Energy of computing: How to optimise your program for something you cannot measure

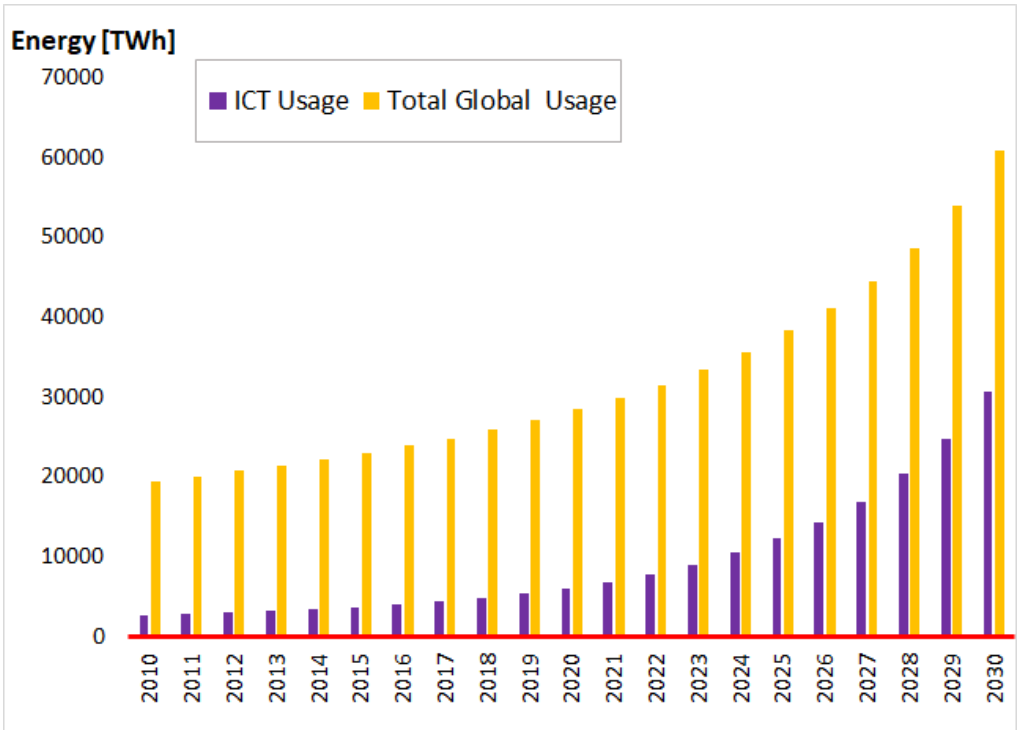
Alexey Lastovetsky

University College Dublin, Ireland

13th International Conference on Parallel Processing and Applied Mathematics (PPAM 2019)
Bialystok, Poland
September 9th, 2019

Outline

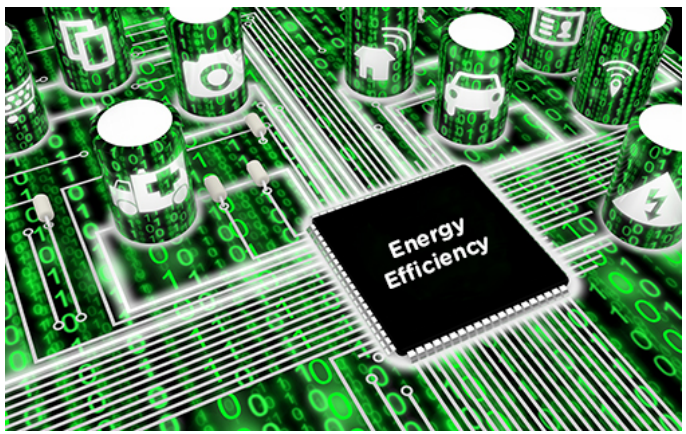
Energy Consumption in ICT



Global Energy Consumption Verses ICT Energy Consumption

Adopted from *A. Andrae, T. Edler (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157.*

Optimization of energy of computing



Produce Energy-Efficient Hardware

Develop Energy-Aware Software Applications

Energy optimization of software applications

Key approaches to energy optimization of applications are:

1. System-level energy optimization such as Dynamic Voltage and Frequency Scaling (DVFS), Dynamic Power Management (DPM), and energy-aware scheduling.
   - Optimization of the executing environment rather than the application – mainstream approach
2. Application-level energy optimization techniques that use application-level parameters [1].
   - Optimization of the application rather than the executing environment

[1] A. Lastovetsky, R.R. Manumachu (2017). New model-based methods and algorithms for performance and energy optimization of data parallel applications on homogeneous multicore clusters. IEEE Transactions on Parallel and Distributed Systems, 28(4), 1119-1133.

Optimization for energy in the ideal world

No room for application-level optimization for energy in the ideal HPC world.

The ideal HPC world:

- Linear
- Homogeneous

This means that

- **Any distribution of workload between identical *linear* parallel processors will consume the same dynamic energy**
- **Load-balanced distribution will always be optimal for performance and energy**

Linearity of performance

Execution time is a linear function of workload.

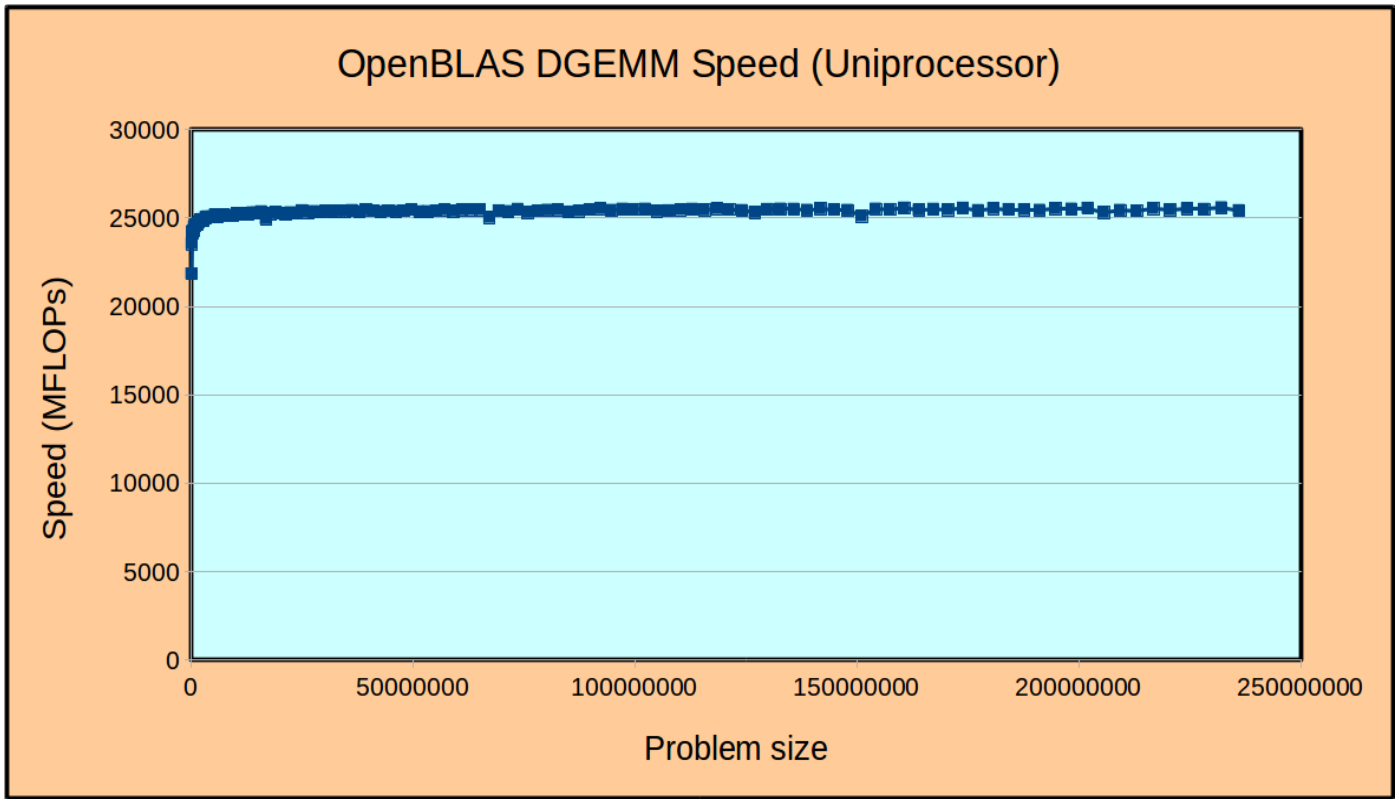


Figure: Speed function of OpenBLAS DGEMM application executed on a single core on the Intel Haswell workstation.

