An architecture to build high performance infrastructures on cloud computing for telecommunications organizations

Omar Antonio Hernández Duany[1], Caridad Anías Calderón[1], Roberto Sepúlveda Lima[2], Fernando de la Nuez García[1], Cornelio Yáñez-Márquez[3]

[1]Telecommunications and Informatics Studies Center, Telecommunications and Electronics School, Technological University of Havana, Havana, Cuba
[2]Computer Engineering School, Technological University of Havana, Havana, Cuba
[3]Centro de Investigación en Computación, Instituto Politécnico Nacional, Ciudad de México, México

Article Info

Article history:

Received Feb 3, 2025
Revised Jun 13, 2025
Accepted Aug 1, 2025

Keywords:

Architectures
Cloud computing
High performance infrastructures
Parallel programming
Telematics

ABSTRACT

Nowadays, many small and medium organizations of the telecommunication sector must solve intrinsic heterogeneous problems in their own environments that have been associated with high computational complexities of their algorithms. These class of problems require to use high performance computing (HPC) infrastructures for their executions. Therefore, these must be accelerated to reduce significantly the execution times, included many problems that should be solved in real time: like the processing of multiples video streams, the pattern recognition in big volumes of data, the traffic analysis in cybersecurity solutions and among others. The building of HPC infrastructure permits to organize the technological platform to increase the productive and business indicators of the organizations. This paper describes an architecture as reference model and ecosystem for the building and systematic improvement of HPC infrastructures based on practical experiences from successive process of HPC infrastructure building on cloud deployment. That's processes have been useful for the organizations permitting the integration of emergent hardware and software components launched to the international market. This landscape vision is pertinent for academics, scientifics and business organizations compelled to implement scientific and engineering applications to diverse fields that have a high impact in the society digital transformation.

Corresponding Author:

Omar Antonio Hernández Duany
Telecommunications and Informatics Studies Center, Telecommunications and Electronics School
Technological University of Havana
La Habana, Havana, Cuba
Email: omar.hd@tele.cujae.edu.cu

1. INTRODUCTION

High performance computing (HPC) is a relevant technology that permits to process data and tasks much faster than a single server or computer devices, usually denominated in this context as computing nodes [1]. The applications should execute complex calculations at high speeds over a collection of computing nodes connected in clusters HPC or in distributed computing environments [2]. Tasks that usually could take weeks or months on a standard computing system consume much less time in HPC environment using most efficiently the hardware resources [3]. Also, it is possible to solve real time processes associated to the problems of the telecommunications field in proportions of times close to milliseconds.

The state of the art in HPC is shaped by advancements in hardware, software, and interdisciplinary applications, driven by the demands of the high cost computational problems. Early foundations (1995-2010) Department of Energy (DOE) leadership computing facilities in 1995 established at Argonne and Oak Ridge National Laboratories for large-scale scientific simulations, early HPC networks such as the energy sciences networks, message passing interface (MPI) became the facto standard for distributed memory parallel programming. Graphics processing unit (GPU) Adoption for scientific computing recognizing their energy efficiency much better of central processing unit (CPU). In (2011-2020) have been developed the exascale and accelerators launched in 2016, this U.S initiative to integrate hardware, software and application for exascale systems. Arm architecture in HPC demonstrated scalability and efficiency for superior performance and power optimization. Benchmarking evolution for CPUs/accelerators, supporting MPI, open multi-processing (OpenMP) and hybrid models for cross platforms, artificial intelligence (AI) and machine learning (ML) integrations on HPC.

The period (2021-2025) is considered the exascale. The first international research center to officially break the exascale computing barrier was Oak Ridge National Laboratory (ORNL) in the Unite States. The Frontier supercomputer was certified as the first true exascale machine on May 30, 2022, by the TOP500 list. Argone 2025 Intel based system targeting AI driven research Xeon CPUs and data center GPUs. Lawrence Livermore National Laboratory (LLNL) 2024 world's fastest supercomputer dedicated to nuclear security simulations. Energy efficiency reflecting advances in power-aware design. Power and edge HPC cloud system more than 550 petaflops expanded access to HPC resources. Global collaborations exascale efforts emphasized regional self-support in HPC infrastructure. Emerging trends (since 2025 and beyond) are heterogeneous architectures mixing, CPU, GPU and domain specifics accelerators. Memory and networking innovations, sustainability reducing the energy consume and democratization the alliance between university and enterprise.

