Data Intensive Scalable Computing (DISC)

Mario A.R. Dantas[1,2]

[1]Departamento de Ciência da Computação (DCC) – Centro de Ciências Exatas (ICE)
Universidade Federal de Juiz de Fora (UFJF)
Rua José Lourenço Kelmer - Martelos, Juiz de Fora - MG, 36036-330 – Brasil

[2] INESC P&D - Brasil

mario.dantas@ice.ufjf.br

Abstract. *This work presents an introduction to the Data Intensive Scalable Computing (DISC) approach. This paradigm represents a valuable effort to tackle the large amount of data produced by several ordinary applications. Therefore, subjects such as characterization of big data and storage approaches, in addition to brief comparison between HPC and DISC are differentiated highlight.*

Resumo. *Este trabalho apresenta uma introdução à abordagem de Computação Escalável Intensiva de Dados (DISC). Este paradigma representa um esforço valioso para lidar com a grande quantidade de dados produzidas por vários aplicativos comuns. Por isso, assuntos como caracterização de big data e abordagens de armazenamento, além de uma breve comparação entre High Performance Computing (HPC) e DISC são focados diferencialmente.*

1. Introduction

Nowadays the ordinary scenario of several applications is to have a large amount of data, which is created by mobile devices (e.g. mobile phones, smart bands and several type of sensors). Usually, this data is neglected in some dimensions due to the lack of understanding procedure on how to treat this data, such as how to properly storage this data. Examples are found in fields of engineering, health, and smart cities.

On the hand, the large experience from those involved in grand challenges applications related to the *High-Performance Computing (HPC)* could bring their mature knowledge to this new type of ordinary data, which is intensive and reaches sometimes level of those HPC applications.

In figure 1, it is illustrated a typical new challenge scenario concerning to the data demands. What are the ***areas*** that can ***benefit*** from this type of ***research***? The answer is *all those data-intensive context-oriented applications*, which could be an HPC or a DISC application.

Therefore, in this work we tackle some aspects related to big data, storage, and HPC/DISC characteristics. Aspects are observed in the terms of on a day-by-day basis trying to demonstrate the necessary change on how to see the data and threat it. The approach could be similar on how we treat a sustainable water lake, where the water comes sometimes in a huge volume and must be appropriated store to be processed.

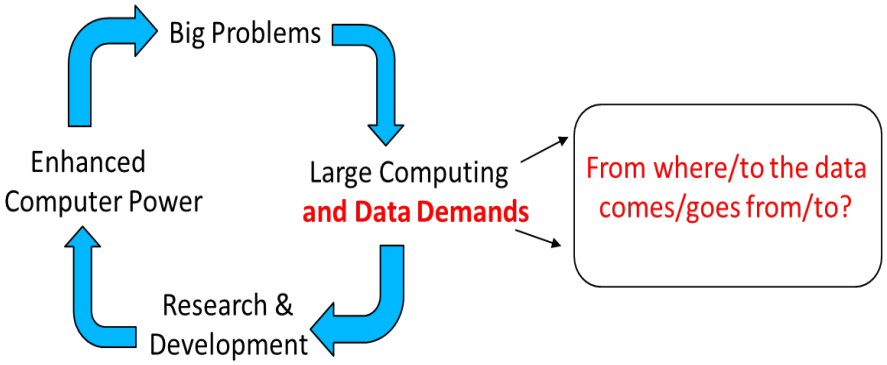

Figure 1. The challenge scenario of data demands

2. Big Data

The increasing synergy of Internet of Things (IoT) and big data are creating a demand of new computational requirements and converge to the knowledge exiting in the HPC communities. It is not possible to ignore the infrastructure where the data which forms the big data environment comes from. Necessary approaches, such as how to collect, cleaning, and storage the data are vital. This scenario is based upon the amount of this mobile data and mandatory requirements for the *data intensive scalable computing*.

The big data approach considers parameters such as volume, velocity, variety, veracity, and value, those known as 5Vs. The challenges of the big data approach can be measured by basic questions(a) Can my organization store / manage large amounts of data?; (b) Can we guarantee, for example, velocity under a specific period (time constraint)?; (c) How to treat such variety technologically; (d) As for veracity, can I check?,(e) Can I quantify the value; can I get it?

