Introduction to the Usage of High Performance Systems, Visualisation, Compute Cloud and Grid Facilities at LRZ

LRZ, 26.2.2015

Presenters

- Presenters:
  - Dr. Reinhold Bader (Group Leader HPC)
  - Dr. Ferdinand Jamitzky (Deputy Group Leader APP)
  - Dr. Anupam Karmakar (Application Support APP)
  - Dr. Helmut Satzger (Application Support APP)
  - Dr. Volker Weinberg (HPC)
  - Dr. Helmut Heller (Group Leader VER)
  - Dr. Christoph Biardzki (Group Leader ITS)
  - Dipl. Inf. Stephan Peinkofer (DAT)
  - Dr. Christoph Anthes (Team Leader V2T)

Time Table

10:00 - 10:30 Welcome and Introduction to LRZ (Jamitzky)
10:30 - 11:00 Access, Documentation, Service Desk (Karmakar)
11:00 - 11:45 User Interface (Bader)
11:45 - 12:45 Programming Environment (Weinberg)
Lunch
13:45 - 14:15 HPC Application Software (Satzger)
14:15 - 15:15 Visualisation and Virtual Reality (Anthes)
Coffee Break
15:30 - 16:00 Filesystems and Archive (Peinkofer)
16:00 - 16:30 Dedicated Resources: Housing and Virtual Servers (Biardzki)
16:30 - 17:00 Grid and Cloud (Heller)

What is a Supercomputer?

- It has overclocked high-speed processors? **NO**
- It has a big internal RAM? **NO**
- It runs Windows? **NO**
- CPU runs at higher frequency than a desktop PC? **NO**
- It will run my software without changes? **NO**
- It will run my software with millions of threads? **NO**
- It will run my old trusted executable? **NO**
- It will run Excel spreadsheets? **NO**

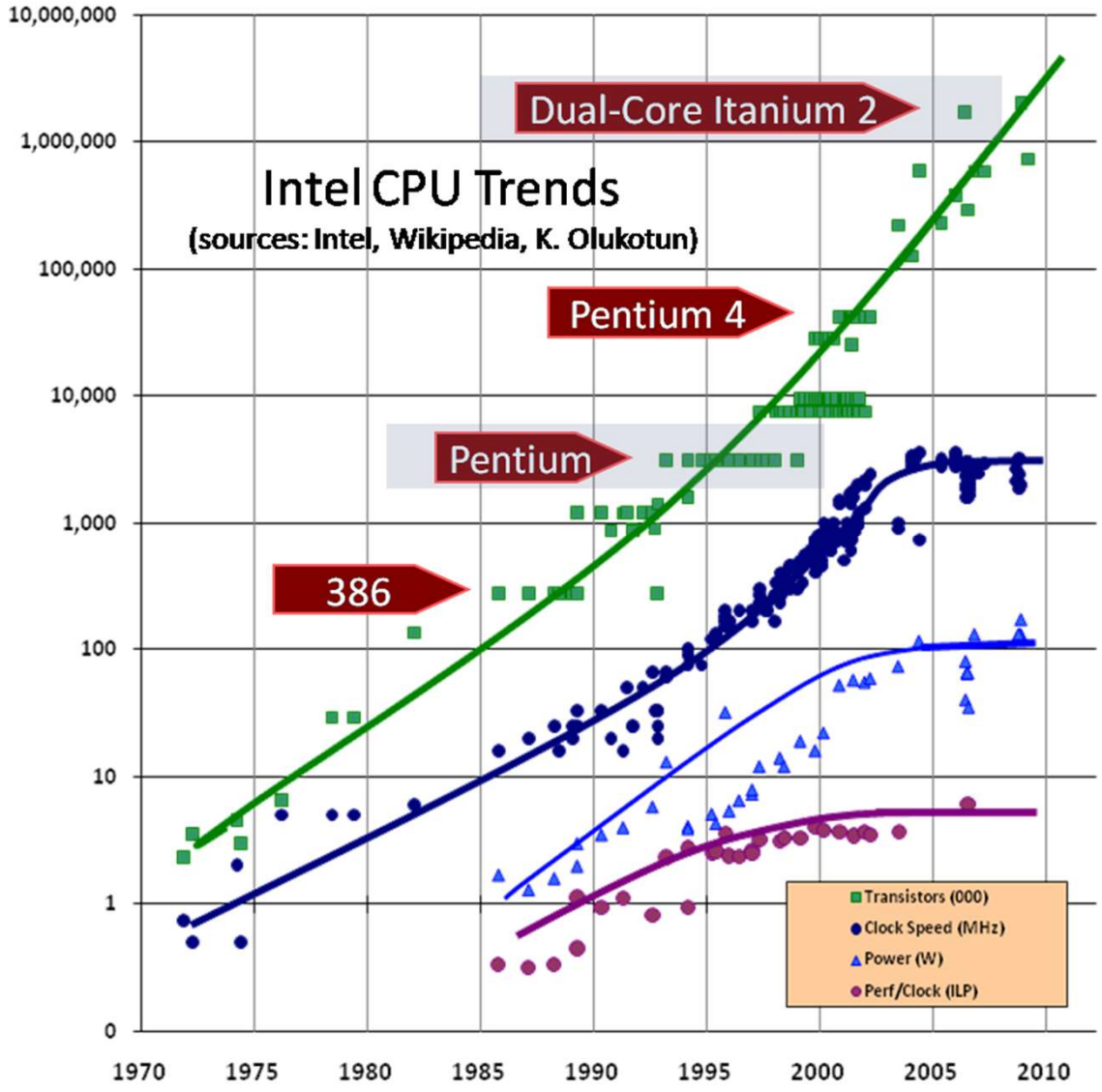
Why parallel programming?

