Exa Scale Innovation Center

JÜLICH FORSCHUNGSZENTRUM

# Realising active storage concepts for today's and future HPC systems

ScalPerf 2013 | 26. September 2013 | Dirk Pleiter (JSC)

Mitglied der Helmholtz-Gemeinschaft

# I/O Challenges

- **Diverging compute vs. I/O performance**
- **Need for I/O operation rates**
  - Trend towards non-traditional, data intensive applications
- **Data transport challenge**
  - Expensive in terms of energy
  - Exascale bandwidth requirements cannot be achieved using JBODs within power budget

R. Stevens and A. White et al. (2010):

| Systems | 2009 | 2018 | Difference Today & 2018 |
|---|---|---|---|
| IO Rates | 0.2 TB | 60 TB/s | O(100) |

# Architecture: Active Storage

- **Originally proposed as "Active Disks"**
  Anurag Acharya et al., "Active Disks: Programming Model, Algorithms and Evaluation" (1998)
  - Motivating observation: Compute power doubles every 18 months, data store capacity every 5 month
    - J. Gray's formulation: **What happens if processors are infinitely fast and storage is free?**
  - Proposed a stream-based programming model
    - Examples: SQL select, image convolution
- **Opportunity: Processing capabilities close to data**
- **Limitations of the original proposal**
  - Too simple processing devices
  - Lacking interconnect

Mitglied der Helmholtz-Gemeinschaft

# Non-Volatile Memory Technologies

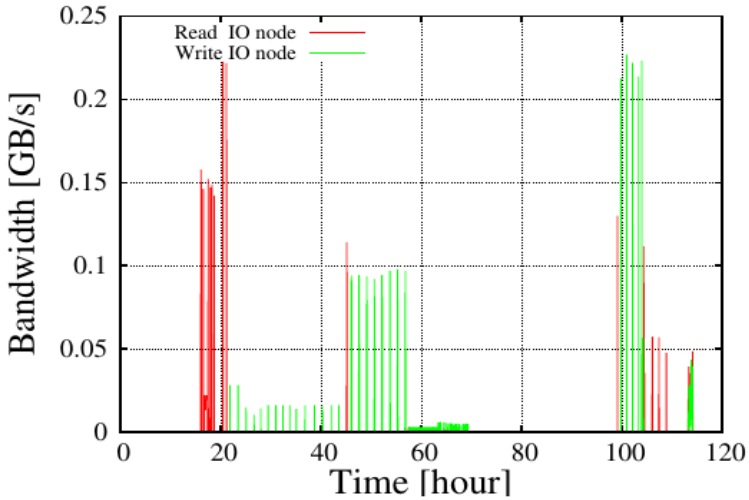
- **Today: NAND flash memory**
  - For HPC: SLC NAND flash
- **Advantages**
  - High bandwidth
  - Low latency → high IOPS
  - Power efficiency
  - Packaging/integrateability
- **Disadvantages**
  - Limited endurance (~100k write cycles for SLC)
  - High costs

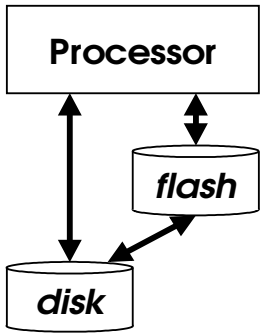
Mitglied der Helmholtz-Gemeinschaft

# PRACE Prototype Experience

[El Sayed et al., ISC'13]

- **Observation:**
  Bursty access to storage
- **Hierarchical storage architecture**
  - Extra storage layer used as
    - write buffer
    - read pre-fetch buffer
- **Realisation: JUNIORS**
  - Storage server with flash memory cards
  - Attached to compute system via 10 GbE
  - GPFS Information Lifecycle Management
- **Critical review**
  - Limited opportunities to improve write bandwidth
  - High speed-up potential in case of I/O rate limitations