It is important to consider that many daily heterogeneous processes in small and medium size organizations of the telecommunication field must be resolved from the integration of new results developed in affine scientific fields must be executed as fast as possible and it should impact in the organizations productivity. The international HPC and parallel applications developer's community is intensely working in the designing and systematical implementation of new models and infrastructures to support HPC scalable technological solutions. Tens, hundreds, thousands, or hundreds of thousands computing nodes may be linked in clusters HPC, computing nodes: servers, personal computers (PCs), laptops, field-programmable gate arrays (FPGAs), and among others [4]. Small and medium telecom organizations are integrating digital innovation and cost-effective service models, while challenges like technological disruption have increasing prices in the market formulating a strategy to take advances of the opportunities in core connectivity, ecosystem collaboration as resource to impulse the society digital transformation.

HPC not only is useful to solve big scientific simulations and modelling of complex specialized problems, as was originally [5]. The telecommunications organizations require right now to solve a wide and diverse range of heterogeneous problems with high computational complexity. These intrinsic problems belonging to the telecommunication sector must be accelerated, such as: adoption of software-define networks, self-service portals implementing, AI-driven support, e-commerce, optimization of heterogeneous data streams processing in real time, operations of call center, visual processing to make decision, real time data analysis, data and video streams processing; concurrent processing of clients request, patterns recognizing in big volume of data, automatic translating in real time and among others. The generality of current telecommunication processes requires to be accelerated to reinforce the efficacy of the capture, processing, transmission and storing of big data to manage many heterogeneous streams: data, audio, video, multimedia, radio signal, and among others. reducing the cost-effective using parallel hybrid work environment,

The elements previously introduced constitute a motivation to formulate the scientific question: how to design an architecture to build HPC infrastructures on cloud for the small and medium-sized organizations of the telecommunications field and beyond to solve heterogeneous complex problems in academic, scientific and business environment? this scientific question is useful for the formulation of the scientific problem of this research: how to design an architecture as reference model to build HPC infrastructures based on the integration of processors advances, management environments, development tools, parallel applications and technological solutions of the telecommunication field, optimizing the cost/performance of the investment and reducing the time and financial resources associated to the designing and deployment processes in small and medium size organizations?

From the need to solve the scientific problem enunciated the authors have defined the following objectives: to evaluate the relation cost/performance, financial resources and return of investment (ROI) of the HPC infrastructure design. To optimize the extreme computational performance of the HPC infrastructures. Design parallel algorithms and applications to elevate the efficiency and performance of the built HPC infrastructures. Process and analyze data stream in real time using artificial intelligence models. In the telecommunication field there are many complex problems whose complexity is increasing in the organizational environment [6], but in the TeleHPC the focus has been oriented in the current stage to accelerate the video processing from the development of parallel applications.

The use of HPC on cloud offers a unified experience for each one of the workloads with a validated methodology that focuses on developing an operational model to accelerate the transformation to a sustainable information technology (IT) strategy for telecommunication organizations [7]. An inform [8] published in 2023 refers that the HPC computing market size was valued at USD 48.51 billion in 2022 and is expected to expand at a compound annual growth rate (CAGR) of 7.5% from 2023 to 2030. The growing demands for high-efficiency computing, advancements in virtualization, continued diversification and expansion of the IT industry, and the increasing preference for hybrid HPC solutions are the factors that are expected to drive the growth. Now, the scientific community have arrived to the exascale computing as a level of supercomputing that is capable to achieve one quintillions of floating point operations per seconds. Exascale systems are characterized by their ability to manage massive dataset, complex simulations and processing of big volume of data in real time.

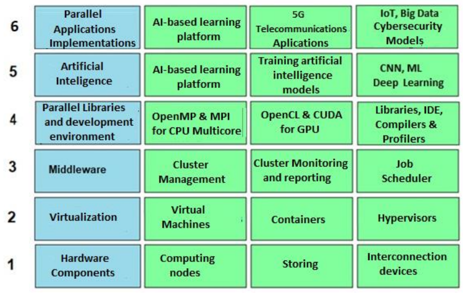
All cases have the purpose to elevate the efficacy of the making decisions process. Many organizations haven't to access to the exascale performance (over 1 quintillion operations per second) through massive parallelization, but there are able to dispose to financial resources for development of their own HPC infrastructures in function of increase their productive and business indicators [9]. These processes stream usually can contain relevant information associated to the intrinsic processes of the organizations that are very useful for the decision making [10]. In many cases these processes should be processed in real time forming part of the critical mission technological solutions. The key finding with the scientific question is to propose a transversal architecture [11] as an ecosystem to facilitate the design and building of full HPC infrastructures or parts of them based on the application of the research methods in Telecommunications organizations [12].

