Towards Elastic and Sustainable Data Stream Processing on Edge Infrastructure

Marcos Dias de Assuncao
assuncao@acm.org
ETS Montreal
Montreal, Quebec, Canada

ABSTRACT

Much of the data produced today is processed as it is generated by data stream processing systems. Although the cloud is often the target infrastructure for deploying data stream processing applications, resources located at the edges of the Internet have increasingly been used to offload some of the processing performed in the cloud and hence reduce the end-to-end latency when handling data events. In this work, I highlight some of the challenges in executing data stream processing applications on edge computing infrastructure and discuss directions for future research on making such applications more elastic and sustainable.

CCS CONCEPTS

• **Networks → Cloud computing**; *In-network processing*.

KEYWORDS

edge computing, data stream processing, cloud computing, resource management

ACM Reference Format:
Marcos Dias de Assuncao. 2021. Towards Elastic and Sustainable Data Stream Processing on Edge Infrastructure. In *Companion of the 2021 ACM/SPEC International Conference on Performance Engineering (ICPE '21 Companion), April 19–23, 2021, Virtual Event,France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447545.3451902

1 INTRODUCTION

Much of the "big data" produced today is created as continuous data streams that are most valuable when processed quickly. Under several emerging application scenarios, such as in smart cities, operational monitoring of large infrastructure, voice assistants, Internet of Things (IoT), and wearable computing, continuous streams of data events are generated and must be processed under very short delays in order to provide feedback to end users. Modern solutions use the edges of the Internet (*i.e.*, edge or fog computing) for performing certain data processing tasks and hence: (i) reduce the end-to-end latency and communication costs, (ii) enable services to react to events locally, or (iii) offload processing from the cloud.

A Data Stream Processing (DSP) application is a directed graph whose vertices are operators that execute transformations over the incoming data (*e.g.*, filtering, projection, aggregation, convolution, or other user-defined functions), whereas edges define how the data flows between operators [6]. Cloud computing is often the target infrastructure for deploying such applications due to its scalability, virtually unlimited number of resources, pay-as-you-go business model, and resource elasticity. While resource elasticity enables an application or service to scale out/in (*i.e.*, allocate/release additional resources dynamically) according to fluctuating demands, attaining fully elastic stream processing services on edge computing environments is challenging for a number of reasons. The problem of placing DSP applications on heterogeneous infrastructure has been shown to be NP-hard [3]. Reconfiguring an application dynamically in order to use newly added resources requires modifying the application graph, which can result in exporting and storing current state, migrating operators, adjusting the number of partitions of a data flow, among other requirements.

In previous work we have focused on algorithms for the (re) configuration and elasticity of data stream processing components on edge computing while optimising metrics such as end-to-end latency, monetary cost and energy efficiency. We have worked on the problem of placing or scheduling data stream processing applications on largely distributed infrastructure comprising cloud and edge computing resources. We have devised a queueing-theory model for data stream processing applications and algorithms for splitting DSP graphs and placing their operators on cloud and edge resources [9, 11]. With the sustainability of applications in mind, we have also investigated the energy consumption of constrained devices while running DSP applications, and such results were used to calibrate a discrete-event simulation tool created to model and simulate DSP applications [1, 2]. In my talk, I highlight some of the challenges faced on making DSP applications in edge computing environments elastic and sustainable and discuss some of the approaches we aim to use to tackle these challenges.

2 ELASTIC AND SUSTAINABLE DSP

The following paragraphs summarise some of the key areas that, in our opinion, require effort in order to achieve elastic DSP applications. The proposed topics revolve around *investigating machine-learning techniques for addressing resource management issues in data stream processing applications on highly-distributed environments*, fault tolerance, and DSP mechanisms to *assist the execution of machine-learning systems on edge-computing environments*:

DSP Elasticity and Deep Reinforcement Learning: Reconfiguring an application during runtime is important as the environment conditions can change over time, and more resources may be required during peak load, whereas previously allocated resources can be released under low demand. Previous work has shown that

ICPE '21 Companion, April 19–23, 2021, Virtual Event, France

ACM ISBN 978-1-4503-8331-8/21/04.
https://doi.org/10.1145/3447545.3451902

reinforcement learning can be used for discovering new configurations for a running application [5]. Devising application reconfiguration plans using tabular reinforcement learning methods such as Q-learning is not scalable since the state and action spaces explode as the number of resources and the orders of the application graphs increase. This scenario can be exacerbated when considering multiple, and often conflicting, performance metrics such as resource utilisation, migration overhead, latency and monetary cost. Hence the main question to be addressed is how to enable application reconfiguration in a scalable manner. We believe that deep-reinforcement learning can shed some light on deriving policies that assist in this regard. This research effort hence aims firstly to model the application reconfiguration scenario as a (Semi)Markov Decision Process and employ deep neural network techniques (*e.g.*, DQN [8], dual DQN [10]) as function approximators. To reduce the action space one can consider hierarchical approaches where edge and cloud computing sites are given a high-level representation [4]. A second stage consists in exploring parallel methods, such as actor critic schemes [7] and deep neural networks to approximate policy and value functions for reconfiguring data stream processing services.

