPC entry level incl. education
- Universities are Shareholders
- Federated operations, multilevel fairsharing

| | Thin | Fat | *In Preparation* |
|---|---|---|---|
| # nodes | 512 | 8 | *352* |
| Core/node | 16 | 32 | *28* |
| Processor | 2.6 GHz (Sandy Br.) | 2.4 GHz (Sandy Br.) | *2.0 GHz (Broadwell)* |
| Main Mem | 64 GiB | 1024 GiB | *128 GiB* |
| Local Storage | 2 TB HDD | 7 TB HDD | *480 GB SSD* |
| Interconnect | InfiniBand 4x FDR | | *InfiniBand FDR/EDR* |
| Blocking | 1:1 (50%), 1:8 (50%) | | *1:1* |
| PFS – HOME | 427 TB Lustre | | |
| PFS – Workspaces | 853 TB Lustre | | |

# bwForCluster JUSTUS

*Federated HPC tier 3 resources*



Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

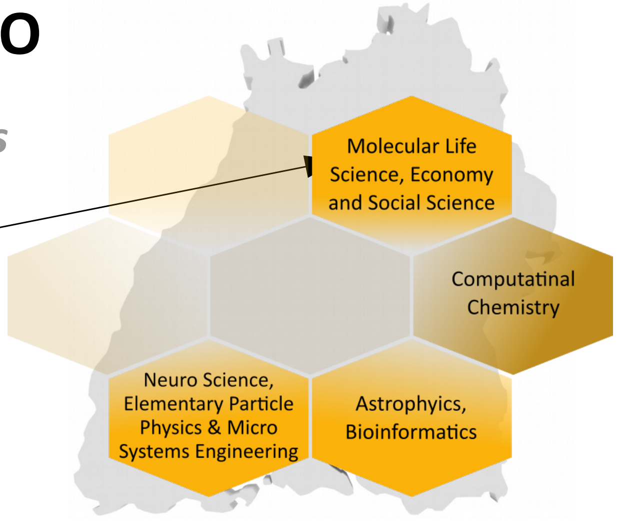
Selected characteristics:

- Dedicated to **computational chemistry**
  - High I/O, large MEM jobs
- User and software support by *bwHPC competence center*

| | Diskless | SSD | Big SSD | Large Mem SSD | Visual |
|---|---|---|---|---|---|
| # nodes | 202 | 204 | 22 | 16 | 2 |
| Core/node | 16 | 16 | 16 | 16 | 16 |
| Processor | 2,4 GHz (Xeon E5-2630v3, Haswell) | | | | |
| Main Mem | 128 GiB | | 256 GiB | 512 GiB | 512 GiB |
| Local Storage | - | 1 TB SSD | 2 TB SSD | | 4 TB HDD |
| Interconnect | InfiniBand QDR | | | | |
| Blocking | 1:8 | | | | |
| HOME | 200 TB NFS | | | | |
| PFS – Workspaces | 200 TB Lustre | | | | |
| Block storage | 480 TB (local mount via RDMA) | | | | |
| Special feature | | | | | NVIDIA K6000 |

# bwForCluster MLS&WISO

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophysics, Bioinformatics

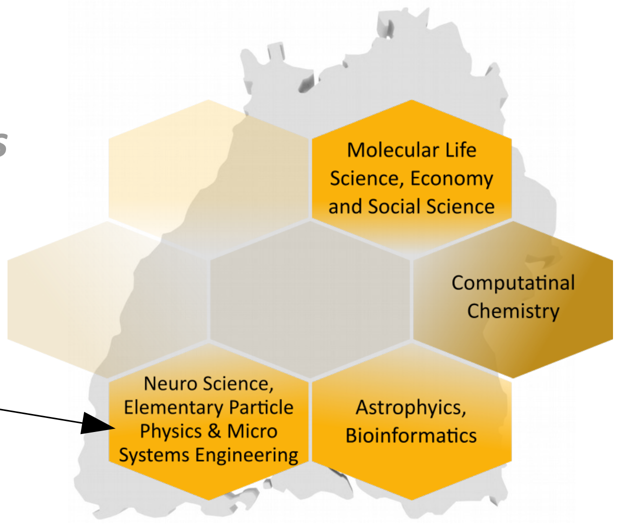
Selected characteristics:

- Dedicated to **molecular life science, economics and social science + cluster for method development**
- User and software support by *bwHPC competence center*

| | Standard | Best | Coprocessor (GPU) | Coprocessor (MIC) | Fat | Fat (Ivy Bridge) |
|---|---|---|---|---|---|---|
| **Node Feature** | standard | best | gpu | mic | fat | fat-ivy |
| **Quantity** | 476 | 148 | 18 | 12 | 8 | 4 |
| **Processors** | 2 x Intel Xeon E5-2630v3 (Haswell) | 2 x Intel Xeon E5-2640v3 (Haswell) | 2 x Intel Xeon E5-2630v3 (Haswell) | 2 x Intel Xeon E5-2630v3 (Haswell) | 4 x Intel Xeon E5-4620v3 (Haswell) | 4 x Intel Xeon E4-4020v2 (Ivy Bridge) |
| **Processor Frequency (GHz)** | 2.4 | 2.6 | 2.4 | 2.4 | 2.0 | 2.6 |
| **Number of Cores** | 16 | 16 | 16 | 16 | 40 | 32 |
| **Working Memory (GB)** | 64 | 128 | 64 | 64 | 1536 | 1024 |
| **Local Disk (GB)** | 128 (SSD) | 128 (SSD) | 128 (SSD) | 128 (SSD) | 9000 (SATA) | 128 (SSD) |
| **Interconnect** | QDR | FDR | FDR | FDR | FDR | FDR |
| **Coprocessors** | – | – | 1 x Nvidia Tesla K80 | 2 x Intel Xeon Phi 5110P | – | – |

