lrz Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

SuperMUC Next Generation

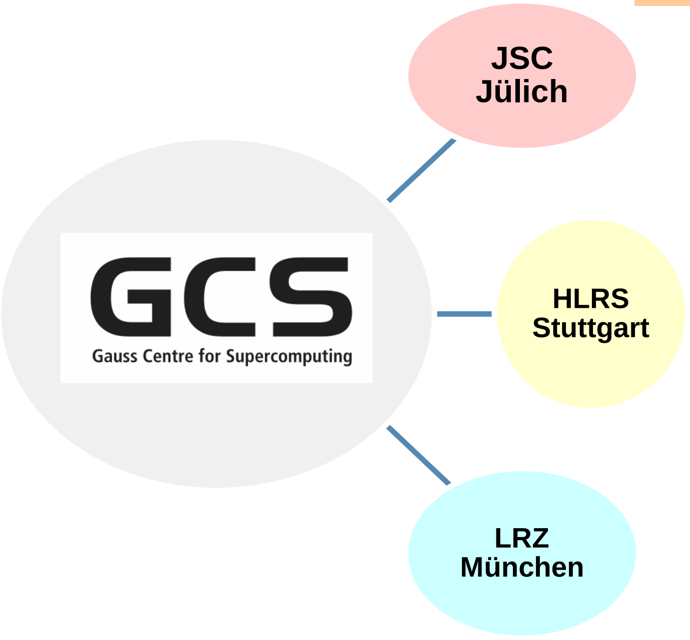
Updating the top tier computational resources at LRZ

