LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MNM TEAM
MUNICH NETWORK MANAGEMENT TEAM

lrz

# SuperMUC – PetaScale HPC at the Leibniz Supercomputing Centre (LRZ)

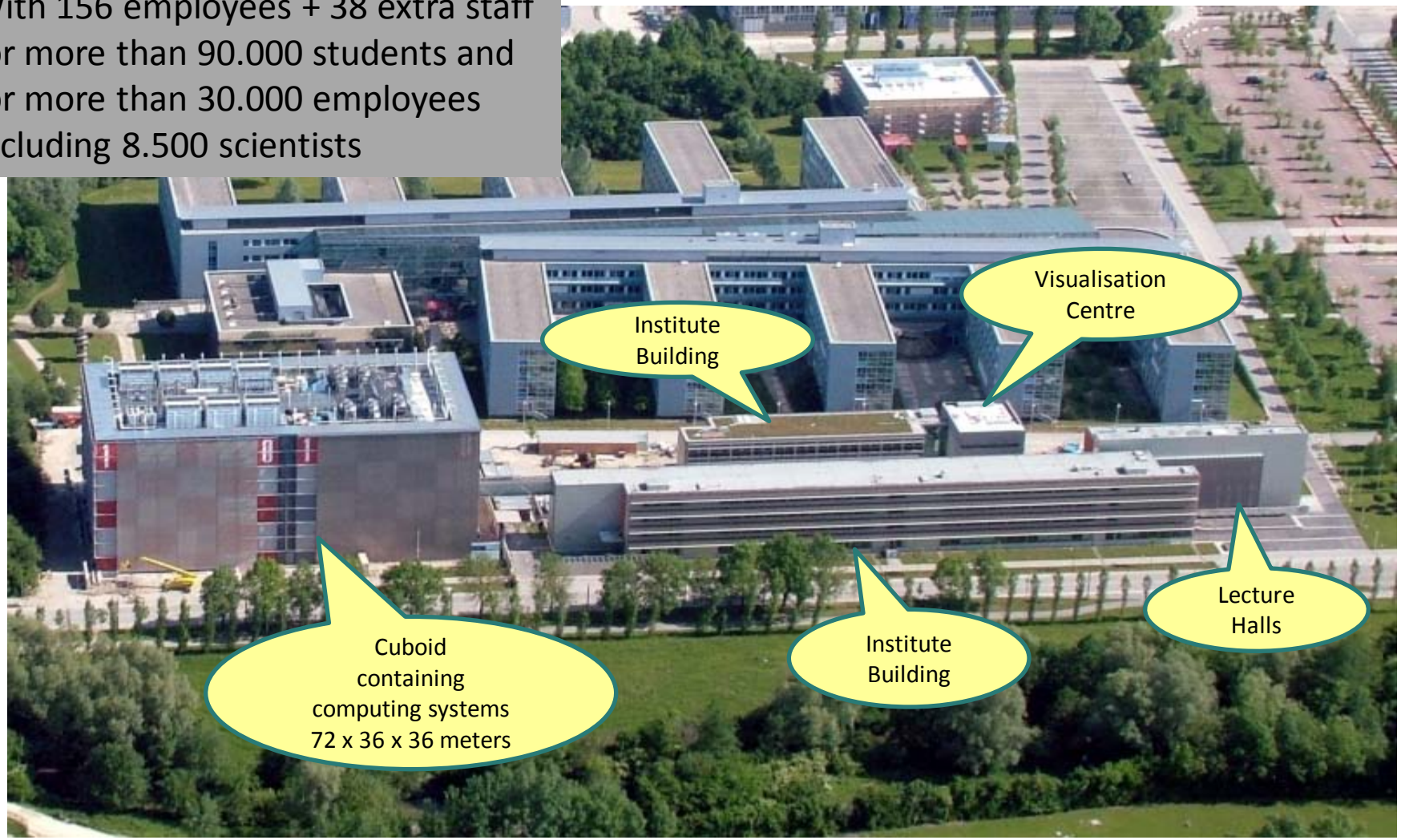
Dieter Kranzlmüller

Munich Network Management Team
Ludwig-Maximilians-Universität München (LMU) &
Leibniz Supercomputing Centre (LRZ)

# SGI UV @ Leibniz Supercomputing Centre

# Leibniz Supercomputing Centre
## of the Bavarian Academy of Sciences and Humanities

With 156 employees + 38 extra staff for more than 90.000 students and for more than 30.000 employees including 8.500 scientists