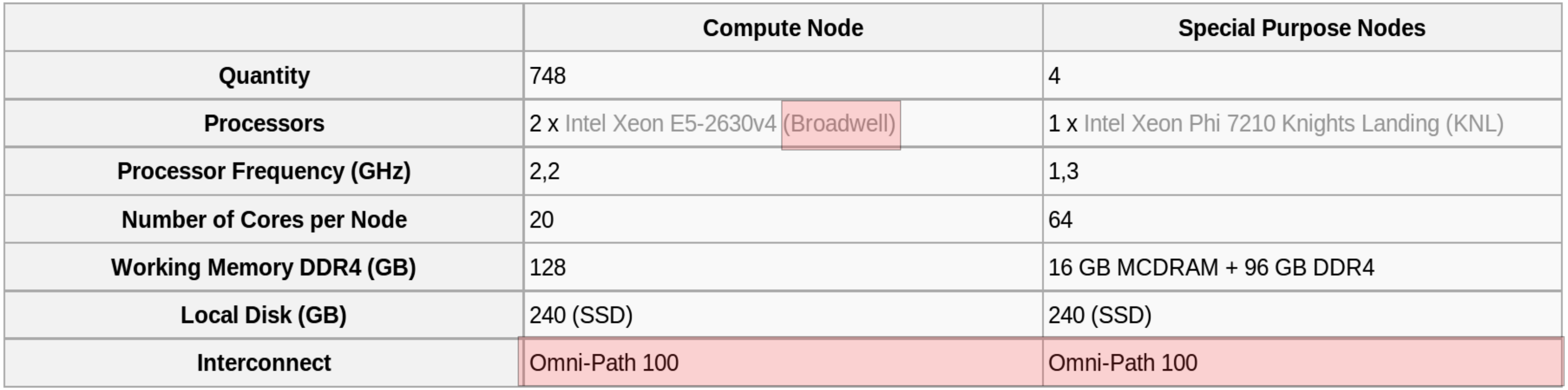
# bwForCluster NEMO

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

- Dedicated to **neuro science, elementary particle physics, micro systems engineering**
  - Virtual machine images deployable
- User and software support by *bwHPC competence center*



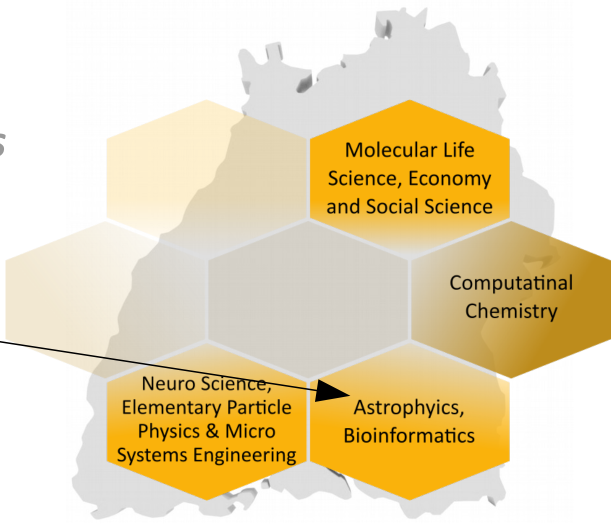
| | Compute Node | Special Purpose Nodes |
|---|---|---|
| **Quantity** | 748 | 4 |
| **Processors** | 2 x Intel Xeon E5-2630v4 (Broadwell) | 1 x Intel Xeon Phi 7210 Knights Landing (KNL) |
| **Processor Frequency (GHz)** | 2,2 | 1,3 |
| **Number of Cores per Node** | 20 | 64 |
| **Working Memory DDR4 (GB)** | 128 | 16 GB MCDRAM + 96 GB DDR4 |
| **Local Disk (GB)** | 240 (SSD) | 240 (SSD) |
| **Interconnect** | Omni-Path 100 | Omni-Path 100 |

# bwForCluster BinAC

*Federated HPC tier 3 resources*

<u>Selected characteristics:</u>

- Dedicated to **astrophysics, bioinformatics**
  - Dual GPU systems
- User and software support by *bwHPC competence center*

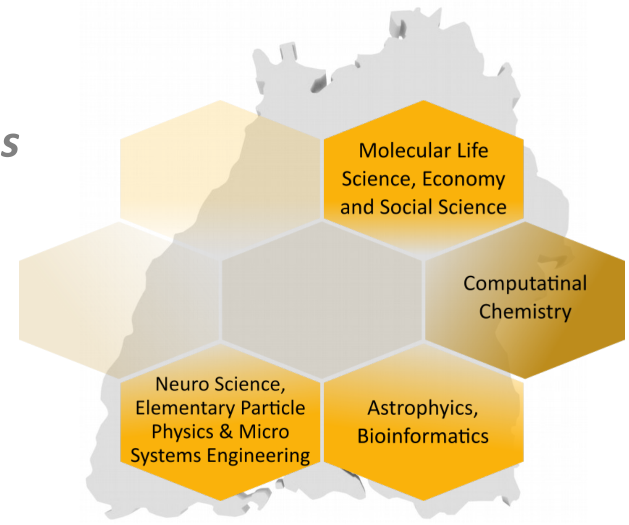


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophycis, Bioinformatics

| | Standard | Fat | GPU |
|---|---|---|---|
| **Quantity** | 236 | 4 | 60 |
| **Processors** | 2 x Intel Xeon E5-2630v4 (Broadwell) | 4 x Intel Xeon E5-4620v3 (Haswell) | 2 x Intel Xeon E5-2630v4 (Broadwell) |
| **Processor Frequency (GHz)** | 2.4 | 2.0 | 2.4 |
| **Number of Cores** | 28 | 40 | 28 |
| **Working Memory (GB)** | 128 | 1024 | 128 |
| **Local Disk (GB)** | 256 (SSD) | 256 (SSD) | 256 (SSD) |
| **Interconnect** | FDR | FDR | FDR |
| **Coprocessors** | – | – | 2 x Nvidia Tesla K80 |

# ForHLR I

*Federated HPC tier 2 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophysics, Bioinformatics

