Distributed virtualized infrastructure for data stream processing in the context of maritime traffic monitoring

Thesis advisors:

- Christine Morin (Senior scientist, Inria)
- Cédric Tedeschi (associate prof., University of Rennes 1, IRISA Lab)

Keywords:

- Maritime traffic monitoring
- AIS
- Data stream Processing
- Virtualized infrastructures
- Resource management
- Dynamic adaptation

Context and scientific challenges

Building and maintaining an updated report of the maritime traffic in a zone of interest or at the global scale is today an important strategic artifact in the surveillance of maritime territories. It is especially true for a country such as France, whose maritime area is the second one in terms of surface, in a context of growth of the traffic worldwide. These reports are crucial for the surveillance of borders, intrusion detections, illegal fishing activities, or for the detection of on-going environmental crimes.

Due to the scale of the areas to consider and the rapidly increasing traffic density, keeping this monitoring efficient requires to develop the ability to automate, at least partially, the detection of illegal activities, and, more generally of any unusual activity, to help operators to raise alerts and trigger their associated procedures. The recent development of different satellite systems, and of the AIS (Automatic Identification System) technology represents a significant step towards the gathering of information on the vessel traffic at the global scale. The AIS streams provide tens of millions of vessel positions each day and are growing in volume. By contrast, the computing resources at the disposal of the analysts of maritime traffic monitoring centers allow them to process only a very limited part of these streams.

The SESAME project, started in 2017, is a French research project funded by the Direction Générale de l'Armement and scientifically coordinated by the Institut Mines Telecom Atlantique. The IRISA Lab is one the academic partners involved in the project. The main objective of SESAME is to develop new tools for the quasi real time monitoring and analysis of traffic, for an increased efficiency of the monitoring process, through the combination of several data sources (from both satellites and the AIS streams), and by porting these tools to virtualized computing platforms.

The stream processing engines that were recently proposed (such as Storm or Flink) [1], target the analysis, consolidation, and visualization in quasi real-time of data being constantly produced. They facilitate the programming of such a processing by providing a programming model easing the description of the workflow to be applied on each incoming new data element (commonly under the shape of a graph of operators). They also take care of the deployment of such a workflow over a pool of computing resources such as a cluster of processors. With the advent of virtualized infrastructures, it becomes easier to dynamically scale in and out the platform underlying such computations, as the demand in terms of computing power and storage of the computation evolves, due in part to the changing velocity of data. Several tools have been recently developed to orchestrate virtual environments or containers over distributed environments [2,3]. The issue of dynamically scaling a shared infrastructure for stream-processing applications has been recently addressed in more or less distributed settings [4,5]. A specific solution oriented towards the maritime traffic management is an open issue.

More particularly in the context of the SESAME project, being able to raise different levels of alert requires not only to dynamically adapt the resources allocated for the monitoring process, but also to dynamically spawn new computations, needed to process one or another particular alert level. *When* such an adaptation needs to be made is hard to predict. Moreover, such an adaptation will require some reconfiguration at the platform level, to redesign the workflow dynamically and to adapt the placement of operators.

Methods and tools envisioned

To tackle the scientific issues described above, the different steps envisioned for the thesis are the following:

- Design generic algorithms for the adaptation and dynamic reconfiguration of the stream processing workflow. This work will suppose to keep in mind that the goal is to allow this adaptation while minimizing the reconfiguration time so as to prevent degradation in the responsiveness of the platform and delays in the processing.
- Devise placement strategies of the different operators composing the workflow which optimize the usage of resources, constraints being the limitation in the latency tolerated by the application, and the location of data (that may be either legally controlled or simply restrained by the size of data)
- Conduct an experimental validation for the two previous aspects. The nation-wide Grid'5000 test bed will be used for this part.

References

[1] Guenter Hesse, Martin Lorenz. Conceptual Survey on Data Stream Processing Systems. 21st IEEE International Conference on Parallel and Distributed Systems (ICPADS 2015) : 797-802
[2] Kubernetes. https://kubernetes.io/
[3] Docker Swarm. https://www.docker.com/products/docker-swarm/
[4] Jack Li, Calton Pu, Yuan Chen, Daniel Gmach, Dejan S. Milojicic: Enabling Elastic Stream Processing in Shared Clusters. IEEE CLOUD 2016: 108-115
[5] Christoph Hochreiner, Michael Vögler, Stefan Schulte, Schahram Dustdar: Elastic Stream Processing for the Internet of Things. IEEE CLOUD 2016: 100-107
[6] Grid'5000. https://www.grid5000.fr/

Skills and profile

- Knowledge and/or experience in algorithmics and distributed systems
- Experimental skills
- Fluent French or fluent English mandatory

Information

- SESAME project: http://www.agence-nationale-recherche.fr/?Project=ANR-16-ASTR-0026
- IRISA Lab: https://www.irisa.fr/en
- Myriads project: https://team.inria.fr/myriads/
- Candidates are invited to contact Cédric Tedeschi - cedric.tedeschi@inria.fr