# SuperMUC @ Leibniz Supercomputer Centre

IBM
lrz
SuperMUC
rendered on SuperMUC by LRZ

- Movie on YouTube

# Peak Performance

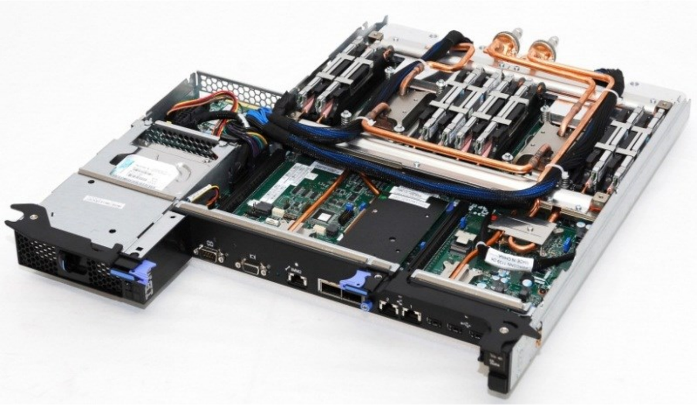
- Peak performance: 3 Peta Flops 3*$10^{15}$ Flops

| | | |
|---|---|---|
| Mega | $10^{6}$ | million |
| Giga | $10^{9}$ | billion |
| Tera | $10^{12}$ | trillion |
| Peta | $10^{15}$ | quadrillion |
| Exa | $10^{18}$ | quintillion |
| Zetta | $10^{21}$ | sextillion |

- Flops: Floating Point Operations per Seconds

# Distributed Memory Architecture

- 18 partitions called islands with 512 nodes
- Node is a shared memory system with 2 processors
  - Sandy Bridge-EP  
    Intel Xeon E5-2680 8C
    - 2.7 GHz (Turbo 3.5 GHz)
  - 32 GByte memory
  - Inifiniband network interface