- Computer Centre for all Munich Universities

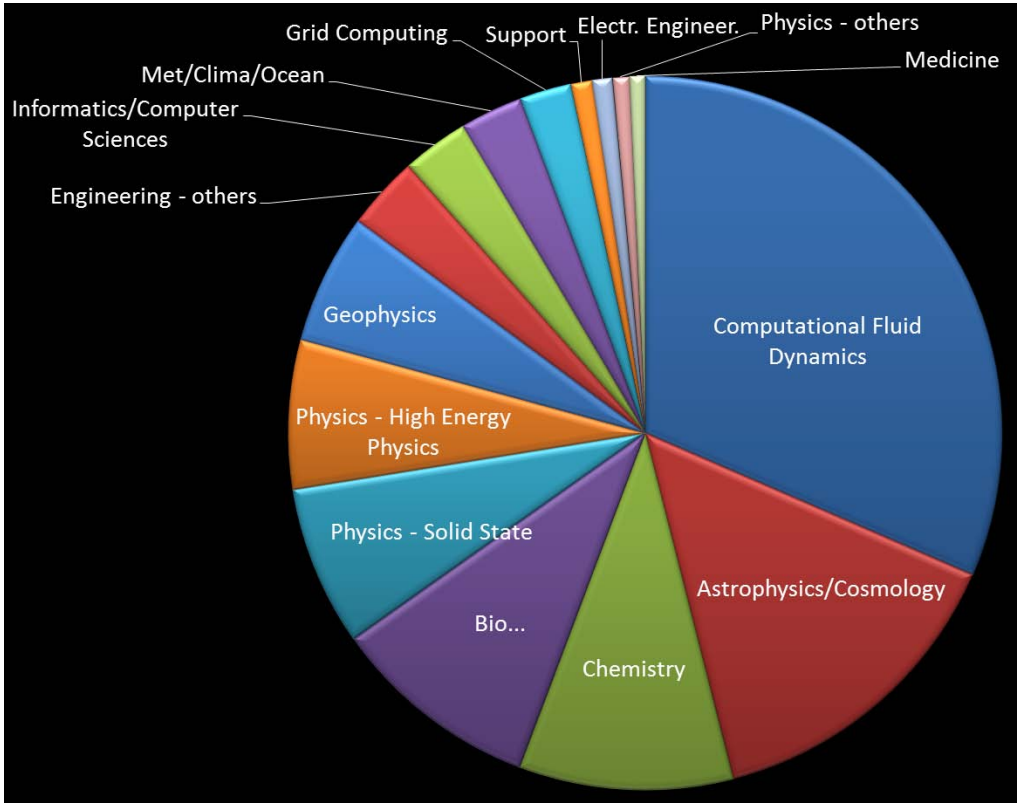
**IT Service Provider**:

- Munich Scientific Network (MWN)
- Web servers
- e-Learning
- E-Mail
- Groupware
- Special equipment:
  - Virtual Reality Laboratory
  - Video Conference
  - Scanners for slides and large documents
  - Large scale plotters

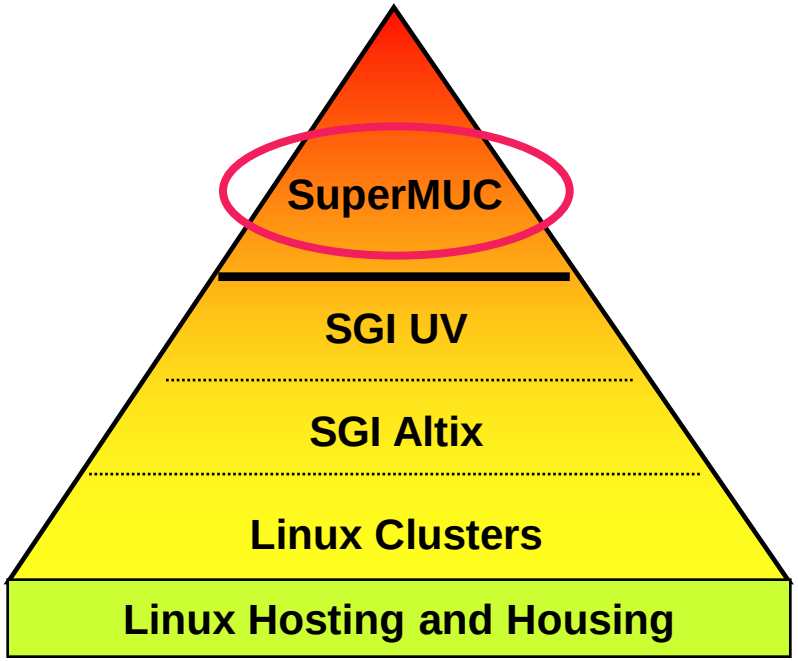
**IT Competence Centre**:

- Hotline and support
- Consulting (security, networking, scientifc computing, …)
- Courses (text editing, image processing, UNIX, Linux, HPC, …)

- European Supercomputing Centre
- National Supercomputing Centre
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities

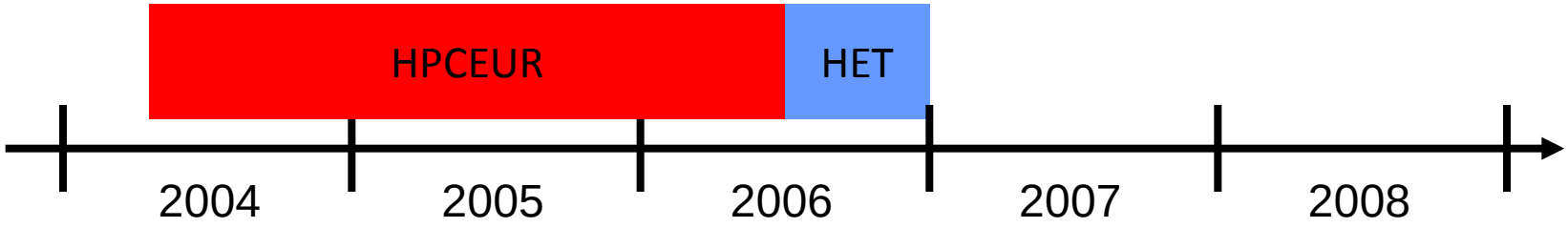


- European Supercomputing Centre
- National Supercomputing Centre
- Regional Computer Centre for all Bavarian Universities
- Computer Centre for all Munich Universities