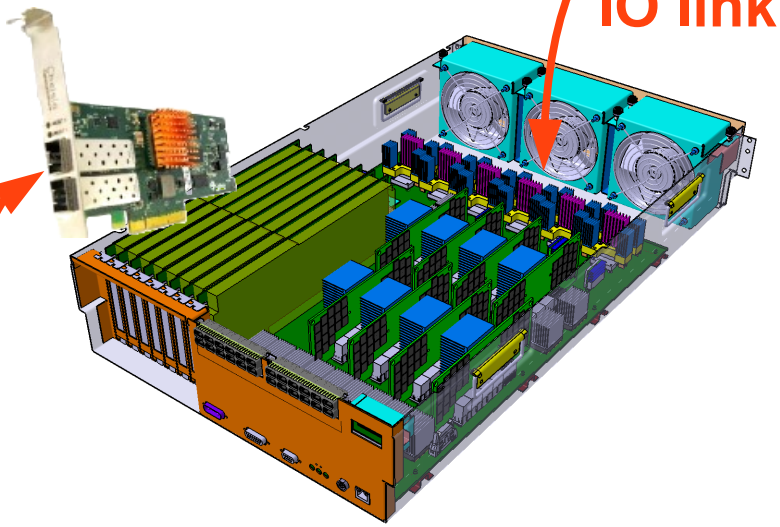

Mitglied der Helmholtz-Gemeinschaft

# Blue Gene/Q I/O Architecture

Data store

IO link

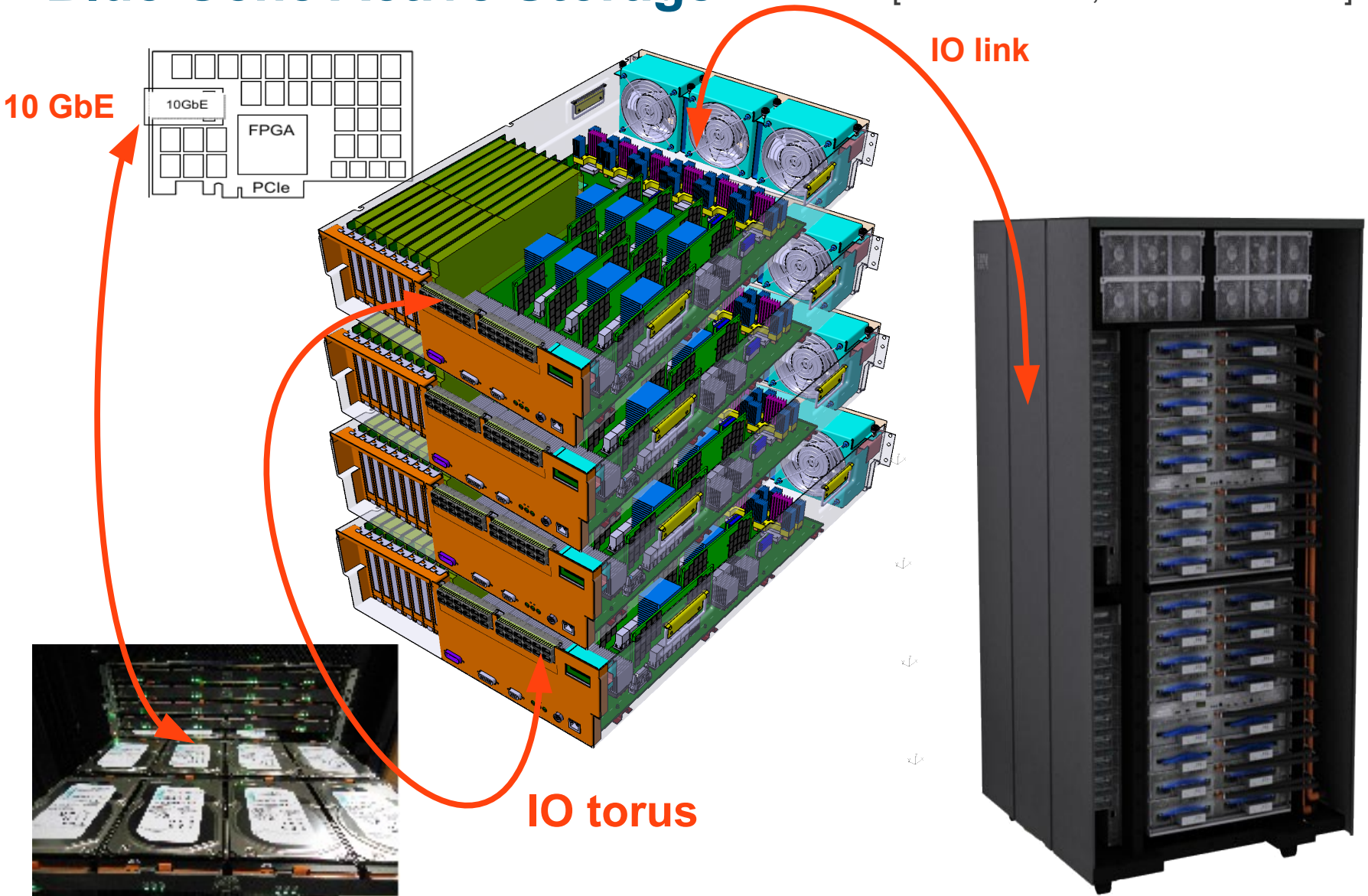
Compute



10 GbE

# Blue Gene Active Storage

[B. Fitch et al., HEC FSIO 2010]



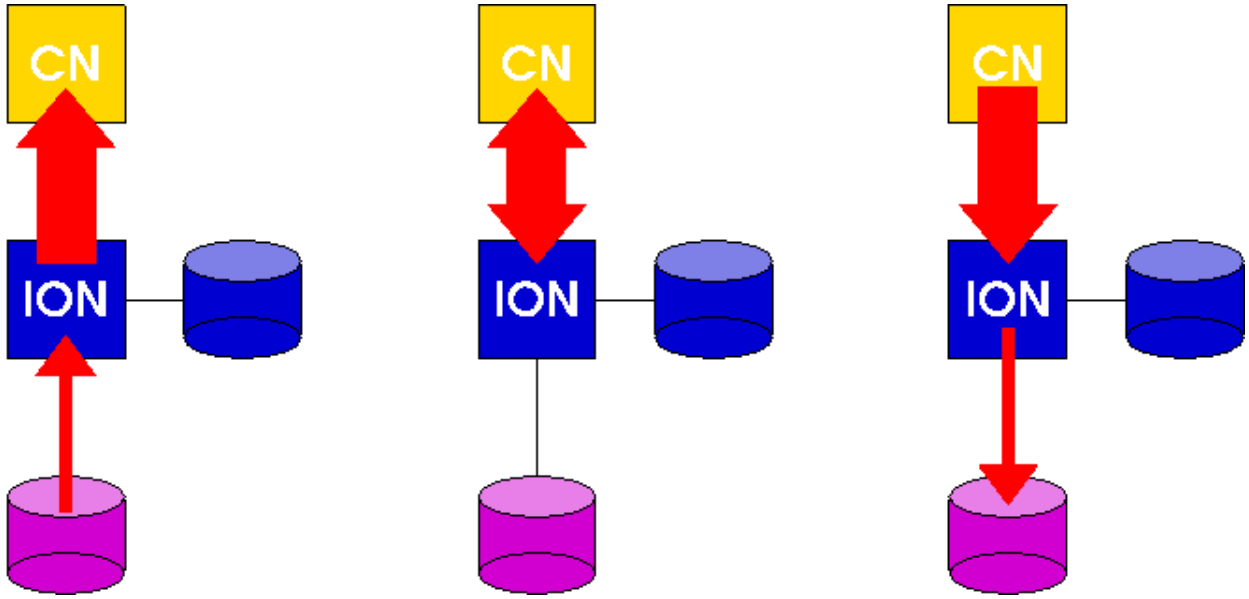

Mitglied der Helmholtz-Gemeinschaft

# Blue Gene Active Storage (cont.)

- **Architecture features**
  - Non-volatile memory integrated into HPC system
  - Attached to many-core processors
  - Interconnected by high-speed torus network
- **Target system specification**
  - 32 Blue Gene/Q nodes (6.6 TFlop/s peak)
  - 64 TBytes flash memory
  - 64 GByte/s from/to flash memory (aggregated)
  - 64 GByte/s bi-section bandwidth
  - 128 GByte/s IO link bandwidth