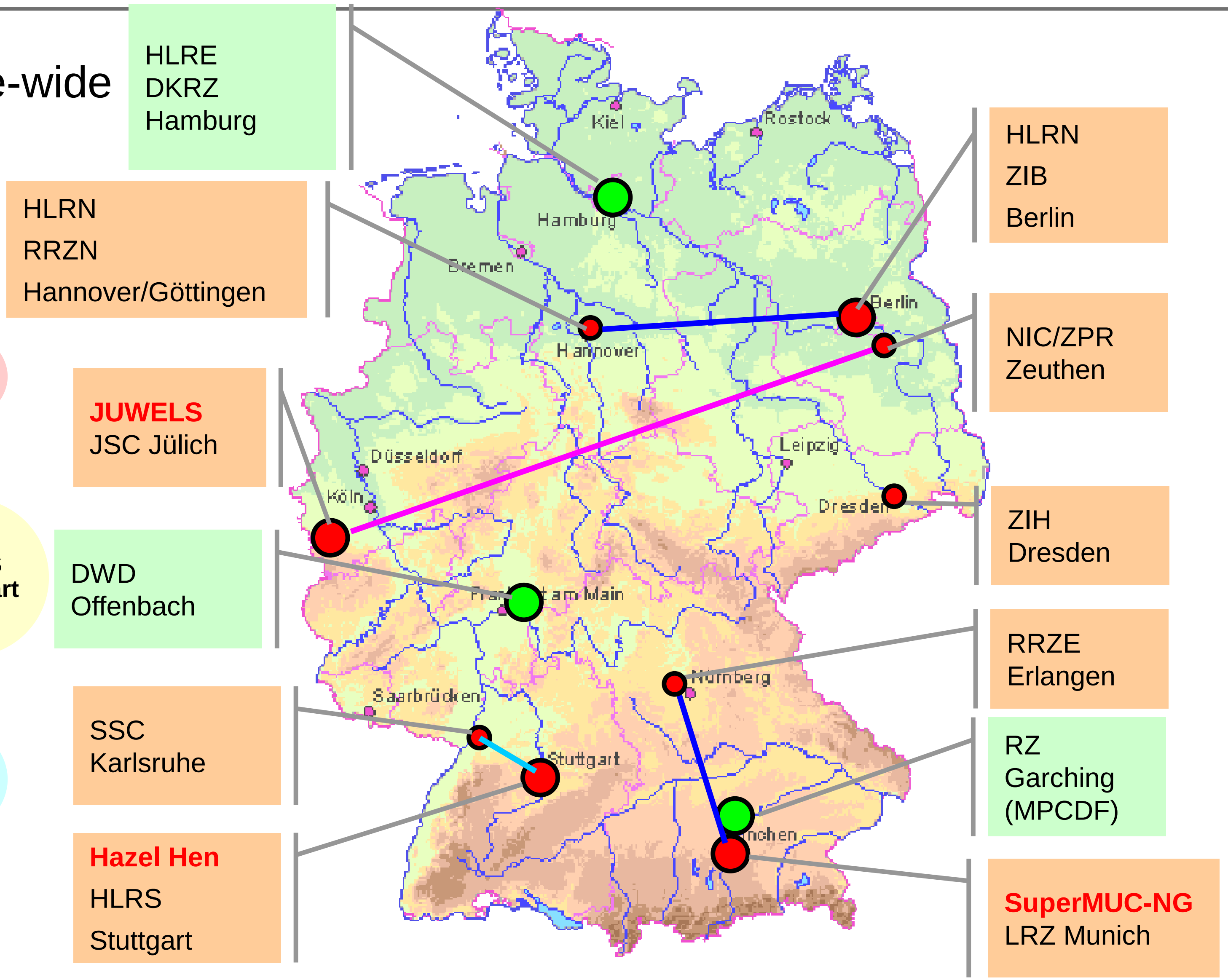
Dr. Reinhold Bader, LRZ

Scientific Supercomputing in Germany

National and/or state-wide services

Thematic Centres





PRACE GCS operates Tier-0/1
HPC systems in Germany

Procurement aims and strategy

- Retain applicability of system to broad application spectrum
- Further improvements for energy efficient operation
- Market diversification:
  - permit vendors of accelerator-based solutions to participate
- LRZ as Big Data competence centre:
  - storage components for long-term/project-specific data need to be integrated
  - cloud components are part of procurement: targets derived services like
    - visualization (possibly using GPUs)
    - computational steering front ends
    - pre/postprocessing (possibly using GPUs)
    - alternative operating environments
    - project-specific data (web) interfaces

  operational concepts for this will take time to mature

security issues ⚠

Benchmark suite – evaluating a weighted overall performance

- MPI benchmarks (24%)
  - latencies, bandwidths and throughput for specific communication patterns
  - bisections are especially important
- Application benchmarks (38%)
  - broad spectrum reflected in codes from
    Astrophysics, Quantum Chemistry, Life Sciences,
    Fluid Dynamics, Geophysics, QCD
  - most of these not ported to GPUs when procurement started
- Kernel benchmarks (38%)
  - evaluate specific system characteristics and data access patterns
  - HPL (Linpack) is still in the list

Procurement process

- Europe-wide procurement (guided by GWB, VgV)
  - initial competitor selection based on financial and technical capability
- Competitive dialogue
  - based on **draft** procurement documents / benchmarks — Jan 2017
  - discussion → clarification of technical issues, remove ambiguities, achieve joint understanding
  - **initial round** with five selected vendors
  - formal bid was evaluated according to published rules — Jun 2017
  - **second round** with the two leading vendors → further sharpening of conditions, **final** procurement documents established
  - final bid evaluated to select vendor for contract negotiations — Nov 2017
- Contract concluded with Intel/Lenovo — Dec 2017