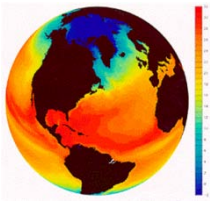
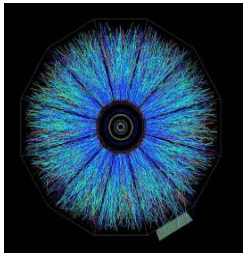
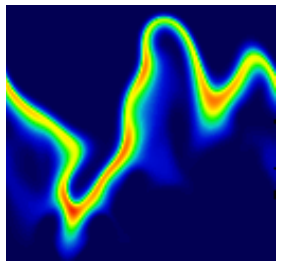
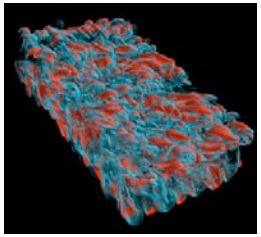
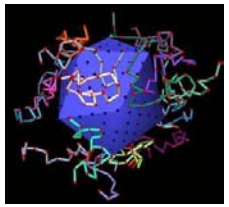
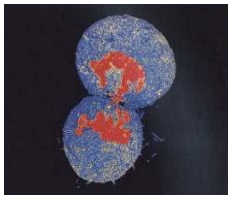
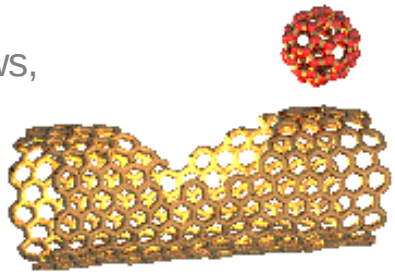
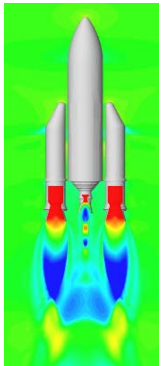
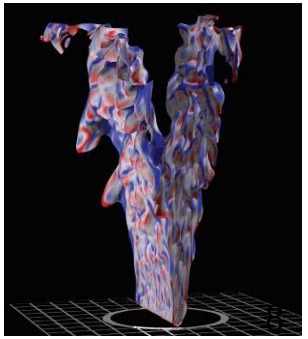


# PRACE Timeline

## PRACE – Partnership for Advanced Computing in Europe



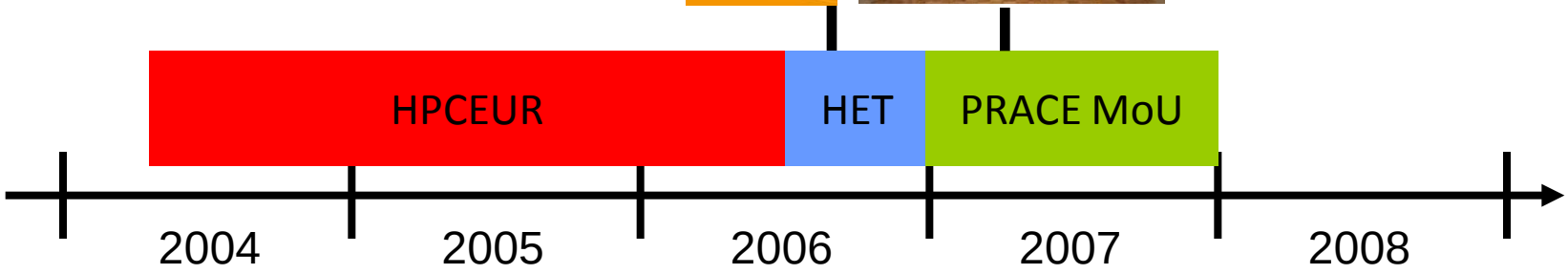
# The European Scientific Case

- Weather, Climatology, Earth Science
  - degree of warming, scenarios for our future climate.
  - understand and predict ocean properties and variations
  - weather and flood events
- Astrophysics, Elementary particle physics, Plasma physics
  - systems, structures which span a large range of different length and time scales
  - quantum field theories like QCD, ITER
- Material Science, Chemistry, Nanoscience
  - understanding complex materials, complex chemistry, nanoscience
  - the determination of electronic and transport properties
- Life Science
  - system biology, chromatin dynamics, large scale protein dynamics, protein association and aggregation, supramolecular systems, medicine
- Engineering
  - complex helicopter simulation, biomedical flows, gas turbines and internal combustion engines, forest fires, green aircraft,
  - virtual power plant