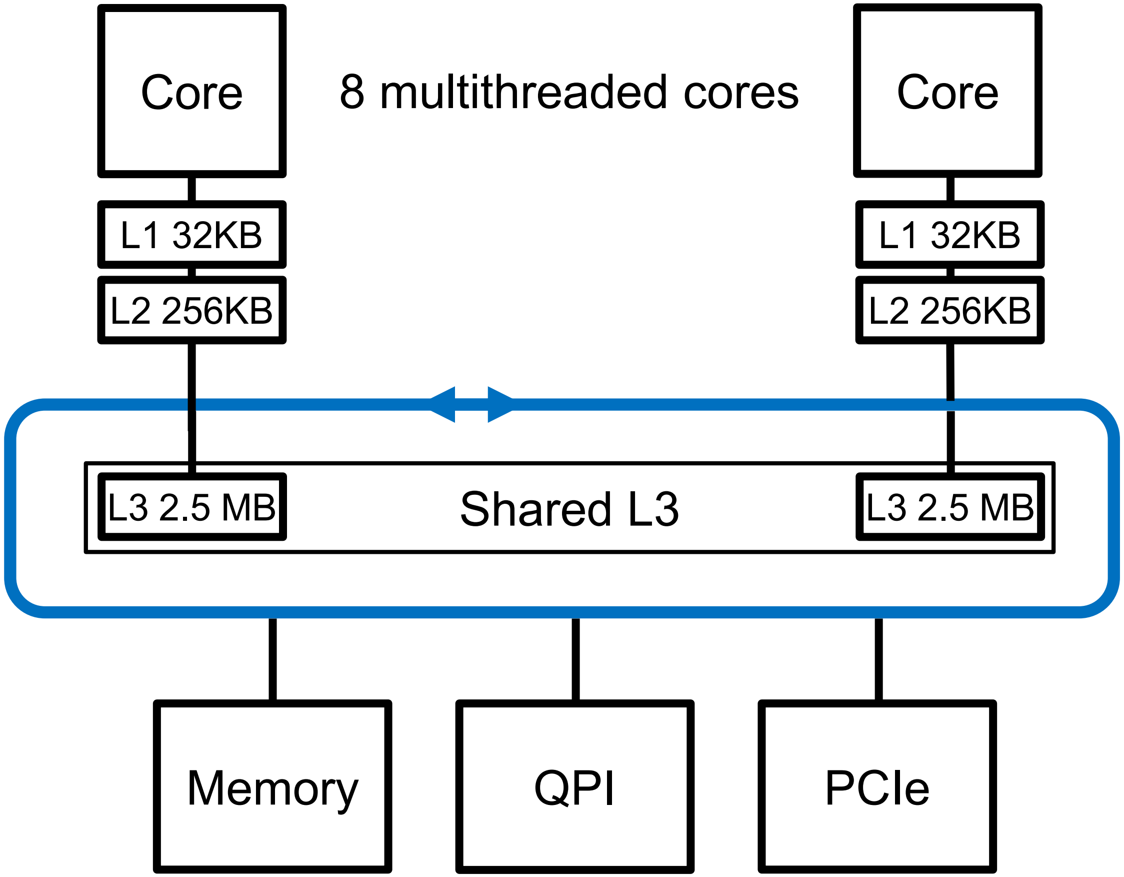


- Processor has 8 cores
  - 2-way hyperthreading
  - 21.6 GFlops @ 2.7 GHz per core
  - 172.8 GFlops per processor

# Sandy Bridge Processor

Latency:

- 4 cycles
- 12 cycles
- 31 cycles



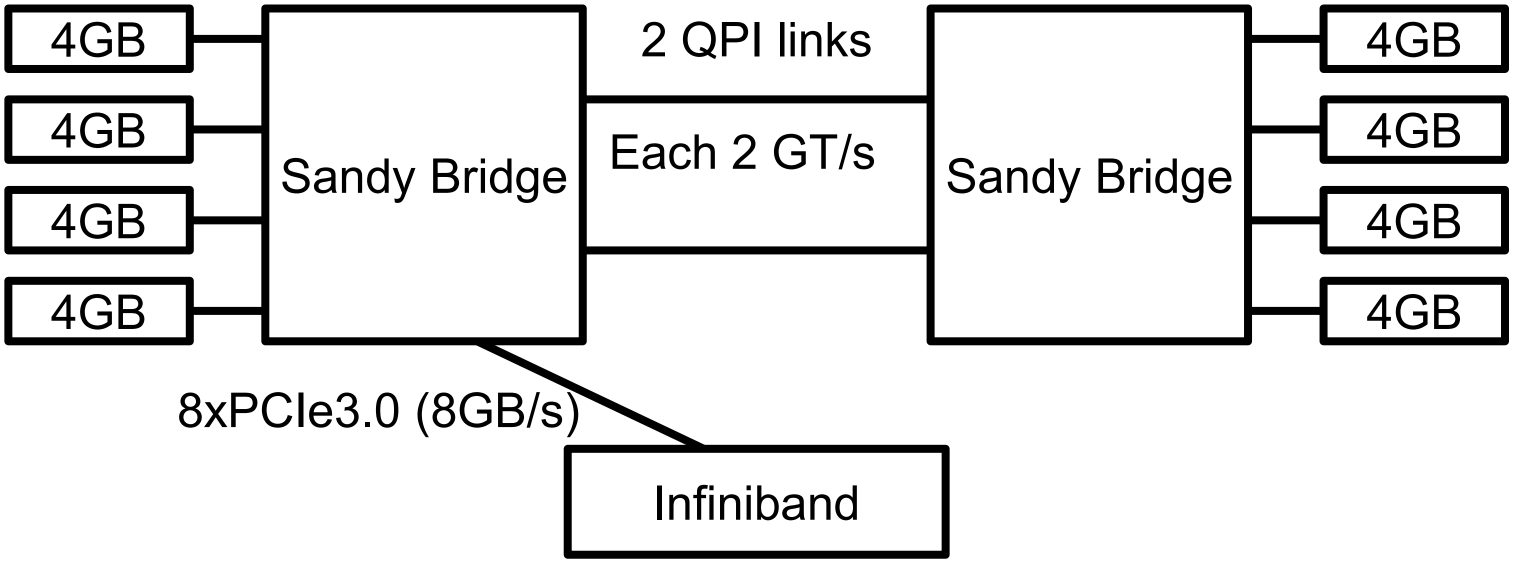
Bandwidth:

- 2*16/cycle
- 32 / cycle
- 32 / cycle

Network frequency
equal to core frequency

- L3 cache
  - Partitioned with cache coherence based on core valid bits
  - Physical addresses distributed by a hash function