The authors of this paper wish to promote a better understanding of the scope and knowledge based on previous practical implementations during the last 25 years in small and medium size organizations of the telecommunication field with low or medium budgets. It has the purpose to propitiate a referential model to design and to implement new HPC infrastructures in different organizations with relatively low cost, from the systematic integration of emergent hardware industry advances and the structuration of a technical recommendations set, described in each one of the layers of the architecture. It is very important to apply the processes belonging to each one of the layers. The precedent computational complexity analysis of the problems helps to the developers to understand how much time or space consume an algorithm and determine what is the proportion of the inputs grows.

2. METHOD

HPC is an essential technology with the purpose of minimize the execution times in the telecommunications field because usually this class of organizations should solve many emergent technological problems using real time and critical mission applications [13]. Therefore, it is necessary to maximize the efficiency of the hardware architecture available [14]. The HPC architecture is a relevant conceptualization to build a HPC infrastructure on the cloud computing environment from the management of hardware and software components by mean of a middleware for the orchestration of each one of these components. This paper presents the main challenges to deploy the mentioned architecture by levels to solve high complex computational problems. The architecture permits to create leverage hybrids systems combining CPUs, GPUs, and specialized accelerator such as tensors cores for AI workloads to optimize performance for difference tasks.

The HPC architecture presents an ecosystem to build HPC infrastructure for an organization describing step by step from bottom to top level. Each one of the level contains a functionality component that are very useful for the entire deployment process, as is showed in the Figure 1.



6	Parallel Applications Implementations	AI-based learning platform	5G Telecommunications Aplications	IoT, Big Data Cybersecurity Models
5	Artificial Inteligence	AI-based learning platform	Training artificial intelligence models	CNN, ML Deep Learning
4	Parallel Libraries and development environment	OpenMP & MPI for CPU Multicore	OpenCL & CUDA for GPU	Libraries, IDE, Compilers & Profilers
3	Middleware	Cluster Management	Cluster Monitoring and reporting	Job Scheduler
2	Virtualization	Virtual Machines	Containers	Hypervisors
1	Hardware Components	Computing nodes	Storing	Interconnection devices

Figure 1. HPC Architecture on cloud computing by levels

The HPC architecture is a multilayer reference model that contain a set of essential ideas well-structured about the process of designing and building of HPC infrastructures. It covers hardware and software components required for the deployment process specifically in the small and medium-size organizations of the telecommunication sector. It establishes a direct connection with other own process of the telecommunication field. The architecture has the purpose to achieve more efficiency and effectiveness for the HPC infrastructures based on cloud computing for the resolution of heterogeneous complex problems in different research subfields [15]. This structured approach ensures a robust HPC architecture to satisfy the client's demands with flexibility for future growth and involving workloads associated to the execution of parallel applications [16].

It is necessary to consider there are a set of no functional requisites that are out of the core of HPC but these have repercussion in the HPC infrastructures and it is very important to be take in consideration for example. The hot generation due to the functioning of hardware components, so it is necessary to guaranty the efficient of cooling systems to maintain optimal operating conditions inside the data center using air/liquid/immersion. Also is important to achieve energy efficiency by mean of minimize energy consumption. There are many other aspects such as cybersecurity, perimeter protection, and among others considering that the importance of the infrastructure as core of the telecommunication organizations.

2.1. Hardware components

2.1.1. Define hardware HPC requirement

First of all, the designers must estimate the problems computational complexity that should be solved in the context of the organization needy to build its high performance infrastructure. It is appropriate to establish the compatibilities with the behavior of the performance metrics associated to the process parallel algorithms design [17]. Designing and deployment of HPC infrastructures from zero should identify relevant intrinsic processes for the organizations and their computational complexity. This class of analysis permits to estimate the approximated time required for their solution. Therefore, it is possible also estimate the computational performance required and the hardware infrastructure that should be built to solve the object of studio problems. This process it isn't trivial. It demands the expertise of designers and developers and the economics cost estimation [18].

The small and medium-size organizations in telecommunications sector must analyze meticulously their financial availability to design the HPC infrastructures [19]. The financial resources should be enough to acquire the most advanced hardware technologies available at the market, taking in consideration the behavior of the cost/performance relation within the projects [20], and the technical characteristics of the problems in order to achieve a rapid ROI [21]. The investment analysis is the starting point for the continuously development of the HPC infrastructures and it will do possible the ulterior estimation of the scaling necessity in function to increase successively the potentiality of the hardware infrastructure with the purpose to resolve every time more complex problems.

By other hand, its important a workload analysis to identify the type of computations (CPU/GPU/ tensor processing unit (TPU)), memory capacity required and input/output (I/O) intensive. Performance goal define the performance required: floating point operations per second (FLOPS), latency, throughput, and scalability needs. Budget analysis consider the investment cost and available financial resources. There are a set of non-functional required that must be exhaustively analysis, such as: power consumption, physical space conditioning, mechanical design. It is necessary to make compatible the possibilities to the demands. The hardware infrastructure must be considered as an investment.

