Towards Approximate Computing for Achieving Energy vs. Accuracy Trade-offs

Aleksandr Ometov and Jari Nurmi
Electrical Engineering Unit, Tampere University
Tampere, Finland
Emails: aleksandr.ometov@tuni.fi, jari.nurmi@tuni.fi

***Abstract*—Despite the recent advances in semiconductor technology and energy-aware system design, the overall energy consumption of computing and communication systems is rapidly growing. On the one hand, the pervasiveness of these technologies everywhere in the form of mobile devices, cyber-physical embedded systems, sensor networks, wearables, social media and context-awareness, intelligent machines, broadband cellular networks, Cloud computing, and Internet of Things (IoT) has drastically increased the demand for computing and communications. On the other hand, the user expectations on features and battery life of online devices are increasing all the time, and it creates another incentive for finding good trade-offs between performance and energy consumption. One of the opportunities to address this growing demand is to utilize an Approximate Computing approach through software and hardware design. The APROPOS project aims at finding the balance between accuracy and energy consumption, and this short paper provides an initial overview of the corresponding roadmap, as the project is still in the initial stage.**

***Index Terms*—Computing, approximation, hardware design, communications, EU projects**

I. Introduction

It is expected that already in 2040, modern devices with computing capabilities would require more energy than the energy resources can provide [1], see Fig. 1. As soon as in five years, the development and heavy use of the data centers alone is expected to consume up to a quarter of all generated electricity [2]. Moreover, from the communications side, the energy consumption tendency of mobile broadband networks and smartphones would be comparable to data centers. Finally, the development of the Internet of Things (IoT) paradigm is also expected to bring up more than 50 billion interconnected devices [3] creating even more pressure on the networks and data centers [4].

Horizon 2020 (H2020) European Union project Approximate Computing for Power and Energy Optimisation (APROPOS) aims to train 15 Early Stage Researchers (ESRs) to tackle the challenges of future embedded and high-performance computing by using disruptive methodologies (http://www.apropos-itn.eu). APROPOS has only started in November 2020 and has just passed the recruitment phase. Therefore, it is still in the *initial phase* of the ESRs starting their literature reviews to support the research demands. The beneficiaries are very dispersed and located in Finland, Sweden, The Netherlands, Austria, Italy, Switzerland, the UK, Spain, and France. Additionally,

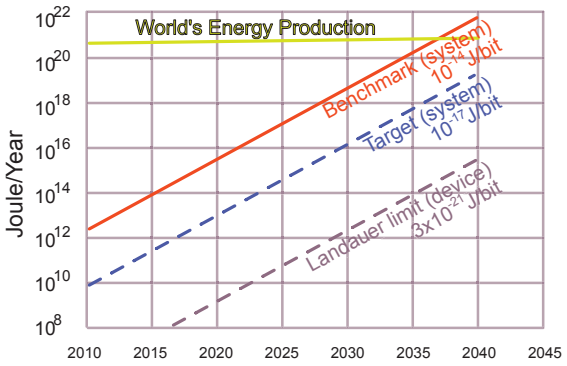

Fig. 1. Energy consumption trend in computing vs. the world energy production.

there are a dozen industrial partners from these countries, plus Ireland and Poland.

In order to alleviate the energy issues, the research executed by the APROPOS ESRs will decrease energy consumption in both distributed computing and communications for Cloud-based cyber-physical systems by introducing an adaptive energy-aware Approximate Computing overlay. Coupling Approximate and Adaptive Precision Computing paradigms with application-specific processing structures is critical in achieving the required energy efficiency improvements [5]. Since the energy consumption is a product of (computing or communication) time and average power consumption of the device while carrying out an operation, these two factors, time and energy, must be addressed for decreasing the power dissipation [6]. The precious and novel third dimension, accuracy adjustment for reducing time and power, is the APROPOS network's main contribution.

Research-wise, APROPOS proposes adaptive Approximate Computing techniques to optimize energy-accuracy trade-offs [7]. Luckily, the possibility to exchange the accuracy to either less power or less operational time consumed could be found for various applications, e.g., in many parts of the global data acquisition, transfer, computation, and storage systems [8]. By introducing a new dimension, accuracy and design optimization, energy efficiency can be improved by 10x-50x.

