Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 97 (2016) 131 – 134

Procedia Computer Science

CLOUD FORWARD: From Distributed to Complete Computing, CF2016, 18-20 October 2016, Madrid, Spain

A holistic approach for high-level programming of next-generation data-intensive applications targeting distributed heterogeneous computing environment

Emanuele Carlini[a], Patrizio Dazzi[a,]*, Matteo Mordacchini[b]

[a]*CNR-ISTI, Area della Ricerca di Pisa, 56124 Pisa, Italy*
[b]*CNR-IIT, Area della Ricerca di Pisa, 56124 Pisa, Italy*

Abstract

The intrinsic richness and heterogeneity of large amount of data is paired with the extreme complexity in its storing and processing, as well as with the heterogeneity of their processing environments, ranging from super computers to federations of Cloud data-centres. This makes the conception, definition and implementation of software tools for programming applications dealing with very large amount of data really challenging from different perspectives, ranging from technological issues to economic concerns. We propose an approach focused on data-intensive applications that goes beyond the state of the art allowing a seamless exploitation of heterogeneous and distributed resources and satisfying users' needs on data processing providing a dynamically determined set of features, depending on the running environment, the application, the user requirements.

Peer-review under responsibility of organizing committee of the international conference on cloud forward: From Distributed to Complete Computing

Keywords: Cloud Computing; Cloud Federation; Resource Management; Data-intensive Applications; High-level Programming Models

* Corresponding author. Tel.: +39-050-315-3074; fax: +39-050-315-2040.
E-mail address: patrizio.dazzi@isti.cnr.it

Peer-review under responsibility of organizing committee of the international conference on cloud forward: From Distributed to Complete Computing
doi:10.1016/j.procs.2016.08.290

1. Background

Data-intensive applications are one of the largest customers of high throughput computing environments. As clearly shown by the BigData raise, there is a widely recognised urge for processing large amount of data, universally considered the "new oil", from which to extract wisdom, knowledge and, more in general, information of very different kinds, ranging from commercial to financial, from medical to political hints. This extreme intrinsic richness and heterogeneity of data is paired with the extreme complexity in its storing and processing. In this light, today's storage and processing platforms are highly heterogeneous and specially characterised in terms of deployment environments, which include super computers, clusters of commodity or specialized hardware, private and public Cloud data-centres, up to federations of Cloud data-centre. Even more, the heterogeneity is not limited to the deployment at-large but also impacts on the type of resources exploited within a single machine: multicore processors, GPU, FPGA, etc. Cloud technology demonstrated to be an effective solution as it is able to elastically adapt the computing capacity with respect to the (dynamic) needs of applications and customers. In any case, this wide offers in platforms and technologies needs to be properly mastered and leveraged to allow an effective and efficient exploitation of data and resources. As a consequence, the conception, definition and implementation of software tools for programming applications dealing with very large amount of data still represent a hot research topic. This is due to several factors, including technological issues, economic concerns and legal constraints. One of the main aspects is that data is more and more decentralised and localised[1,2], thereby moving it is not always efficient in terms of cost and performance, and cannot even be possible, for example for legal aspects. Further, modern processing platforms are characterised by high levels of dynamicity, in which there is no a static set of resources available for the computation, but rather their kind and numbers depends on the context of a particular application and relative data. As matter of fact orchestrating a computation in such environments is complex, time-consuming, error-prone. Even more, the dynamic nature of users[3], applications and computational resources makes this task extremely complex and almost unfeasible to be conducted without the support of automatic tools. In the last years many approaches, models and solutions have been proposed to tackle these issues.

The problem has been faced from very different perspectives. Some solutions focused to the creation of infrastructures enabling a seamless exploitation of very different kind of resources, both in terms of their hardware heterogeneity[4], and in terms of deployment (infrastructures enabling a transparent exploitation of data-centers, cloudlets as well as resources located at the extreme edge of the network)[5,6]. Other approaches focused on smart brokerage solutions aimed at easing the task of finding the most suitable resources to run an application, depending on several factors including: user and application requirements, location, volume of requests, etc. In addition, have been also proposed approaches aiming at simplify the exploitation of heterogeneous and dynamic resources by leveraging proper programming model abstractions[7]. More recently, both the research and industrial cloud communities are trying to define holistic approaches[8] aimed at providing vertical solutions that, on the one hand ease the programming of applications targeting heterogeneous environments and on the other hand simplify the management of large computational infrastructures. In these approaches users specify their requested service level (e.g. a given throughput), and the supporting environment ensure their satisfaction by adopting the proper resources.

In this position paper we propose our vision for a programming ecosystem for organizing the computation of data-intensive applications in heterogeneous platform that goes beyond the state of the art and try to solve the aforementioned issues. In particular, we are envisioning an ecosystem whose ability is not limited to a seamless exploitation of a set of heterogeneous and distributed resources but it is able to address users' needs about data processing by adopting solutions providing a dynamically differentiated set of features, depending on the actual running environment, the hosted application, the user requirements.

2. Our envisioned Ecosystem

What we envision is an ecosystem in which the functional logic realising the applications self-adapts with respect to the available resources, properly defined user requirements and the actual context. Basically, our idea is to go beyond the traditional application development approach. In fact, a general data-intensive application is realised according to a well-defined process: problem definition, algorithm selection, software implementation, data preparation and application deployment. According to this schema, the application management support, provided