2.1.2. Design the HPC architecture

The designing of HPC infrastructures require to analyze specialized hardware components [22], software frameworks and often large facilities to host it [23]. HPC also comes with relatively high price tag, which can act as a barrier to many organizations hoping to implement. An important objective of the organizations is to achieve that the cluster HPC built is continually busy, in this cases the investment can be considered as successful [24] and probably will generate utilities for the organizations. A HPC infrastructure have three main hardware components: compute, network and storage [25]. In nowadays, the computing node can contain different types of processors: CPU multicores, GPU or TPU [26], local random access memory (RAM) memory and interconnection devices of the nodes to the performing computations. All the hardware components must be as high performance and reliability as possible to guaranty the effectiveness of the parallel applications executions also based on the network data transference rate between the master node [27], computing nodes and shared storage to manage high volumes of data and the rapid access and retrieval of the information to increment the applications performance.

In the level 1 each one of the components of the cluster cooperate with the remaining components to enhance computing power. The computing nodes are the result of the integration of hardware components of highest performance possible included performance of the processors, the storage and its access time should be devices of high capacity of storing and minimum hard drive access time possible [28]. It is recommendable the latest in performance-enhancing networking. All these components working together permits to achieve a high performance infrastructure with a superior performance.

The Figure 2 presents the general idea of the physical hardware HPC infrastructure designing based on a federate cluster [29] integrated by nodes of high performance suitable in racks and an indeterminate number of disperse computing nodes on premise or remote. This implementation combines two different focus of deployment the cluster in an organization: rack mount cluster and tier pack cluster. This mechanical design depends mainly of the available financial resources but the manager can use the same management tools for both deployments. All computing nodes are managed using a centralized by means of the high data transfer interconnecting device in a datacenter or on premise of the organization if the developer propose to reuse the computational capacity available.

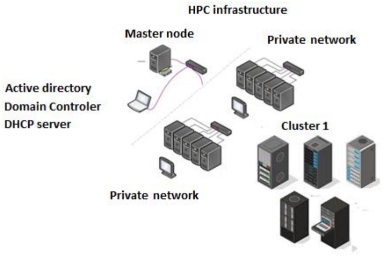


Figure 2. HPC physical infrastructure scheme

In this paper, it isn’t realized a specific analysis about the hardware components available in the hardware market, because the main objective is offers an overview that permits to achieve the generalization of the method as a reference for Telecomm sector organizations and beyond. Always, it’s recommendable to study the tendencies published in the TOP500 list launched semi-annually [30] and the new hardware products commercialized by the most important manufacturing companies at international level in the precise moment in what the organization try to build the HPC infrastructure. The access to these technological components are dependent of the financial resources availability in the organization. In the worst case though the performance doesn’t be so elevated probably will be useful for the solution of organization intrinsic process.

2.1.3. Compute

The compute is formed essentially by two type of nodes: master node and computing nodes. Master node is a special type of node that serves a number different of functions and performs additional responsibilities beyond of client or common computing nodes [31]. The master nodes have a set of important functionalities and it is necessary to add others that can differ in correspondence with the specific purpose the developers need to achieve in function of a specific organization. From the hardware point of view, it is necessary to take into consideration that the requirements of the server node or master node and the quantity depend of the cluster dimensions independently of the number of nodes, because these contain their most important management functionalities.

The nodes are the most relevant hardware component to achieve the computational performance based on their potentialities and quantity of the computing nodes integrated in the HPC infrastructure. The performance of the nodes must be estimated previously to the building of the infrastructure. It is desirable the performance of the computing nodes can be as elevated as possible, but this facility depends of the available financial resources. The hardware requirements of the master node and computing nodes meanwhile higher is better, but it is necessary to consider the relation cost-performance of the hardware that usually is expensive, so the hardware should be acquired after the selection of the more convenient choice available at the market. There are three objectives for hardware procurement and assembly: vendor selection, node assembly and network/storage setup.

2.1.4. Processors, motherboard and memory selection

The processors are the more important hardware components of the nodes and therefore the entire HPC infrastructures. This can effectively manage intensive processes in heterogeneous scientific and engineering applications. These are the main responsible to manage the workload on the infrastructure. The hardware industry is very dynamic and are developing continually new processors models with superior performances and potentialities. The designers must explore continually the main characteristics of the new processor models: performance, prices, caches and among others.