# NUMA Node

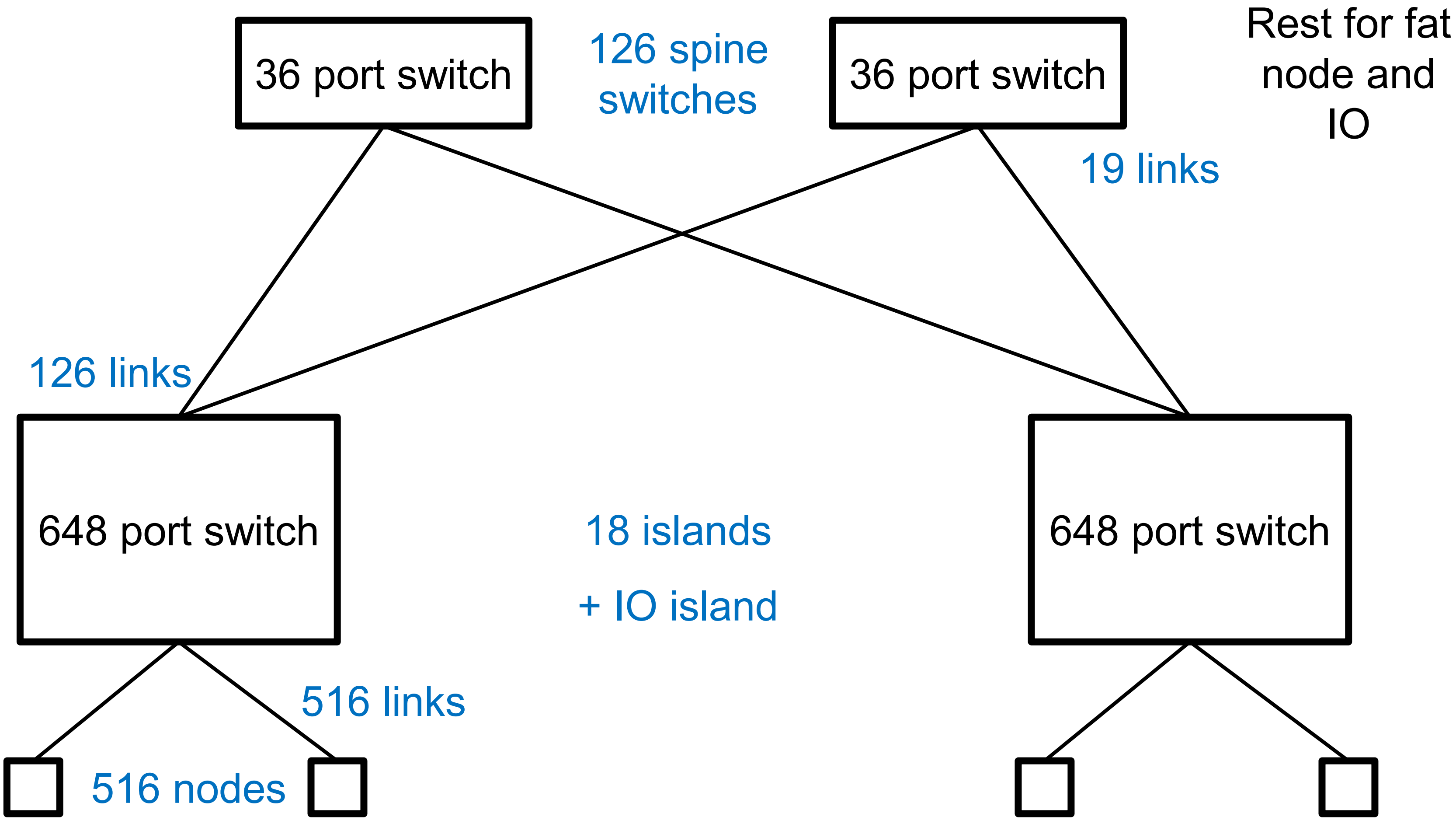


- 2 processors with 32 GB of memory
- Aggregate memory bandwidth per node 102.4 GB/s
- Latency
  - local ~50ns (~135 cycles @2.7 GHz)
  - remote ~90ns (~240 cycles)