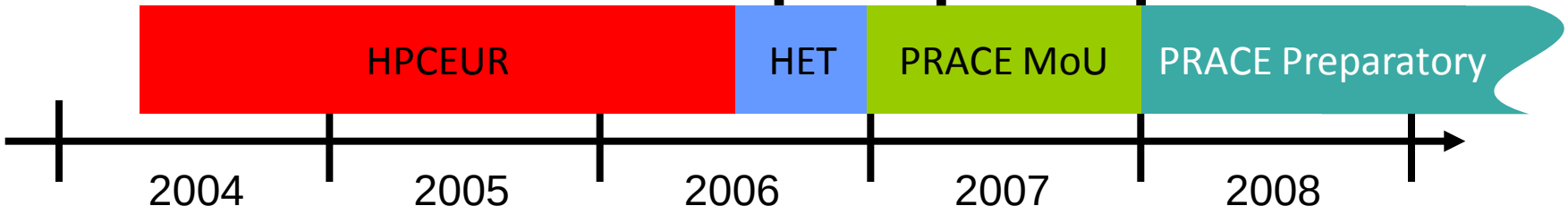
# PRACE Timeline

## PRACE – Partnership for Advanced Computing in Europe

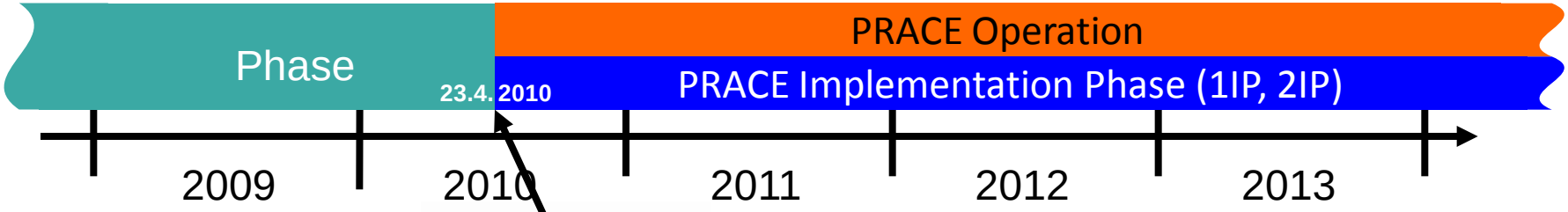



# PRACE Timeline

## PRACE – Partnership for Advanced Computing in Europe


EU-Grant: INFSO-RI-211528, 10 Mio. €





PRACE

**PRACE (AISBL), a legal entity with (current) seat location in Brussels**

# PRACE Research Infrastructure Created

- Establishment of the legal framework
  - PRACE AISBL created with seat in Brussels in April (Association Internationale Sans But Lucratif)
  - 20 members representing 20 European countries
  - Inauguration in Barcelona on June 9

# Gauss Centre for Supercomputing (GCS)



- Joint activity of the **3 German National HPC Centres**
  - John von Neumann Institut für Computing (NIC), Jülich
  - Leibniz Supercomputing Centre (LRZ), Garching near Munich
  - Höchstleistungsrechenzentrum Stuttgart (HLRS), Stuttgart

- **Largest and most powerful supercomputer infrastructure in Europe**

- Foundation of GCS (e.V.) April, 13th, 2007.
- Principal Partner in PRACE
  (Partnership for Advanced Computing in Europe)

# PRACE Research Infrastructure Created

- Establishment of the legal framework
  - PRACE AISBL created with seat in Brussels in April (Association Internationale Sans But Lucratif)
  - 20 members representing 20 European countries
  - Inauguration in Barcelona on June 9

- Funding secured for 2010 - 2015
  - 400 Million € from France, Germany, Italy, Spain
    Provided as Tier-0 services on TCO basis
  - Funding decision for 100 Million € in The Netherlands expected soon
  - 70+ Million € from EC FP7 for preparatory and implementation
    Grants INFSO-RI-211528 and 261557
    Complemented by ~ 60 Million € from PRACE members

# PRACE Tier-0 Systems

- 1st Tier-0 System
  - **Jugene**: BlueGene/P in GCS@Juelich
  - 72 Racks, 1 PFlop/s peak
  - 35% of capacity provided to PRACE