End of the free lunch

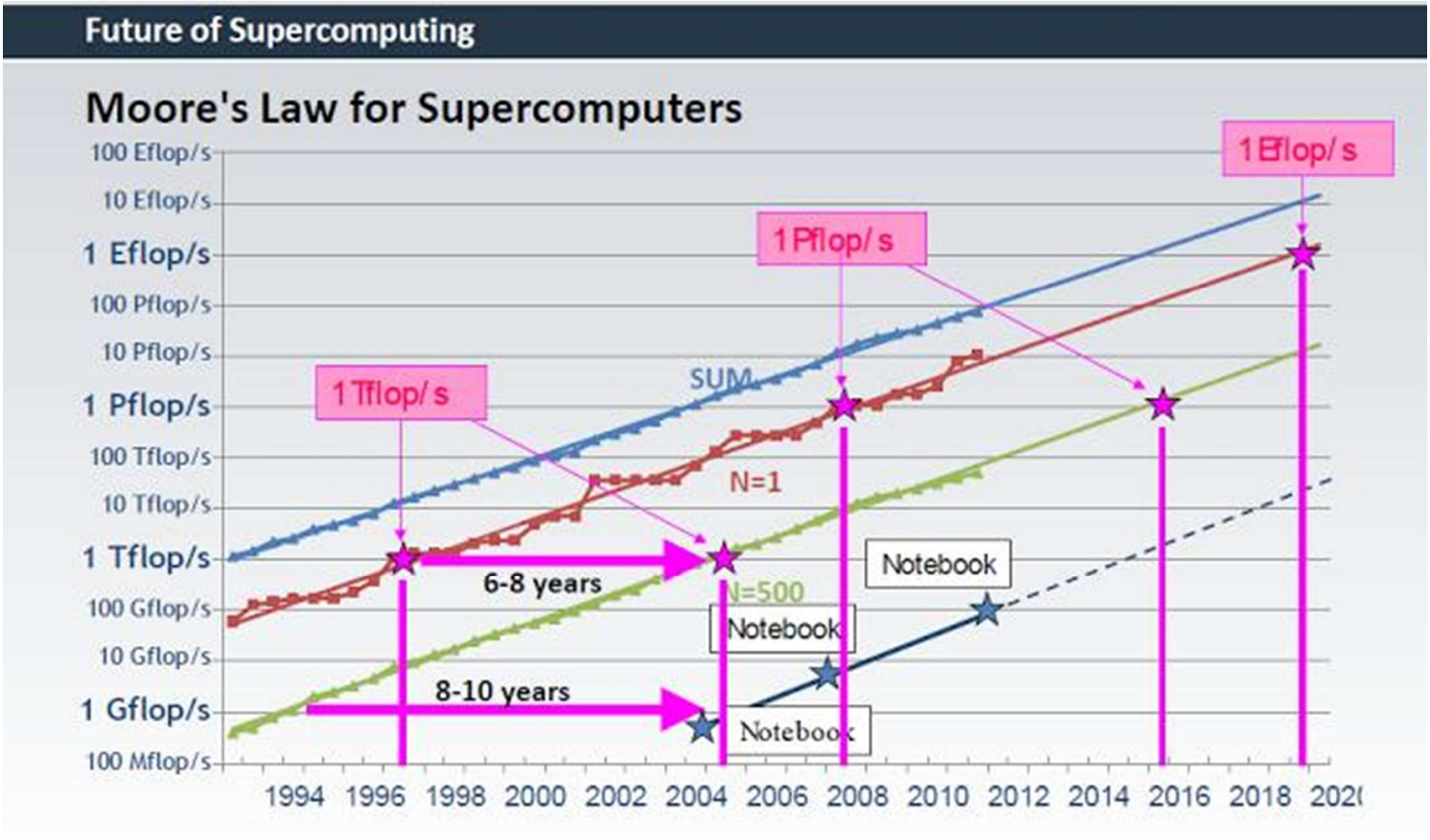
Moore's law means no longer faster processors, only more of them. But beware!