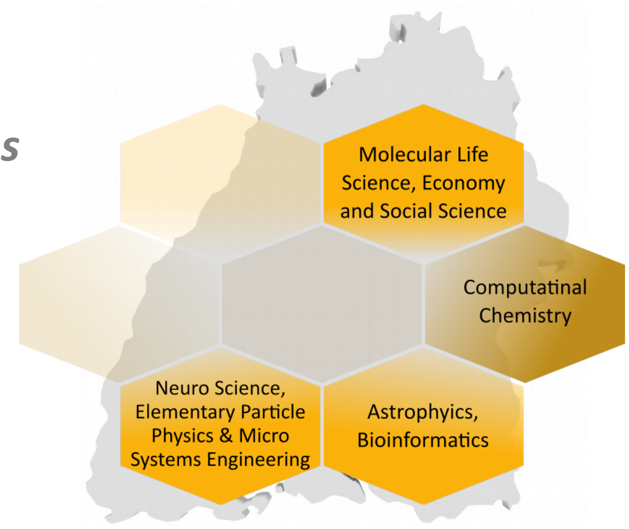
Selected characteristics:

- Next level for advanced HPC users
- Research, high scalability

| | Thin | Fat |
|---|---|---|
| # nodes | 512 | 16 |
| Core/node | 20 | 32 |
| Processor | 2.5 GHz (Sandy Br.) | 2.6 GHz (Sandy Br.) |
| Main Mem | 64 GiB | 512 GiB |
| Local Storage | 2 TB HDD | 8 TB HDD |
| Interconnect | InfiniBand 4x FDR | |
| Blocking | Non-blocking | |
| PFS – HOME | 427 TB Lustre | |
| PFS – Workspaces | PROJECT 427 TB Lustre, WORK/workspace 853 TB Lustre | |

# ForHLR II

*Federated HPC tier 2 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophysics, Bioinformatics

Selected characteristics:

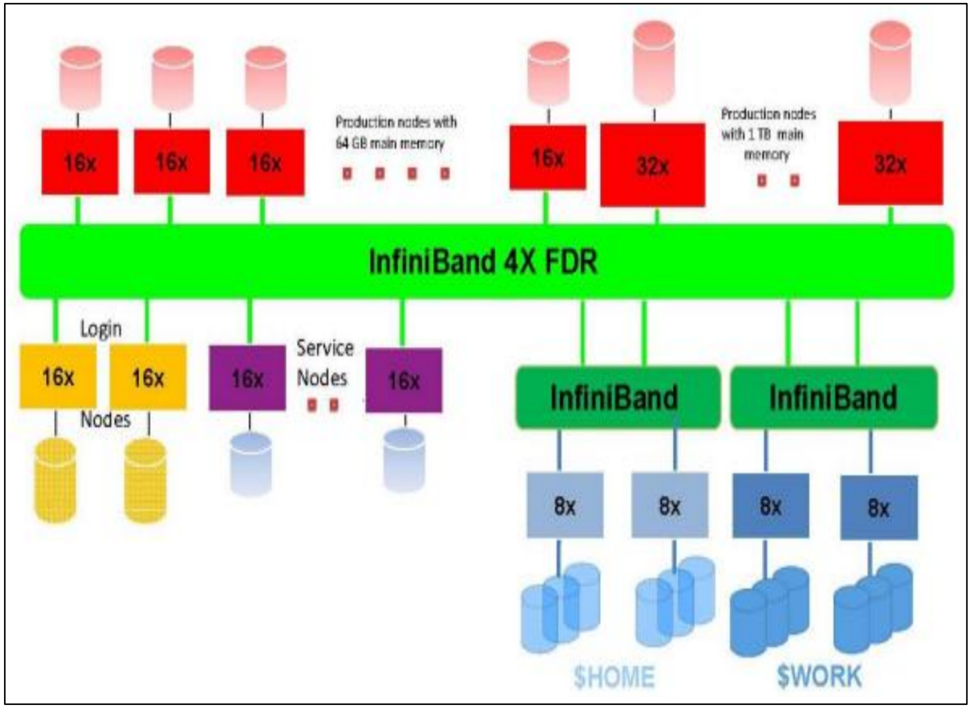
- Next level for advanced HPC users
- Research, high scalability

| | Thin | Fat |
|---|---|---|
| # nodes | 1152 | 21 |
| Core/node | 20 | 48 |
| Processor | 2.6 GHz (Haswell) | 2.1 GHz (Haswell) |
| Main Mem | 64 GiB | 1024 GiB |
| Local Storage | 480 GB SSD | 3840 GB SSD |
| Interconnect | InfiniBand 4x EDR | |
| Blocking | Non-blocking | |
| Graphic cards | | 4 NVIDIA GeForce GTX980 Ti |
| PFS – HOME | 427 TB Lustre | |
| PFS – Workspaces | PROJECT 610 TB Lustre, WORK 1220 TB Lustre, workspace 3050 TB Lustre | |

# Storage Architecture

# System and Storage Architecture (bwUniCluster)

- File Systems:
  - Local ($TMP or $TMPDIR): each node has its own file system
  - Global ($HOME, $PROJECT, $WORK, workspaces): all nodes access the same file system; located in parallel file system

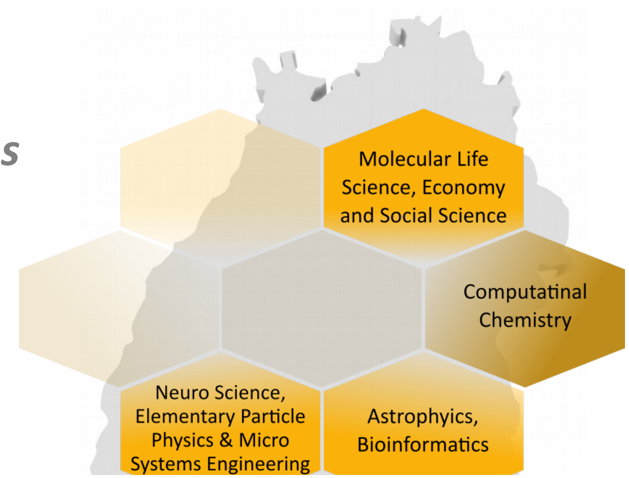


# File Systems

- All Clusters:
  - $TMP or $TMPDIR: local, files are removed at end of batch job, no backup
  - $HOME: global, permanent, backup on most clusters, quota,
    same home directories on ForHLR I+II, bwUniCluster
  - workspaces: global, entire workspace expires after fixed period, no backup, no quota, higher throughput
    HowTo: http://www.bwhpc-c5.de/wiki/index.php/Workspace
- ForHLR I+II, bwUniCluster:
  - $WORK: global, no backup, no quota, higher throughput,
    file lifetime 28 days (1 week guaranteed)
- ForHLR I+II:
  - $PROJECT: global, permanent, backup, quota
    use $PROJECT instead because $HOME quota for project group very small