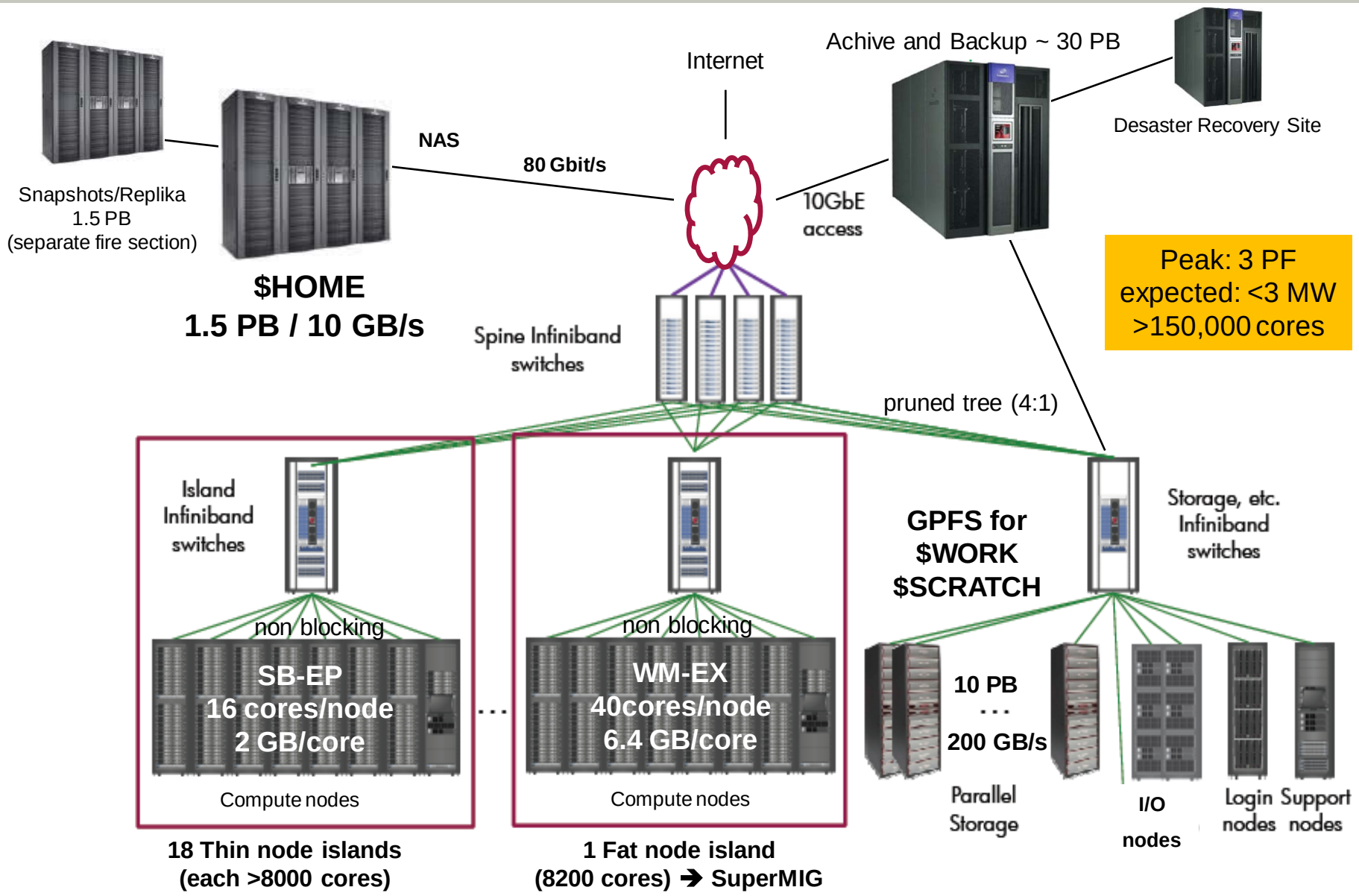

- 2nd Tier-0 System
  - **Curie**: Bull Cluster with Intel CPUs operated by CEA
  - 1.6 PFlop/s peak in Oct. 2011 (1st step in 10/2010)
  - Largest fraction of capacity provided to PRACE

- Next Procurements (in alphabetical order)
  - BSC, CINECA, GCS@HLRS, GCS@LRZ
  - Procurement plan  based on analysis of user requirements and market

# PRACE Tier-0 Access

- Single pan-European Peer Review
- Early Access Call in May 2010
  - 68 proposals asked for 1870 Million Core hours
  - 10 projects granted with 328 Million Core hours
  - Principal Investigators from D (5), UK (2) NL (1), I (1), PT (1)
  - Involves researchers from 31 institutions in 12 countries
- 1st Regular Call closed on August 2010
  - 58 proposals received asked for 2900 million core hours
  - 33 proposals have fulfilled the technical assessment
  - 360 million core hours available
    for a 12 months allocation period starting November 2010
- Further calls being scheduled (every 6 months)
  - 2nd regular call will include both Jugene and Curie