2 x 3 GHz < 6 GHz

(cache consistency, multi-threading, etc)



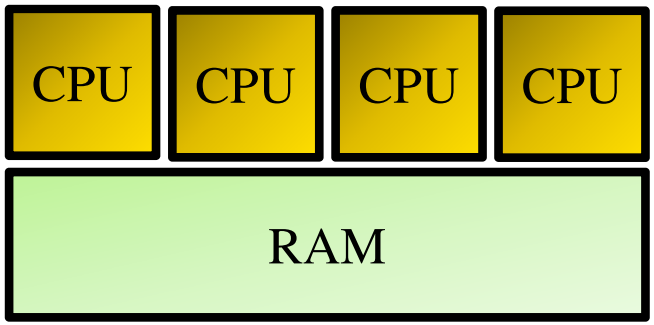
Supercomputer scaling



What is a Supercomputer?

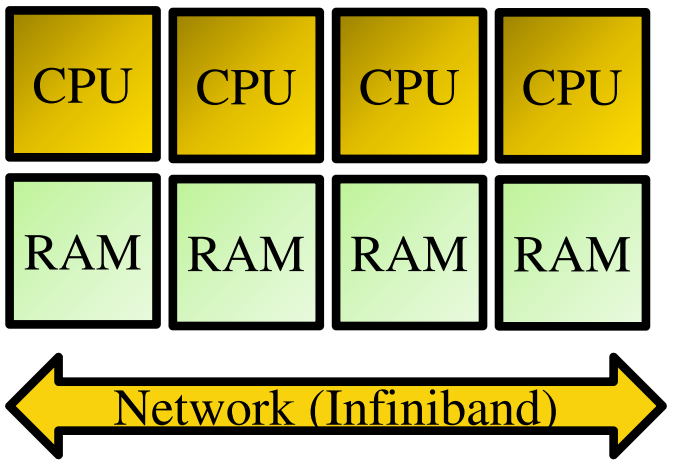
- It has many off-the-shelf CPUs and uses the CPU vector instructions (**AVX**)
- The CPUs are connected by a high-speed internal network (**Infiniband**)
- The compute nodes have to be programmed using **MPI** (Message Passing Interface)
- All nodes are connected to a parallel file system (**GPFS**) which needs special libraries (**MPI-I/O**)
- Programs cannot be run interactively, but have to be submitted to the batch system (**LoadLeveler**)

Shared and distributed memory



Shared Memory System
openMP, pthreads,...
Bandwidth: 96 Gb/s= 12 GByte/s
Latency: 0.06 us

```c
int main(void) {
  #pragma omp parallel
  printf("Hello, world.\n");
  return 0;
}
```



Distributed Memory System
MPI, sockets, GPI,...
Bandwidth: 56 Gb/s = 7 GByte/s
Latency: 0.7 us

```c
int main(int argc, char *argv[]) {
  MPI_Init(&argc,&argv);
  printf("Hello, world.\n");
  MPI_Finalize();
  return 0;
}
```