In the current stage can be selected a wide range of CPU multicores developed for different companies such as: Intel XEON, advanced micro devices (AMD) EPYC. CPU multicores enables te nodes to use the shared memory paradigm that is useful to develop parallel applications for example using C++ language combined with OpenMP with the ability to distribute parallel tasks across cores to low cost. It is necessary to review the compatibility of CPUs and Motherboards sockets and chipset before to acquire these components.

GPUs is a specialized circuit originally designed for digital image processing and to accelerate computer graphics are currently a relevant hardware component for HPC architectures because can have more performance than CPUs and these can support more complex workloads at much faster speeds [32]. This type of computations is known as general purpose GPU (GPGPU). It is a kind of parallel processing on GPU. TPU is a custom developed application-specific integrated circuits (ASICs) used to accelerate machine learning workloads. Nowadays can be taken as reference at least the processors series produced by the most relevant hardware manufacturing companies [33]. The memory subsystem relies on the motherboard for physical connectivity, power and communication infrastructure. The motherboard depends on memory technologies to deliver efficiently to CPUs and peripherals. Optimizations like cache hierarchies and advanced memory controllers are critical for balancing latency, bandwidth, and cost.

2.1.5. Selection of storage subsystem

Data storage support data-intensive parallel applications [34]. Therefore, it should keep storing the data generated during the processing in the infrastructure. The storing is an essential facility features for reliable access to large datasets and high-speed storage systems that seamlessly host all user data, consequently is very important a high-capacity as persistent storage. With the storing devices the organizations can store and analyze their data securely within the HPC infrastructure on premise.

Scalable storage solutions, such as parallel file systems, should be implemented to accommodate the growing data requirements of HPC applications. To ensure high reliability and availability, fault tolerance and redundancy measures should be incorporated into design, including redundant power supplies, backup generators, and disaster recovery protocols. These considerations are essential for creating a successful HPC infrastructure that meets the demands of various scientific, engineering, and business applications.

To validate the deployment, the infrastructure storage has been used a Hewlett-Packard (HP) storage work. This has utilities and features that ease the administration task associated with managing the system. The insight manager is a comprehensive tool designed to be a key component of management systems. The tool permits to monitor the operations of computing nodes providing to system administrators more control through visual interface. The parallel file system like Lustre, general parallel file system (GPFS) can be useful to manage the data storing considering the data redundancy.

2.1.6. Selection of interconnection devices

It is necessary to take in consideration that interconnection devices are important elements for building the HPC infrastructures, because contribute to guarantee the connection between the processing element to join all computational capacity and the cooperation between of computing nodes that can be scaling for the resolution of complex problems.

The network infrastructure is another critical resource because must enable the stable communications between the nodes during the execution of parallel applications. These should provide high-bandwidth, low-latency connectivity between computing nodes and storage systems. This facilitates data transfer efficiently and parallel processing capabilities. The selection of the network interconnection devices is very important because have a direct repercussion in the applications performance. It is desirable to have a high data transfer because this contribute to the inter processes communication when it is used the distributed memory paradigm.

The key of interconnection devices that make up the network are: network interface card (NIC) cards belonging to the nodes, switches, routers, bridges, repeaters, and gateways [35]. All these devices must guarantee in the necessary cases a high internodes communications and have different ranges, based on network requirements of the HPC infrastructures of a specific organization. The selection process of the devices demands an analysis of the behavior cost/performance. The strategic is to choose high data transfer, low latency, high bandwidth fabrics There are different technologies (infiniBand, Ethernet).

2.2. Virtual and physical configuration

The cluster configuration is associated to the setup and maintenance of a HPC computer system. The original configurations were realized physically (bare metal) [36] in at least a master node directly on the hardware. Specifically, in HPC configurations the computing nodes are “diskless”, therefore their operating system are loaded from the master node through of the combination Preboot execution (PXE) and trivial ftp (TFTP) to transfer the operating systems image required to achieve the functionality of the nodes.

Cloud native development has redesigned the software landscape with an important impact in the digital transformations changing the way the organizations build, deploy and manages applications with a biggest simplicity [37]. With these components the managers of the HPC infrastructures of the organizations can include cloud, on premise, or hybrid configurations depending of the organization's needs. Cloud infrastructure is a critical component for some of the most demanding workloads in telecommunications organizations [38].

Virtual technological infrastructure focus has been configured and validated in some of the HPC configurations mentioned above. The integrating of hardware blade servers and PCs as dedicated nodes in the organizations can be powerful resources of processing, and potentialities of GPUs, memory to optimize the parallel processing in the applications, which enables these to perform complex calculations at much faster rates than traditional computers reducing the slow access to disk [39].