And the winner is: an Intel / Lenovo supercluster

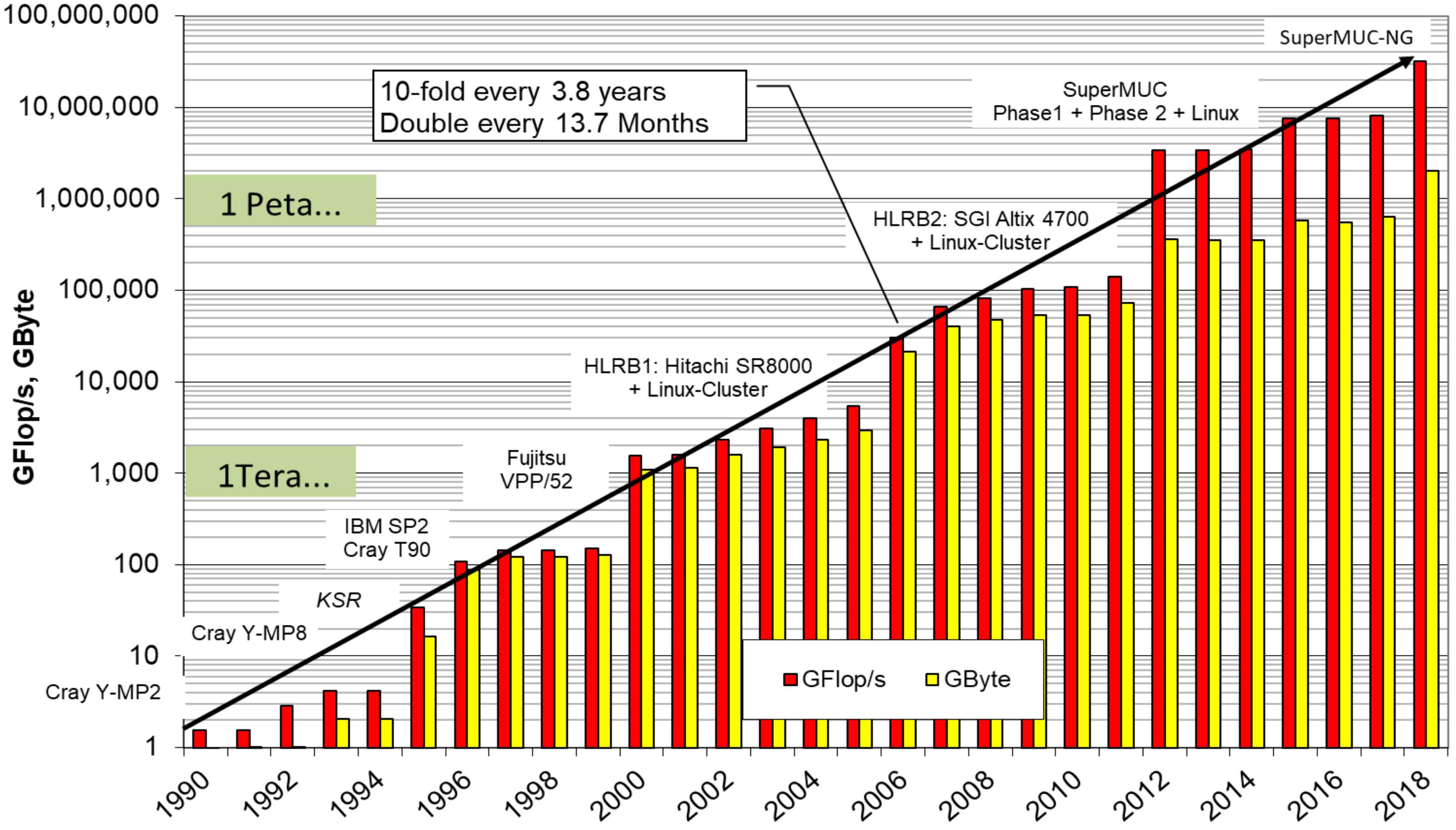
- 6448 compute nodes based on the Skylake architecture
  - SKX 8186 two-socket node (48 cores, mostly 96 GByte DDR4 memory)
    - 144 nodes are "fat" and have 768 GByte DDR4 memory each
  - warm water cooled node design "Lenovo OceanCat"
  - peak performance 26,9 PFLOP/s
- two-level Omnipath generation 1 interconnect fabric
  - pruned fat tree
- storage systems Lenovo DSS-G
  - 50 PByte parallel file system (SCRATCH, WORK)
  - > 10 PByte long-term storage (HOME, PROJECT)

Delivery status

- most of the hardware is delivered
  - that was the easy part, integration and setup will need lots of time
- target milestones
  - HPL run in October '18
  - initial "friendly user" operation in (late) November '18
  - acceptance completed January '19

Evolution of peak performance and memory

(sum over all LRZ systems)



