DOI: 10.2507/29th.daaam.proceedings.144

DATA-INTENSIVE COMPUTING PARADIGMS FOR BIG DATA

Petra Loncar

This Publication has to be referred as: Loncar, P[etra] (2018). Data-Intensive Computing Paradigms for Big Data, Proceedings of the 29th DAAAM International Symposium, pp.1010-1018, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-20-4, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/29th.daaam.proceedings.144

Abstract

The development of ICT has led to enormous growth and data accumulation, and thus created the need for proper storage and processing of large data, known as Big Data. The number of data sources like mobile telephones and applications, social networks, digital television, data of various Internet objects and sensors has increased due to the development of technology and the emergence of the IoT, the evolutionary step in the development of the Internet. Analysis and proper interpretations that take place on the latest distributed platforms are key to data-intensive systems from which feedback can be gained in areas such as industry, finance, healthcare, science and education. Distributed computing paradigms are a fundamental component of research and innovation for e-infrastructures with the intent of providing advanced computing, storage resources and network connectivity necessary for modern and multidisciplinary science and society. The aim of this paper is to provide a systematic overview and address challenges of distributed computing paradigms that marked and brought revolution into computing science and are able to store and process large amounts of data. As an example of a big data source, the world's largest particle detector, CERN's LHC is analyzed.

Keywords: big data; data-intensive; distributed computing; state of art

1. Introduction

In the era of big data, where data is core of everything execution and simulation of applications can be resource demanding. Data is generated by increasing frequency and its quantity is continuously growing. The Internet of Things (IoT), Big Data and other upcoming emerged fields of Industry 4.0 will present a challenge to next decades. The demand for large scale distributed paradigms that can support such data-intensive applications is growing at higher rates than improvements in technology alone can sustain. Moore's law [1] has been used since 1965 to describe the continuous advancement of technological solutions by exponential growth in speed and microprocessor power and exponential price drop in such systems. The applicability of Moore's law on large data sets and the future development of technology is questioned, as data volume more than doubles every 18 months. Large amounts of data and rapid technology development must equally monitor and develop aspects of adequate data storage and processing. New challenges will require new computing models and approach to data.

The distributed computing paradigms that have marked and brought the revolution in computing are: Cluster Computing, Grid Computing, Cloud Computing, Fog Computing and Dew Computing (figure 1.). These modern computing paradigms show potential for handling growing volume of data, providing computing and storage resources for such data-intensive applications.

Their full potential should be analyzed and their interaction need to be coordinated. An example of a large set of data is the CERN which has four large LHC (Large Hadron Collider) experiments that generate large amounts of data on a daily basis. Demanding storage and processing of such data are based on the WLCG (Worldwide LHC Computing Grid) computing infrastructure.

The aim of this paper is to provide a systematic overview of computer paradigms and to analyze their capabilities to handle large amounts of data. The second section provides an overview of the distributed computing paradigms. Definition of big data and CERN's LHC as an example of big data source are presented in the third section. The third section also provides an overview of current challenges for large amounts of data. Then, section four serves as the conclusion.

2. Overview of data-intensive computing paradigms

Autonomous, interconnected computers that make the distributed system and perform multiple tasks simultaneously, communicate, share data, storage resources, networks, files, and services are presented to the user as a coherent system. Advanced, heterogeneous and complex distributed computing systems, often geographically spread, have to meet certain criteria. The system should be interoperable, portable, integrable to other systems, expandable and flexible. Distributed systems should be scalable, meaning they have the ability to operate effectively regardless of the size of the system and transparent in view of access, location, migration, relocation, concurrency, replication and failure. System's performance and QoS are defined by performance, resource allocation, reliability, security, timing, availability, bandwidth, capacity, privacy, ownership costs, and impact on user experience.

2.1. Cluster Computing

Cluster computing system is made up of independent computers locally networked which work cooperatively and coordinated creating unique computing resource. A high degree of integrity is achieved by using a specific software. Using distributed memory and heterogeneous architecture differs cluster from unique multiprocessor system. Depending on the needs, clusters can be realized to support high performance computing (HPC), high throughput computing (HTC) or high availability computing (HAC) tasks. During building of cluster, it is necessary to secure load balancing of nodes that will share the system load to achieve better performance of cluster. Afterwards, directing processing from not-working node to redundant nodes guarantees that processing on cluster will not be stopped in case if an error appears in working node.

Various scientific areas such as mathematics, biology, medicine, computing [2], [3] have a need for HPC or HTC infrastructure. High performance of computing tasks is achieved by parallel processing and using distributed memory for communication. The aim of HPC is fast processing of great amount of data. HPC performance can be compared to performance of supercomputer. Those performances are usually achieved by networking a great number of servers with fast network technology like InfiniBand and message passing interface (MPI). HPC infrastructure features are multi-core processors, bigger amount of memory for processor core, big memory bandwidth, using fast memory, securing parallel access to files, fast and reliable data storage. Cluster computing systems are used for parallel programming in which demanding program is executed on more computers with their own instance of operating system.