Linearity of energy

Energy is a near-linear function of workload

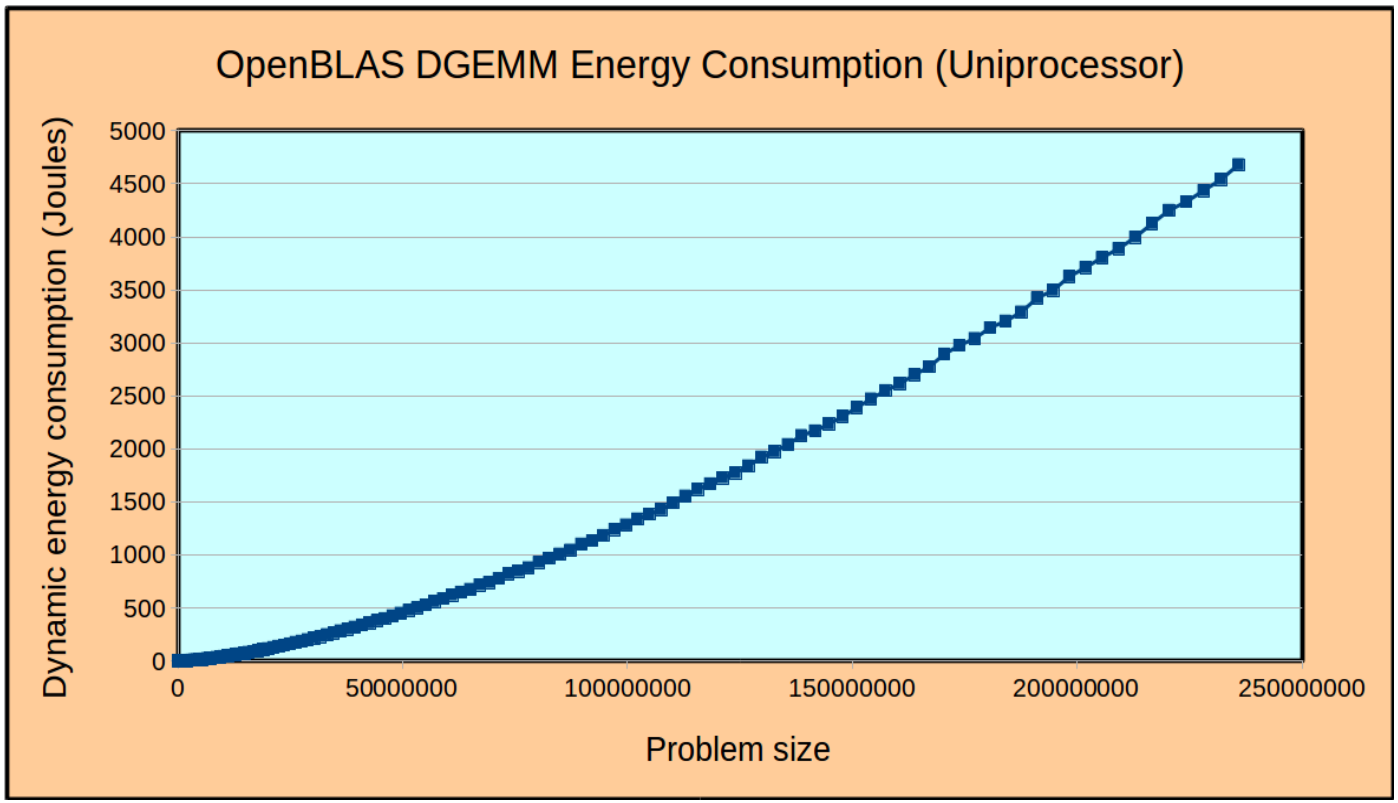


Figure: Dynamic energy consumption of OpenBLAS DGEMM application executed on a single core on the Intel Haswell workstation.

Linearity of energy

Energy function for server is even more linear

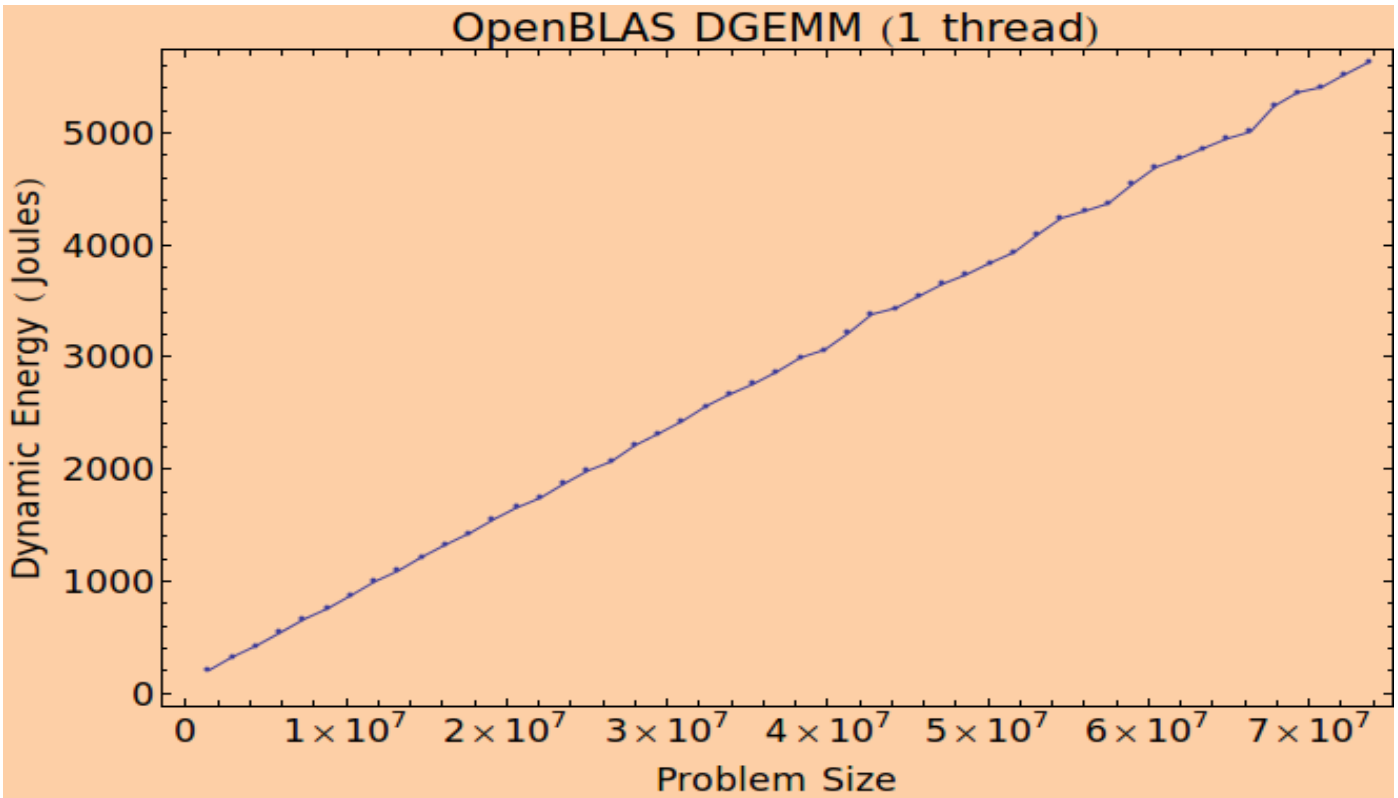


Figure: Dynamic energy consumption of OpenBLAS DGEMM application executed on a single core on the Intel Haswell server.

Energy profile and optimization for energy

- $E(x)$ - energy consumed by processor executing workload x
- Equal distribution of workload between identical processors
  - minimizes energy consumption if $\frac{dE(x)}{dx} \geq 0$ and $\frac{d^2E(x)}{dx^2} \geq 0$.
- If $E(x)$ – linear ($\frac{d^2E(x)}{dx^2} = 0$), all distributions are equivalent
  - $E(x_1) + \ldots + E(x_p) = (p-1) \times E(0) + E(x_1 + \ldots + x_p) = (p-1) \times E(0) + E(w)$ $(x_1 + \ldots + x_p = w)$
- If $E(x)$ is convex ($\frac{d^2E(x)}{dx^2} > 0$), any uneven distribution consumes more energy than the equal load-balanced one.
  - For any x, $E(x) < \frac{E(x-\Delta x)+E(x+\Delta x)}{2}$, and hence, $E(x) + E(x) < E(x - \Delta x) + E(x + \Delta x)$.

