

Sustainability through Massively Integrated Computing: Are We Ready to Break the Energy Efficiency Wall for Single-Chip Platforms?

Partha Pande
School of EECS
Washington State
Univ. Pullman, WA,
USA

Fabien Clermidy,
Diego Puschini and
Imen Mansouri
CEA-LETI / Minatoc
Grenoble, France

Paul Bogdan and
Radu Marculescu
Department of ECE
Carnegie Mellon Univ.
Pittsburgh, PA, USA

Amlan Ganguly
Department of CE
Rochester Institute of
Technology
Rochester, NY, USA

Abstract — While traditional cluster computers are more constrained by power and cooling costs for solving extreme-scale (or exascale) problems, the continuing progress and integration levels in silicon technologies make possible complete end-user systems on a single chip. This massive level of integration makes modern multicore chips all pervasive in domains ranging from climate forecasting and astronomical data analysis, to consumer electronics, smart phones, and biological applications. Consequently, designing multicore chips for exascale computing while using the embedded systems design principles looks like a promising alternative to traditional cluster-based solutions. This paper aims to present an overview of new, far-reaching design methodologies and run-time optimization techniques that can help breaking the energy efficiency wall in massively integrated single-chip computing platforms.

Keywords - *Exascale computing, multicore, small-world, game theory, consensus theory, fractal behavior.*

I. INTRODUCTION

Modern large-scale computing systems, such as data centers and High Performance Computing (HPC) clusters are severely constrained by power and cooling costs for solving extreme-scale (or exascale) problems. The increasing power consumption is of growing concern due to several reasons, e.g., cost, reliability, scalability, and environmental impact. The growing energy demands in data centers and HPC clusters are of utmost concern and there is a need to build efficient and sustainable computing environments that reduce the negative environmental impacts. On the other hand, continuing progress and integration levels in silicon technologies make possible complete end-user systems on a single chip. This massive level of integration makes modern multicore chips widely adopted in multiple domains. The power consumption efficiency (MIPS/mW) as well as the Time-To-Market (TTM) and consequently design time of such embedded systems are orders of magnitude more stringent than HPC. In this context, designing multicore chips for exascale computing while using embedded systems design principles looks like a promising alternative to traditional cluster-based solutions.

Design technologies in the era of massive integration present unprecedented advantages and challenges, the former being related to very high device densities, while the latter to soaring power dissipation and reliability issues. However, given the current trends in terms of power and performance figures, it is hard to imagine the future thousand cores platforms being designed using existing methodologies and tools. Trying to build from existing techniques applied to a relatively low number of processors towards such large-scale systems is a dead-end. Instead, more out-of-the-box, nature-inspired approaches (typically pursued in other areas) seem to be the only way to ensure the needed scalability for performance, power, and reliability figures as we move deeper into the realm of nanotechnology.

In this paper, we elaborate on new, far-reaching design methodologies and run-time optimization techniques that can help breaking the energy efficiency wall in massively integrated single-chip computing platforms where communication happens via the Network-on-Chip (NoC) approach. Specifically, we discuss three nature-inspired approaches - small-world networks, game-theoretic and consensus optimization, and multiscale network modeling - that can have significant impacts on achieving high energy efficiency in future multicore System-on-Chip (SoC).

II. SMALL-WORLD NETWORKS

A. Small-World NoCs with Long-Range Links

The existing methodology for implementing NoCs with planar metal interconnects is problematic due to the high latency and significant power consumption produced by the multi-hop links used in data exchanges. This limitation of conventional NoCs can be addressed by drawing inspiration from the interconnection mechanism of natural complex networks. As an example, the human cerebral cortex consists of approximately 10^{11} neurons linked by 10^{15} connections, thus forming a very complex network. A large proportion of intra-cortical connections are established locally, connecting neurons that are separated by only a few microns. Another large proportion of connections are spread over longer distances, linking neurons that are located in different cortical regions.

These connections ensure that the distant cortical sites can communicate fast enough in addition to the local intra-cortical communications. This kind of network topology, commonly referred to as small-world, can be incorporated in NoCs by introducing long-range, high bandwidth and low power links between distant cores [1].