Dr. M. Brehm

Comments on GPU-accelerated systems

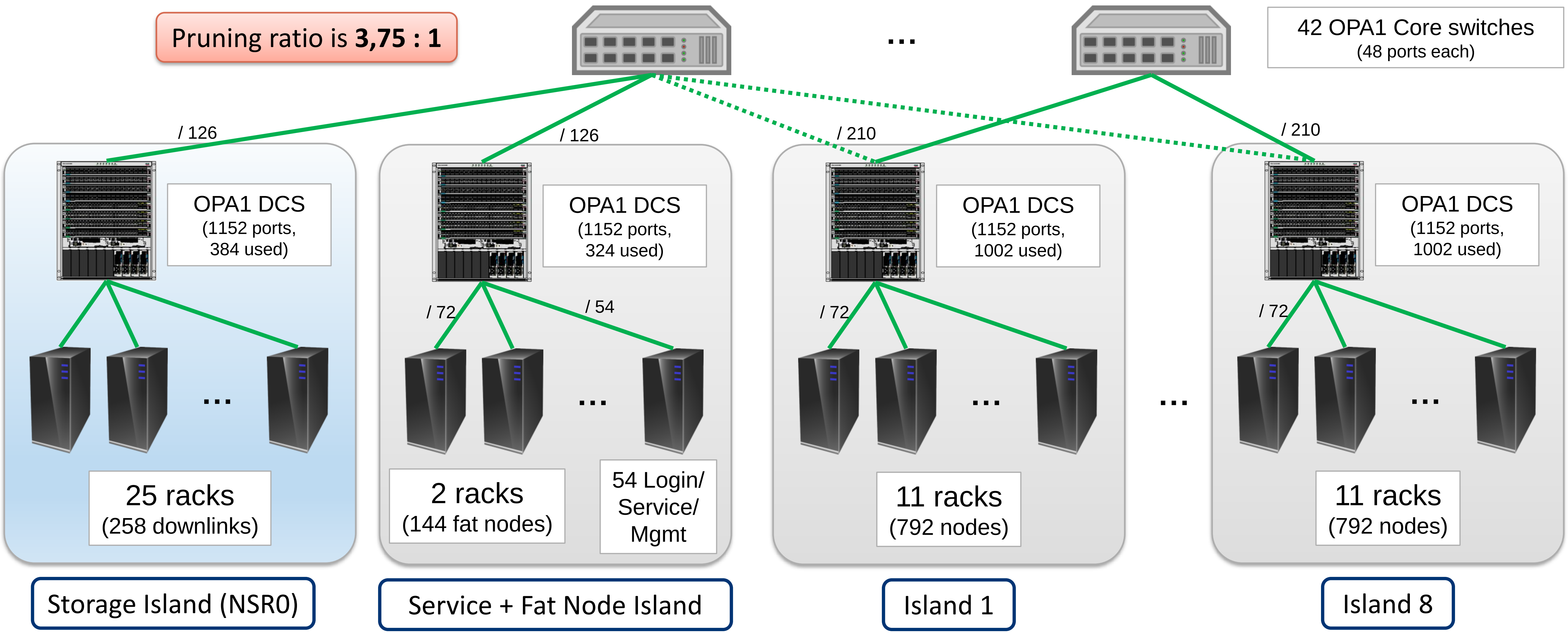
- Pros
  - solid system design (including cooling/power, storage)
  - very high Flops/Watt ratio, assuming efficient node usage can be achieved
  - GPUs can be switched off if not needed
- Cons
  - programming model quite complex, even if directives used
  - use of multi-GPU programming mandatory

  → efficient node usage difficult to achieve
  - potential scaling limitations due to interconnect balance issues

LRZ would have decided in favour of a GPU+CPU „hybrid“ system if the benchmark results had been competitive

The race was relatively tight, though ...