Convex energy profile

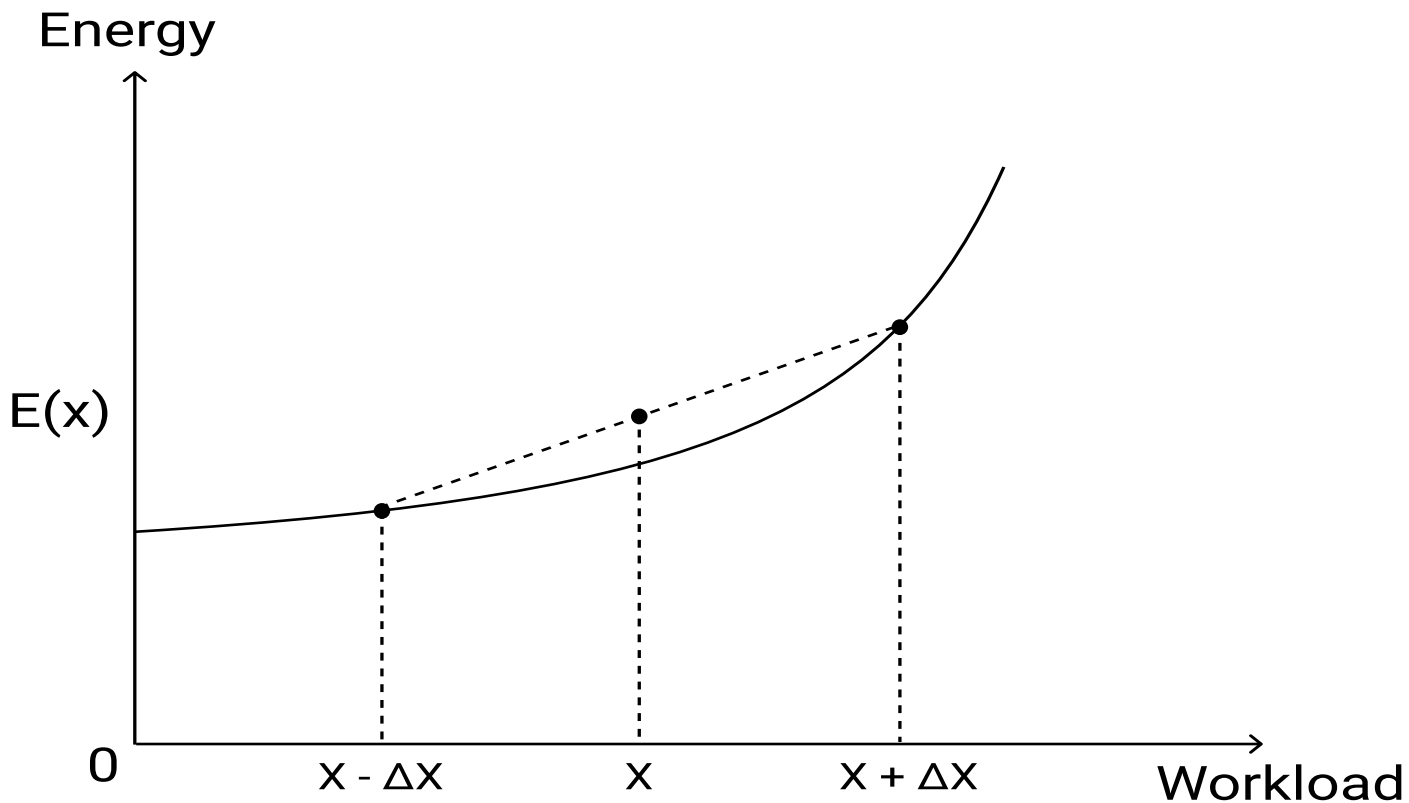


Figure: For any x, $E(x) < \frac{E(x-\Delta x)+E(x+\Delta x)}{2}$, and hence, $E(x) + E(x) < E(x - \Delta x) + E(x + \Delta x)$

Optimization for energy in the ideal world

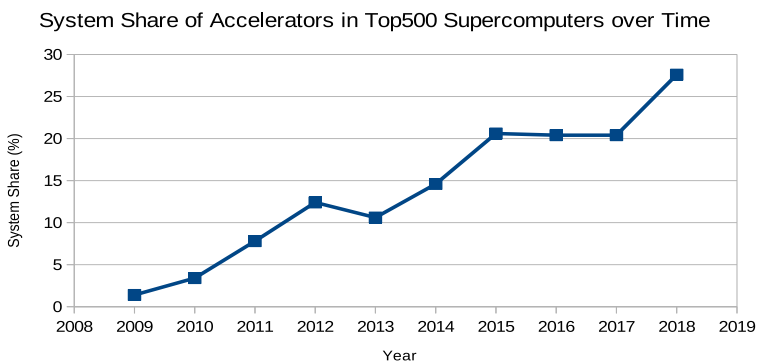
In the linear and homogeneous HPC world:

- Any distribution of workload between identical linear parallel processors will consume the same dynamic energy
- Load-balanced distribution will always be optimal for performance and energy

Optimization for energy in the real world

The real HPC world:

- Increasingly heterogeneous
  - Almost 30% of systems in Top500 are heterogeneous



- Non-linear

Non-linearity of performance

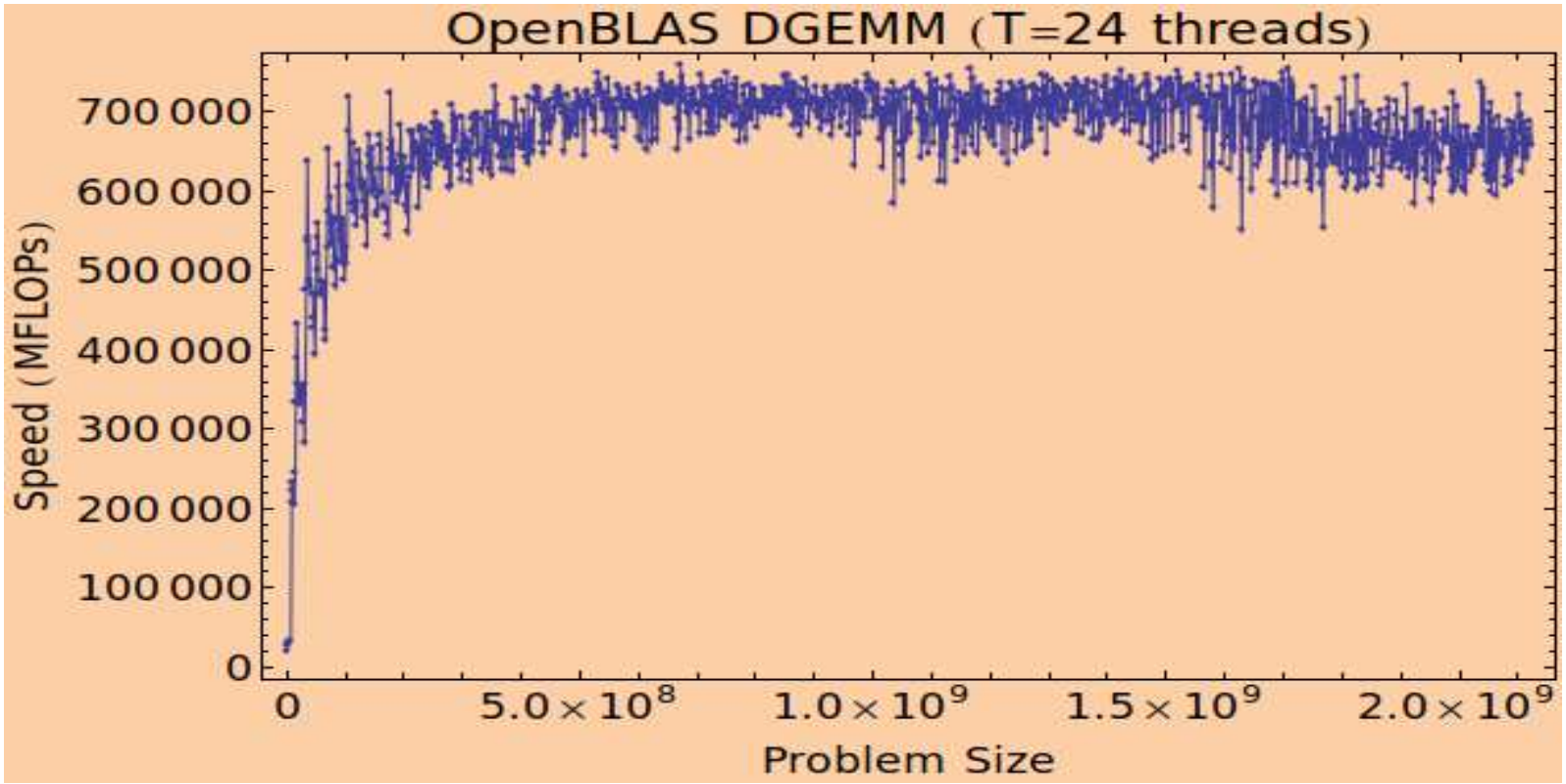


Figure: OpenBLAS DGEMM executing 24 threads on 24-core CPU of the Intel Haswell server.