There have been efforts to improve NoC performance by introducing low-latency long-range links and low power express channels between highly separated nodes, where the performance gain is achieved by bypassing intermediate NoC switches/routers [1][2]. These communication channels are more efficient in terms of power and delay compared to their conventional counterparts, but they are still, basically, metal wires. According to the International Technology Roadmap for Semiconductors, improving the characteristics of metal wires will no longer satisfy the performance requirements. Consequently, different revolutionary approaches for creating low latency, long-range communication channels like optical interconnects, on-chip transmission lines and wireless interconnects have been explored.

Although these emerging methodologies are capable of improving the power and latency characteristics of the traditional NoCs, they need more extensive investigation in order to determine their suitability for replacing and/or augmenting the existing metal/dielectric-based planar multi-hop NoC architectures. One possible innovative and novel approach, which addresses simultaneously the latency, power consumption and interconnect routing problems is replacing multi-hop wired paths in a NoC by high-bandwidth single-hop long-range wireless links [4][7]. The on-chip wireless links facilitate the design of a small-world NoC by enabling one-hop data transfers between distant nodes. In addition to reducing interconnect delay, eliminating multi-hop long distance wired communication reduces the energy dissipation as well. Design of the wireless NoC (WiNoC) involves suitable physical layer design and also architecture-space exploration. In the following subsections we elaborate on these two aspects.

B. Physical Layer Design

Propagation mechanisms of radio waves over intra-chip channels with integrated antennas were also investigated [3]. Indeed, zig-zag monopole antennas of axial length 1-2 mm can achieve a communication range of about 10-15 mm. Depending on antenna configuration and substrate characteristics, achievable frequency of the wireless channel can be within 50-100 GHz. In [4], the feasibility of designing on-chip wireless communication network with miniature antennas and simple transceivers that operate at the sub-THz range of 100-500 GHz has been demonstrated. A relatively long intra-chip communication range facilitates single-hop communication between widely separated blocks. This is essential to achieve the full benefit of on-chip wireless networks for multicore systems by reducing long distance multi-hop wireline communication. Despite all these advantages, in the mm-wave range, the antenna size is still a limitation.

If the transmission frequencies can be increased to THz/optical range, then the corresponding antenna sizes decrease, occupying much less of the chip real estate. Antenna characteristics of carbon nanotubes (CNTs) in the THz

frequency range have been investigated both theoretically and experimentally [5]. Bundles of CNTs are predicted to enhance performance of antenna modules by up to 40dB in radiation efficiency and provide excellent directional properties in far-field patterns. Moreover these antennas can achieve a bandwidth of around 500 GHz, whereas the antennas operating in the millimeter range achieve bandwidths of tens of GHz. Radiation characteristics of multi-walled carbon nanotube (MWCNT) antennas are observed to be in excellent agreement with traditional radio antenna theory [5], although at much higher frequencies of hundreds of THz. Using various lengths of the antenna elements corresponding to different multiples of the wavelengths of the external lasers, scattering and radiation patterns are shown to be improved. Such nanotube antennas are good candidates for establishing on-chip wireless communications links. But unlike the mm-wave antennas, CNTs will face multiple manufacturing challenges.

C. Architecture-Space Exploration

Modern complex network theory provides us with a powerful method to analyze network topologies [6]. Between regular mesh networks and completely random Erdős-Rényi topology, there are other classes of graphs, such as small-world and scale-free graphs. Networks with small-world properties have a very short average path length; this is usually measured as the number of hops between any pair of nodes in the network. The average shortest path length of small-world graphs is bounded by a polynomial in $\log(N)$, where N is the number of nodes, which makes them particularly interesting for efficient communication with minimal resources. This feature of small-world graphs makes them particularly attractive for designing scalable WiNoCs. Most complex networks, such as social networks, Internet, as well as certain parts of the brain exhibit small-world properties. A small-world topology can be constructed from a locally connected network by re-wiring connections randomly to any other node, which creates short cuts in the network. These random long-range links between nodes can also be established following probability distributions depending on the distance separating the nodes. It has been shown that such “shortcuts” in NoCs can significantly improve the performance compared to locally interconnected mesh-like networks with fewer resources than a fully connected system [1].

Our goal here is to use the “small-world” approach to build a highly efficient NoC based on *both* wired and wireless links. Towards this end, we first divide the entire system into multiple small clusters of neighboring cores and call these smaller networks subnets. As subnets are smaller networks, intra-subnet communication will have a shorter average path length than a single NoC spanning the entire system.