Fault-Tolerant DSP Applications on Edge Infrastructure: Edge computing can comprise resources that range from micro data centres with high-end servers and GPUs to more constrained devices such as Intel NUCs and Raspberry Pi's. The network technologies used to interconnect these resources is often less reliable than the solutions employed in cloud computing. Many resources might be powered by intermittent power sources. Despite its clear advantages in improving the response time of services and applications, edge infrastructure is much more failure prone than its cloud counterpart. However, the types of failures in edge computing and their frequency are still not well understood, specially when executing data stream processing applications. We need to investigate failure models by observing edge computing deployments and incorporate them in discrete-event simulation tools [2]. Then application check-pointing and migration mechanisms that respect the network constraints of edge computing environments must be investigated. The data stream processing scenario can then be modelled and such mechanisms put together to enable fault-tolerant applications. These efforts are connected to the first topic since failure poses the need for application elasticity and reconfiguration, and the actions consist partly in using the mechanisms investigated during this phase.

Energy-Efficient Data Stream Processing: The placement and elasticity of DSP applications on edge infrastructure are extremely important as many of the resources on which operators are deployed are often power- or battery-constrained. Several of these resources may be powered by intermittent and renewable energy sources such as solar panels and wind farms. While power consumption can be taken as just another optimisation criterion, better power and performance models on how these applications perform in such environments are needed in order to improve application placement and deployment. Such models can help us understand how scale in/out actions for DSP applications can impact the energy usage at the edge infrastructure and hence better plan under which conditions such actions are viable. By optimising the energy consumed on edge infrastructure we can make an overall DSP ecosystem more sustainable.

BIOGRAPHY

Marcos Dias de Assuncao is an Associate Professor at ETS Montreal, Canada. He was a former researcher at Inria, France (2014–2020), and a former research scientist at IBM Research – Brazil (2011–2014). He holds a Ph.D. in Computer Science and Software Engineering (2009) from The University of Melbourne, Australia. Marcos is interested in multiple aspects of distributed systems and data stream processing, including the use of deep reinforcement learning in the reconfiguration of data stream processing applications.

REFERENCES

[1] Gayashan Amarasinghe, Marcos Dias de Assuncao, Aaron Harwood, and Shanika Karunasekera. 2018. A Data Stream Processing Optimisation Framework for Edge Computing Applications. In *21st IEEE International Symposium on Real-Time Computing (ISORC 2018)* (Singapore). 91–98. https://doi.org/10.1109/ISORC.2018.00020

[2] Gayashan Amarasinghe, Marcos Dias de Assuncao, Aaron Harwood, and Shanika Karunasekera. 2019. ECSNeT++: A Simulator for Distributed Stream Processing on Edge and Cloud Environments. *Future Generation Computer Systems* (2019). https://doi.org/10.1016/j.future.2019.11.014

[3] Anne Benoit, Alexandru Dobrila, Jean-Marc Nicod, and Laurent Philippe. 2013. Scheduling Linear Chain Streaming Applications on Heterogeneous Systems with Failures. *Future Generation Computing Systems* 29, 5 (July 2013), 1140–1151.

[4] Pascal Vincent@umontreal Ca, Larocheh@cs Toronto Edu, Isabelle Lajoie, Yoshua Bengio@umontreal Ca, and Pierre-Antoine Manzagol@umontreal Ca. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research* 11 (2010), 3371–3408. http://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf

[5] Alexandre da Silva Veith, Felipe Rodrigo de Souza, Marcos Dias de Assuncao, Laurent Lefevre, and Julio C.S. dos Anjos. 2019. Multi-Objective Reinforcement Learning for Reconfiguring Data Stream Analytics on Edge Computing. In *48th International Conference on Parallel Processing (ICPP 2019)* (Kyoto, Japan) *(ICPP 2019)*. ACM, New York, USA, 106:1–106:10.

[6] Marcos Dias de Assuncao, Alexandre da Silva Veith, and Rajkumar Buyya. 2018. Distributed Data Stream Processing and Edge Computing: A Survey on Resource Elasticity and Future Directions. *Journal of Network and Computer Applications* 103 (February 2018), 1–17.

[7] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[8] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-Level Control Through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529.

[9] Eduard Gibert Renart, Alexandre Da Silva Veith, Daniel Balouek-Thomert, Marcos Dias de Assuncao, Laurent Lefevre, and Manish Parashar. 2019. Distributed Operator Placement for IoT Data Analytics Across Edge and Cloud Resources. In *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID 2019)* (Larnaca, Cyprus). 459–468.

[10] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-learning. In *13th AAAI Conference on Artificial Intelligence.*

[11] Alexandre Veith, Marcos Dias de Assuncao, and Laurent Lefevre. 2018. Latency-Aware Placement of Data Stream Analytics on Edge Computing. In *16th International Conference on Service Oriented Computing (ICSOC 2018)* (Hangzhou, China), Claus Pahl, Maja Vukovic, Jianwei Yin, and Qi Yu (Eds.). Springer International Publishing, 215–229.