Energy-Aware Sustainable Computing on Future Technology - Paving the Road for Exascale Computing

Exa2Green
energy-aware numerics

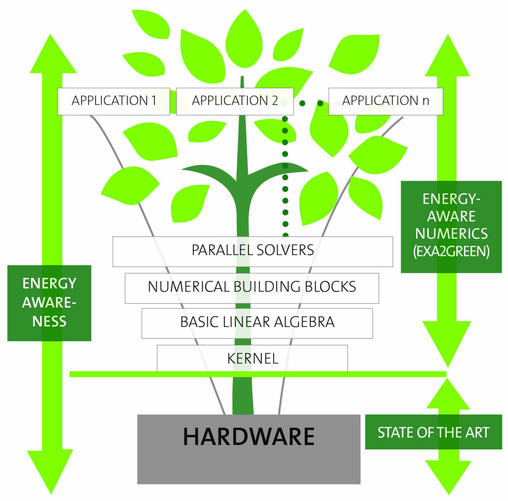
This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 318793.

Motivation

Power wall:
Prohibitive power demand when simply upscaling from current state.

Ecological impact:
Reduce CO2 footprint.

Energy-aware HPC:
Optimize algorithms for energy consumption.
Leverage power saving techniques provided by hardware.



Objectives

New metrics for quantitative assessment and analysis of energy profile of algorithms

Advanced and detailed power consumption monitoring and profiling

Smart algorithms using energy-efficient software models

Power-aware scheduling technology for HPC

Proof of concept using the weather forecast model COSMO-ART

Consortium

Heidelberg University, Germany (Coordinator)

Hamburg University, Germany

IBM Research - Zurich, Switzerland

Universidad Jaume I, Spain

ETH Zurich, Switzerland

Karlsruhe Insitute of Technology, Germany

Steinbeis gGmbh, Germany

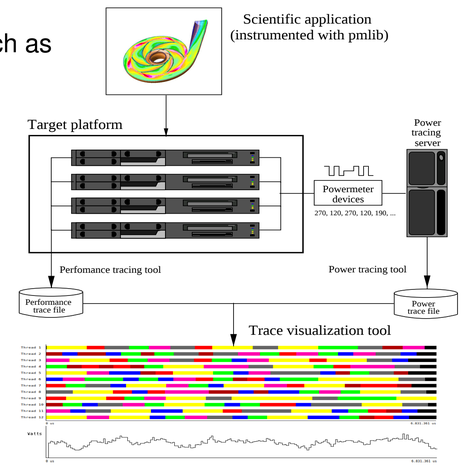
pmlib - A power-performance analysis framework

- Integrated framework to trace and analyse the power and energy consumption of parallel scientific applications

- General interface to utilize a wide range of wattmeters such as
  i) external wattmeters including commercial PDUs, WattsUp? Pro .Net, ZES LMG450
  ii) internal wattmeters attached to the power lines leaving the PSU, including ArduPower
  iii) commercial DAS from National Instruments
  iv) integrated power measurement interfaces including Intel RAPL, Nvidia NVML, IPMI

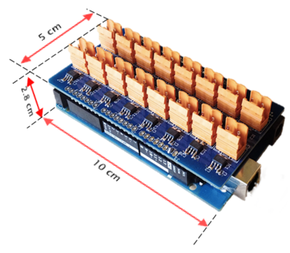
- C library to measure the application code

- Integrating traces into profiling and tracing frameworks including Exatrae + Paraver, VampirTrace + Vampir