Mitglied der Helmholtz-Gemeinschaft

# Use Cases




**Multi-pass analysis**

**Out-of-core / Check-pointing**

**Post-processing / Visualisation**

Mitglied der Helmholtz-Gemeinschaft

# BGAS Programming Models

- **Active messages**
  - Messages forwarded from compute to BGAS system get processed on BGAS system
  - Simple model, decouples simulation and data post-processing
- **Coupled applications**
  - Parallel applications on compute and BGAS system
  - MUSIC: API developed for coupling spiking neuronal network simulators
- **SIONlib**
  - Data aggregation and buffering
  - POSIX like interface

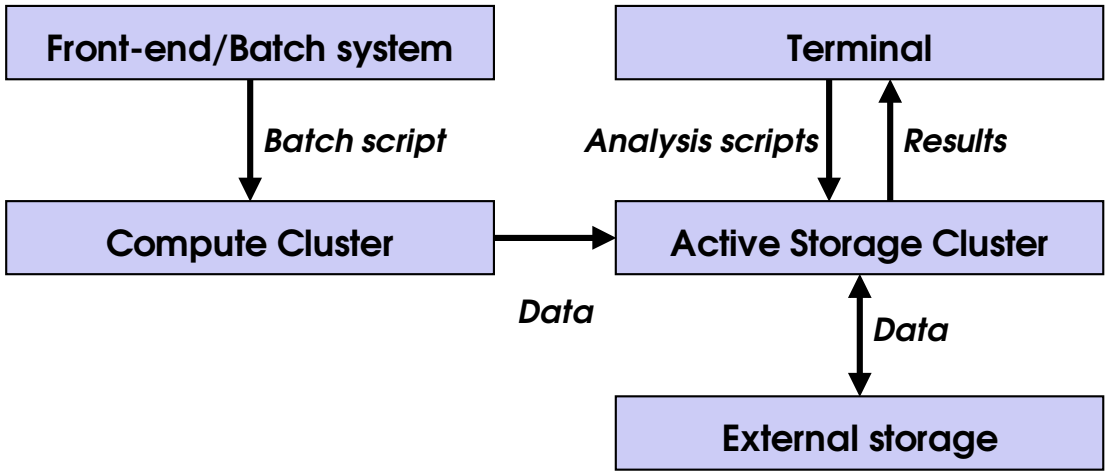
Mitglied der Helmholtz-Gemeinschaft

# Application Experts Workshop (January 2013)

- **Covered application areas:**
  - Climate Research
  - Material Science
    - Density Functional Theory calculations
  - Medicine and Neuroscience
    - Neuronal network simulations
    - Neural tissue simulation
    - Genomics
    - Genetic epidemiology
  - Astronomy and Astrophysics
- **Slides:** `http://bit.ly/10HBUat`

# Neuroscience Simulations: NEST (M. Diesmann, FZJ)

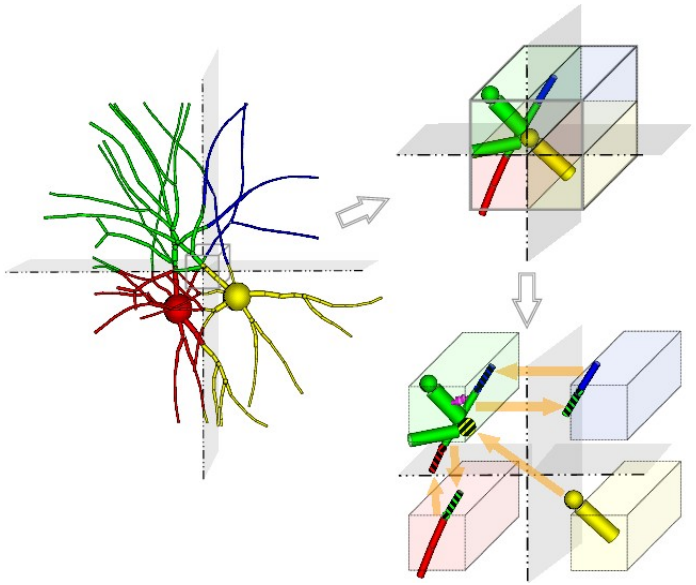
- Simulation of large **spiking neural networks**
- **Data generated for post-processing** (here: $10^8$ neurons)
  - Spiking data for 1 biological second: 1.5 GByte
  - Membrane potential every 10 ms: 149 Gbytes
  - Synaptic weights: $10^4$x more data
- **Near-future vision of the NEST developers:**