The architecture of cluster computing consists of computers, computer network that needs to transmit great amount of data with small latency, operating system (most acceptable is Linux) and clusterware. Clusterware has a task to manage tasks, monitor the system operations, enable data exchange and sharing, and provide a development environment for processes. Homogeneity is a characteristic of cluster computing, which means nodes are identical, with same operating system connected with same local network. Beowulf cluster [4] is the most common architecture of cluster computing which is made of nodes connected in a private network. One of nodes, frontend, presents the center of cluster with located file systems, system for centralized control of working nodes in cluster, system for automatic installation of working nodes and system for managing tasks. One of advantages of cluster computing, in regards to other paradigms of distributed computing, is property and managing cluster by one authority. It provides full access to hardware and remove other needs for association. For accomplishing optimal performances, users can modify cluster and applications. Application for performing uses entire hardware which provides achieving better performance. Disadvantages of cluster computing are need for significant investments and maintenance, as well as possible periods of insufficient utilization of cluster's resources.

2.2. Grid Computing

In 1998, Foster and Kesselman [5] defined grid computing as: "a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities". Grid provides exchanging and sharing of computing power and data storage space. It was created due to increased needs of scientists for computing resources, with the aim of securing almost unlimited space for storing data and power for processing data of e-Science. Grid is realized through computer networking and cluster computing over the Internet. The idea of grid is to provide a simple way of connecting user to such big network of distributed computers like connecting to an electrical network. Hence the name of this concept of distributed computing.

What differs it from cluster is weaker connection of computing nodes (workstations, clusters, servers) which are under different administrative domains, located on different geographical locations, and are heterogeneous (considering different hardware platforms, architectures). Using resources in virtual organizations should be safe, coordinated, flexible, controlled and with appropriate quality of service, QoS. QoS refers to the performance of using resources, their availability and safety of direct use of distributed resources through mechanisms of authentication and authorization. The purpose of grid computing is to offer high performance and high scalability. Computing and data grid are used for storage and processing of great amount of data, scientific simulations etc.

Basic parts of layered grid architecture are distributed resources (memory, processor, storage space, network resources, sensors, etc.), communication protocols, information and control protocols for user interaction with remote resources and services, common services (services for finding resources according to properties, services for distributing requests, environment for development of logic for resources collaboration, service for calculation and billing of resource using) and user applications. Grid intermediate layer is virtualization layer which offers basic functionalities such as managing resources, data managing, safety check, monitoring and detecting failure and attacks.

Due to their primary purpose, there are computational grids that process CPU demanding tasks such as modelling and simulation of complex scientific experiments, subsequently, data grids for saving and processing great amount of data. The advantages of grid computing are ability of solving complex and demanding problems in short time, easier use and coordinate share of resources in dynamic organizations with no central control and less total costs.

2.3. Cloud Computing

Cloud computing is one of the most important paradigms in IT technology. Cloud computing presents a pool of distributed resources and services on computers and other devices that become available at the end user request through the Internet infrastructure. Resources and services do not have to be owned by their end-user. In 2010, NIST [6] defined cloud computing: “Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” The definition describes a concise set of cloud features: on-demand self-service, broad access, resource pooling, rapid elasticity and measured service.

The concept of cloud computing includes achievements of distributed computing and storage, virtualization, Internet technologies and systems management. Cloud computing represents a turning point in way of using technology and optimization of resource use in accordance with changing needs and method of paying only used services. Cloud computing represents a revolution in data storage and scalability it provides. The achievements of cloud computing are significant not only for computer science and industry, but for series of academic disciplines. Cloud computing appears in literature of various fields such as economy [7], bioinformatics [8], medicine [9], [10]. An increasing number of organizations in telecommunications and ICT sector switches to one of the forms of business in cloud to reduce operating costs and improve results.

Cloud computing provides rapid access to flexible IT resources over the Internet when needed by user, e.g. storage containers, servers, databases and wide set of service applications service using pay-per-use principle. User does not need to make big investments in hardware and its managing. The advantages of this paradigms are also elasticity and scalability, wide network access over different heterogeneous devices (mobile devices, laptops), possibility of isolating overloaded or faulty server and migrating tasks on other location in cloud, location independence which can be defined on higher level of abstraction, rapid and flexible provision of resources. However, security issues, data sensitivity in cloud and access to data by unauthorized users pose certain risk. For user of cloud services, it is necessary that service provider guarantees cloud reliability.