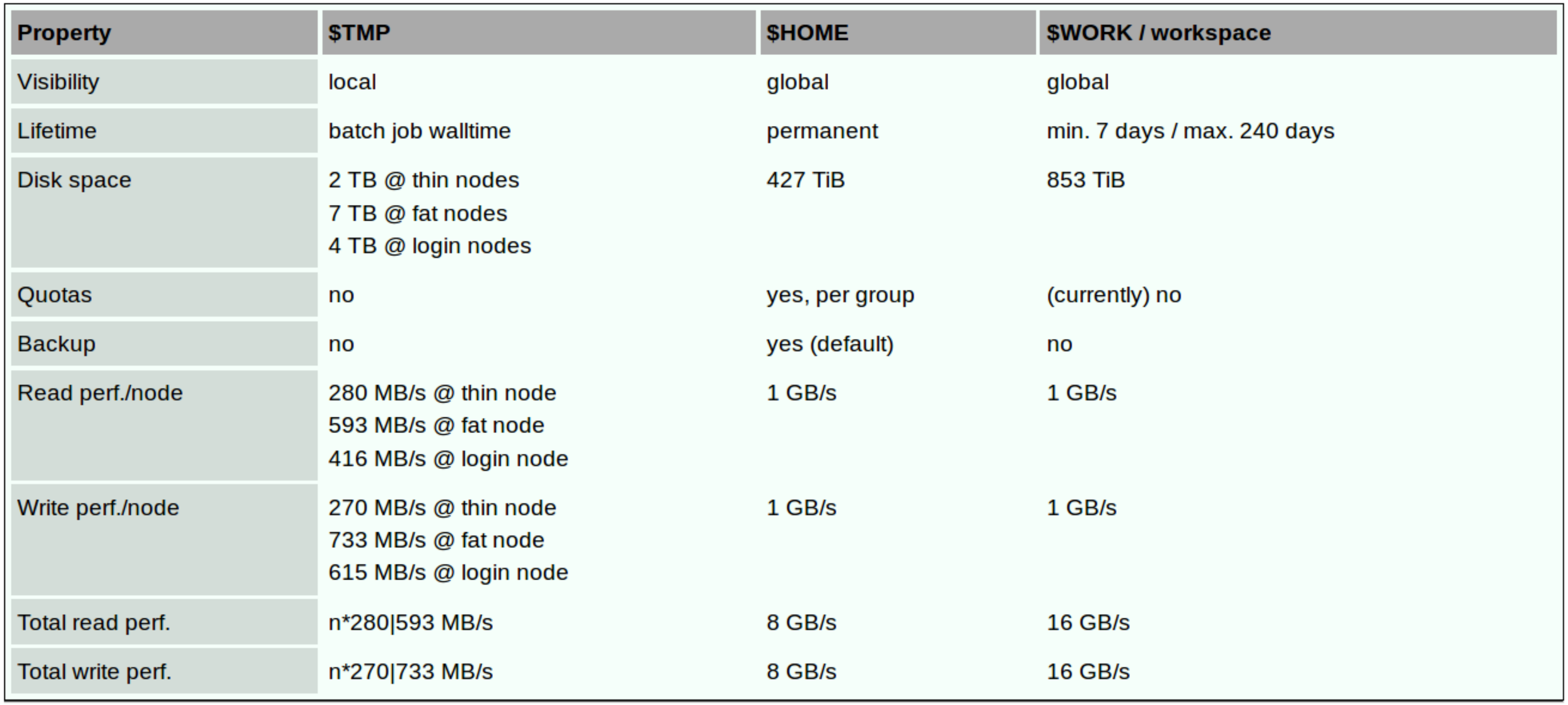
# bwUniCluster

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

## Selected characteristics:

- General purpose HPC entry level incl. education
- Universities are Shareholders
- Federated operations, multilevel fairsharing



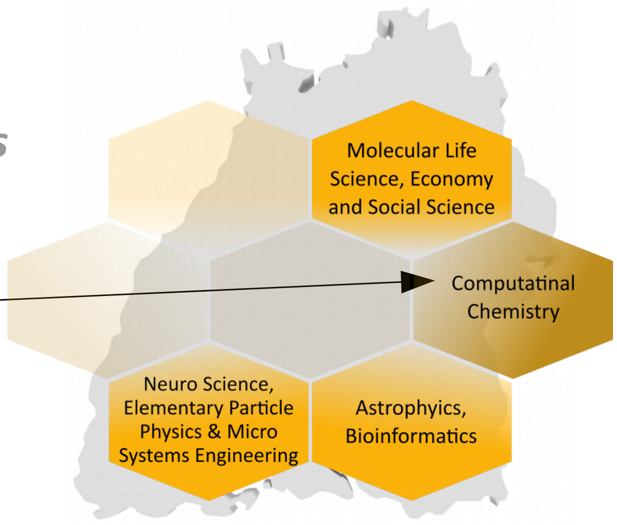
| Property | $TMP | $HOME | $WORK / workspace |
|---|---|---|---|
| Visibility | local | global | global |
| Lifetime | batch job walltime | permanent | min. 7 days / max. 240 days |
| Disk space | 2 TB @ thin nodes<br>7 TB @ fat nodes<br>4 TB @ login nodes | 427 TiB | 853 TiB |
| Quotas | no | yes, per group | (currently) no |
| Backup | no | yes (default) | no |
| Read perf./node | 280 MB/s @ thin node<br>593 MB/s @ fat node<br>416 MB/s @ login node | 1 GB/s | 1 GB/s |
| Write perf./node | 270 MB/s @ thin node<br>733 MB/s @ fat node<br>615 MB/s @ login node | 1 GB/s | 1 GB/s |
| Total read perf. | n*280\|593 MB/s | 8 GB/s | 16 GB/s |
| Total write perf. | n*270\|733 MB/s | 8 GB/s | 16 GB/s |