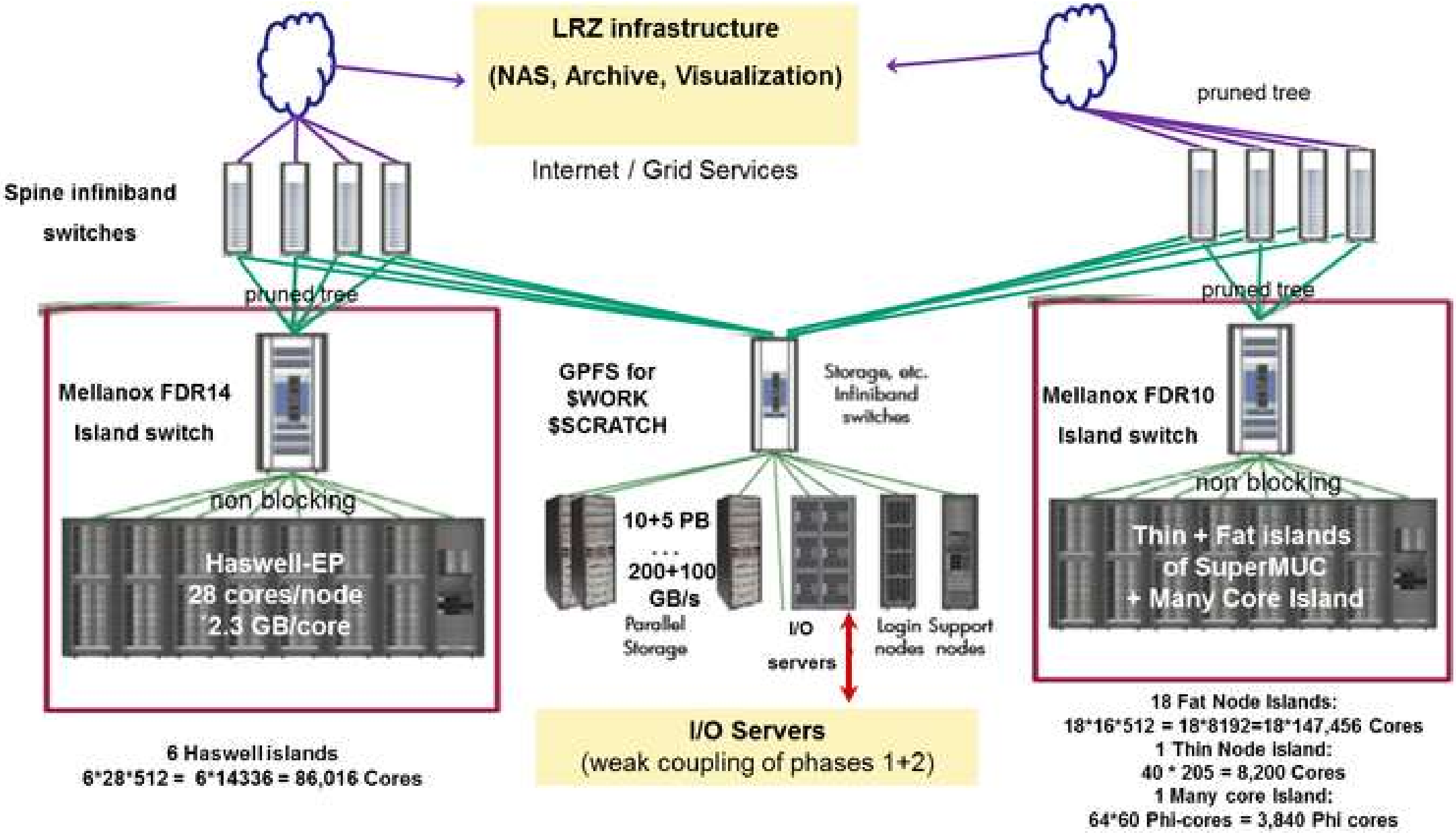
SuperMUC at LRZ

- SuperMUC consists of islands
- The following islands are available:

– SuperMUC fat node island	8,200 cores
– SuperMUC 18 thin nodes islands	147,456 cores
– SuperMIC Xeon Phi Island	3,840 cores
– SuperMUC 6 Phase 2 islands	86,016 cores
Total	**254,512 cores**

- Additionally a remote visualization cluster RVS is available

SuperMUC Phase 1+2



Linux Cluster at LRZ

- Linux Cluster consists of systems
- The following systems are available:
  - CoolMUC
  - SGI UV2/UV3
  - SGI ICE Cluster
  - Serial Nodes
  - Hugemem Nodes
  - GPGPU Cluster
  - Remote Visualisation Cluster GVS

Levels of Parallelism

- Node Level (e.g. SuperMUC has approx. 10,000 nodes)
- Accelerator Level (e.g. SuperMIC has 2 CPUs and 2 Xeon Phi Accelerators)
- Socket Level (e.g. fat nodes have 4 CPU Sockets)
- Core Level (e.g. SuperMUC Phase 2 has 14 cores per CPU)
- Vector Level (e.g. AVX2 has 16 vector registers per core)
- Pipeline Level (how many simultaneous pipelines)
- Instruction Level (instructions per cycle)

Problems: Access Times

Getting data from:	
CPU register	1 ns
L2 cache	10 ns
memory	80 ns
network(IB)	200 ns
GPU(PCIe)	50,000 ns
harddisk	500,000 ns

Getting some food from:	
fridge	10 s
microwave	100 s (ca. 2 min)
pizza service	800 s (ca. 15 min)
city mall	2,000 s (ca. 0.5 h)
mum sends cake	500,000 s (ca. 1 week)
grown in own garden	5,000,000 s (ca. 2 months)