Significance-Driven Computing for Big Data Applications: From *Buttery* Discussions to *Serious* Research

Rishad Shafik
Microsystems Research Group, School of Electrical and Electronic Engineering
Newcastle University, Newcastle upon Tyne, NE1 7RU
e-mail: Rishad.shafik@ncl.ac.uk

Abstract. Sometimes the best of the ideas evolve from informal discussions over coffee in cafes or places of retreat, such as *Buttery*. In this festschrift article we reflect on how one of these ideas inspired serious research on the development of a new generation of intelligent and energy-efficient processors by the Microsystems Research Group (jointly led by Alex and I) at Newcastle University.

1 Pretext

Continued technology scaling and engineering innovations have made digital services ever more affordable, thereby revolutionising the industrial age of data. A number of applications have emerged, which use deeply-embedded sensors to collect data and process them continuously, otherwise known as big data. Examples of these applications include smart computer vision, machine learning, e-governance and financial analytics. However, with widespread adoption of these services and applications the dimensionality and density of data are increasing drastically, rendering an unprecedented resource proliferation. Such proliferation of resources is likely to cause an uncontrolled energy consumption challenge for computing hardware systems, particularly to technology and service providers. For achieving transformational energy efficiency while also coping with increased performance needs a key solution is to design adaptive approximate computing systems. Recently, extensive research efforts have been initiated by the Microsystems Research Group (jointly led by Alex and I) at Newcastle University to design such computing systems. The aim is to learn the data significance during runtime and to process them dynamically adapting to their significance using novel logic design and system-level interactions. This rest of this article will report how these ideas evolved from informal discussions over coffee at Buttery and how they have inspired serious research on the design, implementation and validation of a new generation of intelligent and energy-efficient processors.

2 Prologue: *Buttery* Discussions

Buttery is a retreat space, popular among the students and staff members of the School of Electrical and Electronic Engineering (EEE) at Newcastle University. When I joined the school as a Lecturer in Electronic Systems (in September 2015), this was the first place where Alex invited me to have a coffee with him and discuss my research ambitions. During our rather informal discussions, the words 'adaptive', and 'approximate' were pronounced in very many contexts, mostly within the remits of electronic computing systems. Our discussions made good strides in understanding how the recent computing systems have evolved and how the future systems needed to emerge with magical adaptability features to play the perfect tradeoff game between energy consumption and performance every time.

In *Buttery* we started a little tradition of round-robin payment system for the coffee there – e.g. if Alex paid for the coffee the day earlier day I'd be the one to pay the next day and so on. Many a times, we would not remember who paid in the previous day. In those times we would vaguely try to remember by tagging with some 'significant' ideas that have been discussed in the previous day. We could then find out who paid for the coffee through backtracing techniques, much like *GCC's debugging features*. We both appreciated how our brains tag 'significance' to our everyday activities and process them accordingly. We also recognised how more 'significant' activities are stored by our brain in fast access memory for us to quickly recapitulate them.

Strangely enough, these little appreciations and observations triggered us to think and ask – 'why shouldn't our computing systems and storage subsystems also work likewise – i.e. modulate and adapt computation and storage efforts based on the data significance?' Finding "the" answer meant more work and investigation as outlined in the following sections.

3 Parode: Energy Consumption and Performance Tradeoffs

To find "the" answer it was important to understand the underlying challenges with existing computing systems. The increasing performance demands and resulting energy were the first issues that needed a bit of retrospective analysis without getting into details of big data just yet.

The performance needs of modern embedded computing systems have evolved dramatically over the years due to many emerging applications. According to Koomey's law, the performance per unit watt has doubled every 1.57 years [1], which is faster than the originally predicted 1.8 years by Moore's law [2]. The performance improvement is being enabled by technology scaling and innovative parallelisation techniques. However, the power consumption is also increasing uncontrollably as demonstrated by Dennard's scaling law [3] and numerous experimental observations [4-5]. Indeed, achieving scalable performance improvement with energy-efficiency is highly challenging for current and future generations of computing systems.

To minimise energy consumption, traditional approach is to reduce the supply voltage [6]. However, due to capacitive load imbalance, this also necessitates

lowering the operating frequencies [7]. Such reduction of supply voltage, coupled with the operating frequency is generally known as dynamic voltage/frequency scaling (DVFS), which ensures energy minimisation at the cost of degraded performance [6]. Over the years, significant research works have been carried out to demonstrate energy and performance tradeoffs in computing systems [8-11].

To improve performance at low energy consumption, an effective approach is to operate each processor core at low voltage/frequencies and also parallelise the computation tasks between multiple cores [12]. Significant research has been carried out in the recent past for understanding the best possible schemes and architectures to parallelise application tasks, including both compute- and memory-intensive ones. These works have revealed that the best energy efficiency is exhibited when compute-intensive parallel application tasks are exercised with higher number of cores operating at low voltage/frequency. The memory-intensive parallel workloads tend to favour lower number of cores at higher operating voltage/frequency for energy efficiency [13].