Virtualization technology is key to development of this paradigm. It generalizes physical infrastructure and makes it available and easy to use and manage. The advantages of using virtualization are: on request access to servers and storage resources, energy saving, reduction of required space, rational use of human resources, reduction of capital and operating costs, consolidation, migration, load distribution and isolation. Consolidation of more individual and heterogeneous loads on one physical platform provides better system usability. Workload migration provides hardware support, balances load and enables failure recovery. It is implemented by encapsulation of operating system state in virtual machine which enables migration to other platforms, resume of previous work and saving for later start up. State of virtual machine includes complete image of disk or partition, configuration files and RAM image. In virtualized environments, hypervisor performs abstraction of hardware on virtual machines, supervises and manages the system in complex environment with multiple versions of different IT resources and allows presence of certain applications on multiple systems without physical mapping on each system. The hypervisor can be implemented directly above system hardware or above operating system. It supports multiple virtual machines' performances, adjusts its activities and provides consistent access to the processor, memory and I/O resources of physical machine. More common and technically different performed is container based virtualization, where physical machine is divided to multiple virtual machines with their own subset of physical resources. In container virtualization, there is no hypervisor, no virtualized hardware and it is implemented on operating system level. Performances of hypervisor and container virtualization are compared in [11].

From the standpoint of implementation and use of cloud, there are three different system configurations:

- Platform as a Service, PaaS - user can develop his own application using cloud computing environment, such as operating system, database management system, web server and other software infrastructure required to run program code written in one of the supported development tools. End user can develop its own IT solution using the provided environment. Amazon Web Services, Google App Engine are examples of PaaS.
- Software as a Service, SaaS – the service provider prepares the service in the cloud so that it can be used immediately by the end user. A SaaS service is most commonly accessed through a web browser installed on a computer from anywhere in the world, without the need for special installations on a user's computer. One of the SaaS advantage is easier upgrade and update of services for all users. Google is one of the SaaS providers with products as Google Maps, Google Docs and Gmail.
- Infrastructure as a Service, IaaS – cloud service which provides use of IT infrastructure, that is, computing, network and storage resources and other basic computing resources where cloud user can implement and run desired applications. IaaS offers the highest level of control compared to PaaS and SaaS. Amazon is service provider which offers IaaS and provides it to users to load their own solutions and applications on Amazon infrastructure (Amazon Elastic Compute Cloud, Amazon Simple Storage Service, etc.).

Along with basic services defined with cloud computing, new services are being developed which have suffix "as a

Service (aaS)" for marketing purposes of collective name Everything as a Service, XaaS. Some of these services are: Storage as a Service, Database as a Service, Analytics as a Service, Network as a Service, Backup as a Service, Management as a Service, Firewall as a Service, Voice as a Service, Mobility as a Service, etc. Clouds can be deployed as private (cloud which is used by one organization and managed by an organization or third party), public (cloud provides open use of cloud service provider infrastructure to public), community (cloud provides cloud services to the group of organizations of common interests) and hybrid (cloud is generated by combining and connecting two or more previously mentioned cloud models).

By reviewing basic features of cloud computing, it is necessary to highlight data centralization, flexibility, support for user mobility, scalability, lower costs and democratization of the use of resources as its advantages. The biggest advantage of cloud computing is reducing the cost of IT maintenance and the availability of complex solutions for smaller enterprises. Enterprises can focus more on business rather than spend their resources on handling IT system. Many companies and users were discouraged by insufficient data protection, lack of control over how data is transmitted over a network, their accommodation in the other data cloud. Cloud stimulates innovation, digitization and provides great potential for further development.

2.4. Fog Computing

Fog computing or fogging is a new paradigm of distributed computing that combines the principles of cloud computing and the growing IoT. Cisco has [12] introduced fog as cloud computing extension to the edge of the network, closer to end users. The concept of fog computing has emerged as response to the sudden exponential increase and complexity of IoT data, devices, and networking. The estimation is that by 2020, 50 billion "things" will be connected to the Internet [24]. Fog computing implies virtualization and resource sharing in highly scalable data centers and micro data centers near small devices (smart devices, routers, and other network devices). The basic principles of cloud computing have been moved from the center of network to the edge of the network to provide data, processing, storage, and services to end users. Collaboration with cloud computing that provides on-demand storage and scalability is still needed. It is important to determine which jobs should be processed on nearby computing resources and which should be forwarded to cloud, as well as to optimize resource allocation for such jobs. Delays, issues with system performance, node centralization, data security, power consumption, necessary Internet connectivity and ensuring sufficient bandwidth are some of the cloud computing challenges that fogging can overcome. Processing and analysis closer to the data source decreases latency, reduces energy consumption, unloads cloud and ensures scalability. Sensors and other IoT devices and smart city applications, autonomous vehicles, robots, healthcare applications affect the quality of everyday life and generate large amounts of sensitive data on a daily basis (in milliseconds). It is important to act reliably and provide fast response. Such devices that use machine-to-machine communication, M2M, or human and machine interactions, with improved processing security, can take advantage of the emergence of this concept and work closer to the data source. The advantages of fog computing over other paradigms are described through the term SCALE which stands for Security, Cognition, Agility, Latency and Efficiency. Transactions are performed with improved security and are more client oriented to provide autonomy and better performance. Sensitive sensor data must be interpreted rapidly to react efficiently and should be processed in real-time. Selected data is sent to the cloud for long-term archiving and detailed analysis. Since the fog node can be connected to other fog nodes or clouds, communication levels that needs to be considered are cloud to fog, fog to fog and fog to edge devices regarding share of storage, computer, and network resources and data management. Fog computing facilitates the management and programming of network, computer and storage services between data centers and end devices, business agility and distribution of data analysis to meet the requirements of widely distributed applications that require low latency. Low latency is a result of using fog nodes to reduce the distance between IoT devices and clouds. Location aware fog nodes (controllers, routers, servers, switches, surveillance cameras) run applications and process sensitive data within the core network which also reduces latency and task execution time.

