

Seminar Aspects of Distributed Systems

Implementation of Energy Efficiency in HPC

Introduction, Impulse Presentation (07.04.2016)

About Us

Prof. Dr.-Ing. Gabriele von Voigt

- Managing Director Leibniz Universität IT Services
- Institute of Distributed Systems



M.Sc. Fabian Pflug

- Research Assistant [DCSec](#)

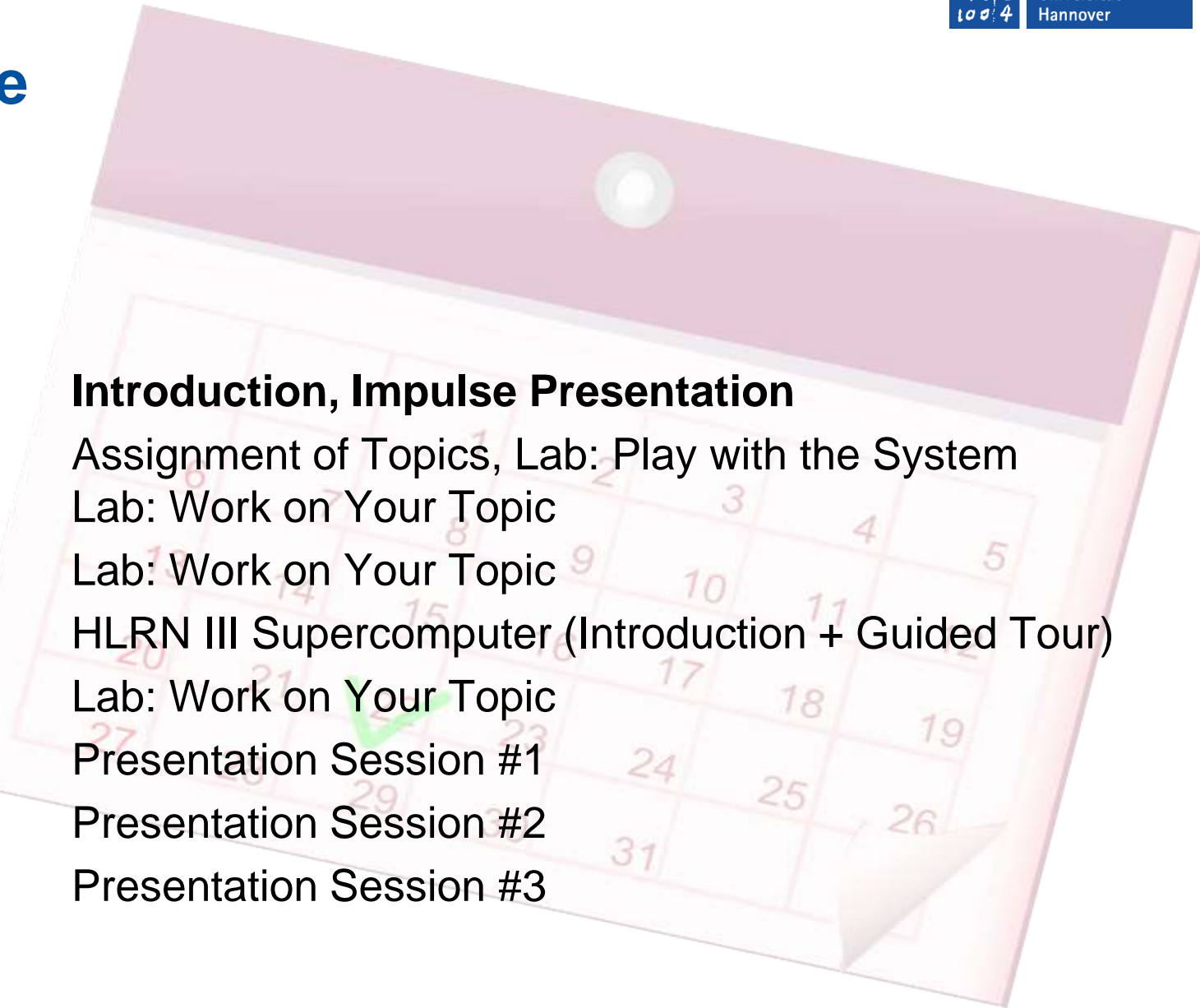


Dipl.-Inform. Hinrich Tobaben

- Relevant industry experience
 - Several years a software developer
 - More than 20 years IT-Architect & IT-Consultant
- Research Manager DCSec



Schedule

- 
- **07.04.16** **Introduction, Impulse Presentation**
 - 14.04.16 Assignment of Topics, Lab: Play with the System
 - 21.04.16 Lab: Work on Your Topic
 - ... Lab: Work on Your Topic
 - xx.xx.16 HLRN III Supercomputer (Introduction + Guided Tour)
 - ... Lab: Work on Your Topic
 - 30.06.16 Presentation Session #1
 - 07.07.16 Presentation Session #2
 - 14.07.16 Presentation Session #3

Points / Area of Expertise etc.