However, modern application workloads cannot always be statically labelled as CPU- or memory-intensive. Workloads change dynamically in these applications all the time. Often the same task or parallel thread can exhibit compute- and memory-intensive contexts at different times [13]. To address such dynamic variations, continuous runtime adaptation approaches have also been demonstrated recently by researchers. These approaches use feedback from the processor performance counters to adjust number of parallel threads, cores, architectural configurations and/or operating voltage/frequency to achieve energy-efficiency [9-10].

Existing system-level approaches have established the relationships between application workloads and power control knobs to achieve energy efficiency. However, modern big data applications are posing new challenges energy efficiency at required performance levels. This is because these applications are typically characterised by high volume and velocity (i.e. real-time processing needs) of data, which will require unprecedented resource allocations (for both computation and storage) using the existing approaches [14].

It is clear that existing computing approaches will hardly be enough to meet the growing performance needs for big data applications. Specific details of this are highlighted next.

4 Agôn: Key Challenges of Big Data Computing

Achieving energy efficiency for computing with big data applications is highly challenging using the existing approaches due to the following three major reasons:

Data proliferation: Existing big data applications have been characterised with a volume growth of several hundreds of petabytes per day. It is envisioned that such expansive growth will continue for the foreseeable future, generating many orders of magnitude higher volume of data. Current research suggests that the typical energy consumption of computing these data will soon approach the complexity of $O(N^3)$ or higher, where N is the number of data samples [15].

Undue performance scaling: To compute such a large volume of data at the required performance, currently existing computing systems exploit system-level controls, e.g. increased number of parallel cores and high operating frequencies through DVFS. However, these controls eventually cause diminishing returns in terms of increased energy consumption and complexity in the systems design with large area. In some cases, to meet the high performance demands custom designed accelerators are also used, which less flexible in terms of design, adaptability and programmability.
Indiscriminate Data Processing: The raw data of these applications acquired from the sources (e.g. sensors or humans) are processed identically in existing computing systems, ignoring the underlying informational value, i.e. significance, of the data. However, in reality the significance of acquired data varies dynamically over time and space depending on the application [16]. As a result, existing computing systems exhibit indiscriminate efficiency in data processing, resulting in large energy costs.

To foster the growth of this technology with the required energy reductions a paradigm shift is much needed from the existing significance-agnostic computing to significance-driven computing. Our research efforts in this direction is briefly outlined in the next section.

5 Parabasis: Significance-Driven Computation Research

Recently, *extensive* and *serious* research works have been initiated by the Microsystems Research Group at Newcastle University, led by Alex and I, to design such computing systems, including logic-level and system-level approaches. Our works at these levels are aimed at achieving holistic energy-efficiency through adaptive computation approach that can intelligently infer the significance of underlying information (i.e. bits and/or data). We give a brief account of the summary of our research to date in the following key areas, as follows.

5.1. Significance-Driven Low-level Logic Design:

In existing data processing logic design, there is no notion of modulating processing effort based on their bit-level significance. All bits are treated equally to generate a precise output. However, many emerging applications are inherently tolerant to imprecisions in less signigicant bits, such as computer vision, data mining and machine learning. This gives a unique opportunity to design next-generation processing logic such that computation efforts can adapted to the significance at bit-levels for achieving energy efficiency.

To this end, our low-level logic design is aimed at developing novel arithmetic and logical data processing subsystems. The more significant bits are treated with progressively higher precision through traditional computation, while bits with lower significance are compressed using variable clustering (i.e. vertical grouping of partial terms). As a result of such logic design, complexity of computation in terms of logic cell counts and length of the critical paths are drastically reduced.

A number of multipliers using this approach have recently been designed using SystemVerilog and synthesised using EDA tools. Our post-synthesis experiments with a 128-bit multiplier showed that up to 60% less energy consumption and 53%

performance improvement can be achieved, when compared with traditional Dadda and Wallace multipliers. These gains are achieved at a low loss of accuracy due to significance-driven bit processing - with up to 30% inaccuracies for small valued operands and exponentially reduced imprecision for higher values. We are currently designing real application demonstrators using this approach to show the comparative advantages of our approach.

5.2. Extracting Data Significance at System-level:

Data are crucial parts of big data applications. However, not all data carry the same informational value. Traditional computing systems are agnostic of such values as all data are processed indiscriminately and equally. As a result, a large energy cost is incurred for processing a large volume of data, much of which have little or no significance.

To extract the informational value of data at operational time we have taken initiatives to modify processor architectures underpinning the theory and practices of approximate computing and machine learning. These involve designing a data inference engine as middleware, which will use application domain-specific knowledge to evaluate measurable significance of data that are being processed in parallel. The aim is to use the measured significance to dynamically scale the computation efforts, e.g. data with higher significance will be processed using accurate data processing unit and standard data flow (prefetch, decode and execute), while those with less significance will be processed using low-complexity and inaccurate data processing unit using already existing cache data (i.e. avoiding further prefetch flow). This will eventually result in significant energy reductions for data-intensive applications.

