Green Computing for Big Data and Machine Learning

Conference Paper · January 2022

DOI: 10.1145/3493700.3493772

CITATIONS
0

READS
40

3 authors, including:

Hrishav Bakul Barua
Tata Consultancy Services Limited
34 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Dr-Kartick Chandra Mondal
Jadavpur University
51 PUBLICATIONS **162** CITATIONS
SEE PROFILE

Green Computing for Big Data and Machine Learning

Hrishav Bakul Barua
hbarua@acm.org
Robotics & Autonomous Systems, TCS Research
Kolkata, West Bengal, India

Kartick Chandra Mondal
kartickjgec@gmail.com
Department of Information Technology, Jadavpur University
Kolkata, West Bengal, India

Sunirmal Khatua
skhatuacomp@caluniv.ac.in
Department of Computer Science & Engineering, University of Calcutta
Kolkata, West Bengal, India

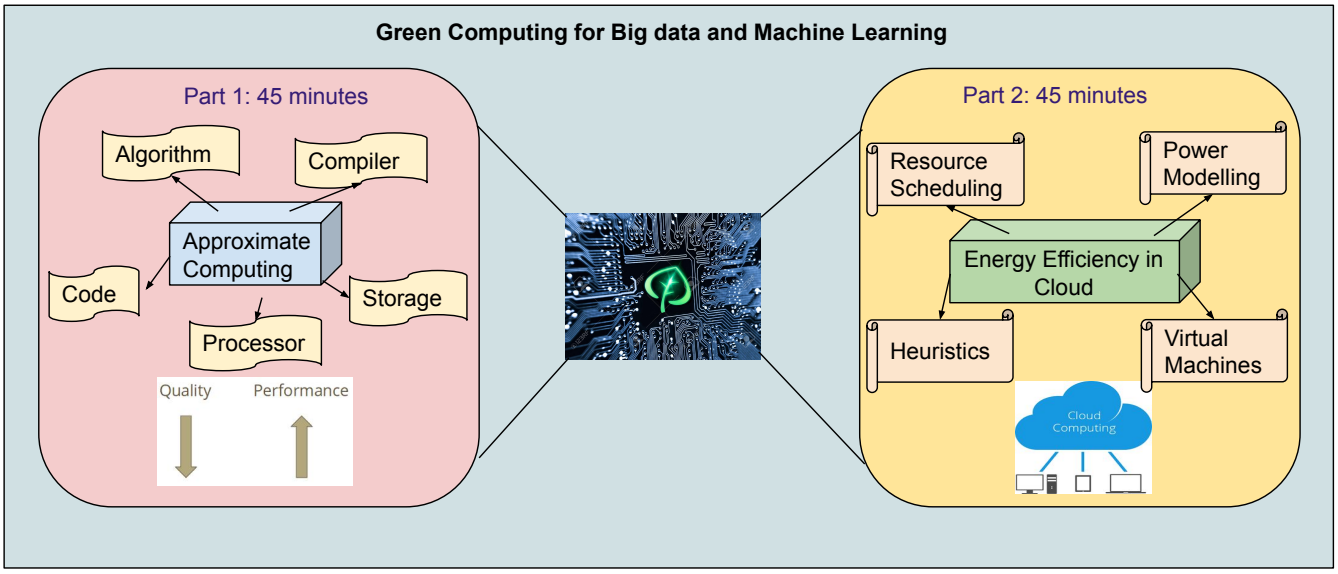

Figure 1: An overview of the topics for the proposed tutorial.

ABSTRACT

The current decade has beheld a tremendous spike in data **volume, velocity, variety** and many other such aspects which we call as **Big Data** and which gave birth to a new kind of science commonly known as "*Data Science*". With the "*Data Apocalypse*" in progress, it is evident that the conventional methods to handle these data would not suffice. We need distributed and parallel architectures like **Cloud services** (IaaS, PaaS, SaaS, STaaS, etc.). But is that enough to satisfy our needs? Here, we propose a tutorial in a very different direction when we are talking about Data Science, that is, bringing greenness in Big Data and Machine Learning (ML). We divide the tutorial into two parts primarily assuming that we are using cloud backbone for analytic and prediction tasks. The first part speaks about the techniques and tools to bring energy efficiency/greenness in the algorithmic and code level for Big Data and ML using Approximate Computing. The second part talks about the green techniques and power models at the infrastructural level for the cloud.