- Credit Hours 2, 3 Credit Points (SWS 2, 3 LP)
- Area of Expertise (Kernkompetenzbereiche)
 - M.Sc. Computer Science: KKB_NVS and KKB_Sys
 - M.Sc. Technical Computer Science: KKB_Sys
- Bachelor Students
 - are allowed to participate
 - Master Students have priority to participate
 - Certificate (Schein) can be used later

What We Expect

- Active participation
- Each student has to work on an assigned topic
 - Presentation (German or English)
 - 30 minutes + 15 minutes Q&A
 - Paper (Abstract 1-2 Pages), longer paper is an option
- Contact us with questions or problems at an early stage
- Goals and objectives

The seminar participants will gain in-depth knowledge in their allocated topic. Furthermore, the presentation should give the other participants an overview of the selected topic. Participants improve their ability to work academically. The procedure is similar as when you publish in a scientific environment.

What We Offer



- ODROID-XU3 Lab Environment
- References
- Very good and easy supervision
- [Presentation Template](#)

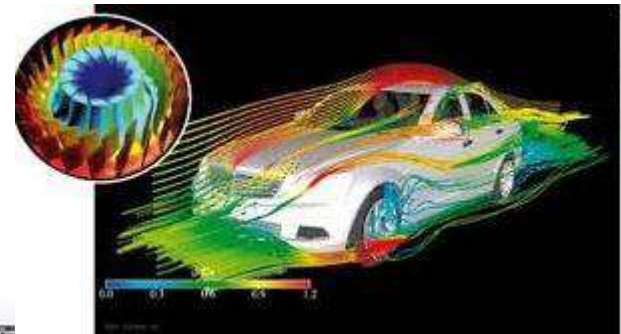
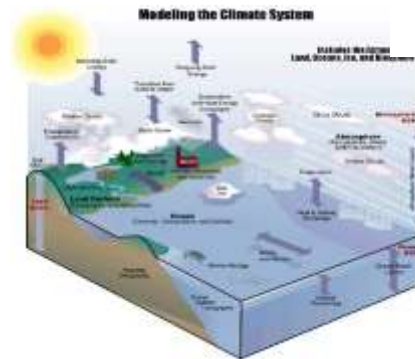
■ Grading:



- Paper and presentation (slides, presentation skills, etc.)
 - Classification of the assigned topic, useful structure
 - Clear statements, easy to understand explanations and helpful illustrations
 - Appropriate number of quotes
- Lab work, active participation

Why HPC?

- Simulations & Calculations
- Climate Models
- Chemistry
- Risk Analysis
- Crash Simulation
- And much more

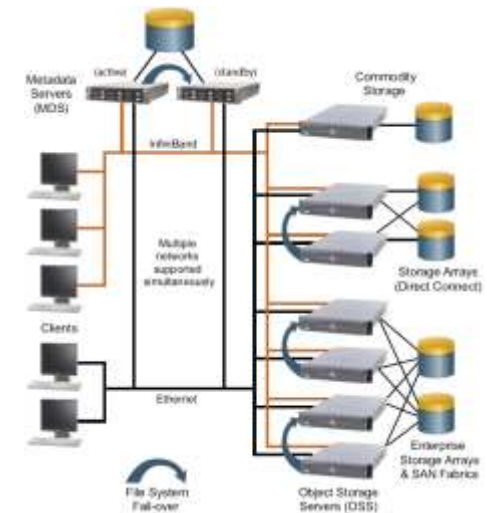
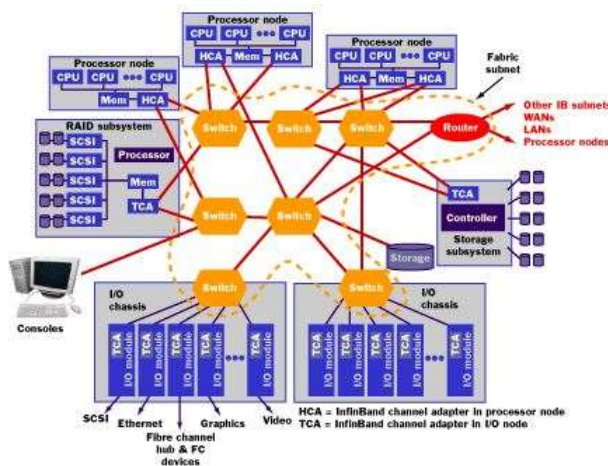