2.2.1. Virtual machines

Virtual environment refers to how virtual machines (VM) resources are arranged and deployed to allow them to operate with effectiveness on a HPC hardware infrastructure, functioning as a virtual computer system with their own processors, RAM memory, storage and network interfaces, created on physical hardware system. These can be located on premise and content the configuration of the cluster HPC. VM allow the configuration of multiples HPC management environment and each one of the software component to deploy the cluster management.

Each VM runs in the same way an HPC environment and applications usually on the host hardware. So, the final users don't know if their applications are running on a physical or virtual machines. The virtualization technology has permitted to impulse the configuration of an experimental platform to validate dissimilar HPC management environment in the master node of the cluster. For this reason, is necessary that the hardware associated to this master node has the highest performance possible [40].

If it's realized a rapid comparative between physical machines and virtual machines from point of view of performance both have approximated similar performance and can be configured to have identical specifications [41], but for the scalability of the systems virtual machines are a lot more flexible for the adding, removing, or updating resources that is very important to guarantee the scalability of the HPC systems reducing the deployment cost.

From the security point of view, if the virtual machine is on premise the cybersecurity have equivalent conditions that physical machines after the exhaustive analysis by cybersecurity expert's groups of the organizations. The virtual machines add new functionalities more easily and to adjust the configurations in more optimal way are created the possibility to migrate the configurations of master nodes between organizations facilitating the master nodes configurations between cluster reducing the deployment cost. The virtualization permits to organize all knowledge, tools, technics, and maturity resultant of successive processes of deployment and practical validation. The master node was configured in the project using virtual box and VMware machines with the HPC configuration. These are appropriate virtualization products for the organizations.

2.2.2. Containers

The containers are portable software instances that can be executed on a physical or virtual machines. These are images of lightweight, standalone and an executable package of software that includes everything needed to run application: code, runtime system tools, system libraries and settings [42]. Containers permit the decomposition of a monolithic applications in micro services. This is very useful for the modular design and implementation of parallel applications. The containers are the standard and independent units of software that packages up code and all its dependencies so the parallel application runs quickly and reliably from one computing environment to another. Containers are more portable than VM, and is an adequate form to deploy more quickly and easily a HPC stable configuration. In the deployment of the infrastructure have been used Linux containers (LXC) containers keeping the focus of free software also with the purpose of reduce the cost of the deployment [43].

2.2.3. Hypervisor

Hypervisor is a software isolated of the machine's resources from the hardware and provisions them appropriately so they can be used by the VM. A full virtual service-oriented infrastructure is usually managed by a technology that provide resource aggregation, management, availability and mobility, the foundational core of virtual infrastructure is the hypervisor. As principle to develop the HPC infrastructure have been used the hypervisor ProxMox virtual environment. This is a complete-source platform for enterprise virtualization and it is appropriate to deploy a not proprietary and free distribution solutions+.

With Proxmox [44] is also possible to manage easily VMs and containers, software-defined storage and networking, high availability clustering, among others for enterprise datacenters should help both IT decision makers and end users to choose the right virtualization hypervisor for their datacenters. ProxMox server delivers production-ready performance and scalability needed to implement an efficient and responsive datacenter.

In this level has been developed diverse virtual machines containing different middleware. That is an iterative and incremental process that have permitted to create a repository with installations of management tools, job schedulers and monitoring tools independently of the hardware architecture configured in the components. The integration of virtual environment is very useful to reproduce relatively fast the cluster management.

Hypervisors are indispensable for modern HPC, offering scalability and resourece efficiency. These facilitate cluster scaling by enabling rapid VM deployment across nodes. This is critical for HPC workloads that require parallel processing across a big number of nodes. Their adoption requires careful tuning to mitigate performance penalties.

2.3. Middleware

The designers must considerate the requirements of the management tools demand and futures applications. Both aspects can have hardware requirements without limits predefined. Middleware in HPC environment is a software spatially located between operating system and the parallel applications [45]. In the HPC also functioning as hidden translation layer. Middleware enables the communication and data management for parallel and distributed applications that will be running in superior levels. As result of the research have been validated a wide range of HPC middleware in the TeleHPC lab. In this space have been possible to improve systematically the HPC environment configured in every one of the virtual machines.

The authors also are improving systematically by mean of technological vigilance and adjust: management tools, deployment and monitoring of cluster HPC that resume the process of configuration and adjustment on the virtual machines and containers with the middleware that also can be a referential guide to select the appropriated middleware to configure for a determinate hardware infrastructure and the type of problems to resolve. In successive processes of designing and deployment have been built and improved continuously the HPC environment on the middleware configured of the master nodes of the hardware infrastructures.