Non-linearity of performance in multicore era

FFTW application computing $2D$ discrete Fourier transform of size $n \times n$:

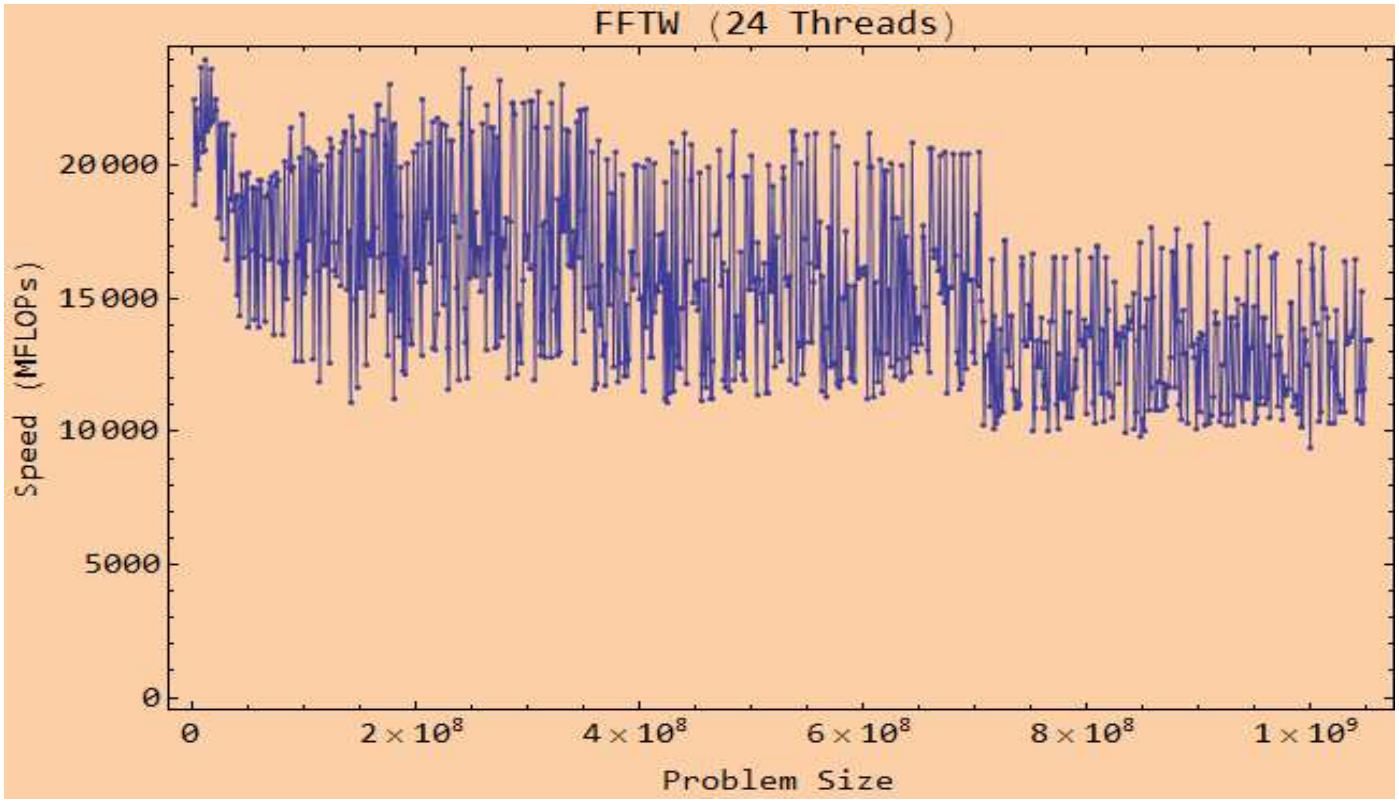


Figure: FFTW executing 24 threads on 24-core CPU of the server.

Non-linearity and heterogeneity of performance

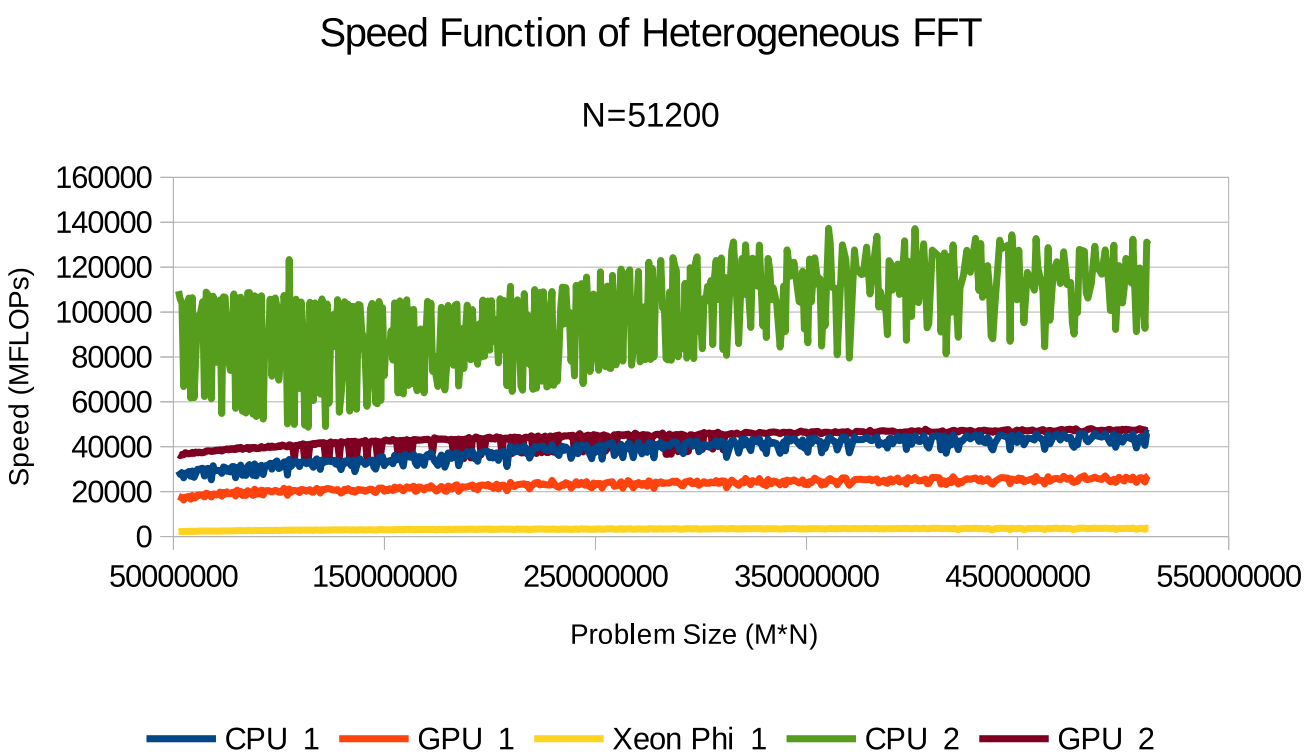


Figure: 2D FFT speed functions.