Fig. 1 shows a hybrid (wireless/wired) NoC architecture with heterogeneous subnets. Besides being connected among themselves, the cores in a subnet are connected to a centrally located hub through direct wireline links and the hubs from all subnets are connected in a 2nd level network forming a hierarchical architecture. This upper level of the hierarchy is designed to have characteristics of small-world graphs. To reduce the overhead arising out of the deployment of wireless

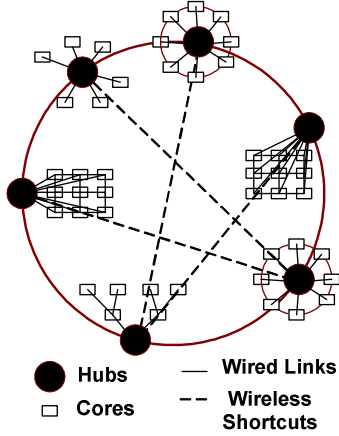


Figure 1. A conceptual hybrid (wireless/wired) NoC architecture with heterogeneous subnets following the small-world principle

links, neighboring hubs are connected by traditional wired links forming a bi-directional ring and a few wireless links are distributed between hubs separated by relatively long distances. Reducing the long-distance multi-hop wired communication is essential in order to achieve the full benefits of on-chip wireless networks for multicore systems.

D. Case Study

To demonstrate the energy efficiency of the WiNoC architectures, we briefly present a case study here. Using CNT antennas, the energy dissipation of the longest possible wireless link on the chip was found to be 0.33 pJ/bit [7]. Table 1 shows the packet energy dissipation of WiNoC, a flat mesh architecture and another hierarchical small-world NoC where the long-range links are implemented with G-lines. It is evident the WiNoC offers orders of magnitude improvement in energy efficiency compared to a flat mesh. Also, when the long-range links are designed with low swing, and ultra-low-latency communication wires, then also the WiNoC improves the energy efficiency by at least an order of magnitude. The largest contribution to packet energy in WiNoC is from the wireless and wireline link traversals combined in the upper level small-world network. This is because, on average, large portions of packets travel through the upper level of the WiNoC to reach other subnets. However as this level has a very small average path length (due to its small-world nature and due to the low power wireless channels), the absolute value of this energy dissipation is very small.

This clearly demonstrates how an hierarchical NoC

Table1. Packet energy dissipation for flat wired mesh, WiNoC and hierarchical G-line NoC architectures

System Size	Subnet Size	No. of Subnets	Flat Mesh(nJ)	WiNoC (nJ)	NoC with G-Line (nJ)
128	16	8	1319	22.57	490.30
256	16	16	2936	24.02	734.50
512	16	32	4992	37.48	1012.81

architecture with long-range wireless links can significantly improve the energy dissipation profile of the traditional wired counterparts. Such on-chip interconnection infrastructure will play a significant role in making single-chip massive multicore platforms energy efficient and hence sustainable.

III. DISTRIBUTED POWER MANAGEMENT

A challenge for massively parallel multicore systems is the performance optimization considering hundreds of cores. Techniques based on centralized and/or off-line optimizations suffer of significant limitations due to multiple factors: (1) High flexibility requirements. Multiple applications, unknown at design-time must be mapped on-line on programmable multicore platforms. One example is the Software Defined Radio (SDR) where the radio access can be uploaded on the fly. (2) Overall variability. Both intrinsic technological and environmental (e.g. battery behavior) variability emphasize the need for on-line, individual chip optimization. (3) On-chip complex communication. As discussed in the previous section, communication is the bottleneck of multicore systems. Consequently, centralized optimization is becoming increasingly complex and must be avoided. Instead, a distributed, on-line optimization process offers better scalability. In this section, we discuss two techniques, namely consensus and game theory, which can be used to deal with large-scale systems in a distributed manner.

A. Multicore Platform Overview

The generic multicore platform template is shown in Fig. 2. Built around a NoC framework, the platform includes homogeneous or heterogeneous functional units. Each of these tiles can be managed independently using sensors for monitoring the processors and actuators aiming overall performance optimization.