978-3-9819263-6-1/DATE22/©2022 EDAA

APROPOS is training the future generation's spearheads to cope with the technologies, methodologies, and tools for successfully applying Approximate Computing to power and energy saving. In this first-ever Innovative Training Network (ITN) addressing Approximate Computing, the training is to a large extent done by researching energy-accuracy trade-offs on the circuit, architecture, software, and system-level solutions, bringing together world-leading European organizations.

In a compact form, the main Research Objectives of APROPOS in the domain of Approximate Computing are:
RO1: To develop a solid understanding of algorithmic formulations and the nature of mathematical models and foundations that enable approximation.
RO2: To develop system software support and runtime support to enable and efficiently exploit error resilience characteristics for performance enhancement, power and energy savings, and other system parameters such as lifetime extension and reliability, and functional safety constraints.
RO3: To design and implement various approximate hardware classes, including transprecision circuitry, inexact and simplified logic blocks, and modules to support extensions from preceding layers of system and application software.

II. Towards accuracy vs. energy efficiency trade-offs

The inevitable growth in the number of devices is tightly related to Cloud computing and the emerging next-generation (5G-6G) communications infrastructure [9]. Already today, IoT sensor nodes are responsible for collecting the data for analysis, Machine Learning (ML), and long-term storage and deployment in the Cloud. However, Cloud computing has introduced some drawbacks compared to local data processing, i.e., exceptionally high latency and energy consumption due to the communication overheads and inefficient general-purpose processing engines.

A recent trend is to bring the computations closer to the endpoint devices by applying Edge or Fog Computing at the edge of the network [10], [11], with the advantages of lower total energy, lower latency, reduced communication needs, and location-awareness [12]. As a drawback, those devices located at the network edge usually have a limited amount of resources: computational power, memory, and energy storage [13], [14]. These limitations have an impact on the capability of those devices to process data locally.

In addition to Cloud computing, the mobile broadband network is a significant energy consumer. There are predictions that communications technologies can consume up to half of the produced electricity by 2030 [15]. Future cellular networks will support energy-saving when data transfers can be minimized in a manner resembling the stop-and-start functionality of modern car engines. The main challenge here is to allow for such energy-saving periods in data transfers and to render the transfer bursts as low-energy as possible. Also, the (IoT and mobile) devices and communications in the periphery of future cellular networks need to be optimized for lower energy consumption [7]. The battery-operated devices are the most dependable on low energy dissipation, although not that big contributor to energy consumption on the global scale.

Radical efforts are needed to improve the energy efficiency of computing and communications. As pointed out, energy consumption is the product of time and power. APROPOS postulates that we can achieve less power consumption and shorter computing time by decreasing processing accuracy at different parts of the computing chain. Trading off accuracy for improved power consumption and/or performance is generally known as Approximate Computing. Transprecision Computing [16] is its subset and extension, where adaptive precision is applied to different parts of the computation, possibly without affecting the system's end-to-end accuracy.

III. Research Directions

Overall, the main Approximate Computing future development directions could be classified into three major components: algorithms to be developed and executed, software-based approximation in order to provide smooth integration, and, most significantly, approximate hardware components base. The main identified research directions are as follows.

A. Algorithmic component

For the timely and efficient development of the Approximate Computing paradigm, firstly, there is a need to provide a solid understanding of algorithmic formulations and the nature of mathematical models and foundations that enable approximation in the field of *algorithms and applications*. Those from different emerging domains should be extensively experimented with for profiling and benchmarking to quantify the effect of error resilience in each class of applications.

Insights on algorithmic formulation and applications' workload characteristics should be formalized as a unified entity that can be used to guide the system software and hardware development phases. Finally, new information security and data privacy aspects should be carefully considered while developing the distributed systems [17]. The aspects of the user trust of delegating their data should become a basis for the novel systems development. The main directions from the algorithmic perspectives lie in the following research lines.

First, efficient Machine (including Deep) Learning inference models for low-energy applications are expected to support the development of new algorithms for neural network compression for efficient inference on IoT devices as well as the automation of synthesis of neural network models targeting IoT. Notably, it is essential to understand the impact of variable precision on the application error in offline and mixed Machine Learning approaches, i.e., to provide a fast and accurate prediction of the impact of approximate operators on complex computation environments. This road should also be supported by developing novel stochastic approaches to quickly estimate the impact in terms of precision and other

design dimensions such as power, area, code complexity, and execution time.