Non-linearity of performance: Analysis

More threads/cores => bigger variations:

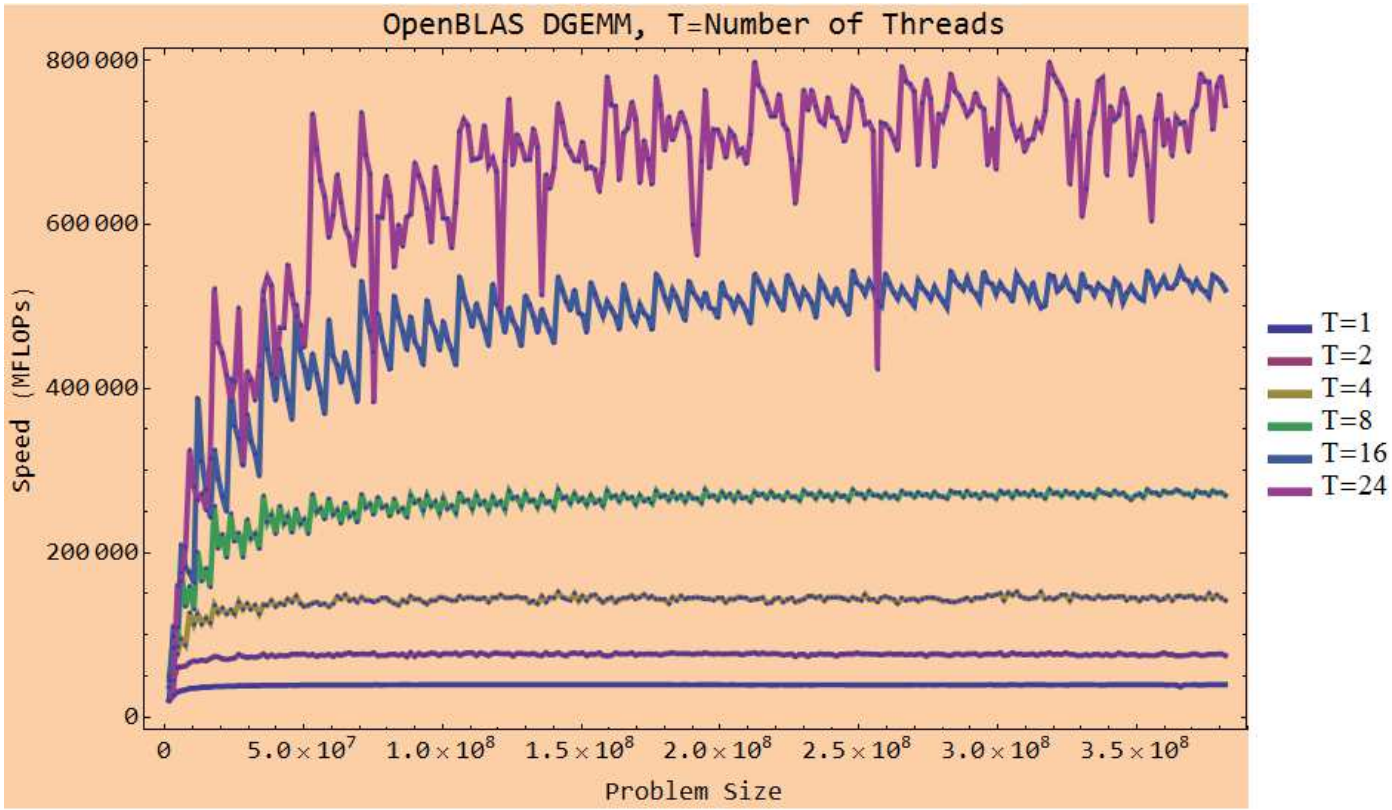


Figure: Speed function of OpenBLAS DGEMM application executing varying number of threads (T) on the Intel Haswell server.

Non-linearity and heterogeneity of energy

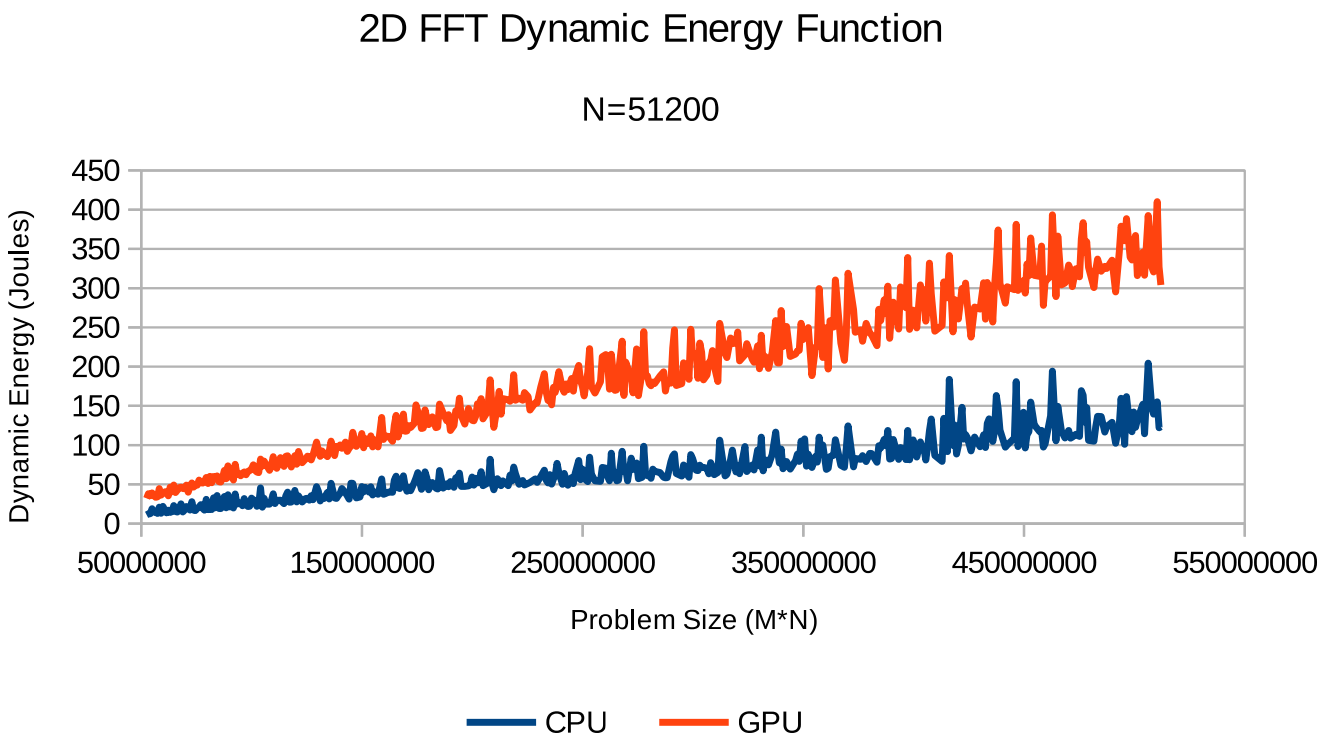


Figure: 2D FFT energy functions of CPU and GPU on the Intel Haswell HCLServer01.