B. Game-theory Based Approach

Game theory is usually considered as a branch of applied mathematics. It aims at modeling the interaction between rational agents or decision makers. Osborne and Rubinstein define the game theory as “a bag of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact” [8]. The basic assumptions that underline this theory are that decision-makers are rational in their objective choices and take into account their knowledge or expectations of other decision-makers’ behavior (i.e., they reason strategically). A game is basically a scenario composed

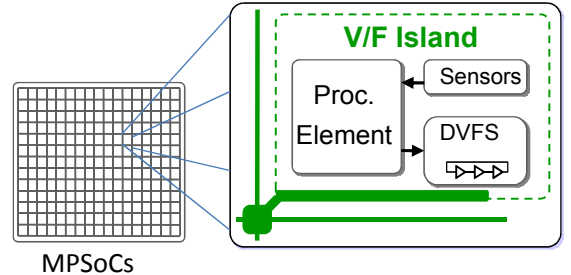


Figure 2. MPSoC voltage and frequency islands tiles

of rational players, each one having a set of possible actions. With respect to the entire system (including interactions), these actions produce consequences quantified as outcomes. In such a scenario, each player makes strategic choices (over the set of possible actions) to execute the best one, pursuing some predefined objective in terms of outcome. The individual success of each player (the quantitative outcome) depends not only on its own choice but also on the actions chosen by other players in the game.

The basic concepts of game theory can be applied to multicore architectures by considering each tile as an actor trying to optimize its outcome by varying its frequency and voltage knobs. Each tile has a dedicated objective function depending on the global optimization objective in terms of energy, temperature, performance, latency or a mix of the above (Fig. 3). The game is then organized as a three-cycle procedure: (1) Each unit decides on a frequency/voltage value maximizing its objective function, (2) the resulting values are communicated between the units (3) for each tile, choices of other units are integrated in the objective function for the next cycle. The game ends once the frequency/voltage values remain unchanged between two cycles. Starting a game cycle can be triggered by the physical sensors (e.g. local temperature fluctuation) or application modifications (e.g. new application mode requiring more/less computing).

C. Consensus Technique

Consensus is derived from the research on cooperative control theory. It was developed mainly for data processing in sensor network and multi-agent coordination. The consensus is defined as “an iterative process utilizing a predefined message-passing protocol, leading a set of communicating elements to an agreement on a value or a common behavior”. The application to distributed multicore optimization is relatively straightforward: Each unit has to interact with its neighborhood to reach an overall global “consensus” on the optimal settings corresponding to a given application. However, consensus aims at distributing the decision between different actors but does not focus on an optimal decision. Similarly to Johansson et al. [9], we propose to combine the consensus formalism with a sub-gradient method [10]. Consensus will ensure convergence in a distributed manner, while sub-gradient will push the

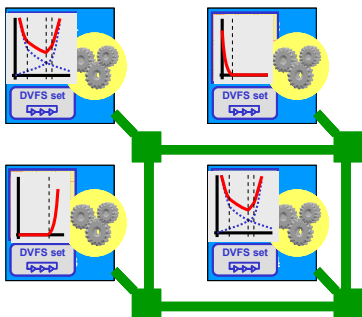


Figure 3. Application of game theory for MPSoC optimization

system towards an optimum

Fig. 4 illustrates the general consensus scheme with three actors (i.e., three VFIs). The global optimization goal is distributed to individual functions in each unit based on the consensus theory. Each individual function follows the sub-

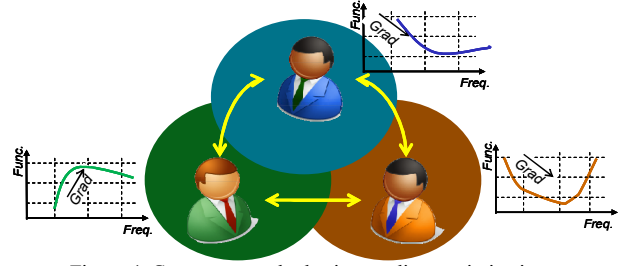


Figure 4. Consensus method using gradient optimization

gradient method by merging different optimization criteria in a convex equation. Similarly to game theory, the global optimization is performed based on successive iterations. The main difference compared to the game theory is the communication step which is based on a local neighborhood instead of a global scheme. Stopping as well as starting criteria are similar to those used in the game theory.