Mitglied der Helmholtz-Gemeinschaft

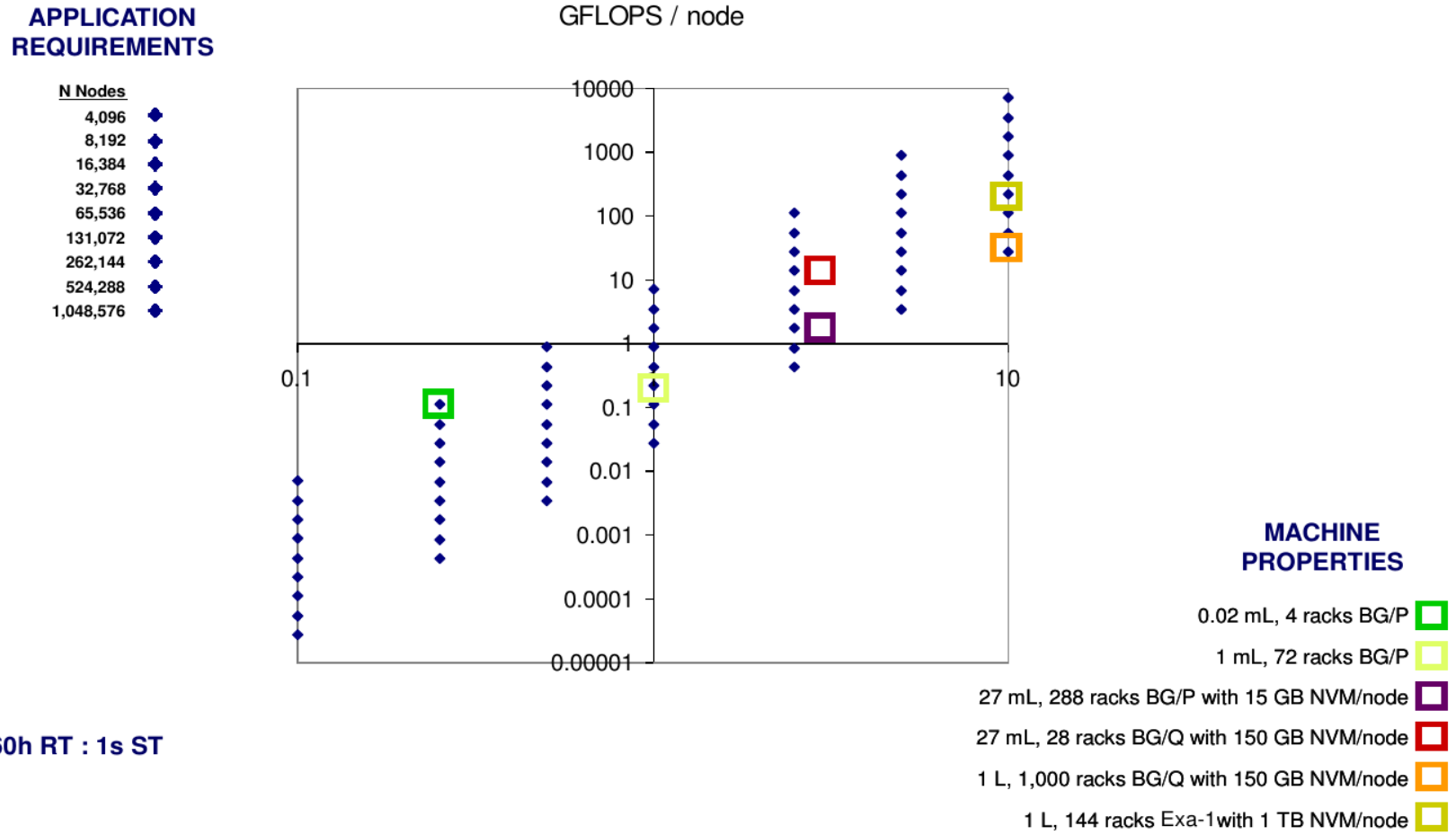
# Neural Tissue Simulations (J. Kozloski, IBM)