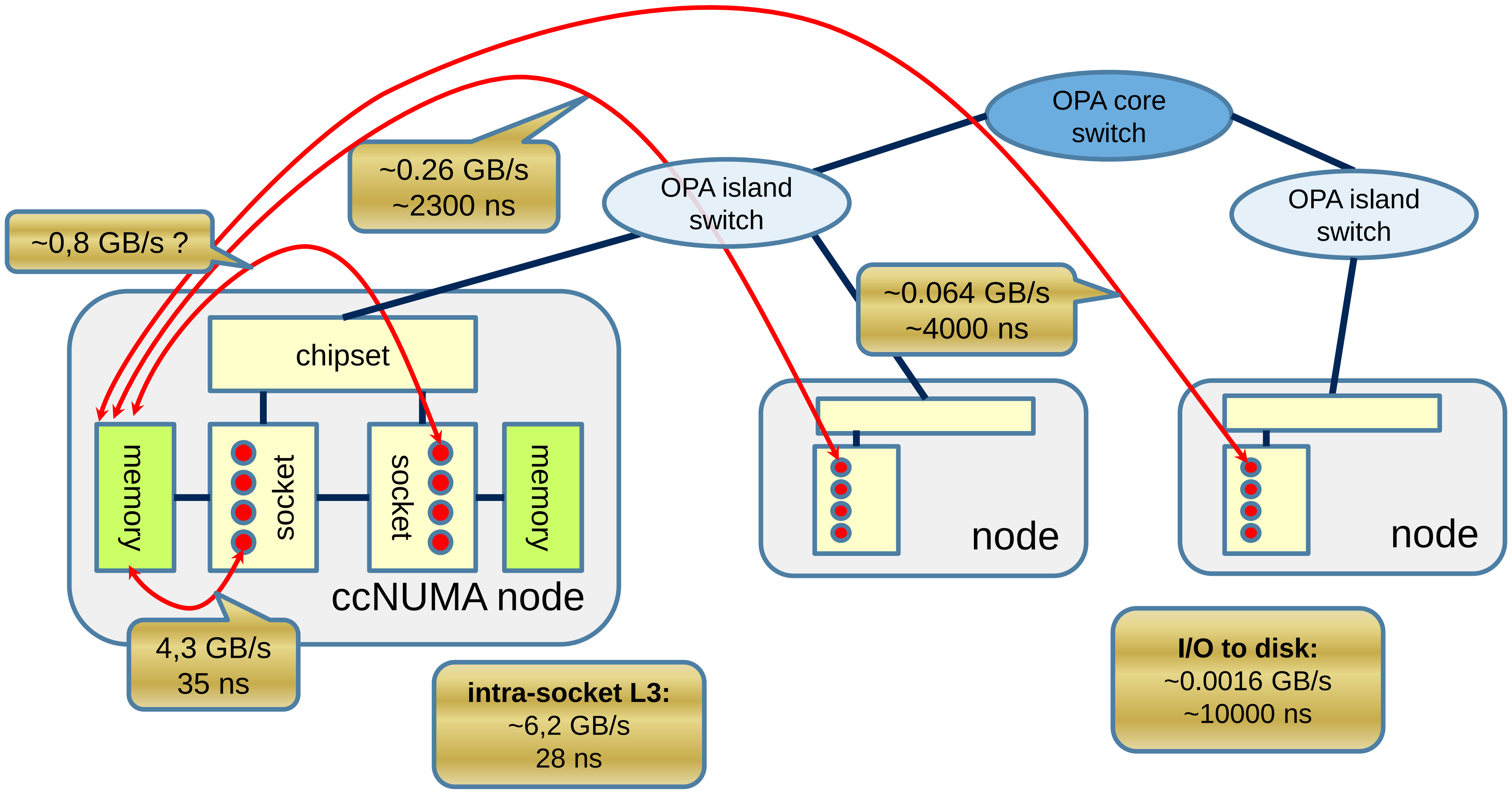
High level system architecture: Omnipath 1 Fat Tree Fabric