D. Application to Power Optimization of Real-Time Systems

The two above mentioned formalisms have been used for optimizing power consumption of multicore architectures having real-time constraints. Details on the formalisms can be found in [11] for the game-theoretic approach and in [10] for consensus-based approach.

The scheme is applied to MAGALI [12], which is a baseband chip consisting of 23 cores connected by an asynchronous NoC, each core having its own clock generator. MAGALI is intended to quickly switch between different Quality-of-Service (QoS) modes which correspond to a telecommunication standard. These modes have particular data-rate/quality characteristics resulting in mode-related processing requirements for different units in the chip. On one hand, optimizing the power consumption for the entire application results in finding, for each unit, the highest frequency required to satisfy the real-time constraints for all the modes. Such an optimization process results in power consumption overhead for low data-rate modes. On the other hand, the large number of modes makes an off-line multi-mode optimization not applicable. Finally, both game theory and consensus employ simple equations, as well as limited messages exchange with the environment. Figs. 5 and 6 present the gains obtained

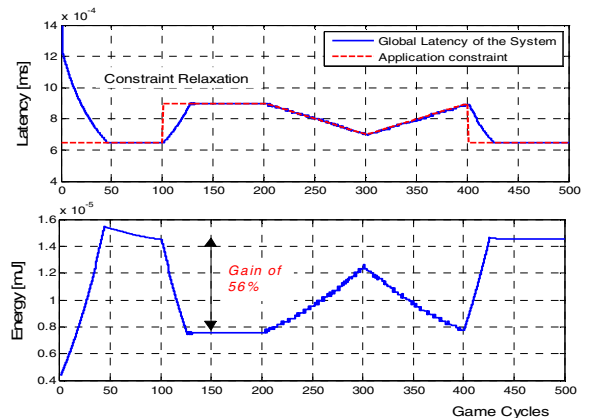


Figure 5. Power gain under moving real-time constraints using Game-Theory

through game theory and consensus approaches, respectively.

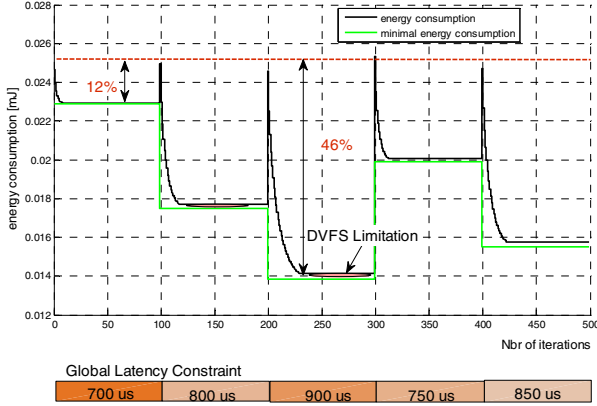


Figure 6. Power gain under moving real-time constraints using Consensus

IV. UNDERSTANDING TRAFFIC FRACTAL BEHAVIOR FOR BETTER MULTICORE OPTIMIZATION

Given the level of uncertainty induced by the dynamics of current and future applications, the optimization of multicore architectures need to rely on appropriate metrics and cost functions. Determining the right metrics and models can not disregard the statistical characteristics of actual workload, particularly when communication happens via the NoC approach [13]. Most current NoC design and optimization approaches rely on either deterministic or stationary probabilistic methods. However, many real world applications are time dependent as their stand alone or aggregated execution is highly dependent on user preferences, memory behavior, etc. This makes deterministic approaches unable to deal with variability and uncertainty levels that can exist in the NoC traffic. At the same time, most of the current probabilistic approaches for multicore optimization are predominantly based on Poisson modeling, general service times, and unlimited buffering capacity. For instance, when modeling the packet arrival process at a buffer, the memory-less assumption implies

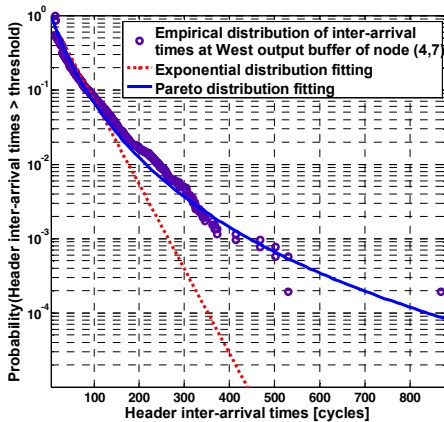


Figure 7. Probability of the header inter-arrival times at the West output buffer of an output node to exceed a certain threshold is better fitted by a Pareto distribution than an exponential law.

that the inter-arrival times are exponentially distributed with constant arrival rates. However, real traffic traces disprove this assumption. For instance, by running a set of multi-threaded applications from the SPLASH-2 suite, one can observe that the empirical header inter-arrival times at a particular output buffer of a node in an 8x8 mesh NoC are better fitted by a Pareto law rather an exponential distribution (see Fig. 7).