Classification

- Growing Market
- Very innovative
- Challenging & Fascinating
- Niche

HPC: What are the Challenges? 1/2

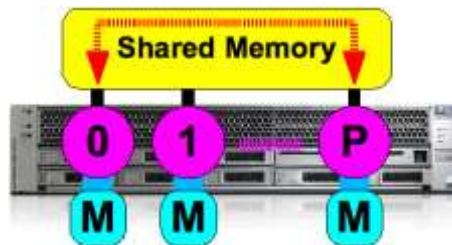
- Just the latest technology is good enough
- Server
 - Heterogeneous (CPU, GPU, etc.)
 - Dark Silicon <http://darksilicon.org/>
- Network, Storage
 - Latency, throughput



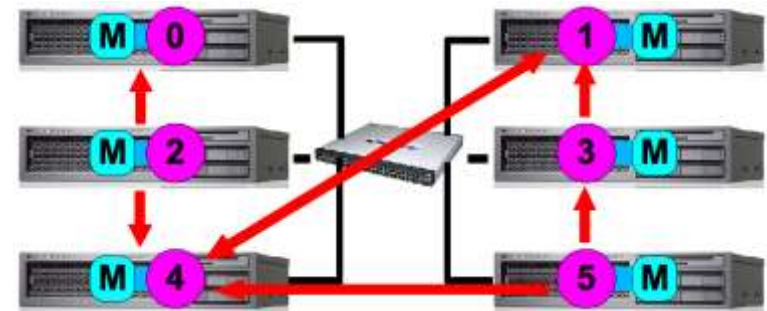
HPC: What are the Challenges? 2/2

- Software
 - Parallel Programming, Scaling

OpenMP



OMPI

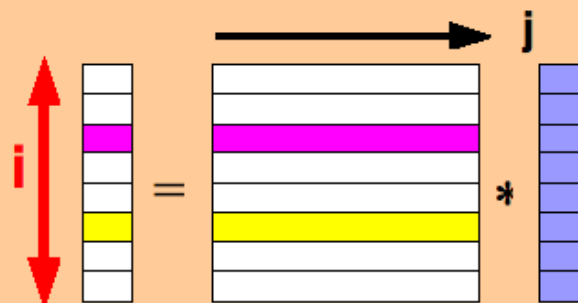


- Large Systems <http://www.top500.org/>
- **Extreme Power Consumption**  **Energy Efficiency**

Example – Matrix * Vector

```
#pragma omp parallel for
for (int i=0; i<m; i++)
{
    double sum = 0.0;
    for (int j=0; j<n; j++)
        sum += b[i][j]*c[j];
    a[i] = sum;
}
```

← parallel loop



Thread 0

for (i=0,1,2,3,4)

i = 0

sum = $\sum b[i=0][j]*c[j]$

a[0] = sum

i = 1

Thread 1

for (i=5,6,7,8,9)

i = 5

sum = $\sum b[i=5][j]*c[j]$

a[5] = sum

i = 6

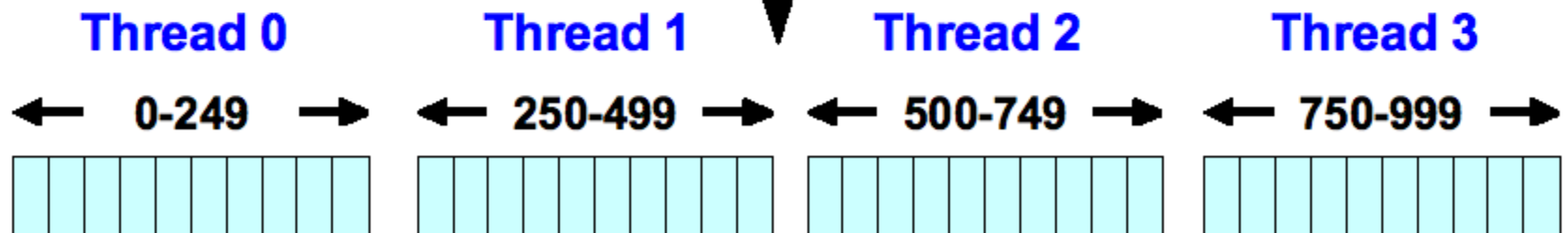
... etc ...

Automatic Parallelization by the Compiler

- Compiler generates parallel code for loops
- Different iterations are executed in parallel
- Binary is independent of the number of threads