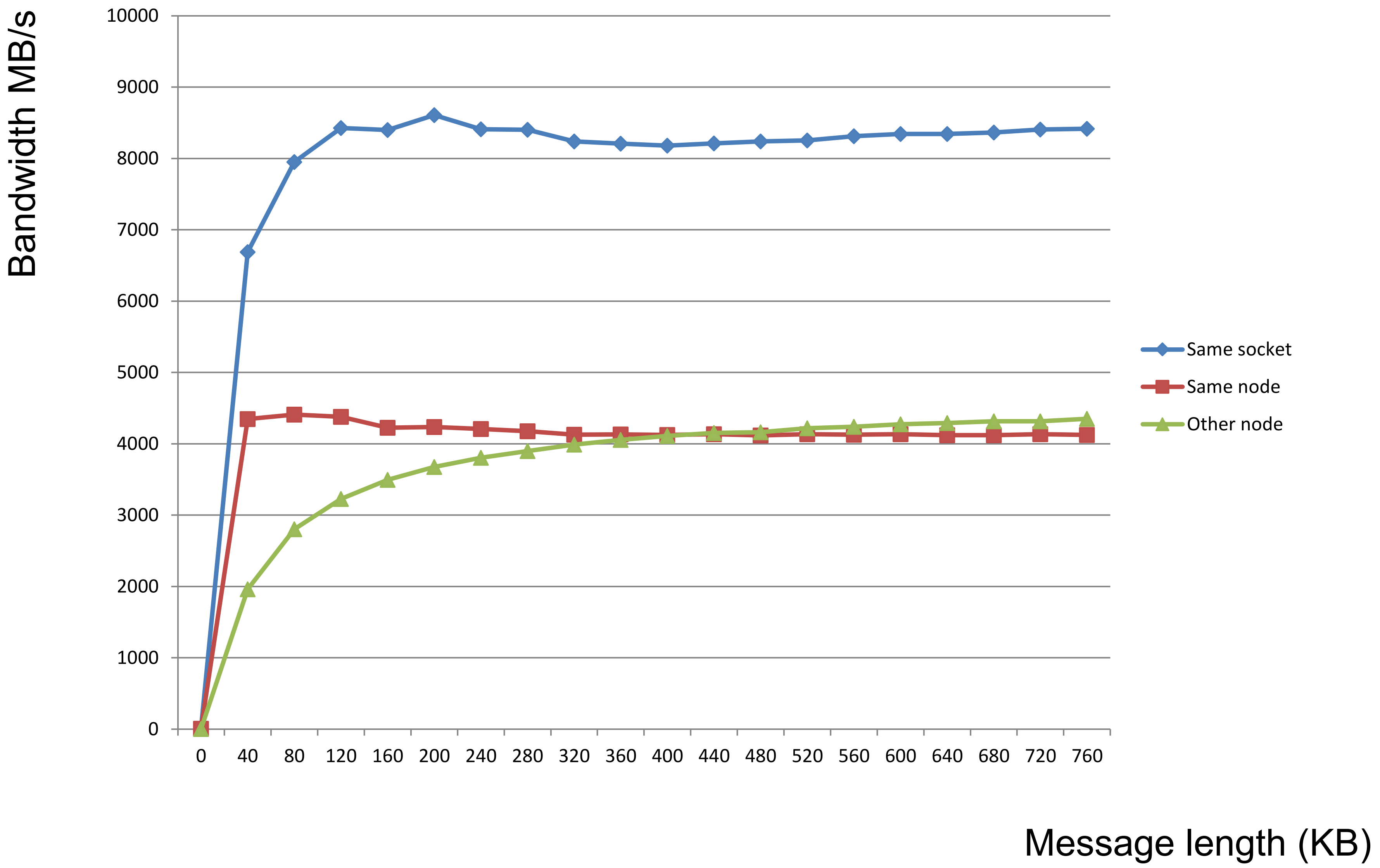
# Interconnection Network

- Infiniband FDR-10
  - FDR means Fourteen Data Rate
  - FDR-10 has an effective data rate of 38.79 Gbit/s
  - Latency: 100 nsec per switch, 1usec MPI
  - Vendor: Mellanox
- Intra-Island Topology: non-blocking tree
  - 256 communication pairs can talk in parallel.
- Inter-Island Topology: Pruned Tree 4:1
  - 128 links per island to next level