In addition, due to the heterogeneous nature of target applications (e.g., variability in the execution times) and generated traffic (e.g., time dependency of arrival processes), there is a stringent need for optimization methodologies that can take into account the dynamic behavior of the network exhibiting both non-zero higher order statistics and non-stationarity. As shown in Fig. 8, running three different applications that start in sequence and overlap their execution has significant implications on traffic behavior. Getting into more details, in this example, the mean, the variance and the kurtosis¹ of the inter-arrival times are computed as a function of the simulation time over time windows consisting of 200 data points. Besides the non-stationary behavior that can easily be observed in the mean and variance plots in Fig. 8, the kurtosis of the header inter-arrival times is non zero; this means that the standard Gaussian approaches are *not* appropriate for accurate traffic modeling. Instead, to account for the exhibited higher order statistics and non-stationary behavior that can result from various sources (e.g., changes in user preferences, feature updates for some applications) we need to develop a multifractal type of formalism that can be used for both traffic modeling, and performance prediction and optimization [14].

Relying on higher order statistics does not necessarily complicate the problem; instead, it can capture more accurately the features of the true distribution characterizing the workload and thus account for its variability during the optimization stage. On one hand, this helps overcome the difficulties related to synthesizing and optimizing NoCs while running full-blown simulations to determine whether or not the performance constraints are satisfied. Obviously, such an approach based on simulation can be too time consuming and eventually not adequate for current time-to-market deadlines. On the other hand, the observed fractal and non-stationary behavior in many

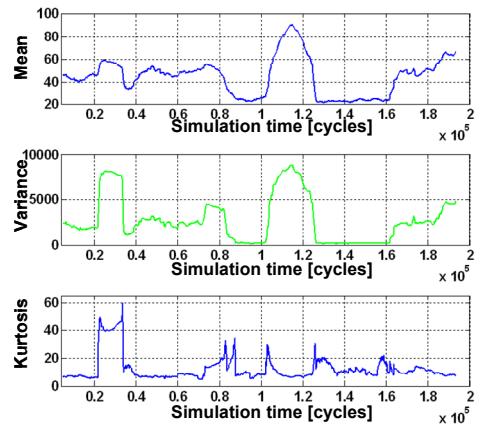


Figure 8. Higher order statistics and non-stationarity exhibited by running dynamic applications on a multicore system.

¹ Kurtosis is used to quantify the degree of non-Gaussian behavior present in a time series.

real world applications call for new optimization strategies that can go beyond the current optimization approaches based on stationary and average value metrics.

To better understand the significance of accounting for higher order statistics, let us discuss a possible optimization procedure that seeks to bring a variable of interest in a stochastic process to a given reference level. If we assume that this stochastic process (e.g., queue occupancy between two voltage and frequency islands as discussed in Fig. 2) is fully characterized by only the first two moments, then it is sufficient to consider the minimization of the square root of the difference between the considered variable and the reference level. This is actually what classical linear theory would do. However, when dealing with multifractal stochastic processes, the optimal control implies to minimize the variability exhibited by higher order moments. Obviously, any kind of power management strategy will be directly affected if such multifractal behavior is properly taken into account during the optimization process.

Simply speaking, for multifractal processes characterized by heavy tails, the variance is *not* enough to quantify the chances of seeing large values for the variables of interest (e.g., large source to destination latencies), as these very large values can be averaged out by many realizations of smaller magnitude. In such cases, we need to investigate the behavior of higher order moments (e.g., fourth order moment). In a way, the purpose of higher order moments is to exacerbate the deviations from the average behavior. Alternatively, the higher the order moment analysis is, the larger the contribution of fluctuations is on the overall behavior of the metric of interest. For instance, by considering the fourth order moment, we can obtain a more accurate picture about the probability of seeing rare events with respect to the latency; in turn, this may adversely impact the overall power management in the system. Fig. 9 illustrates this aspect by showing the comparison between the fourth and second power of the difference between a given latency and the overall mean latency between a two blocks in an MPEG4 decoder application.