Widespread geographical distribution of heterogeneous IoT devices, device mobility and real-time interaction for time-sensitive applications should be supported. Privacy and security of IoT data should be guaranteed, both on fog nodes and in transfer to the cloud. In addition, management of fog computing resources is also challenging. The fog computing infrastructure has to provide support of internal components, ensuring compliance with standards, protect them from the environmental factors they will be exposed (temperature, humidity, vibration), manage cooling system, protect people and things, ability to expand and platform servicing. Virtualization is used to implement the fog platform, allowing entities to share the same physical system, and contributes to system security.

Fog computing will also play an important role in the development of 5G mobile networks and services [13] and better website performance [14]. The Open Fog Consortium [15] will play important role in further development, research and standardization of fog computing.

2.5. Dew Computing

Further development of computing has led to another hardware-software paradigm for computing, which aims to make full use of on-premises computers, called dew computing [16]. Two features of dew computing are independence and collaboration. The foundation of dew computing is the hierarchical Cloud-dew architecture described in [17], whose key property is collaboration of micro and macro services. The dew device (PC, tablet, smartphone, server, etc.) with appropriate software support provides functionality independently of cloud services, and works with them. While fog computing includes sensors, routers and automated IoT devices, dew computing seeks to achieve full potential of resources that are not used in the cloud computing. Although dew computing is not intended for IoT, it can also offer solutions for use in IoT and serves as intercessor between fog and cloud. Dew computing cooperates with cloud services and controls fog devices. The idea of dew computing is to maximize the use of resources before processing is transferred to the cloud server, reduce the complexity and improve the productivity of scalable distributed computing. Dew services use smaller devices that can make compression of image or sound and process small amount of data. Heterogeneous dew devices are programmable and reconfigurable to effectively perform complex tasks using various tools and applications.

Hierarchical architecture of the dew, fog and cloud computing [18] paradigm realizes vertical scalability and expands the client-server model. In addition to the cloud server, a dew server is also introduced. The Dew server is located in a local computer and has the task to serve only one client with services that cloud server offers. The dew server database needs to be synchronized with the cloud server database. Dew server has capacity smaller than cloud server and stores only user data. The main advantage is that it can be simply reinstalled with the help of cloud data backup and can be accessed independently of the Internet connection because it runs on local computer. One of the disadvantages of computing in fog and cloud is the need for permanent Internet connection and the availability of a cloud server. Under such conditions, the user cannot access the data and can hardly synchronize the computer tasks. Dew applications are not entirely online and they must use cloud computing services and automatically exchange information with them while performing. An example of dew applications are Dropbox, OneDrive, and Google Drive Offline. User can use their services regardless to the Internet connection and can be synchronized with cloud services. These services belong to the Storage in Dew (STiD) category. The potential use of dew computing is also in Web in Dew (WiD), which allows search without Internet connection but with the support of a special Web server on on-premises computers that would locally support duplicate web addresses to be synchronized with web pages in the cloud once the Internet connection is established [16]. WiD could also be used to control the IoT device and exchange information with the cloud servers. Other services in computing are: Database in Dew (DBiD), Software in Dew (SiD), Platform in Dew (PiD), Infrastructure as Dew (IaD), Data in Dew (DiD) [16]. To support development of this new paradigm some challenges need to be solved, including specialized hardware, software, network, operating systems, databases, servers and search engines.