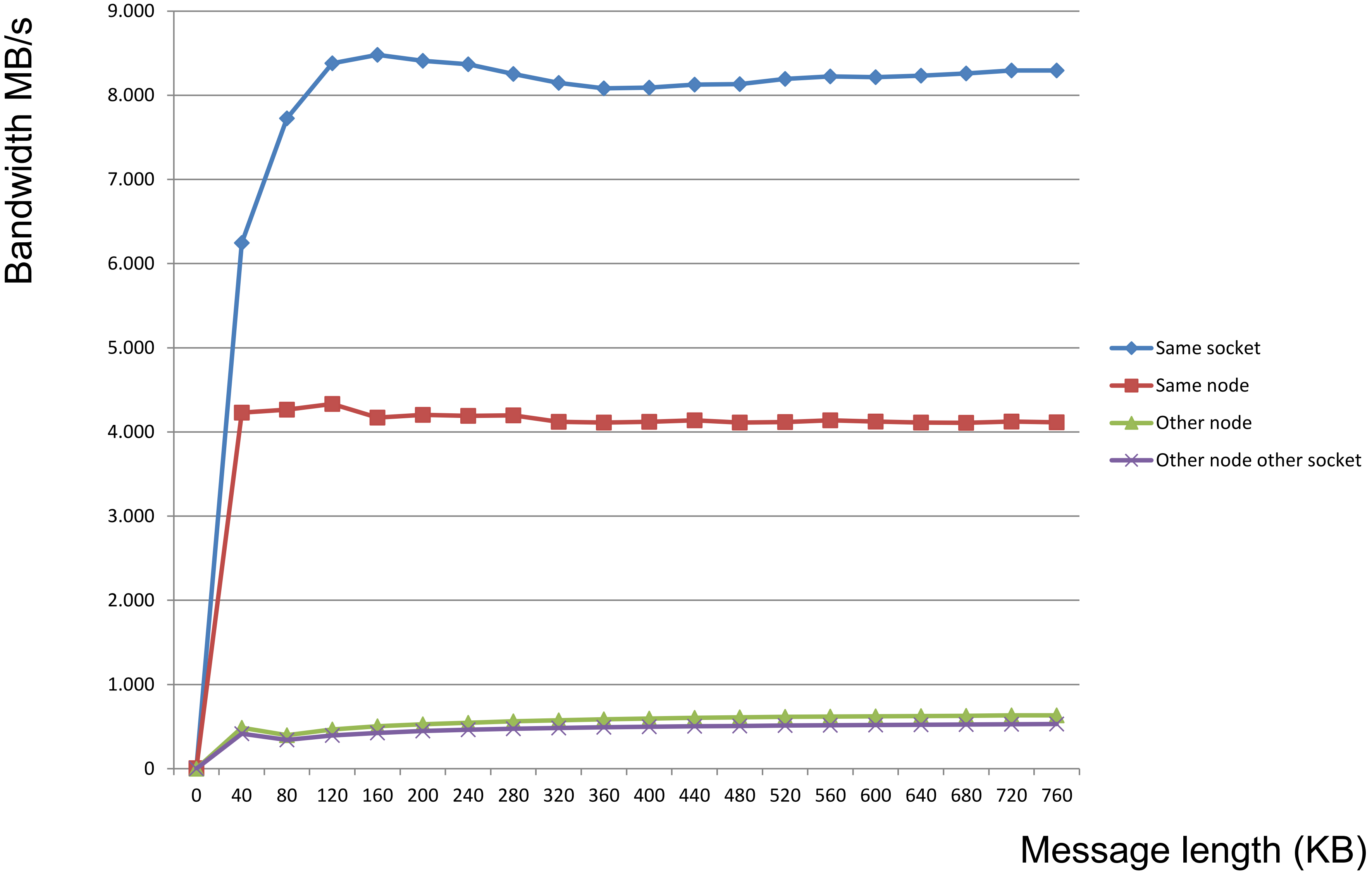
# Peak Performance



# MPI Performance – IBM MPI over Infiniband



# MPI Performance – IBM MPI