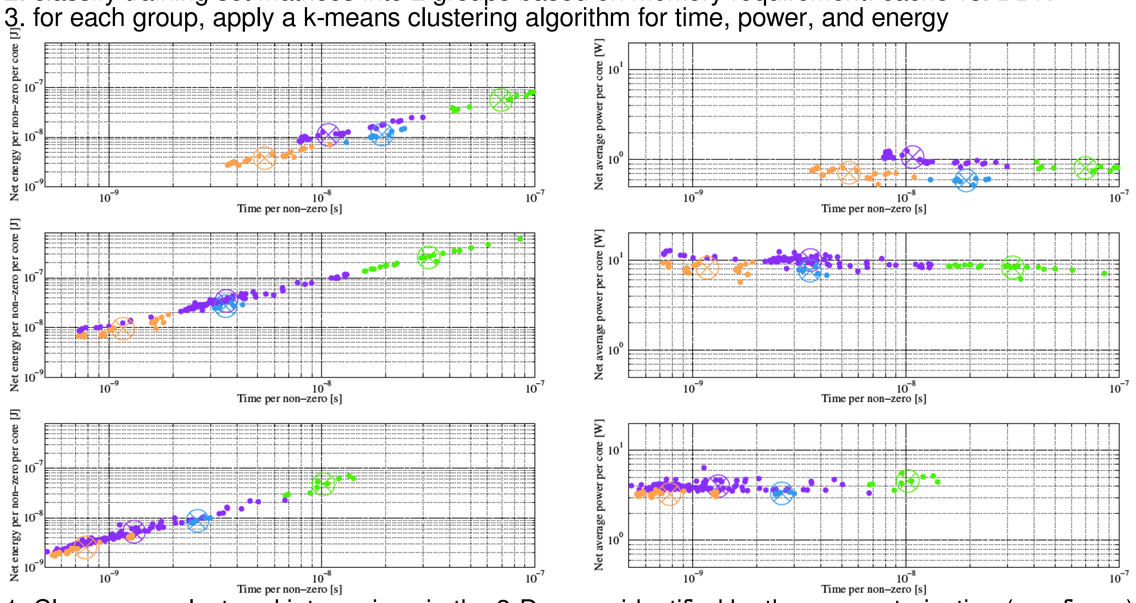
ArduPower - A low cost wattmeter to improve efficiency of HPC applications

- Internal DC wattmeter
- Measuring instantaneous power consumption of internal components of computing systems
- Small size to be deployed in a variety of server machines
- Low cost DAS device, total production cost approx. 100 EUR
- Fine spatial granularity, 16 channels concurrently
- Fine temporal granularity, up to 5,880 samples / second



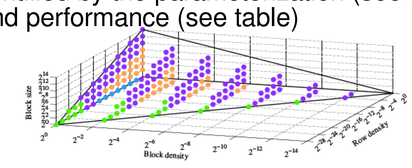
SpMV classification according to parameters and performance

Classification Algorithm:
1. define a reference training set of sparse matrices with a “uniform” (though sparse) non-zero pattern
2. classify training set matrices into 2 groups based on memory requirement: cache vs. DDR
3. for each group, apply a k-means clustering algorithm for time, power, and energy



1. Classes are clustered into regions in the 3-D space identified by the parameterization (see figure)
2. Straight correlation exists between parameterization and performance (see table)

Class	Color	Memory	Time	Power	Energy
1	●	Cache	Low	Low-Medium	Low
2	●	Cache	Medium	Low	Medium
3	●	DDR	Low-Medium	Medium-High	Low-Medium
4	●	DDR	High	Medium	High

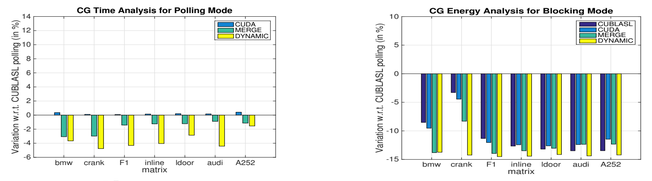


Development of energy-aware numerical linear algebra libraries

- Integrate energy-efficiency policies and techniques into run-times for linear algebra libraries
- Mantain performance scalability as well as iso-efficiency
- Explore state-of-the-art multi-core processors as well as accelerators

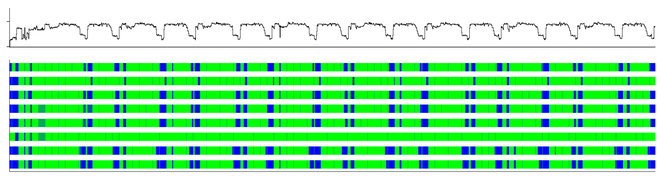
Improve energy-efficiency of CPU-GPU codes for linear algebra