V. CONCLUSIONS

In this paper, we have discussed three forward-looking ideas addressing new approaches for achieving energy efficiency in future multicore system-on-chips. By adopting a small-world interconnection infrastructure inspired by natural complex networks, where long distance communications will be predominantly achieved through high performance specialized single-hop wireless links, communications can be made significantly more energy efficient. Leveraging on modeling used in economics can help scaling up the optimization processes for multicore. Finally, performance analysis and power optimization based on higher order statistics have the promise of providing an accurate picture of the overall behavior of multicore systems in presence of highly variable workloads and user behaviors.

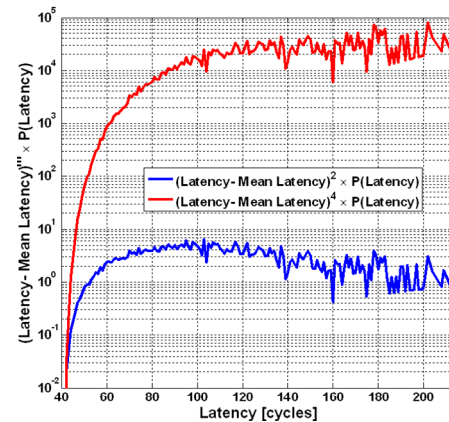


Figure 9. Comparison in terms of size of fluctuations between the second and fourth order moments.

REFERENCES

- [1] U. Y. Ogras and R. Marculescu, "It's a Small World After All": NoC Performance Optimization Via Long-Range Link Insertion", IEEE Transactions on VLSI Systems, Vol. 14, No. 7, July 2006
- [2] A. Kumar et al., "Toward Ideal On-Chip Communication Using Express Virtual Channels," IEEE Micro, Vol. 28, Issue 1, January-February 2008
- [3] Y. P. Zhang et al., "Propagation Mechanisms of Radio Waves Over Intra-Chip Channels with Integrated Antennas: Frequency-Domain Measurements and Time-Domain Analysis", IEEE Transactions on Antennas and Propagation, Vol. 55, No. 10, October 2007
- [4] S. B. Lee et al., "A Scalable Micro Wireless Interconnect Structure for CMPs", Proceedings of ACM Annual International Conference on Mobile Computing and Networking (MobiCom), 20-25 September, 2009
- [5] K. Kempa, et al., "Carbon Nanotubes as Optical Antennae," Advanced Materials, vol. 19, 2007
- [6] R. Albert and A.-L. Barabasi. "Statistical mechanics of complex networks," Reviews of Modern Physics, 74:47-97, January 2002
- [7] A. Ganguly et al., "Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems", IEEE Transactions on Computers (TC), August, 2010 (in press)
- [8] M. J. Osborne and A. Rubinstein, A Course in Game Theory. MIT Press, 1994.
- [9] Johansson, B. et al., "Subgradient methods and consensus algorithms for solving convex optimization problems,". Decision and Control, CDC, Dec. 2008
- [10] I. Mansouri, F. Clermidy, P. Benoit, L. Torres, "A Run-time Distributed Cooperative Approach to Optimize Power Consumption in MPSoCs", SOCC'10, Sept. 2010
- [11] D. Puschini, F. Clermidy, P. Benoit, G. Sassatelli, and L. Torres, "Dynamic and distributed frequency assignment for energy and latency constrained MP-SoC," DATE '09, March 2009
- [12] F. Clermidy et al. "A 477mW NoC-Based Digital Baseband for MIMO 4G SDR", ISSCC'10, San-Francisco, USA, Feb. 2010
- [13] P. Bogdan, M. Kas, R. Marculescu and O. Mutlu, "QuaLe: A Quantum-Leap Inspired Model for Non-stationary Analysis of NoC Traffic in Chip Multi-processors," NOCS'10, May 2010
- [14] P. Bogdan and R. Marculescu, "Workload characterization and its impact on multicore platform design," CODES+ISSS, Oct. 2010