Non-linearity and heterogeneity of energy

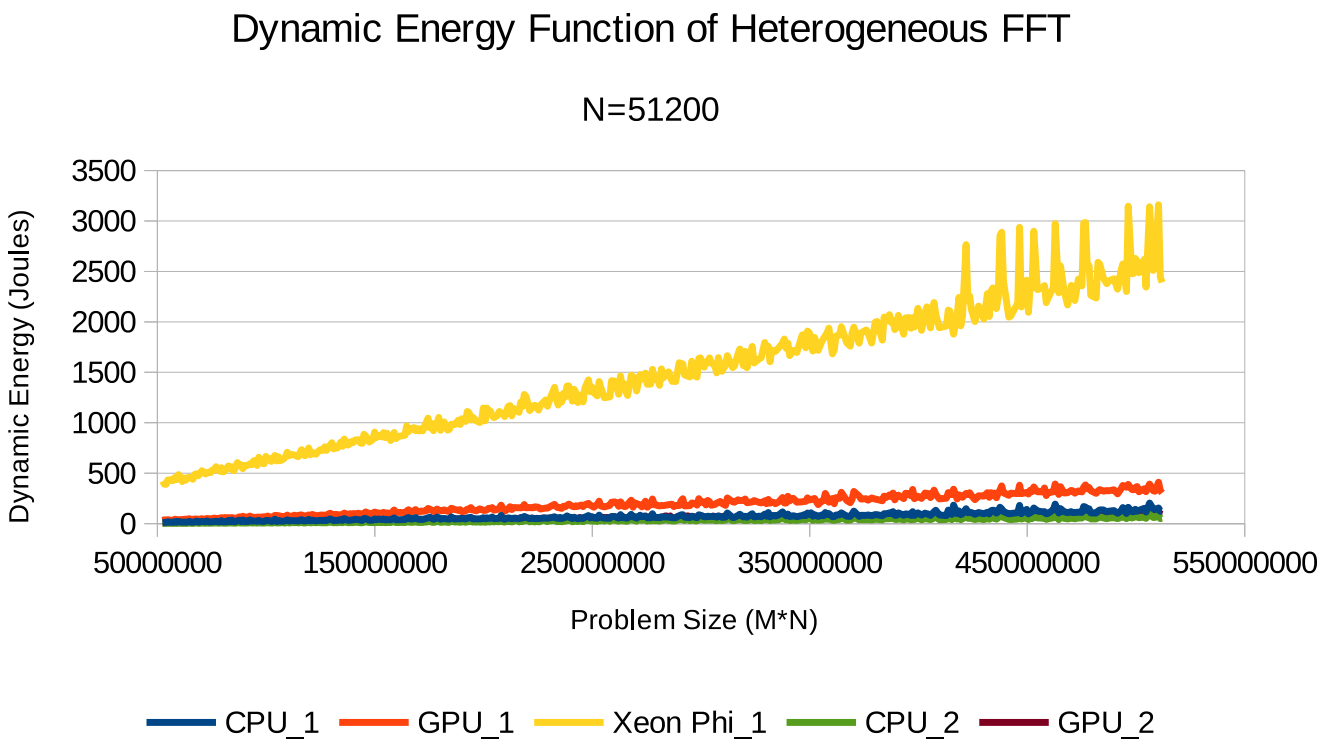
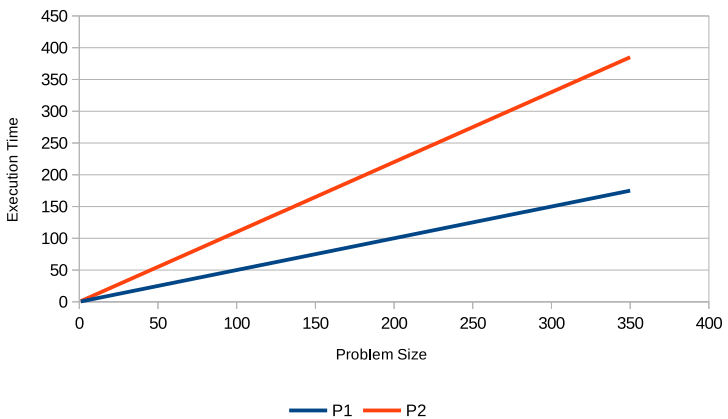


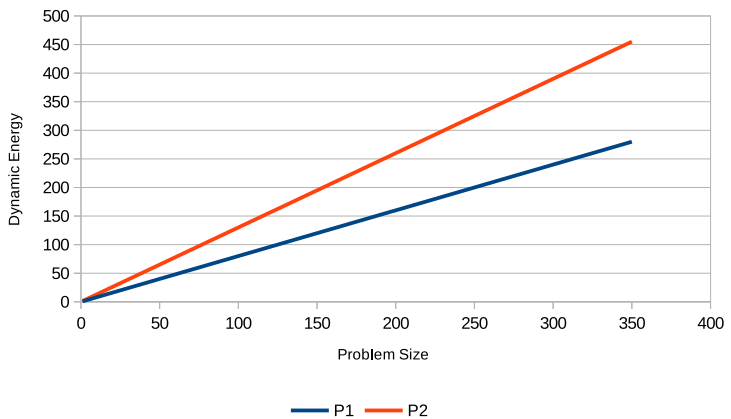
Figure: 2D FFT energy functions of five compute devices across Intel Haswell HCLServer01 nad HCLServer02.

Implications of heterogeneity

Performance and energy profiles of two heterogeneous processors



Linear performance profiles



Linear dynamic energy profiles

Implications of heterogeneity

For *p* linear heterogeneous processors [1]:

- Load-balanced solution minimizes time
- Solution using the single most energy-efficient processor minimizes dynamic energy consumption
- There is infinite number of globally Pareto-optimal solutions
  - Can be found analytically

[1]. H.Khaleghzadeh, M.Fahad, A. Shahid, R. Manumachu, and A. Lastovetsky, "Bi-objective Optimization of Data-parallel Applications on Heterogeneous HPC Platforms for Performance and Energy through Workload Distribution", revision under preparation for IEEE Transactions on Parallel and Distributed Systems.

Implications of non-linearity

For p homogeneous non-linear processors:

- Performance-optimal and energy-optimal solutions are not necessarily load-balanced [1]
  - There are exact efficient algorithms of complexity $O(m^2 \times p^2)$, where m is the cardinality of discrete performance and energy profiles, finding performance-optimal and energy-optimal solutions [1]
- The number of globally Pareto-optimal solutions for performance and energy is significant
  - There is an exact efficient algorithm of complexity $O(m^2 \times p^2)$ solving the bi-objective optimization problem [2]

[1]. A. Lastovetsky and R. Reddy, "New Model-based Methods and Algorithms for Performance and Energy Optimization of Data Parallel Applications on Homogeneous Multicore Clusters", IEEE Transactions on Parallel and Distributed Systems 28(4):1119-1133, 2017.

[2]. R. Reddy and A. Lastovetsky, "Bi-Objective Optimization of Data-Parallel Applications on Homogeneous Multicore Clusters for Performance and Energy", IEEE Transactions on Computers 67(2):160-177, 2018.

Implications of heterogeneity AND non-linearity

For p non-linear heterogeneous processors [1]:

- Performance-optimal and energy-optimal solutions are generally load-imbalanced [1]
  - There are exact efficient algorithms of complexity $O(m^3 \times p^3)$ finding performance-optimal, and dynamic energy-optimal, and total energy-optimal solutions [1]
- The number of globally Pareto-optimal solutions for performance and energy is quite significant
  - There are exact efficient algorithms of complexity $O(m^3 \times p^3)$ solving the bi-objective optimization problems for performance and dynamic energy and for performance and total energy [1]

[1]. H.Khaleghzadeh, M.Fahad, A. Shahid, R. Manumachu, and A. Lastovetsky, "Bi-objective Optimization of Data-parallel Applications on Heterogeneous HPC Platforms for Performance and Energy through Workload Distribution", revision under preparation for IEEE Transactions on Parallel and Distributed Systems.

Challenges of practical optimization for energy

To apply the energy-optimization methods in practice, we need

- Energy profiles of individual components of a hybrid parallel application
- Their performance profiles (for bi-objective optimization)

Therefore, we need:

- Methods for component-level measurement of the execution time
  - All processing units equipped with clocks
  - While not trivial for tightly coupled units but doable
- Methods for **component-level measurement of energy consumption**
  - No reliable equivalent of clocks
  - **Real challenge**
- In the absence of such methods, the problem becomes intractable
  - We will have to experimentally build the energy profile of cardinality m^p instead of $m \times p$