CODS-COMAD 2022, January 8–10, 2022, Bangalore, India

ACM ISBN 978-1-4503-8582-4/22/01…$15.00
https://doi.org/10.1145/3493700.3493772

CCS CONCEPTS

• **Hardware** → **Power and Energy**; **Power Estimation and Optimization**; • **Computing Methodologies** → *Machine Learning*; • **Information Systems** → **Data Mining**; *Data Management Systems*; • **Computer Systems Organization** → **Cloud Computing**.

KEYWORDS

Green Computing, Approximate Computing, Cloud Computing, Big Data, Data Science, Machine Learning, Resource Scheduling, Power Modelling, Heuristics

ACM Reference Format:
Hrishav Bakul Barua, Kartick Chandra Mondal, and Sunirmal Khatua. 2022. Green Computing for Big Data and Machine Learning. In *5th Joint International Conference on Data Science and Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD 2022), January 8–10, 2022, Bangalore, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3493700.3493772

1 DURATION OF TUTORIAL

Our proposed tutorial will be of 1.5 hours. The proposers will distribute the time among themselves appropriately to cover the topic efficiently.

2 OUTLINE

Motivated from the fact that green computing is gaining prominence for its energy efficient techniques and/or accelerating heavy computing intensive processes, we want to show the utility of it in the data science domain. We intend to present a tutorial on energy efficient and green data processing and analytics. Our tutorial can be seen as a combination of two different parts.

(1) *Use of Approximate Computing in Big Data and ML:* (45 minutes approximately)
   - Basics and concepts of Approximate Computing [3]
   - Techniques, methods and tools [2, 7, 13]
   - A compiler tool chain for green computing in Big Data and ML [14, 15]
   - Cloud based green computing frameworks for huge data processing and ML [1, 5, 11, 12]

(2) *Green techniques and power models for cloud:* (45 minutes approximately)
   Use of cloud services increase exponentially over the last 10 years. As of 2020, 93% of businesses adopt multi-cloud infrastructures to run their business. Consequently, data center power consumption account for roughly 3% of global electricity generated on the planet [19]. The situation is alarming with the adoption of 5G and IoT devices [10]. So energy-efficient scheduling of data center resources becomes an important research challenge among the researcher for the last few years. In the literature, the energy-efficient scheduling problem for data center resources are modelled as a multi-dimensional bin packing problems which are NP-hard in nature. Therefore, most of the solutions are heuristic based or approximation algorithms to solve the scheduling problem [16–18]. Moreover, the efficiency of such a resource scheduling algorithm heavily depends on proper power modelling of the heterogeneous physical resources used in the data center to deliver different cloud services. In this tutorial we'll discuss various heuristics and power modelling used for providing green solution to the scheduling of data center resources.

The Figure 1 shows a holistic overview of the tutorial as discussed with time distribution.

3 DESCRIPTION

This section briefly describes the details of the tutorial in a concise manner.

(1) *Use of Approximate computing in Big data and ML:* The paradigm called 'Approximate Computing' [3] unleashes a whole new set of possibilities in the age of Data science. Approximate Computing (AC) is the idea of reducing the computational time and energy requirements by allowing a tolerable amount of error or *in-exactness* in the system. This system can be a piece of code, an algorithm, a hardware set, an arithmetic or logic circuit, a storage system, and even a data communication system [2, 7, 13]. Consider the following simple cases:

$$Y = \frac{1}{X} \quad \text{and} \quad Y = 2.823 - 1.882 \times X$$

where Y in the right-hand side equation gives an approximate (in-exact) result very close to the left-hand side one for a range of $X = [0.5, 1]$. But, the catch here is that, the right-hand side equation takes much lesser time (almost 2 times faster) to execute compared to the left-hand side one because division operation requires much higher machine cycles to execute than mere addition/subtraction or even multiplication. This also guarantees energy savings by reducing the machine cycles in the CPU.
Since, Data science, Big data, Machine learning, etc., are inherently tolerant towards approximate results or in-exact outcome, we can leverage the power of AC in these domains effectively. There are many methods and techniques to implement approximate computing in software as well as hardware level [2, 13]. Some famous techniques involve loop perforation, data sampling, task skipping/dropping, memoization, and precision scaling [13]. In hardware level, approximate logic circuits, approximate memory units, and approximate processors are also available [2, 13]. Scientists have also proposed full compiler tool-chains to facilitate AC both in software and hardware level in a more cohesive manner. Some examples of such compiler tool-chains are *Accept* [14] (based on C programming) and *EnerJ* (based on Java programming) [15]. Since, cloud is an integral part of Big data processing environments now-a-days [4], researchers have also come up with cloud based AC frameworks to facilitate AC in distributed and parallel environments, mostly over the cloud [1, 5]. Some examples of such frameworks are *ApproxHadoop* [11] and *IncApprox* [12]. These are built on top of famous Big data processing frameworks like Hadoop and Spark.

(2) *Green techniques and power models for cloud:* The data centers contain a large number of hosts or servers that consume enormous amount of electrical power leading to high operational costs and emission of greenhouse gases. Therefore, efficient power modelling of data center hosts becomes an important research challenge that may help scheduling virtual resources in cloud data center in an energy efficient way. Most of the works in the literature consider a simplified power and energy model based on CPU utilization as shown below:

$$P(u) = P_{idle} + (P_{max} - P_{idle}).u$$

$$E_h = \int_t P(u(t))dt$$

where P_{idle} and P_{max} represent the power consumption of an idle and a fully utilized host respectively, u is the current CPU utilization and E_h is the energy consumption by the host h during the time t.
However, the running jobs in the data center may be CPU-intensive, memory-intensive and disk-intensive in nature. Clearly, this simplified power model can not properly estimate the energy consumption leading to wrong scheduling of virtual resources. In order to incorporate multiple parameters in the power model linear interpolation, multivariate

Linear Regression, Support Vector Regression (SVR) and Artificial Neural Network (ANN) based power model may be considered.

These power and energy models provide a number of energy efficient VM consolidation techniques for data centers. For example, greedy approaches like Modified Best Fit Decreasing [6], evolutionary approaches like Ant Colony System based VM Consolidation (ACS-VMC) [9], heuristic based approaches, Dynamic Voltage and Frequency Scaling (DVFC) based approaches [8] and many more. In this tutorial, we discuss various power models and the corresponding VM consolidation techniques used during the last decade.

4 GOALS OF THE TUTORIAL

The tutorial aims at giving the attendees and participants a holistic overview of some green computing techniques which can be used for big data scenarios and ML applications. Specifically the techniques like approximate computing and cloud power models will be discussed. The attendees can get an idea about the basics and applicability of these techniques for processing and performing learning and prediction in huge amount of data. Also, they can learn about implementing these kind of big data tasks in an energy efficient way in cloud platforms. Since, the primary theme of CODS-COMAD speaks about data science at large and management of huge amount of data along with applicability of the same in real-life scenarios, we feel that our tutorial theme matches tremendously with it.

5 TARGET AUDIENCE

The targeted audience can be anyone from computing or related background. A basic knowledge of data mining, big data and ML is desired. The knowledge of classical data mining techniques and the cloud related architectures and frameworks for big data and ML can be advantageous.