In the TeleHPC lab has been configured a variety of virtual machines that are continuously improved. The conception of the solution idea, design and deployment of a HPC infrastructure are these conditions many more fast. The main efforts have been oriented to the development of open source management tools. The deployment of the management system of HPC infrastructures should be configured integrating heterogeneous nodes with different potentialities to maximize the total computational power. The infrastructure HPC is designed for propitiate the scalability, allowing easy expansion of computing resources to accommodate the growing demands of computational performance managed by mean a software middleware, distributed file systems. The levels 3 to 6 of Figure 1 usually must be configured on the master nodes.

Management and monitoring of HPC infrastructure involve a comprehensive set of processes and tools to ensure the optimal performance, reliability, and efficiency of the facility. This includes monitoring the health and performance of computing hardware, cooling systems, network infrastructure, and storage solutions, as well as tracking energy consumption and environmental conditions. Advanced monitoring systems collect and analyze real-time data on various parameters, such as temperature, power usage, and network traffic, to identify potential issues and facilitate proactive maintenance [46].

The system administrators must manage the automation and orchestration tools to simplify the deployment, configuration, and management of computing resources, software, and workloads. These tools help maintain high levels of system availability, optimize resource utilization, and minimize downtime, all while ensuring the HPC data center operates within the desired performance and efficiency parameters [47].

The building process of cluster HPC on cloud require software components to deploy and manage the infrastructure, including provisioning tools, development tools and scientific libraries. At least a master node must control the functioning of all infrastructure's computing nodes of the cloud, and it can be escalated relatively easy. All HPC infrastructure aren't equal because the systems can take on several different forms depending of the necessities and their capabilities. In all cases, the concept used is quite generalizable and can help to their administrators to assume the deployment role and future maintenance.

The virtual machines with cluster configuration in the master nodes permits to deploy the cluster rapidly only adjusting the characteristics of the hardware infrastructure, previous structuration of the system building with simplified and fast cloud with file systems to combine multiple data sources into a single HPC system. Resource aggregation refers to the capability to improve pool, memory, processing power, transfer rate of the network, and storage across computing node instances with the capability to perform live migrations of running virtual machines or containers from one physical server to another in response of the hardware resource availability [48]. This process is transparent to final users.

2.3.1. Configuration of master node

The master node is the most relevant component of the HPC hardware infrastructure because carry out the fundamental cluster's functionalities: configuration, system management, supervision, job schedulers administration, database servers, load balancers, data management software from the file systems, networking and web application for the access to the infrastructure. The entire system is managed and controlled by at least a dedicated master node. The authors also recommend to use free software. Users can access to the cluster HPC only via the master node using the system's frontend.

The jobs on the cluster computing doesn't simply work out of the box. It is necessary to use a specialty software serving as the orchestrator of shared computing resources will actually drive nodes to work efficiently with modern data architecture denominated HPC schedulers. These must distribute the jobs using the balanced way on the computing node of the infrastructure, that tool of HPC system will always use cluster computing scheduling software in place. This software will often direct system management operations like moving data between long-term cloud storage and the clusters HPC or optimizing cloud file systems [49].

The HPC management tool is a software system optimized for the administration of the HPC infrastructures should be selected considering the specific needs of the organization. The architecture permitted the validation and continue enhancement of different environment of cluster HPC management. These have approximately similar capabilities and have been validated on ProxMox using VM: virtual box and containers LXC and can be useful to different organizations in dependence of their necessities in heterogeneous scenarios.

The HPC system is running on the distribution of Linux Ubuntu 20.04 and simple Linux utility for resource management (SLURM) workload manager for job scheduling. The cluster can contain an undetermined general purpose (GP) computing nodes, special purpose (SP) computing nodes and 2 frontend nodes interconnected, this aspect also depends of financial resources. Intensive I/O is supported by HP storagework, while home directories are serviced by NFS file system with global access. All inter-node communication (OpenMPI) is through a low-latency network.

It is important to consider that OpenHPC (www.openhpc.community) is a collaborative community that develop and share common packets to aggregate different common functionalities required to deploy and manage High Performance Computing. The community works to integrate different components that are commonly used in HPC systems, and this are freely available for open source distribution. This community is a source of valuable resources to developers and maintainers that provide key components used in HPC around the world today. The authors recommend to review before to begin the building of a new modules the existence of similar solutions in OpenHPC. This focus can help to reduce the cost in diverses resources.

2.3.2. Configuration of computing nodes

The client or computing nodes are responsible to realize the calculation that is sent from the master node. Each one have been configured as diskless nodes, which employ network booting to load their operating system images received from the server (master node). This method is so appropriated and useful to obtain a centralized management way that reduce the deployment cost and contribute to the successive actualizations of the HPC infrastructures and their scalability more easily. The only configuration required for the computing node is to set preboot execution environment (PXE) in the basic input/output system (BIOS) as a client-server interface in the cluster network to boot from master node [50]. After preparing the target computing node the OS image and the HPC configuration the work can start. The computing nodes are stateless, so these don't store any information or session state about previous request in a local store. All information about the state is stored in the share storage management by the master node.