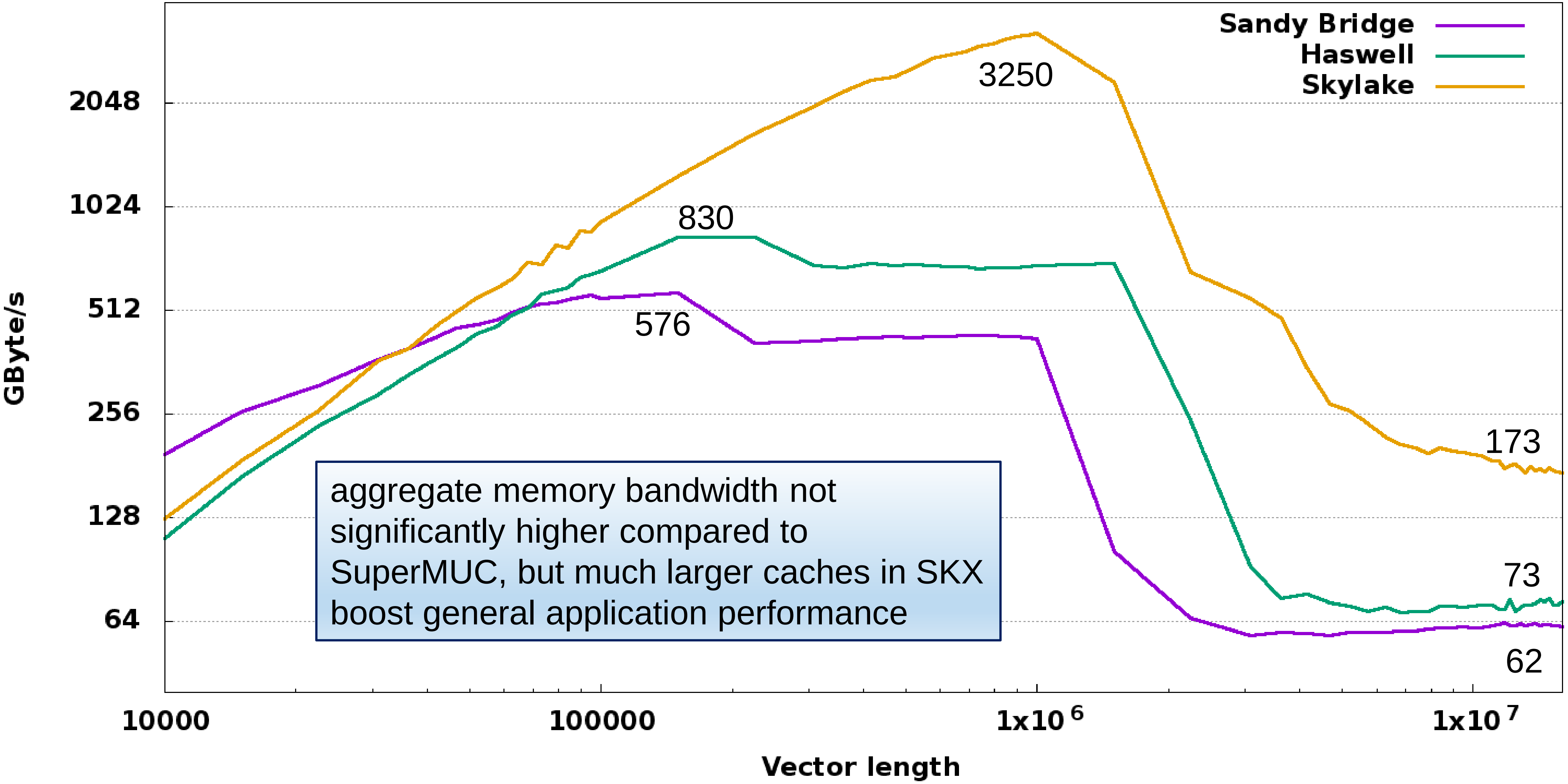
Some architecture characteristics

- Indicating relevant parameters:
  - latencies, moderately saturated bandwidths **per core**
  - values give impression of general magnitude