Nonetheless, Approximate Computing could be used to make future communications more efficient by, e.g., identifying the energy-accuracy trade-offs in approximate Medium Access Control (MAC) layer and applying Approximate Computing to low-energy adjustable radio protocol design for both broadband and positioning technologies.

Another separate niche corresponds to the dynamic data-dependent precision scaling, e.g., to implement novel low-cost schemes for enabling dynamic approaches and probabilistic dynamic error confinement techniques.

There are many more architectural challenges to be addressed in the Approximate Computing field while the developed solutions may become an underlay for the overall ecosystem's software and hardware components.

B. Software component

Indeed, most developers perceive computing as a black box where the data is fed, and the result is delivered as soon and as precisely as possible. However, it becomes very different while trying to balance computational expenses while executing software on various energy-constrained devices.

Today, there is a must to develop the system software- and runtime support to efficiently exploit error resilience characteristics for performance enhancement, power, energy savings, and other system parameters such as lifetime extension, reliability and functional safety constraints. Moreover, it is necessary to translate the notion of applications' performance requirements and error resilience characteristics into system-level control parameters to fulfill those requirements in the formal guidelines.

In particular, enabling novel industrial and scientific applications for benchmarking Approximate Computing strategies would require creating those as open-source solutions. One of the most promising directions in the software niche is related to the automated driving and approximation of computation at the network edge to achieve low latency, resiliency, and better quality of experience for end-users while keeping energy efficiency and growing load in mind.

To this point, while developing such benchmarks, one should keep in mind the design space exploration related to accuracy-aware computing. It is of utmost importance to define and implement a framework for accuracy-aware computing to be used by the researchers.

To summarize, software tools aim to provide easy-to-use Approximate Computing tools for the developers and integrators. Indeed, those systems would mainly focus on executing conventional architectures and hardware, while the next subsection mainly focuses on lower-level Approximate Computing strategies.

C. Hardware component

Finally, various novel classes of approximate hardware, including transprecision circuitry, inexact and simplified logic blocks, and modules to support extensions from preceding system and application software layers, are to meet the physical world. Indeed, the Approximate Computing paradigm implies accuracy as an extra parameter to be considered during the overall production flow and even through mission time. Let us suppose that an application cannot effectively communicate its accuracy requirements to the underlying hardware platform and monitor the computational quality at runtime.

In that case, it is unrealistic to expect the system to achieve the desired amount of energy-efficiency gain under quality constraints. Vice versa, the hardware layer has to be able to communicate to the software layer any modification of the accuracy or precision of the computation. For this reason, a strong interaction between those three directions is present: the application characteristics, both performance and error resilience, are translated through system software to hardware layer, which enables approximate execution to support and leverage the hints provided from the application level down to implementation level.

To start with, novel compiler technologies should be developed to support the execution of Approximate Computing at the compiler level, balancing the need to minimize user intervention with the ability of the user to understand the trade-offs involved.

Following the autonomous driving scenario, reconfigurable approximating accelerators could be developed to reach a lower level for cases of Edge computing. The need to redesign the dynamically adjustable approximation and vectorization for trade-offs in parallelism, energy, latency, and accuracy is significant for heterogeneous environments that include various IoT devices ranging from autonomous cars to wearables and sensors.

Moreover, the development of novel architectures would require to put fault-tolerance and security as one of the main priorities that may require finding balance for such seemingly incompatible concepts, i.e., how can those be deployed at the system level to improve the fault tolerance of embedded systems, while trying to meet their performance and power/energy consumption constraints by approximation.

IV. Conclusions

The physical limitations in hardware design are pushing towards developing new techniques to prolong arrival to the superior energy consumption limit. One of the solutions to achieve that is by sacrificing the precision to energy savings by utilizing various techniques from the Approximate Computing field.

This research area may be mistakenly considered purely hardware-focused. However, a significant share of related ongoing and future research also covers algorithmic and software-oriented directions. Those include but are not limited to benchmarking, Edge/Fog applications, precision scaling, as well as various prediction techniques, among others. Those, in turn, bring completely new challenges to be resolved by both academy and industry.