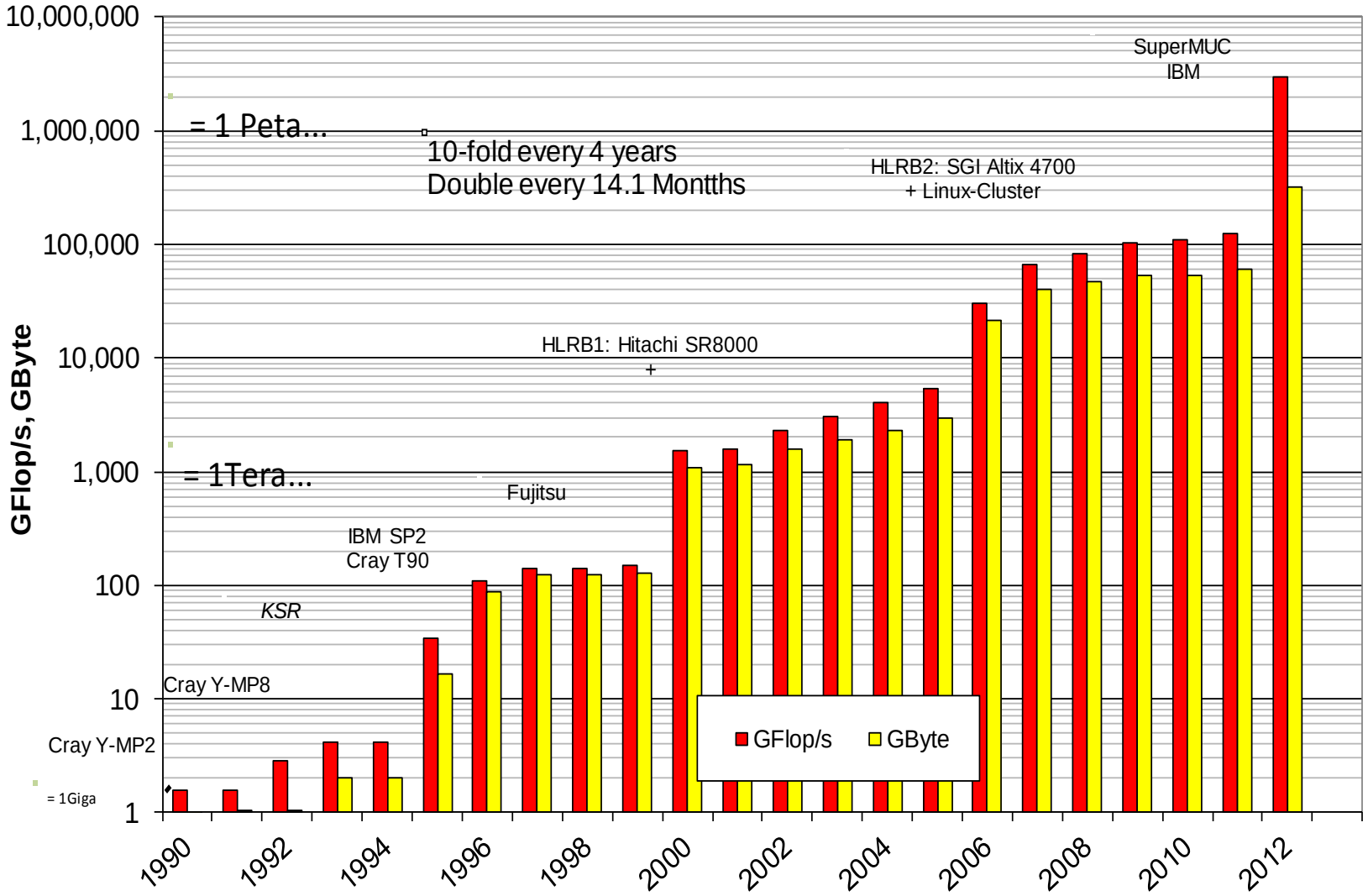
# PRACE Tier-0 @ LRZ: SuperMUC



# System Parameters SuperMUC

| | |
|---|---|
| Number of Islands (thin+fat) | 18+1 |
| Number of nodes | >9000 |
| Number of cores | >150000 |
| Processor types (thin + fat) | SB-EP + WM-EX |
| Total size of memory (TByte) | >300 |
| Expected electrical power consumption of total system (kW) | 2800 |
| Inlet temperature of compute node coolant (°C) | > 30 |
| Outlet temperature of compute node coolant (range) (°C) | 33 to 50 |
| Topology within an island | fully nonblocking |
| Topology between islands | pruned tree (4:1) |
| IB technology | FDR10 |
| Theoretical bisection bandwith of the entire system (GByte/s) | >11000 |
| Parallel file system type | GPFS |
| NAS user storage | NetApp |
| Size of parallel storage (Pbyte) | 10 |
| Size of NAS user storage (PByte) | 2 + (2 for Replica) |
| Aggregate theoretical bandwidth to/from SAN/DAS storage (GByte/s) | 200 |
| Aggregate theoretical bandwidth to/from NAS storage (GByte/s) | 10 |