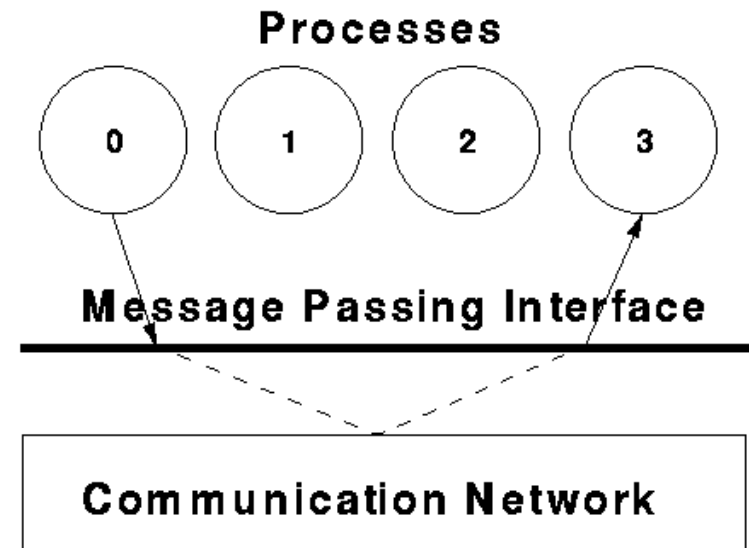
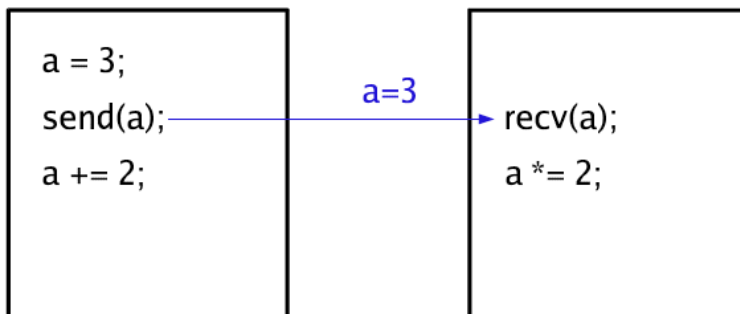
```
for (i=0; i<1000; i++)  
    a[i] = b[i] + c[i];
```

OMP_NUM_THREADS=4

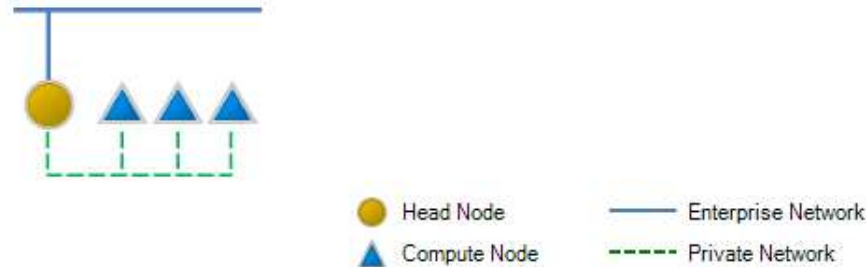


What is

- M P I = **M**essage **P**assing **I**nterface
- Standard interface for writing parallel programs
- Data is exchanged in the form of messages



Architecture



- Head Nodes (or Front End)
- Compute Nodes
 - different job classes / Hardware mapping
 - MPP (**M**assively **P**arallel **P**rocessing)
 - Small Nodes: 2 sockets , 24 Cores, 64GB, High-speed network
 - Scaling across **many** nodes
 - SMP (**S**ymmetric **M**ulti**P**rocessing)
 - Fat Nodes: 4 and more sockets, 32 Cores, a lot of Memory (256GB and more)
 - Parallelization by fork, exec, OpenMP within the node
 - Local Storage (scratch space)

Running Jobs

Workload Management

- Software for managing jobs and resources on SMP and distributed systems
- Examples: Moab, Univa Grid Engine, LSF



- More than one kind of processor
- Gain performance or energy efficiency not just by adding the same type of processors
- Specialized processing capabilities to handle particular tasks



Heterogeneous Computing Challenges

- Programming:
 - Parallel programming more complex than serial
 - Knowledge of architecture needed
 - Performance prediction without actual implementation difficult
- Hardware:
 - Costs for data transfer via PCI Express
 - More complex to build than homogeneous

NVIDIA



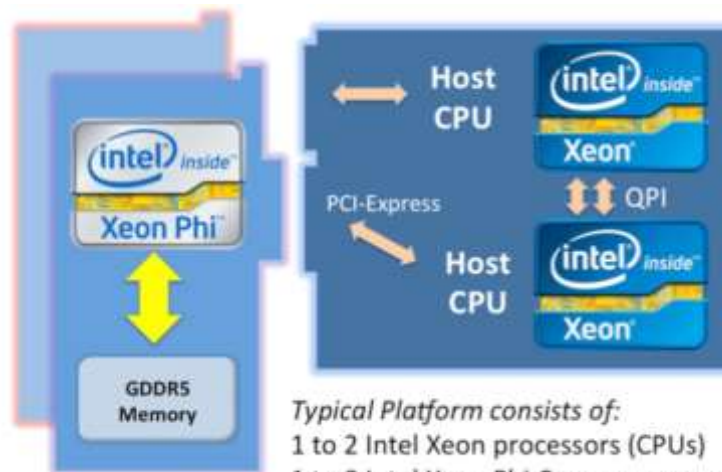
- NVIDIA Kepler™ compute architecture
- K40: up to 2880 Cores & 1.43 TFLOPS
- Programming
 - CUDA
 - NVIDIA's Architecture for parallel computing
 - OpenCL (Open Computing Language)
 - OpenACC (Open Accelerators)
 - Compiler directive `#pragma acc ...`