Note:
Skylake in LRZ system will have **24** cores/socket



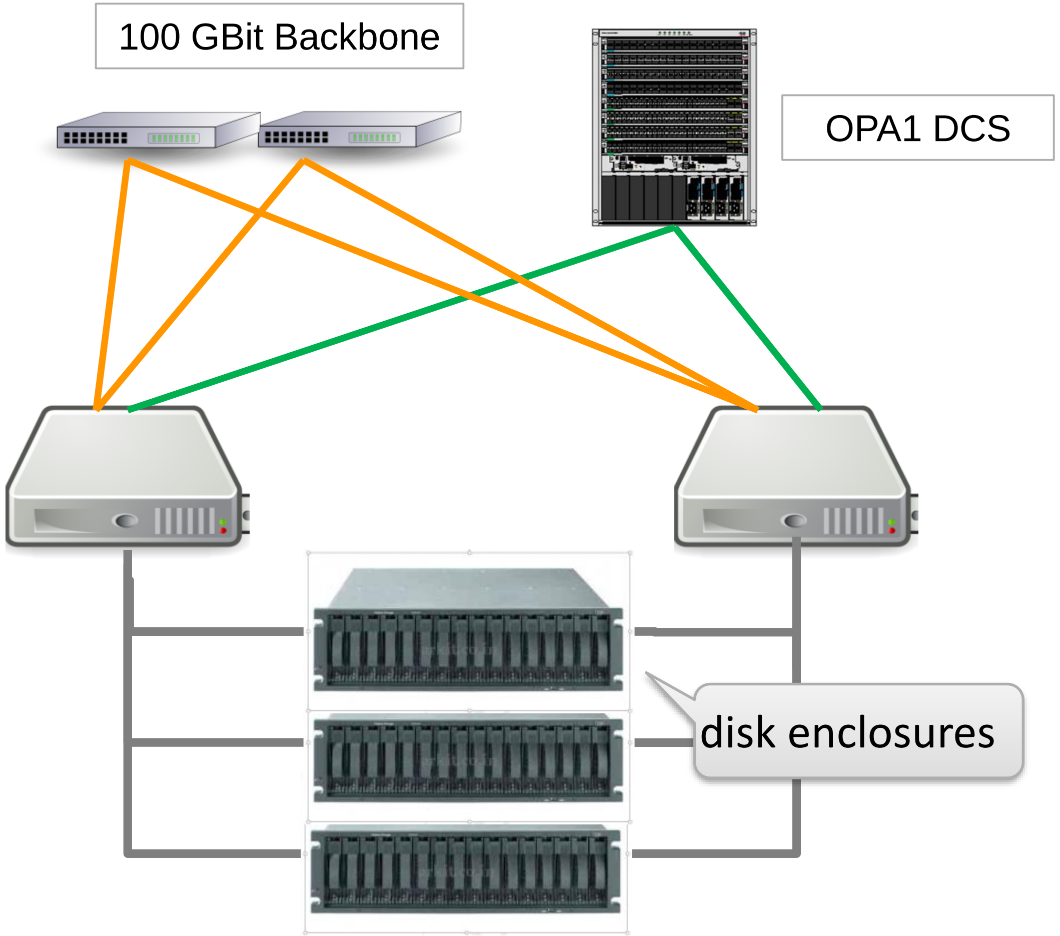
Vector triad A = B * C + D: OpenMP node bandwidth



aggregate memory bandwidth not significantly higher compared to SuperMUC, but much larger caches in SKX boost general application performance

Storage

- generic DSS-G building block



- Setup

NSR0 → separate power/cooling

  - two servers in HA configuration
  - integration with OPA1 → **data access** from system
  - optional integration with 100 GBit storage backbone (HOME/PROJECT only) → **data access** from "outside world" (e.g., LRZ's Linux-Cluster)
- Total of 54 building blocks
  - SSDs are mostly used for metadata, HDDs for data
- Cooling: RDHX on rack level
  - adsorption chillers generate cold water

Phase 2 information and LRZ expectations

- Budget is significantly smaller compared to phase 1
- Installation in timeframe 2021/22
- At least same level of aggregate performance as phase 1
- Additional storage with same capacity and bandwidth

- Technical possibilities
  - future processor (better Flop/Watt ratio)
  - future programming models
  - OPA1 → OPA2 (integration into existing fabric is an option)
  - further advances in cooling technology (power supplies / network components)

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

Questions?