over Ethernet



# 9288 Compute Nodes

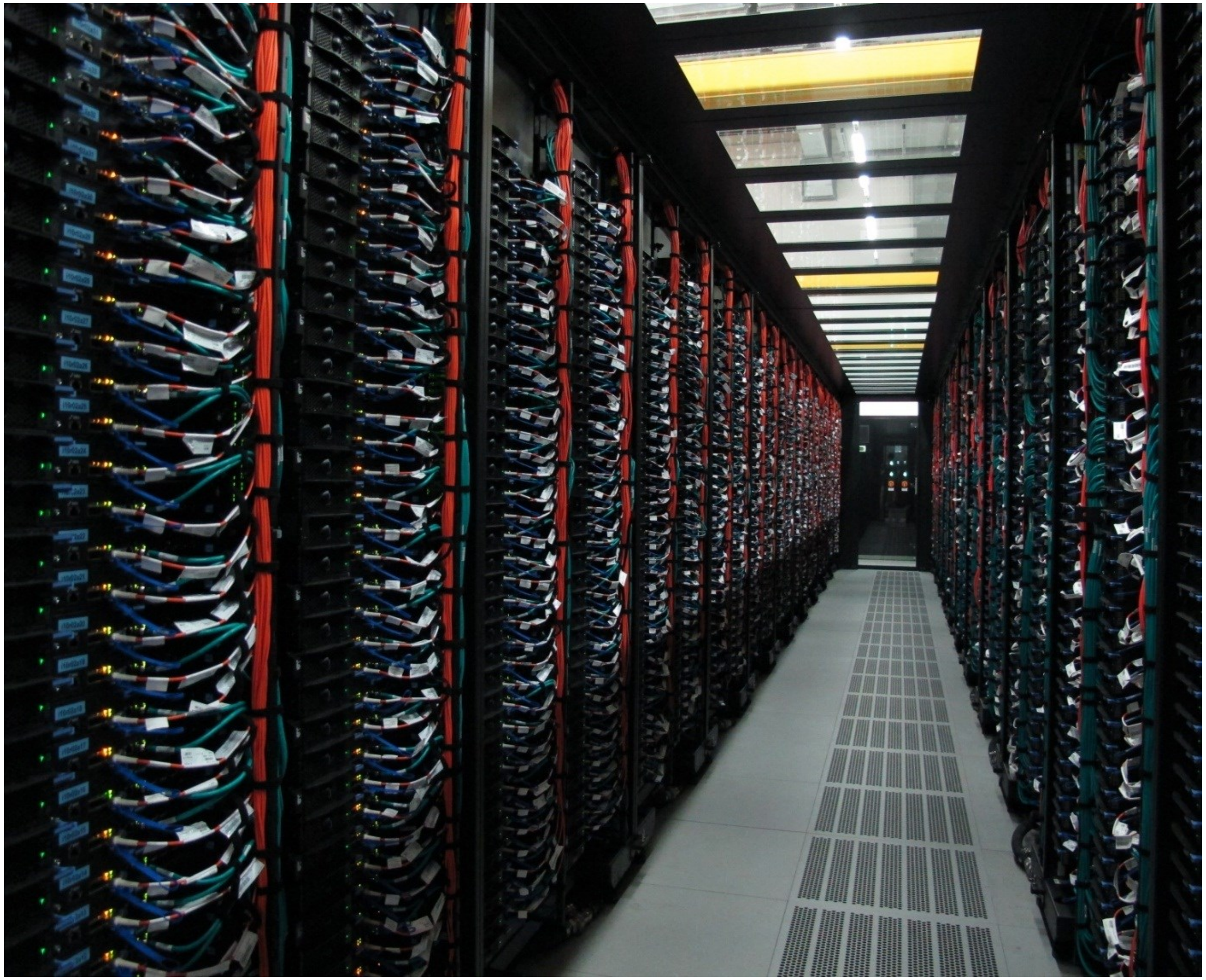
Cold Corridoor

Infiniband (red)

and

Ethernet (green)