Intel Phi

- Intel® Many Integrated Core-Architecture
- Up to 61 Cores, 244 Threads and 1.2 TFLOPS
- Programming
 - Intel Fortran and C /C++ compiler
 - OpenMP



Typical Platform consists of:
1 to 2 Intel Xeon processors (CPUs)
1 to 8 Intel Xeon Phi Coprocessors per host

PEZY-SC Many Core Processor

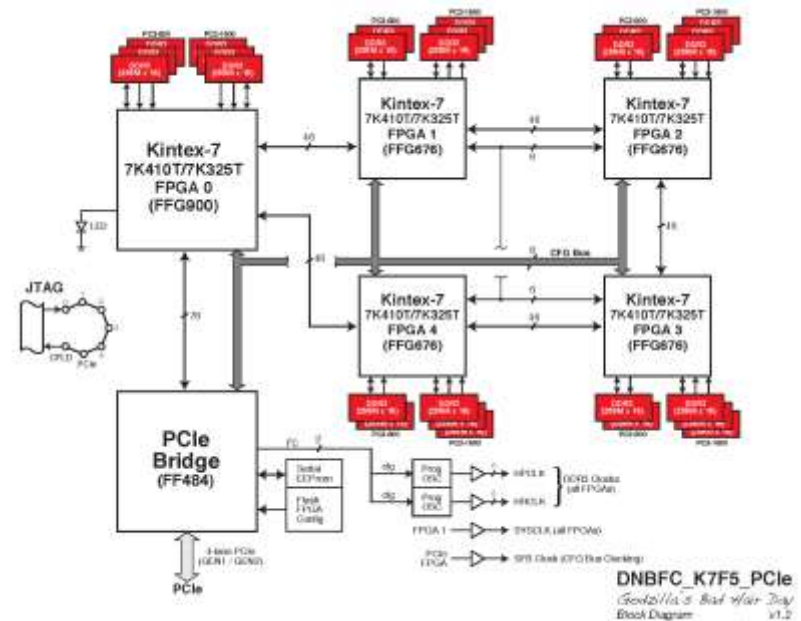
- No. 1 system in the green500
- Startup in Japan
- 1024 Cores & 1.5 TFLOPS
- Programming
 - Subset of OpenCL



- Source: <http://pezy.co.jp/en/products/pezy-sc.html>

FPGA (Field Programmable Gate Array)

- Very special, not in the TOP500 list
- Vendor: Altera, XILINX
- Programming
 - OpenCL



Motivation for Energy Efficiency

- An HPC system has a **very** high power consumption

Reducing energy consumption:

- Saves energy costs
- Helps to generate less CO₂

- Dark Silicon

Energy Efficient Application

- Less energy consumption
- Less heat
- More transistors can be used simultaneously (less Dark Silicon)