Finally, we would like to highlight that it would be impossible to develop Approximate Computing solutions in a timely manner without actual industrial support. Therefore, supported by the European Union, the APROPOS project is expected to push this technology advancement and integration significantly.

ACKNOWLEDGMENT

This work was supported by the funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska Curie grant agreement No. 956090.

REFERENCES

[1] A. Burgess and T. Brown, "By 2040, There May Not Be Enough Power for All Our Computers," *HENNIK RESEARCH*, 2021.

[2] J. M. Lima, "Data Centres of the World Will Consume 1/5 of Earth's Power by 2025," *Data Economy*, 2017.

[3] S. Smith, "IoT Connections to Grow 140% to Hit 50 Billion By 2022, As Edge Computing Accelerates ROI," *Juniper Research* , 2018.

[4] W. B. Qaim, A. Ometov, A. Molinaro, I. Lener, C. Campolo, E. S. Lohan, and J. Nurmi, "Towards Energy Efficiency in the Internet of Wearable Things: A Systematic Review," *IEEE Access*, 2020.

[5] G. S. Rodrigues, J. Fonseca, F. Benevenuti, F. Kastensmidt, and A. Bosio, "Exploiting Approximate Computing for Low-Cost Fault Tolerant Architectures," in *Proc. of the 32nd Symposium on Integrated Circuits and Systems Design (SBCCI)*, pp. 1–6, IEEE, 2019.

[6] A. Bosio, S. Di Carlo, P. Girard, E. Sanchez, A. Savino, L. Sekanina, M. Traiola, Z. Vasicek, and A. Virazel, "Design, Verification, Test and In-Field Implications of Approximate Computing Systems," in *Proc. of IEEE European Test Symposium (ETS)*, pp. 1–10, IEEE, 2020.

[7] A. George and A. Ravindran, "Scalable Approximate Computing Techniques for Latency and Bandwidth Constrained IoT Edge," in *International Summit Smart City 360°*, pp. 274–292, Springer, 2020.

[8] R. Airoldi, F. Campi, and J. Nurmi, "Approximate Computing for Complexity Reduction in Timing Synchronization," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–7, 2014.

[9] R. Florea, A. Ometov, A. Surak, S. Andreev, and Y. Koucheryavy, "Networking Solutions for Integrated Heterogeneous Wireless Ecosystem," *Cloud Computing*, p. 103, 2017.

[10] N. Garg, M. Sellathurai, V. Bhatia, and T. Ratnarajah, "Function Approximation Based Reinforcement Learning for Edge Caching in Massive MIMO Networks," *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2304–2316, 2020.

[11] H. Zahmatkesh and F. Al-Turjman, "Fog Computing for Sustainable Smart Cities in the IoT Era: Caching Techniques and Enabling Technologies: An Overview," *Sustainable Cities and Society*, vol. 59, p. 102139, 2020.

[12] A. Ometov, O. Chukhno, N. Chukhno, J. Nurmi, and E. S. Lohan, "When Wearable Technology Meets Computing in Future Networks: A Road Ahead," in *Proc. of the 18th ACM International Conference on Computing Frontiers*, pp. 185–190, 2021.

[13] W. Yu, A. Najafi, Y. Huang, and A. Garcia-Ortiz, "Combination of Task Allocation and Approximate Computing for Fog-Architecture-Based IoT," *IEEE Internet of Things Journal*, vol. 8, no. 9, pp. 7638–7648, 2020.

[14] J. H. Anajemba, J. A. Ansere, F. Sam, C. Iwendi, and G. Srivastava, "Optimal Soft Error Mitigation in Wireless Communication Using Approximate Logic Circuits," *Sustainable Computing: Informatics and Systems*, vol. 30, p. 100521, 2021.

[15] A. S. Andrae and T. Edler, "On Global Electricity Usage of Communication Technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[16] G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benini, "A Transprecision Floating-Point Platform for Ultra-Low Power Computing," in *Proc. of Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1051–1056, IEEE, 2018.

[17] M. Gao, Q. Wang, M. T. Arafin, Y. Lyu, and G. Qu, "Approximate Computing for Low Power and Security in the Internet of Things," *Computer*, vol. 50, no. 6, pp. 27–34, 2017.