Component-level measurement of energy consumption

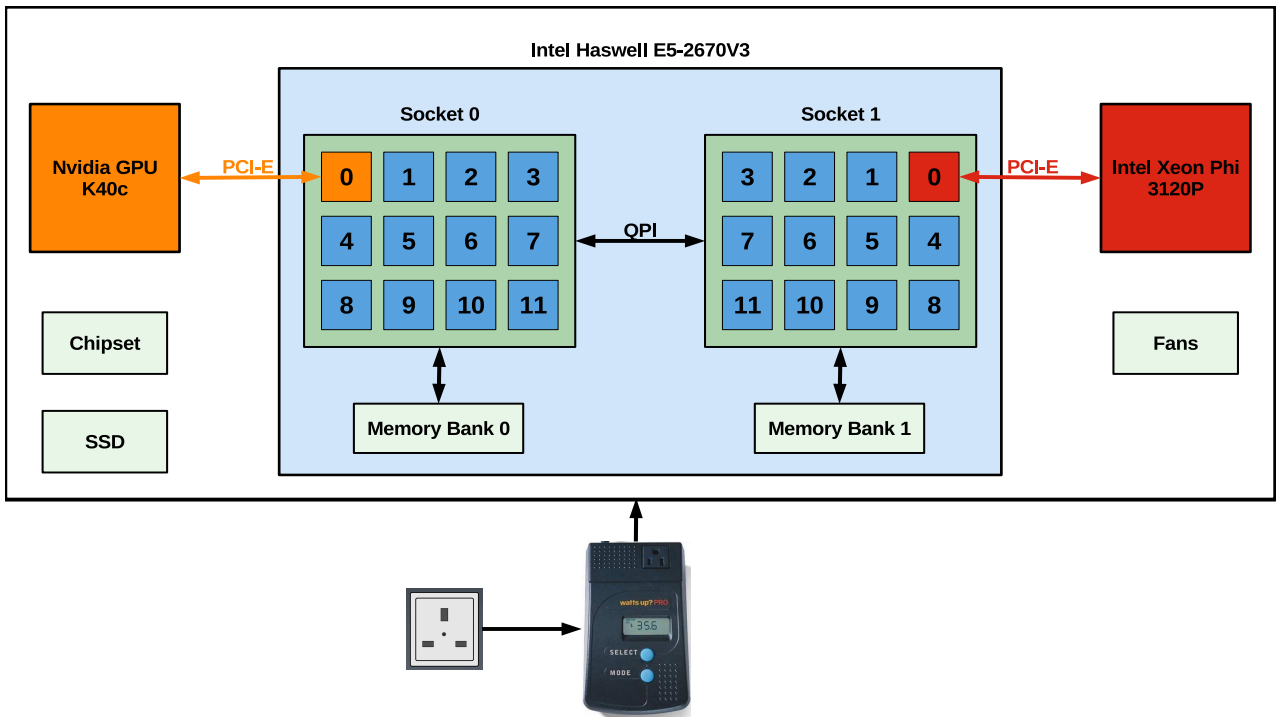


Figure: Hybrid Intel Haswell-based server and typical configuration of a parallel application.

Methods of measurement of energy consumption

Three approaches to component-level energy measurement [1]:

- System-level physical measurements using external power meters
  - Most accurate but expensive
  - We use them as ground truth
- Measurements using on-chip power sensors
  - Intel RAPL (Running Average Power Limit), Intel Xeon Phi SMC (System Management Controller), AMD APM (Application Power Management), Nvidia NVML (Nvidia Management Library)
  - Inaccurate and poorly documented
- Energy predictive models
  - Mostly use performance monitoring counters (PMCs) as predictor variables
  - State-of-the-art models not accurate
  - The only realistic alternative to methods using power meters

[1]. M. Fahad, A. Shahid, R. R. Manumachu, and A. Lastovetsky, "A Comparative Study of Methods for Measurement of Energy of Computing", Energies, vol. 12, issue 11: MDPI, pp. 42, 2019.

Accuracy of on-chip power sensors

Accuracy of RAPL on Intel Xeon Skylake (2 socket, 28 cores each) [1]

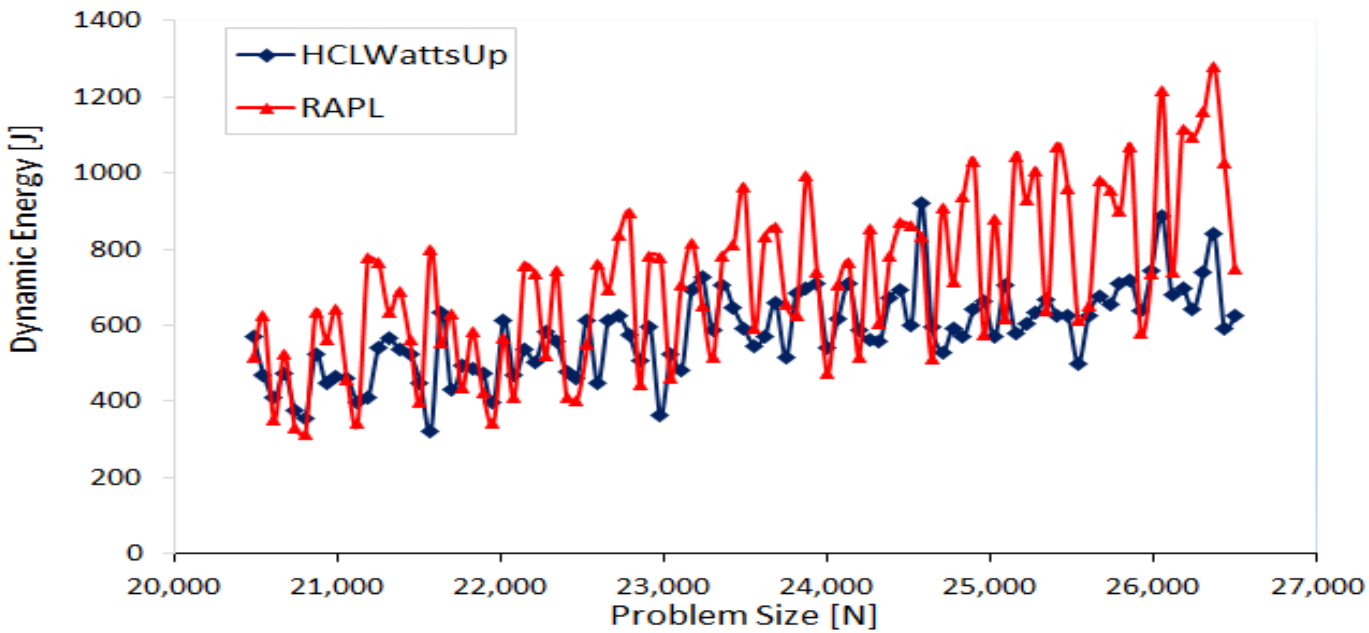


Figure: Dynamic energy profiles of FFTW (N = 20, 480 – 26, 560) by RAPL and HCLWattsUp on HCLServer03.

[1]. M. Fahad, A. Shahid, R. R. Manumachu, and A. Lastovetsky, "A Comparative Study of Methods for Measurement of Energy of Computing", Energies, vol. 12, issue 11: MDPI, pp. 42, 2019.

Accuracy of on-chip power sensors

Accuracy of RAPL, NVML and RAPL+NVML on a hybrid server [1]

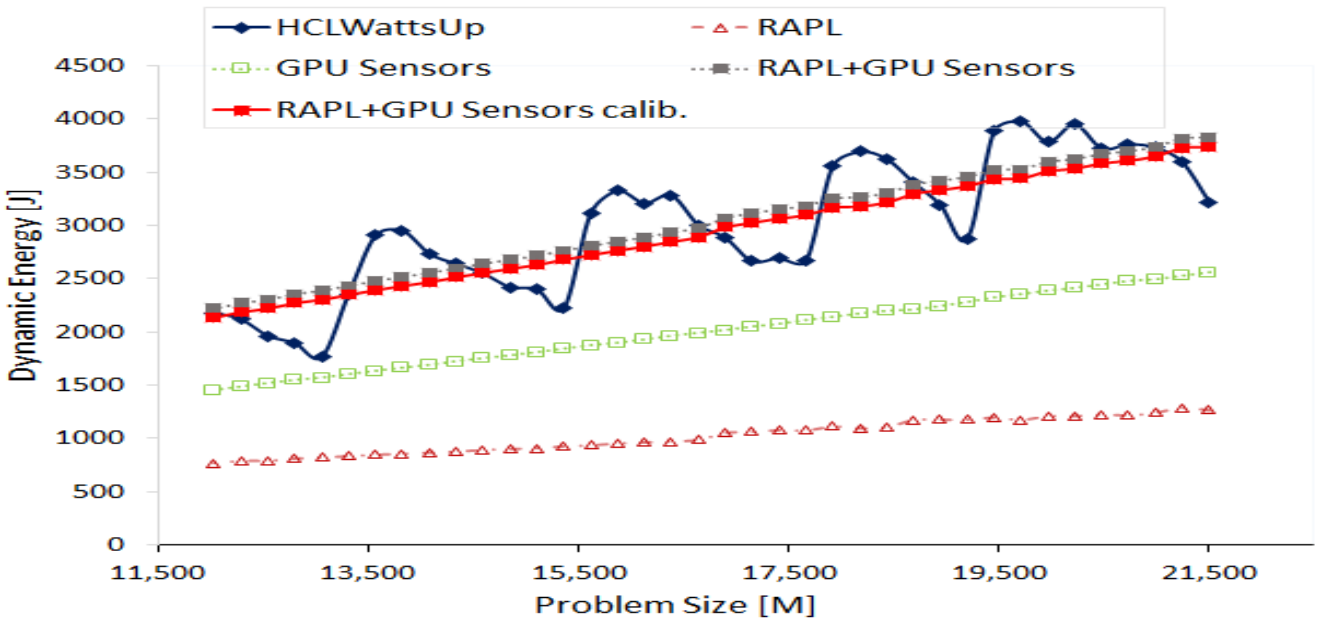


Figure: Dynamic energy consumption profiles of DGEMM on Nvidia K40c GPU on HCLServer01.