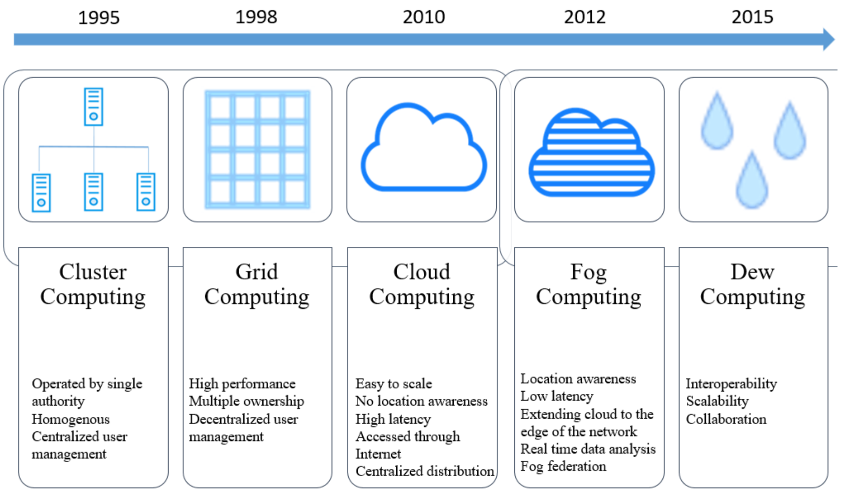

Fig. 1. Distributed computing paradigms timeline

3. Big Data

Big Data is currently among the most important research technologies in science, biology, engineering, medicine, entrepreneurship, astronomy, social sciences and other areas where large amounts of data are generated as a result of experiments and simulations.

3.1. Big Data – definition and challenges

The concept of data science has a long history and has been particularly popular lately because of the growing use of cloud computing and IoT in creating a smart environment, as well as the great use of social networks (Facebook, Twitter) and the appearance of the fourth industrial revolution. Large data is described by features summed up in 5V - volume, velocity, value, veracity and variety (figure 2.). Data of large volumes, generated from different distributed sources in different formats that are created and processed at high speeds, is required to be statistically predictable. Their authenticity and value can vary significantly depending on how they are handled. The data size of one terabyte (1 TB) and greater is considered a Big Data. So large amounts of structured data organized in the form of databases or tables and unstructured data (multimedia content, photos, sound, video, GPS data, user-generated content) require scalability of appropriate resources for their management, storage and analysis. Analysis and correct interpretation give importance to collected data. Large data sets represent a milestone in the development of society and data-based economy.

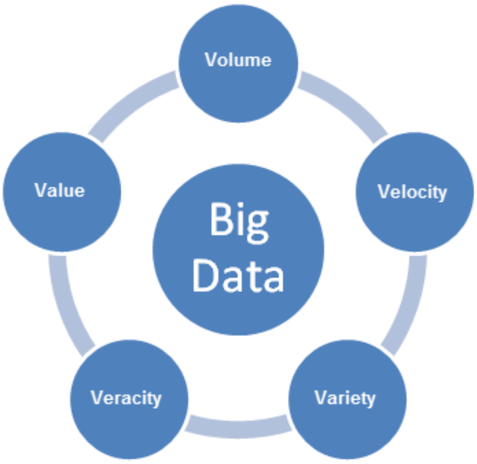

Fig. 2. The 5V features of Big Data [26]

Big data technologies imply the use of highly scalable distributed architectures and parallel processing systems. Cloud computing is highly scalable geographically distributed computing platform that needs to be designed to support the development of big data applications and ensure secure storage, processing, learning and analysis of diverse non-centralized data that presents a major challenge. Big data technology must respond to key issues of security, governance and privacy. Clouds, data centers and IoT platforms must make decisions, discover and anticipate knowledge with help of artificial intelligence, machine learning and data mining algorithms for sorting, organizing and grouping a large number of data and extracting relevant information. Therefore, it is important to ensure data consistency and security, most often through a non-relational database (NoSQL) that corresponds to the big data features. Cloud computing requires some adjustments to support large volume of data, and it can be assumed that these two technologies should be developed in parallel. The basic cloud computing services (SaaS, PaaS and IaaS) are being upgraded by the emergence of large amounts of data and additional models appear as a database as a service (Database as a Service) or large data as a service (Big Data as a service Service). Cloud computing allows the storage of large data in distributed locations with balanced load. Large organizations that use the cloud computing environment to work with large data are Google, Amazon and Microsoft.