6 PROPOSERS

(1) **Hrishav Bakul Barua**,
Robotics & Autonomous Systems, TCS Research, Kolkata, E-mail - hrishav.barua@tcs.com, hbarua@acm.org
Mr. Hrishav Bakul Barua is working as a Robotics and AI Researcher and Consultant at TCS Research Labs (Robotics and Autonomous Systems, Cognitive Robotics Unit), Kolkata. He completed his bachelor's and master's study in Information Technology and Software Engineering from SMIT, Sikkim in 2012 and Jadavpur University, Kolkata in 2018 (first class 2nd position), respectively. He has served as reviewer/ TPC/ referee/ session chair for many International conferences and journals of repute such as IEEE-ICRA'21, IEEE RO-MAN '20-21, Big Data (mary ann liebert inc.), IEI-B Springer, SN Computer Science, IEEE INDICON etc. He is the recipient of Computer Society of India (CSI) - Young IT Professional award 2020 and ICDCN (ACM) 2020 best demo award in industry track and has filed 7 patents (in geographies like US, EUROPE and INDIA) out of which 1 recently received US grant. He has published about 15 papers in journals and conferences of high repute such as ACM, IEEE and Springer. He has delivered about 25 invited talks/Keynotes in many forums, seminars, workshops, and conferences on AI, ML, Big data and Robotics. He has received more than 15 organizational awards from Tata Consultancy Services (TCS) including Best Research Team award for Cognitive Robotics research. He is a member of ACM, IEEE SIGs, IEI, and CSI. He also has been elected to the State Executive Council of The Institution of Engineers, India (IEI) to serve a term of 2 years from October 2021 to 2023. He is also serving as an external guide/mentor for two M.Tech degree students of Data Science & Engineering course (in WILP mode) of BITS Pilani. He is also actively collaborating officially/unofficially with institutes such as Jadavpur University, BITS Pilani, Jamia Millia Islamia University, Myanmar Institute of Information Technology, ISI Kolkata, IIT Kharagpur, etc., for various research projects.

(2) **Kartick Chandra Mondal**,
Department of Information Technology, Jadavpur University, Kolkata, E-mail - kartickjgec@gmail.com
Dr. Kartick Chandra Mondal is an senior grade Assistant Professor in the Department of Information Technology at Jadavpur University, India. Before joining as Assistant Professor, he has completed his PhD from University of Nice Sophia Antipolis, France and served as Postdoctoral Researcher for one year in University of Strasbourg, France. He has more than 13 years of direct teaching and research experience at different universities and institutes. He has also served as adjunct faculty of different national and international universities and institutes. His research interests includes the area of Data Mining, Data Archiving, Bioinformatics, Bio Diversity, and Cloud Computing. He has published more than 45 research publications in different reputed books, international journals, and conferences. He is a member and life member of ACM, IEEE, IEI, and CSI.

(3) **Sunirmal Khatua**,
Department of Computer Science & Engineering, University of Calcutta, India, E-mail - skhatuacomp@caluniv.ac.in
Dr. Sunirmal Khatua is an Assistant Professor in the department of Computer Science and Engineering, University of Calcutta since 2010. He has completed his PhD in Computer Science and Engineering from Jadavpur University. He has received Gold Medal and University Medal from Jadavpur University in 2006. Prior to joining University of Calcutta, he has worked as a R&D Engineer in Tejas Networks India Ltd and as a Lecturer in Jadavpur University. Dr. Khatua has been awarded the Visvesvaraya Young Faculty Fellowship from Digital India Corporation, Govt. of India for the duration of 2018-2023. He has published 66 research papers in various reputed Journals and Conferences. His research interests include Cloud Computing, Wireless Sensor Network and High Performance Computing. He is a senior member of IEEE, ACM, and life member of IETE.

The first part of the tutorial, as mentioned in the outline, will be jointly presented by Hrishav Bakul Barua and Kartick Chandra Mondal. The second part will be presented by Sunirmal Khatua.

All the above mentioned proposers will be present for the tutorial either in offline or in remote online mode depending on the **current pandemic** situation.

REFERENCES

[1] Hrishav Bakul Barua. 2021. Data science and Machine learning in the Clouds: A Perspective for the Future. *arXiv preprint arXiv:2109.01661* (2021).

[2] Hrishav Bakul Barua and Kartick Chandra Mondal. 2018. Green data mining using approximate computing: An experimental analysis with rule mining. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 115–120. https://doi.org/10.1109/GUCON.2018.8675095

[3] Hrishav Bakul Barua and Kartick Chandra Mondal. 2019. Approximate computing: A survey of recent trends—bringing greenness to computing and communication. *Journal of The Institution of Engineers (India): Series B* 100, 6 (2019), 619–626.