High power consumption  Efficiency

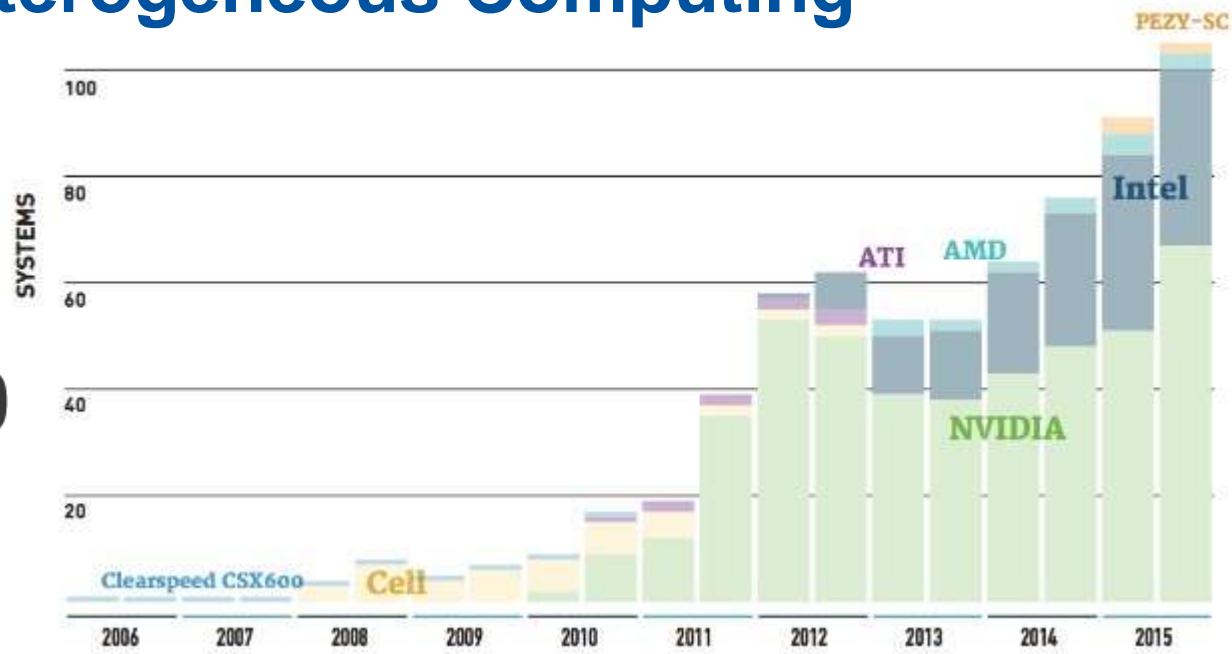
High temperatures  Reliability



Top500 vs. Green500 (November 2015)

Rank Top/Green	Site	Technology	Performance RMAX (TFLOPS)	Power (kW)	Efficiency (MFLOPS/W)
1/90	Tianhe-2 National Super Computer Center in Guangzhou	Intel Xeon E5 + Intel Xeon Phi 3,120,000 Cores	33,862.700	17808 = 17.8 MW	1901.54
136/1	Shoubu Institute of Physical and Chemical Research (RIKEN)	ExaScaler-1.4 80Brick, Xeon E5- 2618Lv3 8C 2.3GHz, InfiniBand FDR, PEZY-SC 787,968 Cores	353.820	50.31872	7031.58
83/162	Gottfried HLRN at Universität Hannover (LUIS)	Cray XC40, Intel Xeon E5-2695v2 12C 2.4GHz/E5- 2680v3 12C 2.5GHz, Aries interconnect 40320 Cores	829.805	787.5	1053.72

Trend Heterogeneous Computing



Source: <http://www.nextplatform.com/wp-content/uploads/2015/04/top500-nov-2015-acelerators.jpg>



Use of Accelerators

- November 2014: Top 23 systems
- June 2015: Top 32 systems
- November 2015: Top 40 systems

Research Exchange

- Energy Efficient High Performance Computing Working Group (EE HPC WG)

- Mostly US DOE and other governmental agencies
- Also participants from industry, academe and international organizations