Privacy and security are expected to be the biggest challenges facing cloud computing, which has an impact on the processing and storage of large data due to the large number of 3rd party services and the use of infrastructure that processes and stores ever-increasing amounts of confidential data. Besides storage, the challenges of large data are [19]:

- data processing - requiring appropriate computer resources (CPU, network, storage) and efficient scalable algorithms for quality assurance, processing sensitive data in real time and data reduction, removing irrelevant data
- data management - visualization, analysis, integration and management of large quantities of unstructured heterogeneous data using new technologies for data organization
- data transfer - in different phases of data lifecycle data has been transferred from sensors to storage, integrated from multiple data centers, transferred to the processing platform and transferred to the site for analysis which requires preprocessing and development of algorithms to reduce data size before transfer
- latency and network bandwidth, information on the data location when using cloud computing

- security and privacy - a fundamental issue for ensuring reliability, integrity of access to unstructured and heterogeneous data seeking new encryption standards, algorithms and methodologies and data management systems that will successfully respond to scalability and required performances and ensure user privacy (health data, social networks, etc.)
- data quality - accuracy, completeness, redundancy and consistency of variable data used by multiple participants
- standardization of applicable solutions and technologies
- reduction of energy consumption in data centers.

3.2. CERN's LHC as a source of Big Data

Scientific research conducted in many areas collects large amounts of data through high-pass sensors and instruments. Scientific achievements are difficult to conceive without modeling and simulation, processing experimental data, or observational instrumentation. Top-level science requires infrastructure for efficient and fast data processing, storing large amounts of data, for transmitting large volumes of data, remote access to resources and systems for the joint work of a large number of geographically remote researchers. Infrastructure needs to be customizable, scalable and easily approachable.

A representative example of big data is the world's largest particle collider LHC [20] at CERN, the European Laboratory for Particle Physics, which produces up to 1PB data per second. There are numerous challenges in the areas of network connectivity, architecture, storage, databases, and clouds due to the growing needs of LHC experiments. Four detectors of the underground LHC ring (figure 3.) at the French-Swiss border, ALICE, ATLAS, CMS and LHCb, observe collisions.

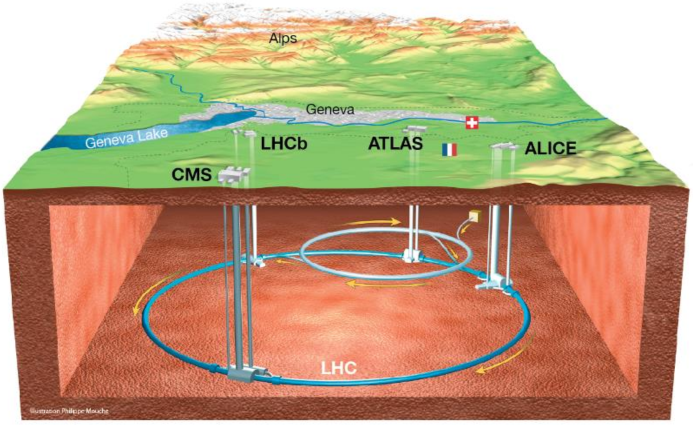

Fig. 3. LHC accelerator at CERN [22]

About 1 billion particle crashes occur every second within the LHC experiments and data is generated at a rate of 1PB/s before filtration. LHC experiments produce tens of PB per year. The storage infrastructure is derived using appropriate software mediation levels that ensure high reliability, performing demanding functionalities such as tiering, data deduplication, and data replication. Grid computing is one of the major paradigms for scientific application. The distributed grid infrastructure of the WLCG [22] has the task to provide resources that should support the successful functioning of the entire experiment. The WLCG is organized in tiers (figure 4.), and today consists of more than 170 computer centers distributed in 42 countries around the world. The grid infrastructure is based on the World Wide Web technology that was created at CERN in 1989. Tier-0 provides the CERN Data Center and the Wigner Research Center for Physics in Budapest, providing only 20% of the required resources. Tier-0 delivers unprocessed raw data from the detector and the reconstructed data to the Tier-1. Also, Tier-0 performs reprocessing when LHC is not running. Tier-1 made of 13 computer centers offers computers and systems to store the proportional part of Tier-0 data and reprocesses the data. Tier-0 and Tier-1 connected with high-bandwidth network links provide disk storage and permanent tape storage. Tier-2 includes resources from about 160 scientific institutions that perform Monte-Carlo simulation and data analysis of data distributed from Tier-1. Part of simulation results is stored on Tier-1 resources. Individual local clusters have access to WLCG resources through the informal Tier-3 level.

Grid as a centralized system has many advantages when analyzing data in this high energy physics experiment. Multiple copies of data are distributed over sites around the world, providing a straightforward access to more than 9,000 scientists regardless of their geographic location. The WLCG consists of network, hardware resources, middleware level that provide connectivity and robust resources and data analysis software framework ROOT [27]. The average data transfer rate this year is 35 Gb/s. It is expected that by end of 2018, WLCG will need to provide 50 PB for data storage. The volume LHC produces make it big data source. The data is generated at high speeds, under varying conditions, and from that large amount of data, data interesting for further analysis has to be extracted. The computer needs of the LHC experiments are growing and different models to meet them have to be considered.