[4] Hrishav Bakul Barua and Kartick Chandra Mondal. 2019. A comprehensive survey on cloud data mining (CDM) frameworks and algorithms. *ACM Computing Surveys (CSUR)* 52, 5 (2019), 1–62.

[5] Hrishav Bakul Barua and Kartick Chandra Mondal. 2021. Cloud Big Data Mining and Analytics: Bringing Greenness and Acceleration in the Cloud. *arXiv preprint arXiv:2104.05765* (2021).

[6] Abawajy Beloglazov and Rajkumar Buyya. 2012. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future generation computer systems* 28, 5 (2012), 755–768.

[7] Filipe Betzel, Karen Khatamifard, Harini Suresh, David J Lilja, John Sartori, and Ulya Karpuzcu. 2018. Approximate communication: Techniques for reducing communication bottlenecks in large-scale parallel systems. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–32.

[8] Liu L Ding Y, Qin X and Wang T. 2015. Energy efficient scheduling of virtual machines in cloud with deadline constraint. *Future Generation Computer Systems* 50 (2015), 62–74.

[9] Plosila Porres Farahnakian Ashraf, Pahikkala Liljeberg and Tenhunen. 2015. Using ant colony system to consolidate vms for green cloud computing. *IEEE Transactions on Services Computing* 8, 2 (2015), 187–198.

[10] Datacenter Forum. 2021. 5G will prompt energy consumption to grow by staggering 160% in 10 years. https://www.datacenter-forum.com/datacenter-forum/5g-will-prompt-energy-consumption-to-grow-by-staggering-160-in-10-years.

[11] Inigo Goiri, Ricardo Bianchini, Santosh Nagarakatte, and Thu D Nguyen. 2015. Approxhadoop: Bringing approximations to mapreduce frameworks. In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, Istanbul, Turkey, 383–397.

[12] Dhanya R Krishnan, Do Le Quoc, Pramod Bhatotia, Christof Fetzer, and Rodrigo Rodrigues. 2016. Incapprox: A data analytics system for incremental approximate computing. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1133–1144.

[13] Sparsh Mittal. 2016. A survey of techniques for approximate computing. *ACM Computing Surveys (CSUR)* 48, 4 (2016), 1–33.

[14] Adrian Sampson, André Baixo, Benjamin Ransford, Thierry Moreau, Joshua Yip, Luis Ceze, and Mark Oskin. 2015. Accept: A programmer-guided compiler framework for practical approximate computing. *University of Washington Technical Report UW-CSE-15-01* 1, 2 (2015), 1–14.

[15] Adrian Sampson, Werner Dietl, Emily Fortuna, Danushen Gnanapragasam, Luis Ceze, and Dan Grossman. 2011. EnerJ: Approximate data types for safe and general low-power computation. *ACM SIGPLAN Notices* 46, 6 (2011), 164–174.

[16] Anurina Tarafdar, Mukta Debnath, Sunirmal Khatua, and Rajib K Das. 2020. Energy and quality of service-aware virtual machine consolidation in a cloud data center. *Journal of Supercomputing* 76, 11 (2020), 9095–9126.

[17] Anurina Tarafdar, Mukta Debnath, Sunirmal Khatua, and Rajib K Das. 2021. Energy and Makespan Aware Scheduling of Deadline Sensitive Tasks in the Cloud Environment. *Journal of Grid Computing* 19, 2 (2021), 1–25.

[18] Anurina Tarafdar, Kamalesh Karmakar, Sunirmal Khatua, and Rajib K Das. 2021. Energy-Efficient Scheduling of Deadline-Sensitive and Budget-Constrained Workflows in the Cloud. In *International Conference on Distributed Computing and Internet Technology*, Vol. 12582 LNCS. Springer, Springer Science and Business Media, Deutschland GmbH, 65–80.

[19] VXCHNGE. 2021. Improving Data Center Power Consumption & Energy Efficiency. https://www.vxchnge.com/blog/growing-energy-demands-of-data-centers.