- SC Conference

- Annual conference in US

- International Supercomputing Conference

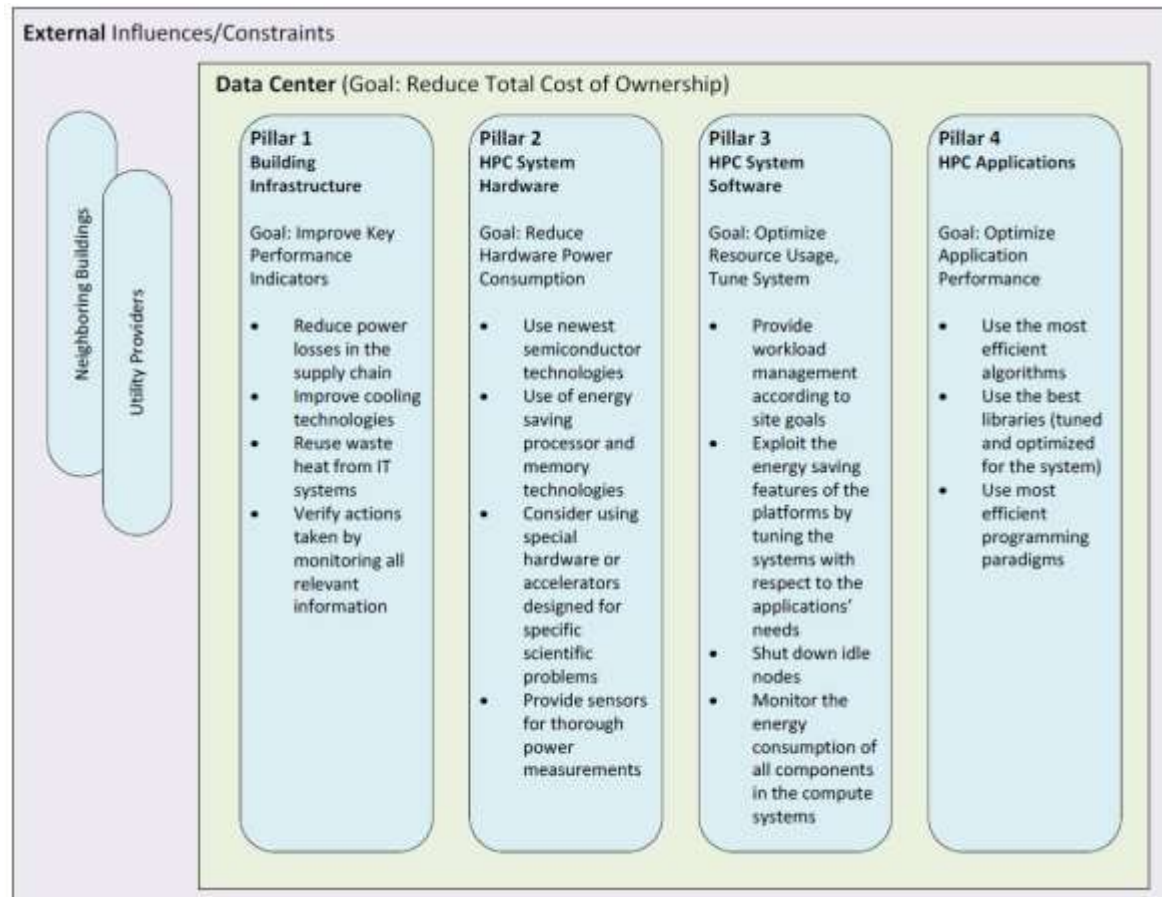
- Annual conference in Germany

- Workshop on Energy-Aware High Performance Computing



Structure: Energy Aware / Energy Efficiency

The 4 Pillar Framework for energy efficient HPC data centers*



* Wilde T., Auweter A., Shoukourian H., The 4 Pillar Framework for energy efficient HPC data centers, In: Computer Science – Research and Development, Special Issue, July 2013, Springer

Topics: Energy Aware / Energy Efficiency

- Building Infrastructure
 - Cooling, reuse waste heat
 - Hardware
 - Heterogeneous Computing
 - At least two different processor (CPU, GPU, Accelerator)
 - Embedded Systems
 - Dedicated function (router, phone), interesting building block for HPC
 - System Software
 - Runtime System, Workload Management, Scheduling
 - Applications
 - Approximate Computing, Energy Aware Algorithms
- **Power Measurement, Monitoring, Benchmarking**

NAS Parallel Benchmarks 1/3

<http://www.nas.nasa.gov/publications/npb.html>

- A small set of programs designed to help evaluate the performance of parallel supercomputers
- Are derived from computational fluid dynamics (CFD) applications
- Problem sizes are predefined and indicated as different classes:
 - S: small for quick test purposes
 - W: workstation size (a 90's workstation; now likely too small)
 - A, B, C: standard test problems; ~4X size increase going from one class to the next
 - D, E, F: large test problems; ~16X size increase from each of the previous classes

NAS Parallel Benchmarks 2/3

Benchmark Specifications (original)

- OpenMP (Fortran/C), MPI (Fortran/C) and Java
- 5 kernels
 - IS - Integer Sort, random memory access
 - EP - Embarrassingly Parallel
 - CG - Conjugate Gradient, irregular memory access and communication
 - MG - Multi-Grid on a sequence of meshes, long- and short-distance communication, memory intensive
 - FT - discrete 3D fast Fourier Transform, all-to-all communication
- 3 pseudo applications
 - BT - Block Tri-diagonal solver
 - SP - Scalar Penta-diagonal solver
 - LU - Lower-Upper Gauss-Seidel solver

NAS Parallel Benchmarks 2/3

Benchmark Specifications (Multi-zone)

- Designed to exploit multiple levels of parallelism in applications
- OpenMP (Fortran/C), MPI (Fortran/C), no Java
- 3 Multi-zone versions of the NPB applications
 - BT-MZ - uneven-size zones within a problem class, increased number of zones as problem class grows
 - SP-MZ - even-size zones within a problem class, increased number of zones as problem class grows
 - LU-MZ - even-size zones within a problem class, a fixed number of zones for all problem classes