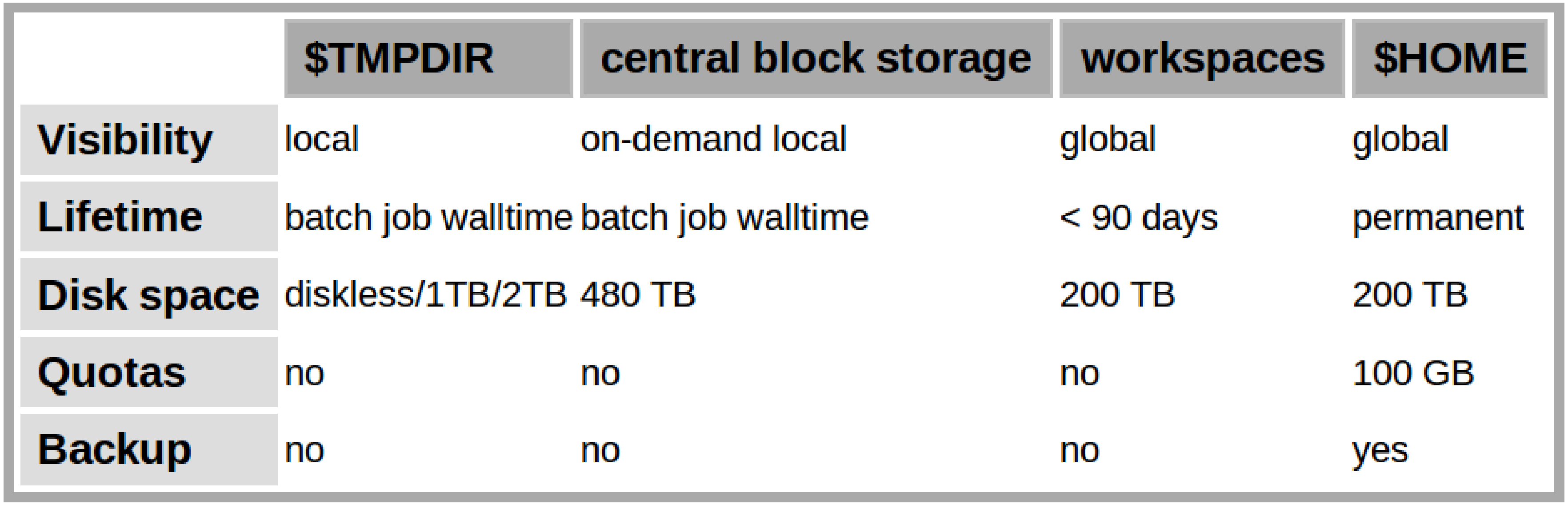
# bwForCluster JUSTUS

*Federated HPC tier 3 resources*

Selected characteristics:

- Dedicated to **computational chemistry**
  - High I/O, large MEM jobs
- User and software support by *bwHPC competence center*

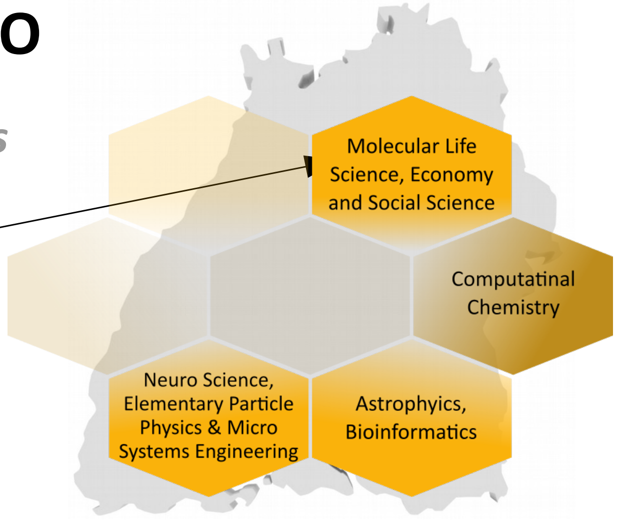

Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics



| | $TMPDIR | central block storage | workspaces | $HOME |
|---|---|---|---|---|
| **Visibility** | local | on-demand local | global | global |
| **Lifetime** | batch job walltime | batch job walltime | < 90 days | permanent |
| **Disk space** | diskless/1TB/2TB | 480 TB | 200 TB | 200 TB |
| **Quotas** | no | no | no | 100 GB |
| **Backup** | no | no | no | yes |

# bwForCluster MLS&WISO

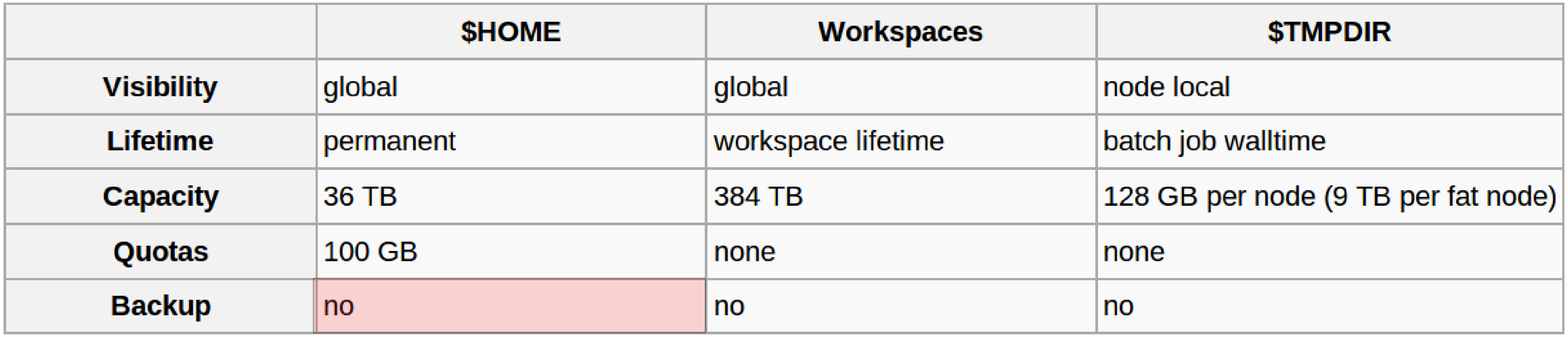
*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