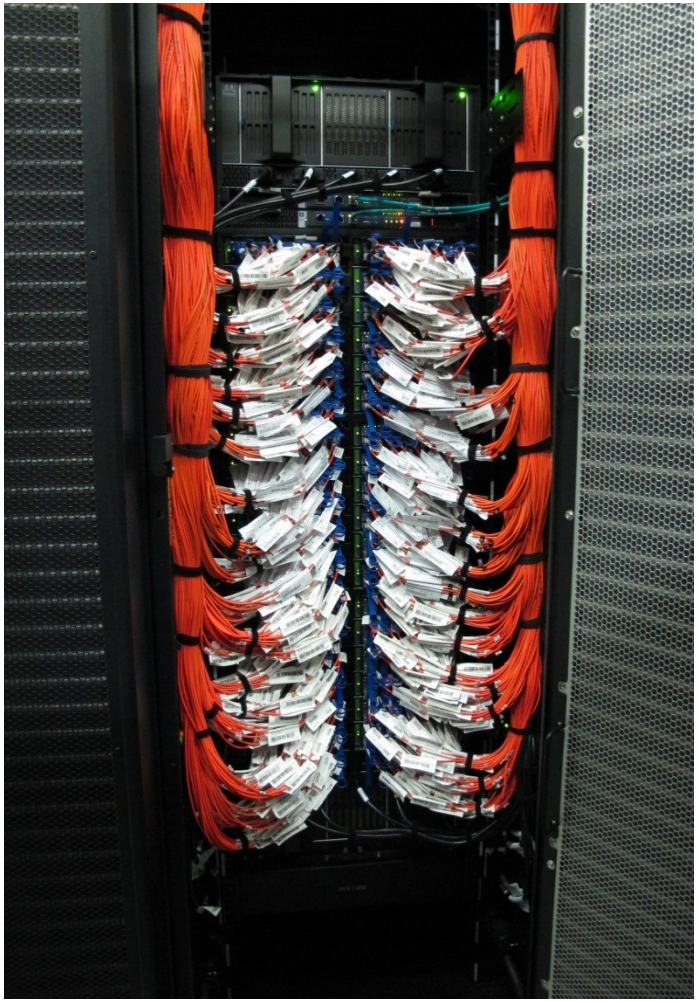
cabling



# Infiniband Interconnect

## 19 Orcas 126 Spine Switches

11900 Infiniband Cables



# Run jobs in batch

- Advantages
  - Reproducable performance
  - Run larger jobs
  - No need to interactive poll for resources

- Test queue
  - Max 1 island, 32 nodes, 2h, 1 job in queue
- General queue
  - Max 1 island, 512 nodes, 48 h
- Large
  - Max 4 islands, 2048 nodes, 48 h
- Special
  - Max 18 islands …

# Job Script

```
#!/bin/bash
#@ wall_clock_limit = 00:4:00
#@ job_name = add
#@ job_type = parallel
#@ class = test
#@ network.MPI = sn_all,not_shared,us
#@ output = job$(jobid).out
#@ error = job$(jobid).out
#@ node = 2
#@ total_tasks=4
#@ node_usage = not_shared
#@ queue
. /etc/profile
cd ~/apptest/application
poe appl
```

- llsubmit job.scp
  - Submission to batch system
- llq –u $USER
  - Check status of own jobs
- llcancel <jobid>
  - Kill job if no longer needed

# Limited CPU Hours available

- Please
  - Specify job execution as tight as possible.
  - Do not request more nodes than required. We have to „pay“ for all allocated cores, not only the used ones.
  - SHORT (<1sec) sequential runs can be done on the login node.
  - Even SHORT OMP runs can be done on the login node.

# Login to SuperMUC, Documentation

- First change the standard password
  - https://idportal.lrz.de/r/entry.pl
- Login via
  - lxhalle due to restriction on connecting machines
  - ssh <userid>@supermuc.lrz.de
  - No outgoing connections allowed
- Documentation
  - http://www.lrz.de/services/compute/supermuc/
  - http://www.lrz.de/services/compute/supermuc/loadleveler/
  - Intel compiler: http://software.intel.com/sites/products/documentation/hpc/composerxe/en-us/2011Update/cpp/lin/index.htm

# Batch Script Parameters

- #@ energy_policy_tag = NONE
  - Switch of automatic adaptation of core frequency for performance measurements
- #@ node = 2
- #@ total_tasks= 4
- #@ task_geometry = {(0,2) (1,3)}
- #@ tasks_per_node = 2
  - Limitations on combination documented at LRZ web page
- Use Intel MPI
  - `module unload mpi.ibm`
  - `module load mpi.intel`