To date we have already carried out proof-of-concept designs, currently also carrying out implementation-ready logic design. We will follow this by integrated processor design and optimisation, including cache localisation for fast middleware routines and DVFS features to corroborate energy reductions. The new processor will be validated using real case study big data applications for practical demonstrations to key industries and academia in the UK and beyond. The overarching goal will be to create a critical mass in this important area.

5.3. Significance-driven Big Data Storage:

Memory constitutes a major component in modern computing systems. The energy efficiency or performance of these systems cannot be achieved in isolation without considering the memory effect. As such, our future research plan will include significance-driven memory management, including cache optimisation and development of fast memory systems for significant data.

6 Exode: Conclusions

Our rather informal *Buttery* discussions over such a short period time have given us the impetus to combine our expertise synergistically to carry out *serious* research on a new breed of intelligent processors. So this *Exode* is really just a beginning rather

than an end. We will continue to engage in further fruitful discussions in the future, potentially involving other interested academics and industrial peers to advance the understanding of research needs further.

We expect that our research will be a small but important contribution to UK's world-leading portfolio in low power systems engineering. A key differentiator for maintaining this portfolio and enhancing it further will be to be able to design a new breed of intelligent processors with control over computation efforts. It is a relatively new area, which combines and advances the theory and practices of traditional approximate computing and machine learning. Our innovations in this space will be crucial to enable many emerging applications of profound impact on our businesses and society.

References

1. Koomey, Jonathan; Berard, Stephen; Sanchez, Marla; Wong, Henry; (March, 2010). Implications of Historical Trends in the Electrical Efficiency of Computing. IEEE Annals of the History of Computing 33 (3): 46−54..
2. Moore, Gordon E. (April 1965). "Cramming more components onto integrated circuits". Electronics 38 (8): 1-4.
3. Dennard, Robert H.; Gaensslen, Fritz; Yu, Hwa-Nien; Rideout, Leo; Bassous, Ernest; LeBlanc, Andre (October 1974). Design of ion-implanted MOSFET's with very small physical dimensions. IEEE Journal of Solid State Circuits SC–9 (5): 668-678.
4. Schaller, R. R. (1997). Moore's law: past, present and future. IEEE spectrum, 34(6), 52-59.
5. Moore, G. E. (1995, May). Lithography and the future of Moore's law. In SPIE's 1995 Symposium on Microlithography (pp. 2-17). International Society for Optics and Photonics.
6. Shafik, R. A., Al-Hashimi, B. M., & Chakrabarty, K. (2010, March). Soft error-aware design optimization of low power and time-constrained embedded systems. In Proceedings of the Conference on Design, Automation and Test in Europe (pp. 1462-1467). European Design and Automation Association.
7. Wang, L., Von Laszewski, G., Dayal, J., & Wang, F. (2010, May). Towards energy aware scheduling for precedence constrained parallel tasks in a cluster with DVFS. In Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on (pp. 368-377). IEEE.
8. Shafik, R. A., Al-Hashimi, B. M., Kundu, S., & Ejlali, A. (2009). Soft Error-Aware Voltage Scaling Technique for Power Minimization in Application-Specific Multiprocessor System-on-Chip. Journal of Low Power Electronics, 5(2), 145-156.
9. Das, A., Kumar, A., Veeravalli, B., Shafik, R., Merrett, G., & Al-Hashimi, B. (2015, March). Workload uncertainty characterization and adaptive frequency scaling for energy minimization of embedded systems. In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (pp. 43-48). EDA Consortium.
10. Shafik, R. A., Yang, S., Das, A., Maeda-Nunez, L. A., Merrett, G. V., & Al-Hashimi, B. M. (2016). Learning transfer-based adaptive energy minimization in embedded systems. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 35(6), 877-890.
11. Kim, S. G., Eom, H., Yeom, H. Y., & Min, S. L. (2014). Energy-centric DVFS controlling method for multi-core platforms. Computing, 96(12), 1163-1177.
12. Shafik, R. A., Das, A., Yang, S., Merrett, G., & Al-Hashimi, B. M. (2015, January). Adaptive energy minimization of OpenMP parallel applications on many-core systems. In

Proceedings of the 6th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architectures (pp. 19-24). ACM.

13. A Aalsaud, R Shafik, A Rafiev, F Xia, S Yang & A Yakovlev (September 2016). Power-Aware Performance Adaptation of Concurrent Applications in Heterogeneous Many-Core Systems. In Proceedings of International Symposium on Low Power Electronics and Design (ISLPED) (in press). IEEE.
14. Burke, D., Shafik, R., and Yakovlev, A. (March 2016). Challenges and Opportunities in Research and Education of Heterogeneous Many-Core Applications, European Symposium on Microelectronics Education (EWME) (in press).
15. Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing, 74(7), 2561-2573.
16. Mohapatra, D., Karakonstantis, G., & Roy, K. (2009, August). Significance driven computation: a voltage-scalable, variation-aware, quality-tuning motion estimator. In Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design (pp. 195-200). ACM.