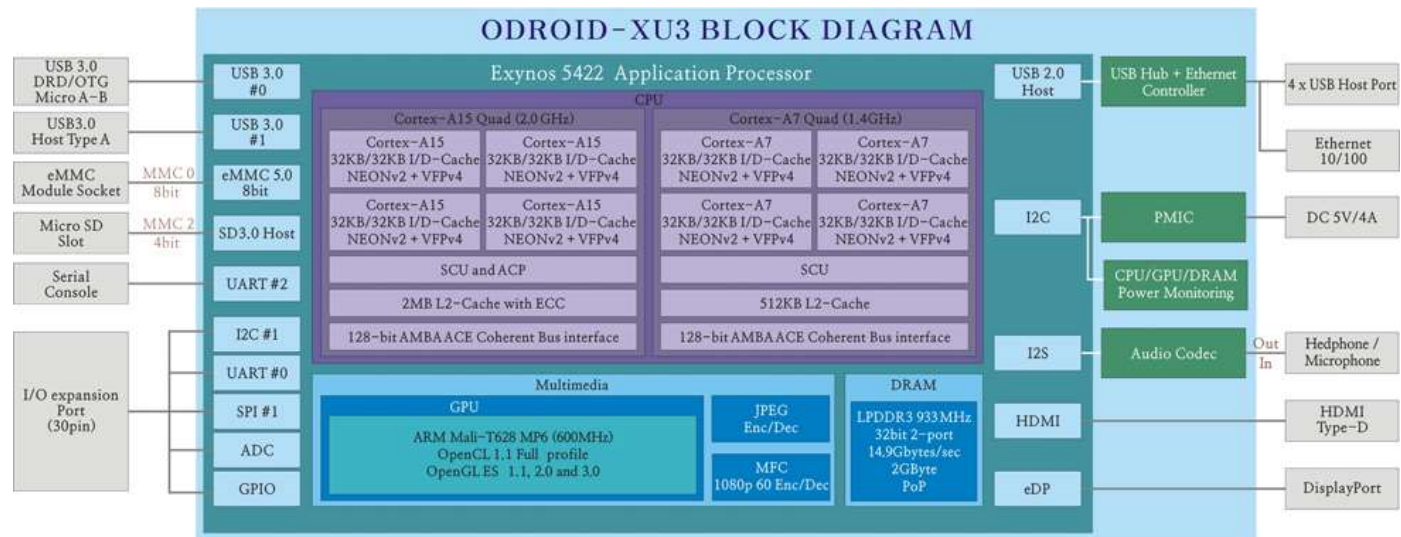
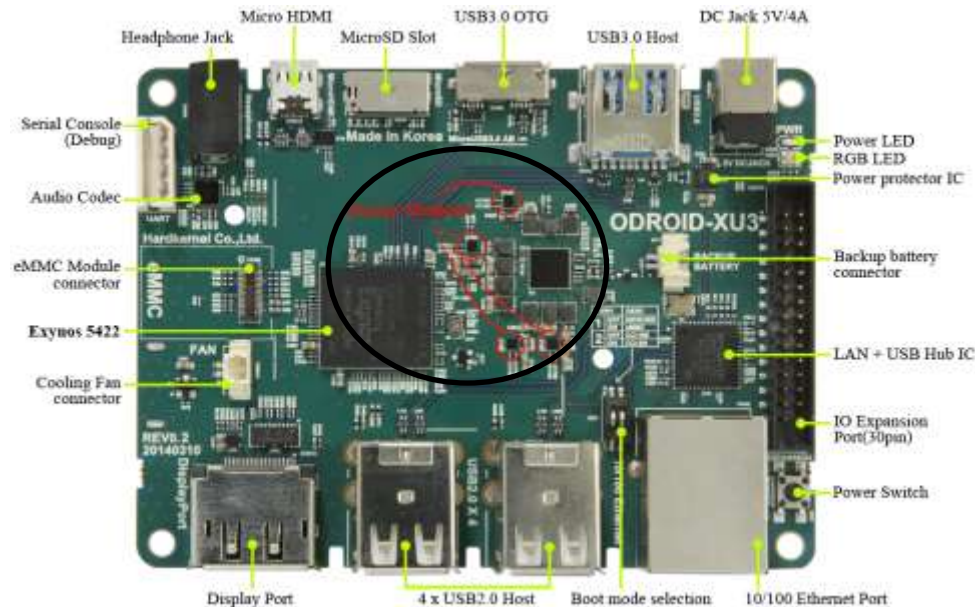
ODROID-XU3 Lab Environment

- 6 Lab Systems
 - Ubuntu 15.10
 - ssh access, no GUI
 - Java, C, C++, Fortran
 - OpenMP, MPI (MPICH)
 - Shared storage via NFS server
 - Demo accounts with sudo rights
- 6 ODROID Smart Power Devices
- 6 USB Gigabit Adapter
- Tools for power measurement



ODROID-XU3

179 \$



Overview of your Main Task

- We will provide a couple of sample programs (NAS)
- We expect for instance:
 - Power measurement
 - Black Box (Smart Power Device)
 - White Box (XU3 Sensors)
 - Compare Performance vs. Energy Efficiency
 - Scaling and Energy Efficiency
 - Compare the Fortran/C and Java versions
 - Compare different compiler options
 - Bind a particular process to a specific processor (core)
 - taskset
- You may define your own task and/or use other sample programs
- **Laptop?**

Links

- Templates
 - <http://www.dcsec.uni-hannover.de/vorlagen.html>
- Presentation Skills (German)
 - <http://www.mobile.ifi.lmu.de/wp-content/uploads/lehrveranstaltungen/seminar-trends-in-mobilen-und-verteiltern-systemen-ws1516/15-10-26-Praesentationstechnik.pdf> (last downloaded: 30.03.2016)
- Hints (German)
 - http://wr.informatik.uni-hamburg.de/teaching/organisatorische_hinweise#seminare

A long, low display cabinet filled with historical documents and artifacts. The cabinet is labeled 'HILRN' and 'CRAY'. It contains a portrait of a woman, a large clock face, and various coats of arms. The background is a blurred image of a modern building interior with large windows and a high ceiling.