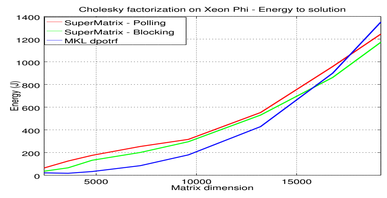
- Exploit DYNAMIC parallelism or MERGE to reduce the number of CUDA kernels
- Allow better usage of energy-efficient C-states in the CPU
- Appealing for sparse linear systems solvers, with low workload mapped to the CPU



Energy-aware runtimes

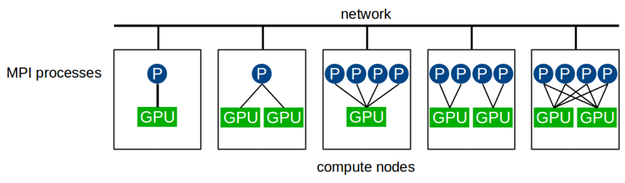
- OmpSs (Barcelona Supercomputing Center) for sparse linear algebra
- libflame+SuperMatrix (The University of Texas at Austin) for dense linear algebra



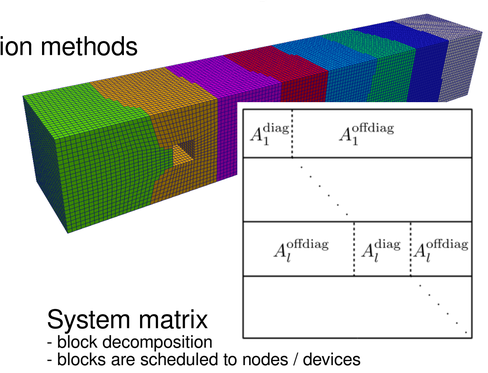


Asynchronous iteration on multi-GPU multi-node HPC systems

- Relax synchronization requirements of classical relaxation methods
- Exploit parallelism of hardware, improve efficiency



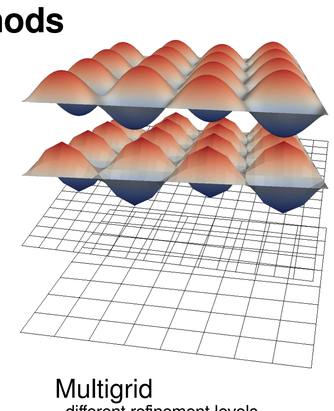
Supported parallel configurations
- MPI, OpenMP & CUDA
- Nvidia HyperQ, Multi-Process Service

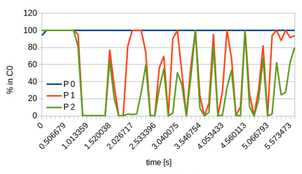


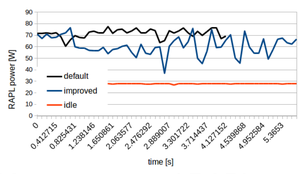
System matrix
- block decomposition
- blocks are scheduled to nodes / devices

Energy-efficient parallel geometric multigrid methods

- Grid hierarchy with different problem sizes
- Individual domain decomposition per level
- Parallel prolongation and restriction operators
- Adjust hardware activity to solver state







Measurement of C-states:
- P0 is active on all grid levels
- P1 is paused on coarsest level
- P2 is only active on finest level

Measurement of RAPL:
- default: average ~70 W, variation ~10 W
- improved: average ~60 W, variation depends on solver state
- idle: average ~28 W, small variation

Multigrid
- different refinement levels
- transfer between grids
- smoothing on finer levels
- error correction on coarsest level

Computational and energy efficieny optimization of the air quality prediction model COSMO-ART

- Energy profiling of COSMO-ART
- Optimal setup for parameters, compilers
- Refactoring for CPUs
- ODE solver algorithmic changes
- Mixed-precision COSMO-ART
- Port COSMO-ART components to accelerators
- Feasibility study of a reduced model for gas phase chemistry
- Investigation of possibilities of ART on ARM