# Energy Efficient Supercomputing at LRZ



# Financing Scheme

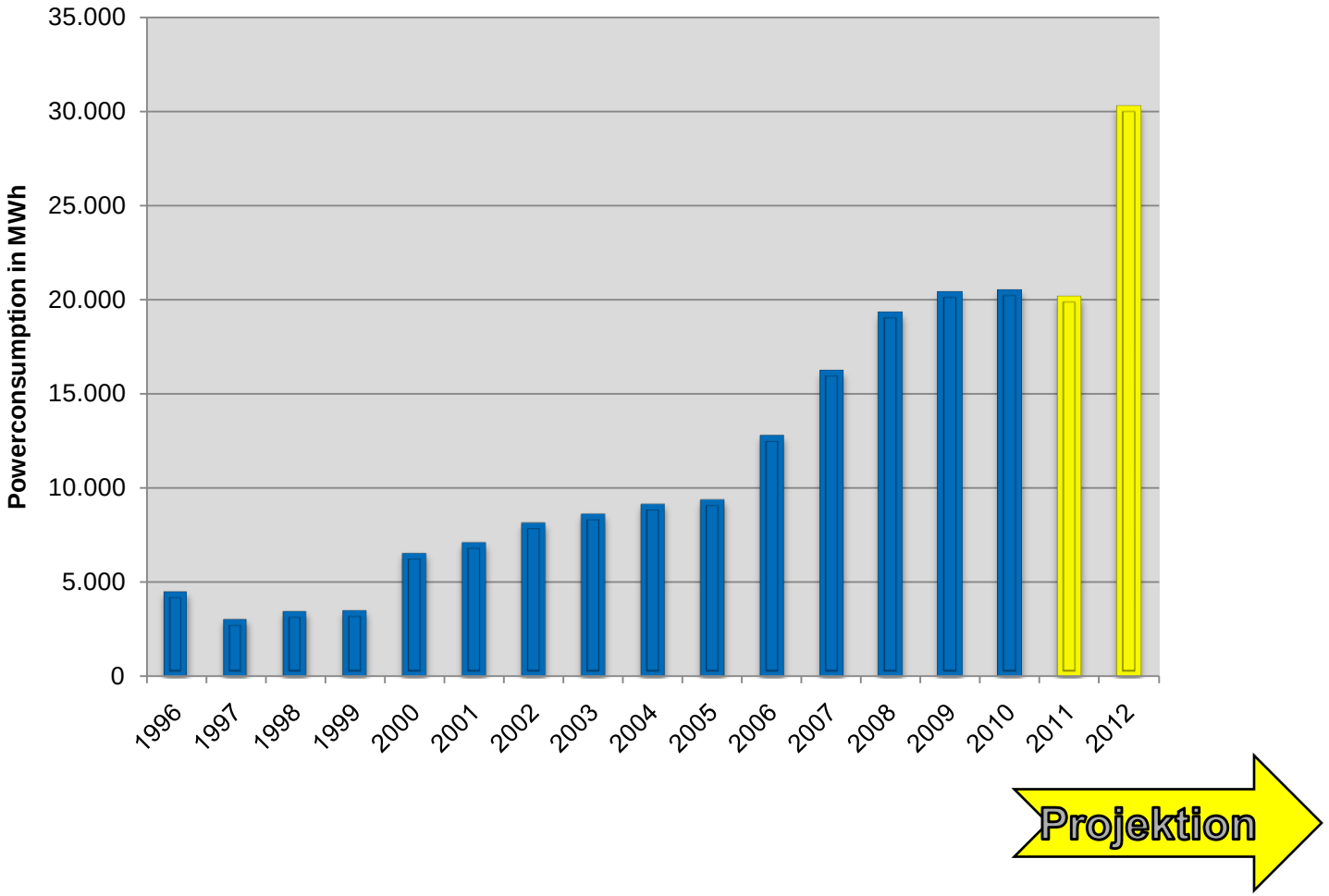


**for SuperMUC Investment and Operating Costs (gross, incl. VAT)**

| | 2010-2014 Phase 1 | 2014-2016 Phase 2 |
|---|---|---|
| High End System | | |
| Investment Costs (Hardware and Software) | 53 Mio € | ~ 19 Mio € |
| Operating Costs (Electricity costs and maintenance for hardware und software, some additional personnel) | 32 Mio € | ~ 29 Mio € |
| SUM | 85 Mio € | ~ 48 Mio € |
| Extension Buildings (construction and infrastructure) | 49 Mio € | |