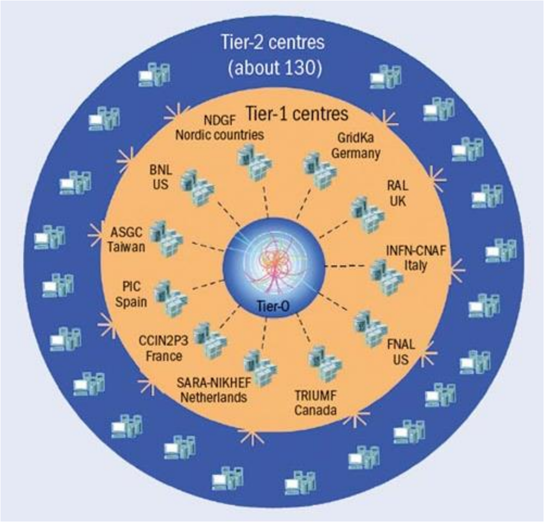

Fig. 4. WLCG tiers [21]

ALICE (A Large Ion Collider Experiment) [23] experiment is formed around the detector for the study of heavy ions and the study of a new state of matter called quark-gluon plasma (QGP) using proton-proton, nucleus-nucleus and proton-nucleus collisions at high energies (up to 14 TeV). It uses 18 systems of specific design and roles. Next year, ALICE will begin with adaptations [25] to achieve maximum scientific productivity of the detector. The upgrade will last two years, and will significantly increase the trigger system rates that filters the data and selects those events that are interesting for further analysis. After the upgrade, in the next Run 3 and Run 4, significant increase of scientific data is expected. The increase of two orders of magnitude will need to be processed and stored. ALICE takes approximately 20% of the total WLCG resources.

The grid computing paradigm became one of the major paradigms for high performance computing systems and is slowly replaced by the scalable cloud computing paradigm. With the development of new mediation systems in the field of cloud computing, grid infrastructures in the world have begun to move to the cloud computing paradigm. CERN increasingly virtualizes Tier-0 resources and introduces private cloud computing infrastructure using open source platform OpenStack. Cloud as a platform can be used for sharing data under appropriate conditions that meet privacy and reliability.

It is been considered to introduce commercial clouds in a hybrid model to ensure maximum efficiency and with a certain budget. Big data technologies, Apache Spark, Apache Hadoop and MapReduce should also be considered to address the challenges of scalability, analysis, storage, data visualization, intensive data processing and data management. New technologies and approaches need to be further explored to optimize resource utilization.

4. Conclusion

Data-intensive science consists of taking, storing and analysis of the great volumes of data that are network accessible from any location by portable devices. Scaling of data represents a challenge for tools and technologies for data management. The CERN’s LHC, is capable of generating several petabytes (PB) of data per day and scaling of its data is particularly important. Data-intensive computing system should be based on storage, computing, and presentation services at every node of an interconnected network. Grid computing became one of the major paradigms for scientific applications. Cloud computing is the technological revolution step of computing science based on large datacenters for processing, long-term storing and analyzing large volumes of multidisciplinary data. In the future, cloud infrastructures will have big role and task to match with large amount of data. It has potential to transfer society in many sectors, health, finance, energy.

Big data poses challenges for scientists and IT and computer science experts that will play key role in enabling future scientific discoveries. Data-based innovations are key drivers of market growth and creating of new technologies. Collection of tools and technologies will need to be developed to support research of the 21st century. Hadoop, a big data open source processing tool, is an example.

This paper contains overview of emerging computing paradigms for data-intensive society and for supporting big data applications. To overcome the computing challenges, the industry is set to face in the coming years, cooperation of data paradigms and new approach to data handling need to be discussed. It can be concluded that particularly important is the safe and adequate data storage. Migration to the cloud is accelerated and it is important for users to understand a number of aspects and to become familiar with the responsibilities of using cloud. Privacy and security concerns arise due to distribution of data and their diversity. Furthermore, adoption of standards is needed to ensure that all manufacturers and operators work on a single product and to reduce the loss of energy and resources to the variety of offered solutions.

Next step of author´s research will be compatibility of these paradigms in providing storage and security of data.

5. Acknowledgments

The author would like to thank the supervisor Prof. S. Gotovac for the encouragement. Also, thanks to the anonymous reviewers for their comments on an earlier draft of this paper.