3. Storage

The storage process, usually, is considered as shown in figure 2. Where in three of those pictures the data is store as a garbage without considering any type of selective approach. This is not a usual in several organizations, mainly because the aspects presented in figure 3 are not well known.

Figure 2. Data storage approaches

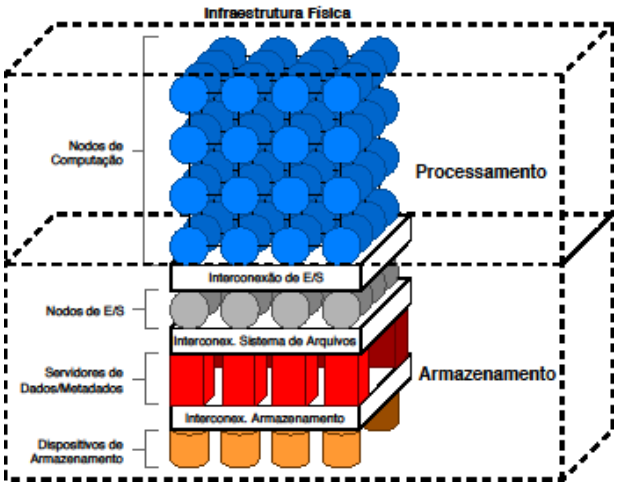

Figure 3. Data processing and storage

Data processing and storage must be understood and figure 3 shows. The upper layer is computation nodes which are connect through a I/O interconnection network to the layer of storage. This storage layer is divided in I/O nodes, metadata servers and finally the data storage device. In the past, the data storage device as seen as the unique element in the storage process.

4. HPC and DISC Characteristics

After presenting a non-conventional view of big data and storage, in this section we present the characteristics of HPC and DISC, targeting to better situated those approaches.

In references [Bryant, 2007, Inacio and Dantas, 2018, Patrick et al., 2018] it is possible to find some more comprehensive material about this topic.

Figure 4 presents dimensions of the focus, applications, target, and objective to provide a wider view of the HPC and DISC paradigms. As a parameter it was chosen processing aspects (i.e. computing and data), type of applications, main goal of the paradigm and finally objective scope.

Dimensions	HPC	DISC
Focus	Computing-oriented	Data-oriented
Applications	Science and Engineering	Web and Business
Target	Simulation	Management and data Analysis
Objective	High Performance	Scalability, F.T., Availability and Cost-Performance

Figure 4. HPC and DISC Characteristics

The comparison presented in figure 4 indicate clearly that both paradigms target to have a differentiated performance, one in the computing dimmension and the other in data processing approach. Other interesting observations is related to the applications, where HPC focus in science and engineering where as the DISC in ordinary conventional web and business.

5. Conclusions

In this work, it was briefly present the *Data Intensive Scalable Computing (DISC)* approach, which is an emergent paradigm conceived to tackle the data intensive challenge from several ordinay applications.

The method to present this paradigm was to draw a comparison with the High-Performance Computing (HPC) a well known approach considered in science and engineering applications.

Our efforts were focusing mainly in topics such as big data and storage. The big data paradigm usually considered as a final stage process where the data is there already to use. On the other hand, storage is commomly see only as a storage device. In both aspects we show new concerns about how to treat these two elements.

Finally, it is important to comment that this new era of data intensive, where large amount of data is generated by mobile devices, it is very important to consider the discussion presented in this work.

References

Bryant, "*Data-Intensive Supercomputing: Intensive Supercomputing: The case for DISC The case for DISC*" Tech Report: CMU-CS-07-128, 2007.

Inacio, Eduardo C., Dantas, Mario A. R., "*Introdução a Sistemas de E/S e Armazenamento Paralelo"*, Mini-Curso WSCAD, 2018

Patrick Valduriez, Marta Mattoso, Reza Akbarinia, Heraldo Borges, Jose J. Camata, Alvaro L. G. A. Coutinho, Daniel Gaspar, Noel Moreno Lemus, Ji Liu, Hermano Lustosa, Florent Masseglia, Fabrício Nogueira da Silva, Vítor Silva Sousa, Renan Souza, Kary A. C. S. Ocaña, Eduardo S. Ogasawara, Daniel de Oliveira, Esther Pacitti, Fábio Porto, Dennis E. Shasha: "*Scientific Data Analysis Using Data-Intensive Scalable Computing: The SciDISC Project*.", Ladas, 2018.