[1]. M. Fahad, A. Shahid, R. R. Manumachu, and A. Lastovetsky, "A Comparative Study of Methods for Measurement of Energy of Computing", Energies, vol. 12, issue 11: MDPI, pp. 42, 2019.

Accuracy of on-chip power sensors

Accuracy of RAPL, NVML and RAPL+NVML on a hybrid server [1]

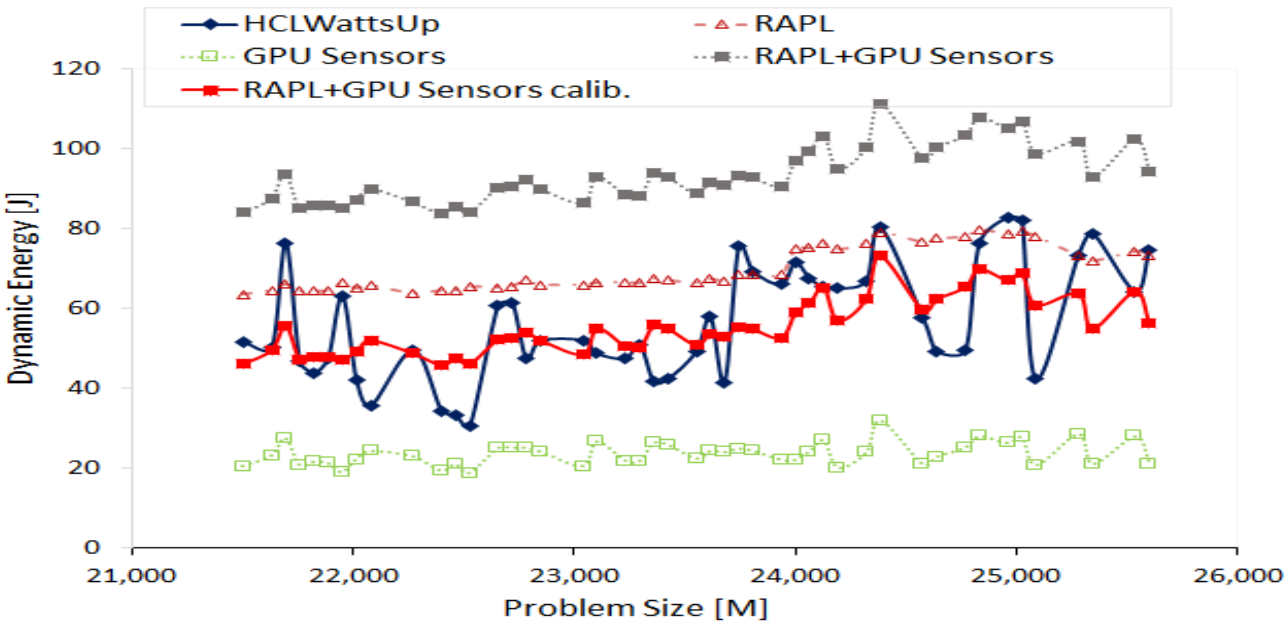


Figure: Dynamic energy consumption profiles of CUDA FFT on Nvidia K40c GPU on HCLServer02.

[1]. M. Fahad, A. Shahid, R. R. Manumachu, and A. Lastovetsky, "A Comparative Study of Methods for Measurement of Energy of Computing", Energies, vol. 12, issue 11: MDPI, pp. 42, 2019.

Energy losses due to inaccurate measurements

Optimization for energy of a data-parallel parallel DGEMM application employing Intel MKL and running on Intel Haswell and Intel Skylake CPUs [1]:

- Matrices A and C partitioned horizontally between the CPUs to minimize the dynamic energy consumption
- Energy profiles of the CPUs are inputs to the partitioning algorithm [2]
- Two solutions are compared
  - Using energy profiles obtained with HCLWattsUp
  - Using energy profiles obtained with RAPL
- Measured energy losses of RAPL-based solutions against HCLWattsUp ranged from 37% to 84% depending on problem size

[1]. M. Fahad, A. Shahid, R. R. Manumachu, and A. Lastovetsky, "A Comparative Study of Methods for Measurement of Energy of Computing", Energies, vol. 12, issue 11: MDPI, pp. 42, 2019.

[2]. H.Khaleghzadeh, M.Fahad, A. Shahid, R. Manumachu, and A. Lastovetsky, "Bi-objective Optimization of Data-parallel Applications on Heterogeneous HPC Platforms for Performance and Energy through Workload Distribution", revision under preparation for IEEE Transactions on Parallel and Distributed Systems.

Performance Monitoring Counters (PMCs)

Performance Monitoring Counters (PMCs):

- Special-purpose registers to store the counts of software and hardware activities
  - Primarily used for low-level performance analysis and tuning
- Large in number $\rightarrow$ Intel Haswell: 164, Intel Skylake: 385
- Can not be collected all simultaneously $\rightarrow$ 3-4 in one application run
- Architecture specific $\rightarrow$ non-portable

Linear PMC-based Energy Predictive Model

Linear PMC-based dynamic energy predictive models are most common:

$$E = \sum_{j=0}^{M} \beta_j \times x_j + \epsilon$$

where x_j are the predictor variables (PMCs) and ϵ is the error term

Dominant PMC groups:

- Cache misses
- Branch instructions
- Floating point operations
- Page faults
- Memory Accesses

PMC-based Models – More Insights

Existing Issues

- Large number of PMCs to consider
- Require tremendous programming effort and time to collect PMCs
- Pure PMC-based model lacks portability

Existing techniques to select PMCs for a model

- Consider all PMCs to capture all possible contributors to energy consumption
- Based on a statistical methodology $\rightarrow$ Correlation, Principal Component Analysis (PCA)
- Using expert advice or intuition to pick a subset

Accuracy of PMC-based Energy Models

- Selection of PMCs for a predictive model is still an open problem
- Numerous PMC-based models proposed but none is sufficiently accurate [1]
  - Publications report excellent accuracy but
    - They typically report the prediction accuracy of **total** energy with very high static power base
    - Most reported results are not reproducible
  - The best verifiable average prediction error of average **dynamic** power by such models for a Intel Haswell platform is in the range of 90-100% [1]

[1]. K. O'Brien, I. Petri, R. Reddy, A. Lastovetsky, and R. Sakellariou, "A survey of power and energy predictive models in HPC systems and applications", ACM Computing Surveys, vol. 50, issue 3, 2017.

Causes of inaccuracy of PMC-based energy models

- One cause of inaccuracy of PMC-based energy models was recently discovered [1]
  - Many popular PMCs are not *additive* on modern multicore processors
- Energy is additive
  - Energy consumption of serial execution of two applications A and B is equal to the sum of their individual consumptions, $E_{AB} = E_A + E_B$
- Therefore, any PMC parameter x in a linear power/energy predictive model should be additive, $x_{AB} = x_A + x_B$
  - Many popular PMCs are non-additive, some with up to 200% deviation from additivity
  - Numbers of non-additive PMCs increases with the increase of cores (very few non-additive PMCs in the case of single core)
- Another cause – violation of basic laws of energy conservation [1]
  - Non-zero intercept in a dynamic energy model
  - Negative coefficients

[1]. A. Shahid, M. Fahad, R. R. Manumachu, and A. Lastovetsky, "Additivity: A selection criterion for performance events for reliable energy predictive modeling", Supercomputing Frontiers and Innovations, vol. 4, pp. 50–65, 2017.

Improving the accuracy of PMC-based energy models

The accuracy of PMC-based energy models can be improved by

- Removing non-additive PMCs from models
- Enforcing basic laws of energy conservation
- By applying this technique, the accuracy of the state-of-the-art models have been significantly improved, bringing it to 25–30% [1]

Still long way to go.

[1]. A. Shahid, M. Fahad, R. R. Manumachu, and A. Lastovetsky, "Improving the Accuracy of Energy Predictive Models for Multicore CPUs Using *Additivity* of Performance Monitoring Counters", HerteroPar 2019.

Summary

- Modern platforms provide significant opportunities for application-level energy optimization and bi-objective optimization for energy and performance
- Efficient optimization algorithms is not a grand challenge
  - They exist and many more to follow
- The real grand challenge is accurate component-level energy measurements

Acknowledgments

The research was conducted in collaboration with:

Muhammad Fahad, Hamidreza Khaleghzadeh, Semyon Khokhriakov, Ravi Reddy (University College Dublin, Ireland)

Thank You!

Questions?