6. References

[1] Moore, G. (1965). Cramming More Components onto Integrated Circuits, Electronics, Vol. 38, No. 8, pp. 114-117

[2] Torelli, J.C. et al. (2010). A high performance 3D exact Euclidean distance transform algorithm for distributed computing, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 24, No. 6., pp. 897-915

[3] Kornobis, E. et al. (2015). TRUFA: A User-Friendly Web Server for de novo RNA-seq Analysis Using Cluster Computing, Evolutionary Bioinformatics Online, Vol. 11, pp. 97-104, doi: 10.4137/EBO.S23873

[4] Sterling, T. L. et al. (1995). Beowulf: A Parallel Workstation for Scientific Computation, Proceeding, International Conference on Parallel Processing, pp. 11-14

[5] Foster, I. & Kesselman C. (1998). The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publishers, 1st edition, ISBN-10: 1558604758, San Francisco, USA

[6] Simmon, E. (2018). Evaluation of Cloud Computing Services Based on NIST SP 800-145, National Institute of Standards and Technology, Information Technology Laboratory, Available from: https://doi.org/10.6028/NIST.SP.500-322, Accessed: 2018-05-09

[7] Gastermann, B.; Stopper, M.; Kossik, A. & Katalinic B. (2014). Secure Implementation of an On-Premises Cloud Storage Service for Small and Medium-Sized Enterprises, Procedia Engineering, 25th DAAAM International Symposium on Intelligent Manufacturing and Automation, pp. 574-583, doi: 10.1016/j.proeng.2015.01.407

[8] D’Agostino, D. et al. (2013). Cloud Infrastructures for In Silico Drug Discovery: Economic and Practical Aspects, BioMed Research International, Vol. 2013, Article ID 138012, 19 pages, doi: 10.1155/2013/138012

[9] Kagadis, G. C. et al. (2013). Cloud computing in medical imaging, Medical Physics, Vol. 40, No. 7, doi: 10.1118/1.4811272

[10] Simjanoska, M.; Gusev, M. & Ristov, S. (2015). Platform’s Architecture for Colorectal Cancer Research in the Cloud, Procedia Engineering, 25th DAAAM International Symposium on Intelligent Manufacturing and Automation, pp. 1099 – 1107, DAAAM 2014, doi: 10.1016/j.proeng.2015.01.472

[11] Morabito, R.; Kjallman, J. & Komu, M. (2015). Hypervisors vs. Lightweight Virtualization: A Performance Comparison, Proceedings of the IEEE International Conference on Cloud Engineering, pp. 386–393, doi: 10.1109/IC2E.2015.74

[12] Bonomi, F.; Milito, R.; Zhu, J. & Addepalli, S. (2012). Fog Computing and Its Role in the Internet of Things, Proceedings of the first edition of the Mobile Cloud Computing workshop, pp. 13–16, ACM

[13] Vilalta, R. et al. (2017). TelcoFog: A unified flexible fog and cloud computing architecture for 5G networks, IEEE Communications Magazine, Vol. 55, No. 8, pp. 36–43, doi: 10.1109/MCOM.2017.1600838

[14] Zhu, J. et al. (2013). Improving Web Sites Performance Using Edge Servers in Fog Computing Architecture, 2013 IEEE Seventh international Symposium on Service-Oriented System Engineering, pp. 320–323, IEEE, doi: 10.1109/SOSE.2013.73

[15] https://www.openfogconsortium.org/, OpenFog Consortium, Accessed on: 2018-05-19

[16] Wang, Y. (2016). Definition and Categorization of Dew Computing, Open Journal of Cloud Computing, Vol. 3, No. 1, pp. 1-7

[17] Wang, Y. (2015). Cloud-dew architecture, International Journal of Cloud Computing, Vol. 4, No. 3, pp. 199-210

[18] Skala, K. (2015). Scalable Distributed Computing Hierarchy: Cloud, Fog and Dew Computing, Open Journal of Cloud Computing, Vol. 2, No. 1, pp. 16-24

[19] Yang, C. et al. (2017). Big Data and cloud computing: innovation opportunities and challenges, International Journal of Digital Earth, Vol. 10, No. 1, pp. 13-53, doi: 10.1080/17538947.2016.1239771

[20] https://home.cern/topics/large-hadron-collider, The Large Hadron Collider, Accessed on: 2018-05-09

[21] https://www.uibk.ac.at/austrian-wlcg-tier-2/background.html, Austrian Federated WLCG Tier-2, Accessed on: 2018-05-09

[22] http://wlcg.web.cern.ch/, Worldwide LHC Computing Grid, Accessed on: 2018-5-09

[23] https://home.cern/about/experiments/alice, ALICE, Accessed on: 2018-05-15

[24] https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf, (2015). Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are, Cisco, Accessed on: 2018-04-25

[25] https://cds.cern.ch/record/2011297/files/ALICE-TDR-019.pdf, (2015). Technical Design Report for the Upgrade of the Online-Offline Computing System, Accessed on: 2018-01-15

[26] https://www2.microstrategy.com/producthelp/10.7/WebUser/WebHelp/Lang_1033/Content/mstr_big_data.htm, Analyzing Big Data in MicroStrategy, Accessed on: 2018-05-10

[27] https://root.cern.ch/, ROOT, Accessed on: 2018-05-09