Funding for Phase 2 is announced but not legally secured

# Energy Consumption at LRZ



# System Parameters SuperMUC

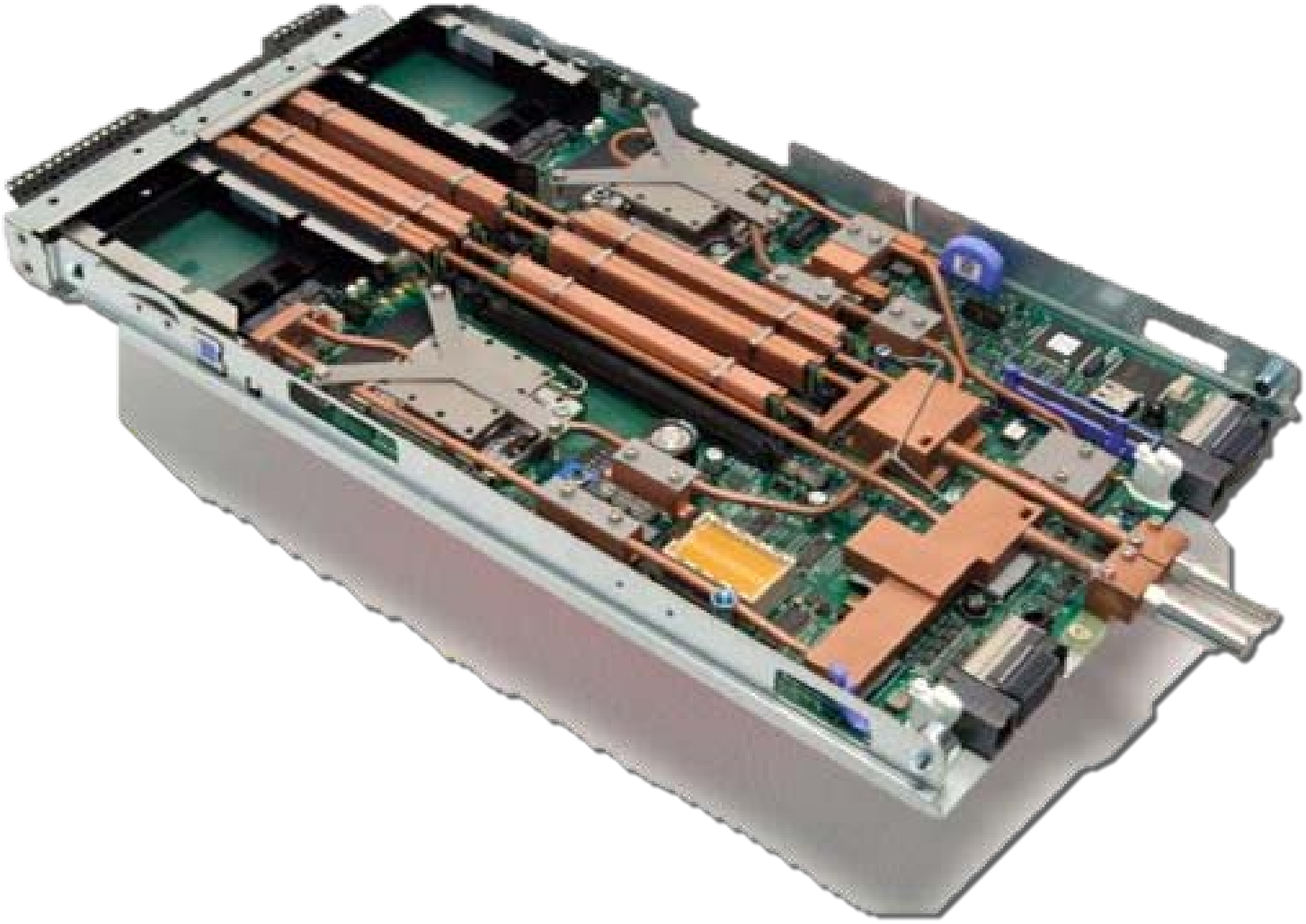
| | |
|---|---|
| Number of Islands (thin+fat) | 18+1 |
| Number of nodes | >9000 |
| Number of cores | >150000 |
| Processor types (thin + fat) | SB-EP + WM-EX |
| Total size of memory (TByte) | >300 |
| Expected electrical power consumption of total system (kW) | 2800 |
| Inlet temperature of compute node coolant (°C) | > 30 |
| Outlet temperature of compute node coolant (range) (°C) | 33 to 50 |
| Topology within an island | fully nonblocking |
| Topology between islands | pruned tree (4:1) |
| IB technology | FDR10 |
| Theoretical bisection bandwith of the entire system (GByte/s) | >11000 |
| Parallel file system type | GPFS |
| NAS user storage | NetApp |
| Size of parallel storage (Pbyte) | 10 |
| Size of NAS user storage (PByte) | 2 + (2 for Replica) |
| Aggregate theoretical bandwidth to/from SAN/DAS storage (GByte/s) | 200 |
| Aggregate theoretical bandwidth to/from NAS storage (GByte/s) | 10 |

# SuperMUC - Energy Efficient Supercomputing



- Probably **most powerful x86-system in Europe** (3PetaFlops peak)
- Use for science in Europe (PRACE), Germany (GCS) and Bavaria (KONWIHR)
- System with >150.000 cores, 324 TeraByte Main Memory
- **Most energy efficient General Purpose Supercomputer** in Europe in 2012
  - Hot liquid cooling
  - Reuse of waste heat
  - Hardware and software tools for clock scaling and optimization („dynamic frequency scaling“, „CPU throttling“ WIKIPEDIA)

# SuperMUC Thin Node Cooling



# Energy Efficiency at LRZ

- Measures around SuperMUC
  - New Contract (Spot Market / Evaluation of alternative Technologies)
  - Optimization of Building and Cooling Infrastructure (additional cooling loop)
  - Hot liquid cooling PUE < 1,1
  - Cooperation LRZ / TUM / LMU / IBM on Tools and Provider / User Strategies
  - Cooperation with Building Management YIT
- PRACE
  - 1 IP – Evaluation Prototype SGI Ultraviolet
  - 2 IP – Evaluation Prototype T-Platforms
- Exascale EU Project DEEP:
  - System based on „Accelerator – Architecture“ (Intel MIC)
  - Cooling and Prototype Evaluation
- Exascale EU Project Mont-Blanc
  - System based on low-power commercially available embedded CPUs
  - Next-generation HPC machine with a range of embedded technology
  - Software applications to run on this new generation of HPC systems

# The MNM-Team AppleTV Cluster



- 4+2 nodes, connected by a 100 MBit Ethernet switch



ATV2 Benchmarks:

- 4 nodes achieve $R_{max}$ of **160.4** MFlops/Watt
- Power consumption: about 10 Watts (all 4 nodes)
- → Energy efficiency of 16 MFlops/Watt
- Green500 list #500: 21 MFlops/Watt

# SuperMUC Procurement Goals

**To establish an integrated system and a programming environment which interact to enable the solution of the most challenging scientific problems from widely varying scientific areas in the least amount of time.**