- Dedicated to **molecular life science, economics and social science**

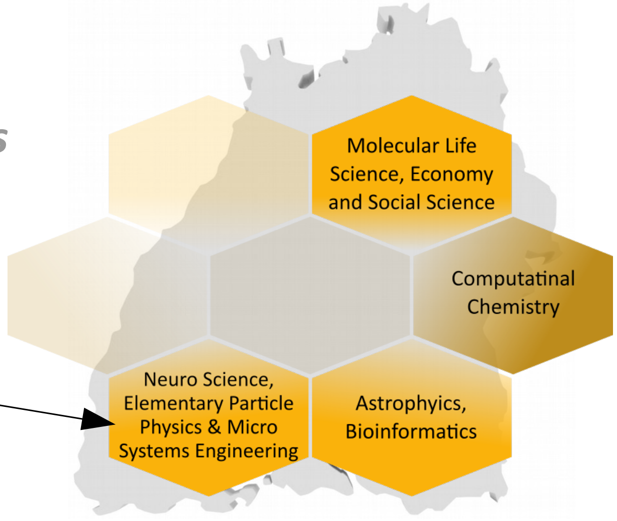
  **+ cluster for method development**
- User and software support by *bwHPC competence center*



| | $HOME | Workspaces | $TMPDIR |
|---|---|---|---|
| **Visibility** | global | global | node local |
| **Lifetime** | permanent | workspace lifetime | batch job walltime |
| **Capacity** | 36 TB | 384 TB | 128 GB per node (9 TB per fat node) |
| **Quotas** | 100 GB | none | none |
| **Backup** | no | no | no |

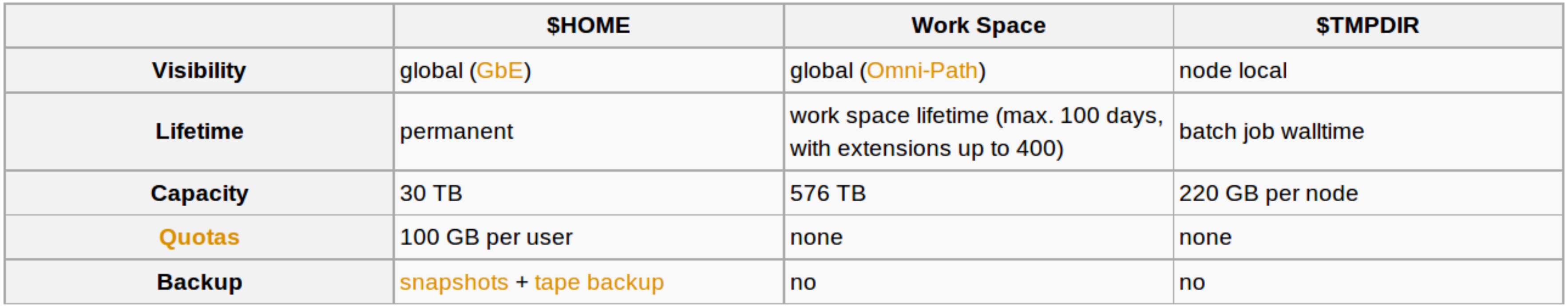
# bwForCluster NEMO

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

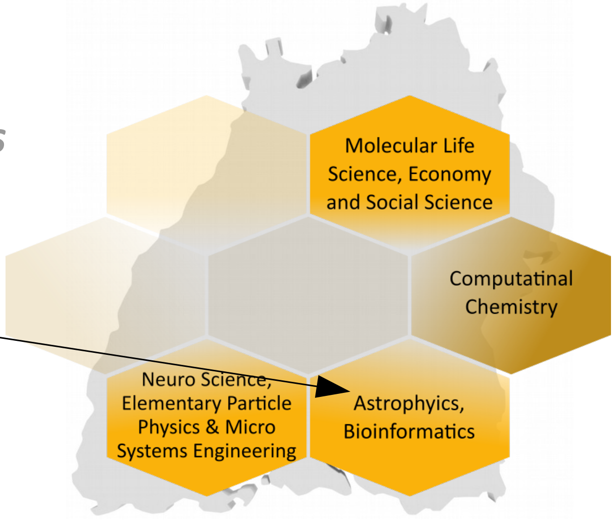
- Dedicated to **neuro science, elementary particle physics, micro systems engineering**
  - Virtual machine images deployable
- User and software support by *bwHPC competence center*



| | $HOME | Work Space | $TMPDIR |
|---|---|---|---|
| **Visibility** | global (GbE) | global (Omni-Path) | node local |
| **Lifetime** | permanent | work space lifetime (max. 100 days, with extensions up to 400) | batch job walltime |
| **Capacity** | 30 TB | 576 TB | 220 GB per node |
| **Quotas** | 100 GB per user | none | none |
| **Backup** | snapshots + tape backup | no | no |