- **Modelling of large scale neural tissue anatomy and physiology**
- **Scalable application**
- **Memory capacity limitations**
  - Example:
    Simulation of human brain
    = Exascale challenge
  - 1 liter of neural tissue in $10^7$ volumes
  - 2.25 GByte per volume → 20-30 PByte total



# Neural Tissue Simulations (cont.)

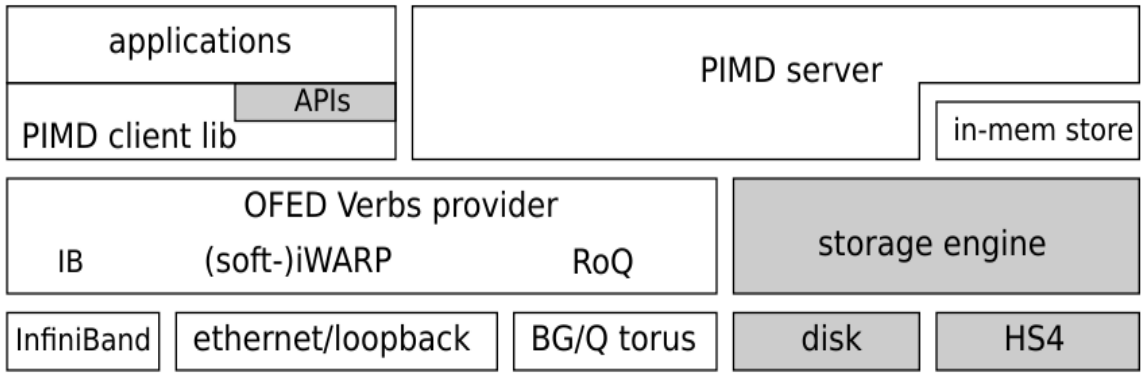
- **Node performance requirements:**




Mitglied der Helmholtz-Gemeinschaft

# Genomics (D. Carrera, BSC/U)

- **Application challenge:**
  Exponential growth of sequencing data
  - Data volume doubles every 5 months
- **Need for online processing →**
  **Challenge:** Random access on large data set
- **PIMD:** Parallel In Memory Database
  - Parallel key-value store engine



# Non-volatile Memory: Future Opportunities

[Lee et al. (2011); Micron 2011]

| | DRAM | SLC NAND | PCM |
|---|---|---|---|
| Minimum access unit [Bytes] | O(10) | O(1000) | O(10) |
| Density [Gb] | 4 | 256 | 1 |
| Active current [mA] | ~130-175 | 15 | 10-70 |
| Idle current [mA] | 35 | ~1 | ~1 |
| Write energy [nJ/b] | ~0.1 | 0.1-1 | <1 |
| Read energy [nJ/b] | ~0.1 | <<1 | <<1 |
| Endurance | $10^{15}$ | $10^{5}$ | $10^{6\sim8}$ |

- **Other technology**
  - Higher density flash memory using VNAND
  - Other non-volatile memory technologies, e.g. STT-MRAM

Mitglied der Helmholtz-Gemeinschaft

# Programming Models: Challenges

- **Programming of a heterogeneous system**
  - How to limit application code modifications?
  - How to address portability?
- **Library approach**
  - Function shipping from compute to active storage node
  - Candidate: HDF5 format or data conversion
- **Existing solutions typically assume sequential computation within active storage**
  - Processing of data streams
  - Identified need for parallel processing

# Summary and Discussions

- **New I/O architectures required**
  - Increasing performance gap
  - Non-traditional, data-intensive HPC applications
  - Power constraints
- **New opportunities**
  - Active storage architectures
  - Non-volatile memory technologies
- **Blue Gene Active Storage**
  - Turns Blue Gene into an architecture suitable for data-intensive applications
- **Potential use by relevant applications**
  - Significant application development efforts needed

Mitglied der Helmholtz-Gemeinschaft