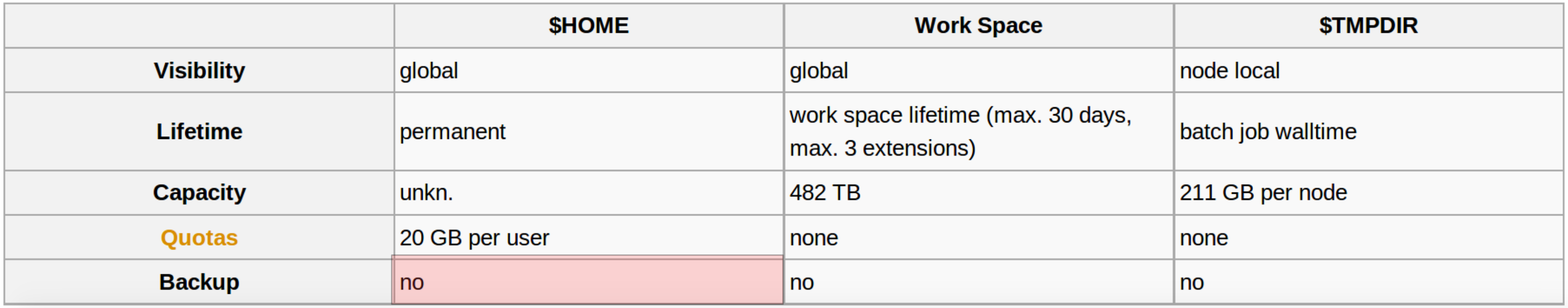
# bwForCluster BinAC

*Federated HPC tier 3 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

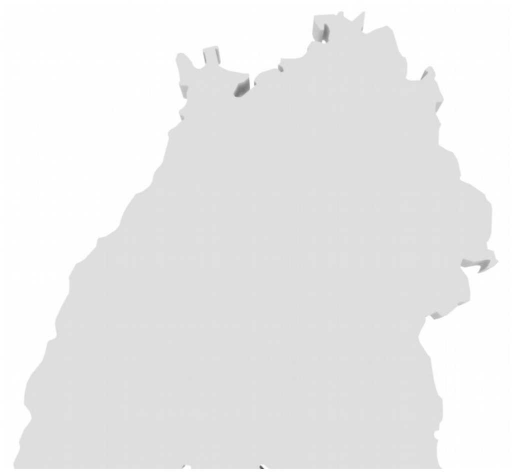
- Dedicated to **astrophysics, bioinformatics**
  - Dual GPU systems
- User and software support by *bwHPC competence center*



| | $HOME | Work Space | $TMPDIR |
|---|---|---|---|
| **Visibility** | global | global | node local |
| **Lifetime** | permanent | work space lifetime (max. 30 days, max. 3 extensions) | batch job walltime |
| **Capacity** | unkn. | 482 TB | 211 GB per node |
| **Quotas** | 20 GB per user | none | none |
| **Backup** | no | no | no |

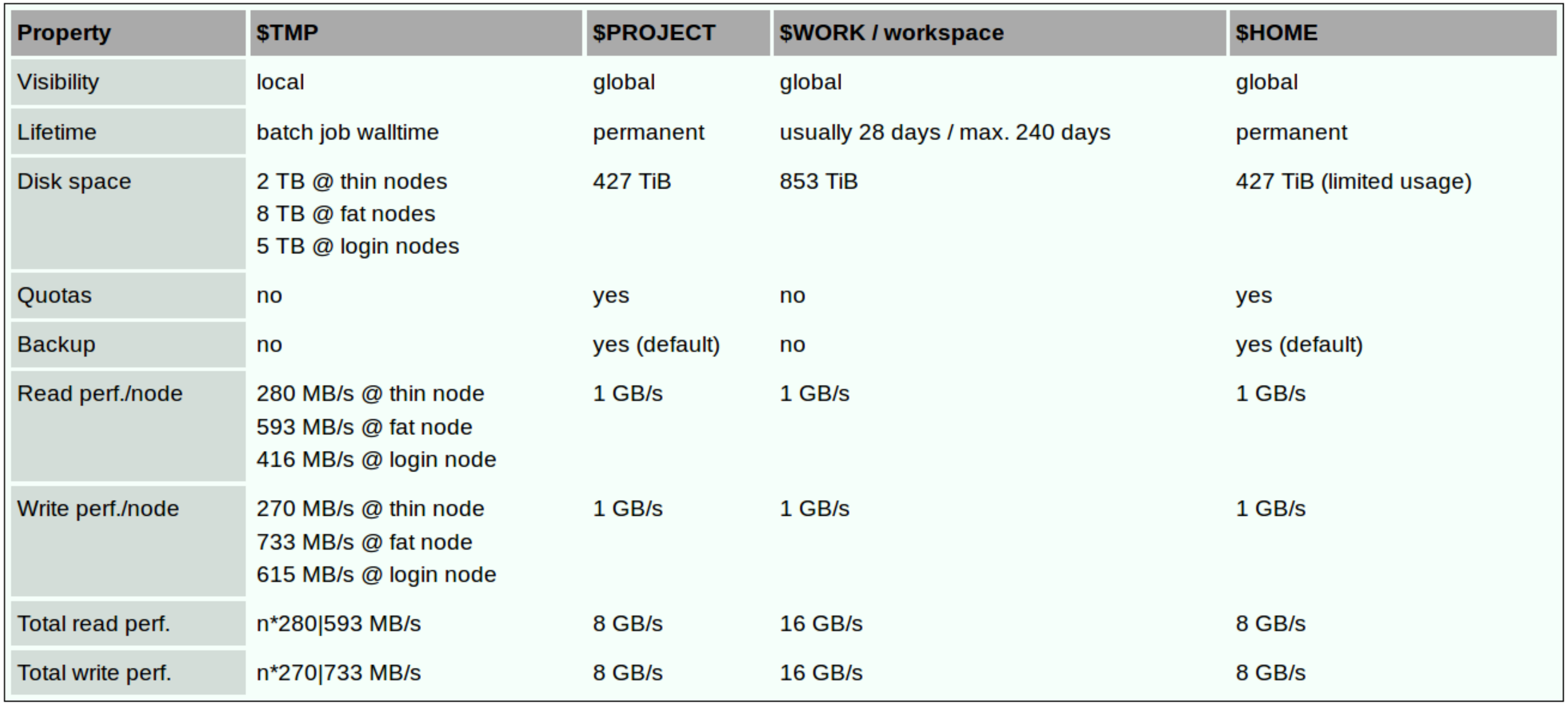
# ForHLR I

*Federated HPC tier 2 resources*



Selected characteristics:

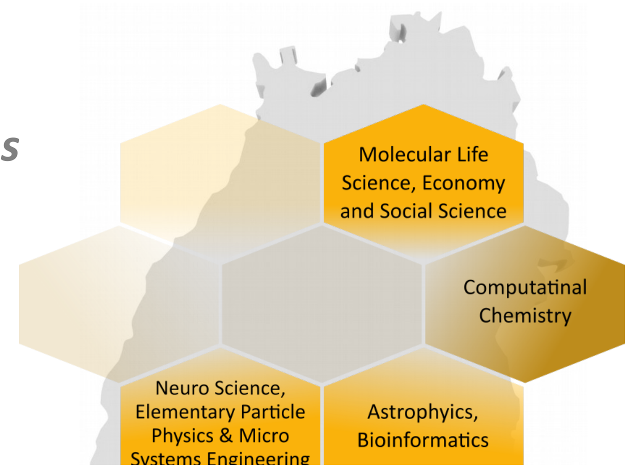
- Next level for advanced HPC users
- Research, high scalability



| Property | $TMP | $PROJECT | $WORK / workspace | $HOME |
|---|---|---|---|---|
| Visibility | local | global | global | global |
| Lifetime | batch job walltime | permanent | usually 28 days / max. 240 days | permanent |
| Disk space | 2 TB @ thin nodes<br>8 TB @ fat nodes<br>5 TB @ login nodes | 427 TiB | 853 TiB | 427 TiB (limited usage) |
| Quotas | no | yes | no | yes |
| Backup | no | yes (default) | no | yes (default) |
| Read perf./node | 280 MB/s @ thin node<br>593 MB/s @ fat node<br>416 MB/s @ login node | 1 GB/s | 1 GB/s | 1 GB/s |
| Write perf./node | 270 MB/s @ thin node<br>733 MB/s @ fat node<br>615 MB/s @ login node | 1 GB/s | 1 GB/s | 1 GB/s |
| Total read perf. | n*280\|593 MB/s | 8 GB/s | 16 GB/s | 8 GB/s |
| Total write perf. | n*270\|733 MB/s | 8 GB/s | 16 GB/s | 8 GB/s |

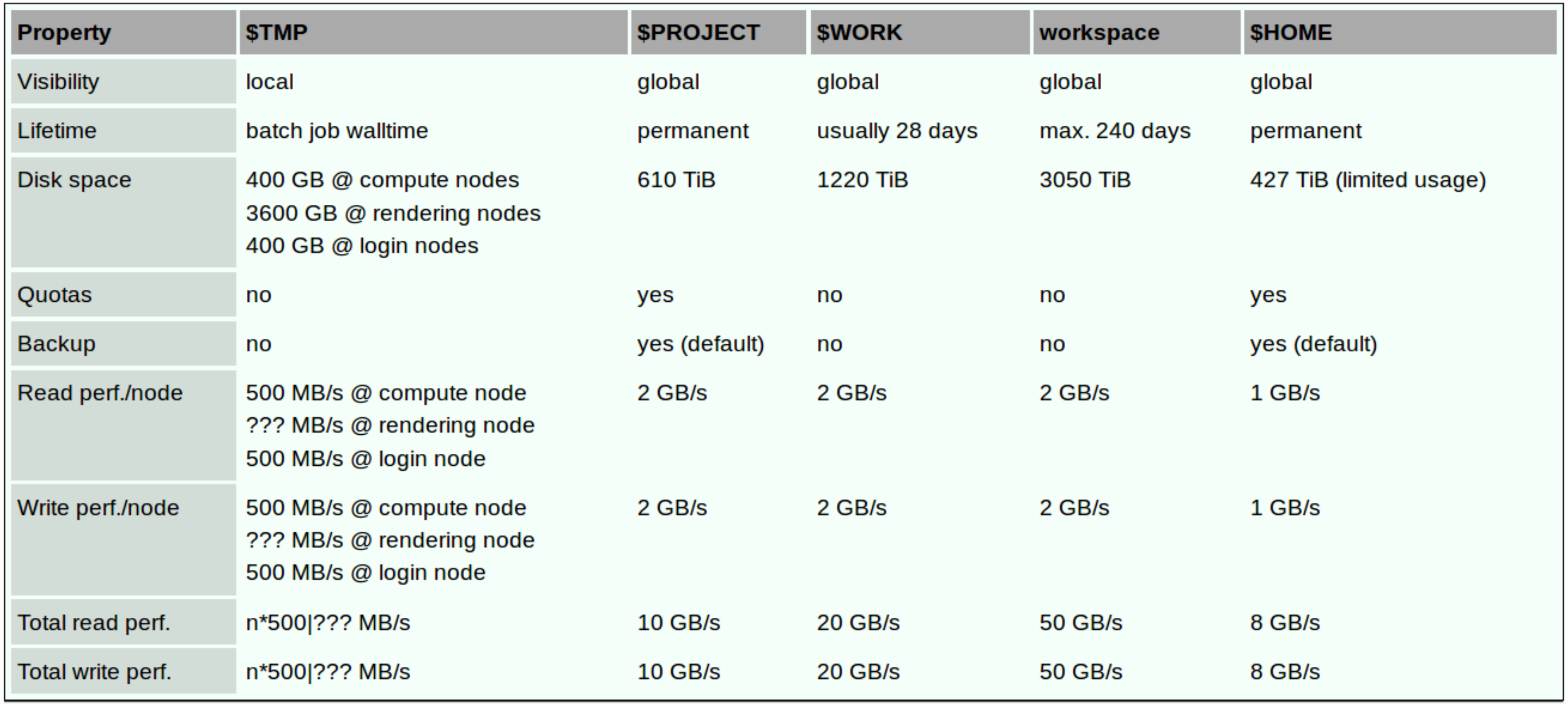
# ForHLR II

*Federated HPC tier 2 resources*


Molecular Life Science, Economy and Social Science
Computatinal Chemistry
Neuro Science, Elementary Particle Physics & Micro Systems Engineering
Astrophyics, Bioinformatics

Selected characteristics:

- Next level for advanced HPC users
- Research, high scalability



| Property | $TMP | $PROJECT | $WORK | workspace | $HOME |
|---|---|---|---|---|---|
| Visibility | local | global | global | global | global |
| Lifetime | batch job walltime | permanent | usually 28 days | max. 240 days | permanent |
| Disk space | 400 GB @ compute nodes<br>3600 GB @ rendering nodes<br>400 GB @ login nodes | 610 TiB | 1220 TiB | 3050 TiB | 427 TiB (limited usage) |
| Quotas | no | yes | no | no | yes |
| Backup | no | yes (default) | no | no | yes (default) |
| Read perf./node | 500 MB/s @ compute node<br>??? MB/s @ rendering node<br>500 MB/s @ login node | 2 GB/s | 2 GB/s | 2 GB/s | 1 GB/s |
| Write perf./node | 500 MB/s @ compute node<br>??? MB/s @ rendering node<br>500 MB/s @ login node | 2 GB/s | 2 GB/s | 2 GB/s | 1 GB/s |
| Total read perf. | n*500\|??? MB/s | 10 GB/s | 20 GB/s | 50 GB/s | 8 GB/s |
| Total write perf. | n*500\|??? MB/s | 10 GB/s | 20 GB/s | 50 GB/s | 8 GB/s |

# Thank you for your attention!

# Questions?