2.4. Parallel applications development

Parallel applications in the telecommunications field have an increasing tendency with a common denominator of accelerate their execution time. Developers can use many of the same techniques used to design the HPC traditional systems [51], accelerate and optimize applications enhancing the efficiency of high performance processors: CPU multicores, GPU many cores and others coprocessors. In this level the first objective is the knowledge about the principles of design of parallel algorithms dividing a computation into smaller computations and assigning them to different processors for parallel execution. The design can be defined as a set of processes or tasks that will be executed simultaneously and may communicate which other in order to solve a big dimensions or complex problem.

By other hand, C++ is also extensible with OpenCL for the parallel applications acceleration using GPU regardless the manufactures. The efficiency of code in parallel applications is always critical regardless the modern compilers usually optimize the code with the purpose of develop applications “close to the hardware”.

Software Stack meet the diverse needs of the Telecommunication research community. It includes a range of scientific computing libraries, parallel processing tools, and programming languages such as C/C++. Additionally, it supports a wide range of scientific software packages and modeling tools using general library functionalities such as: computer vision: OpenCV, Yolo, point cloud library (PCL), robot operating systems (ROS), MATLAB, Mathematic. Digital audio processing: PDS Project Library (PDSPLib), SoLoud, Essentia Library, Ajaakman, Maximilian. RealTime IOT Logic communication: SON, ibm-Watson-iot, POCO. Serial communication TCP/IP: Seriallib, Serialport, Serial communication. Networks management: Superpowered Networking Library, Apache Serf, Boost Asio, cpp-netlib, cpr, dlib and libcurt.

2.5. Artificial intelligence and algorithms design

Currently, many technological solutions of the telecommunications field are being resolved using applications based on AI. HPC on their hardware infrastructures that handle compute-intensive tasks and deeply AI and machine learning can benefit greatly from HPC.

An AI algorithm in the telecommunication field contain tree main components. Data entry, a function or trained model with data, and the output results. It is possible to build many AI models and algorithms to different purposes. The organizations can carry out many functionalities in the telecomm field: manage, maintain infrastructures, customer support operations, network optimizations, predictive maintenance, virtual assistants, cybersecurity, video processing, signal processing and among others.

The AI model training is the process of feeding the algorithms data, examining the results, and tuning the model output to increase accuracy and efficacy. The AI models training is a process with landscape applicability. It is relevant to the most effective executions of the algorithms. AI-based learning platform on HPC infrastructures is a valuable tool for the telecommunications organizations during AI model training as process of feeding curated data to selected algorithms to help the system refine itself to produce accurate responses to queries, considering many different type of AI algorithms for a project depends on scope, budget, resources and goals.

Nowadays, HPC and AI models should be linked to obtain the enhancement of the algorithms more rapidly. The computational power and scalability of cluster HPC are relevant factors for AI-based software. A cluster with more performance could reduce the time it takes to train AI models and also offer potentialities for the generation of more precise models. These generative models are used as computational representation to do predictions, decision making or specifics jobs in this field.

In the current stage the artificial intelligence and HPC infrastructure are relevant tools that are called to solve many problems on HPC infrastructure that provide the massive computational resources necessary for training advanced machine learning, models, development of innovations in AI research, and enabling the development of wide spectrum of algorithms across various knowledge areas of the telecommunication sector.

2.6. Applications

HPC infrastructure enable a wide range of applications across various domains that require a high computational power and high data processing capabilities. At the current stage, some notable applications include the integration telematics, computer vision, real-time signal processing, big data, internet of things (IoT), an AI providing an environment to share knowledge, resources, and expertise about complex scientific problems and perform large-scale simulations and modeling of heterogeneous technological solutions of the telecommunication field [52], to name just a few examples. HPC is useful to solve problems that require to minimize their execution times by optimizing the use of HPC hardware architectures capacity, thus achieving an expansion of the scope of their results depending on the applications [53].

HPC can potentiate the analytical engines and AI behind some of the more innovative telecommunications platforms, with emphasis in the capture, processing and storing of heterogeneous data streams for different applications in this environment. The HPC processes constitute the base to make decisions from analyses the information contained in big volumes of data, specifically in real time, maximizing the computational capacity of the organizations in the path of digital transformation [54].

Among the applications validated in the HPC infrastructure that are common and transversal to various organizations in the telecommunications sector are the following: processing of video streams in real time for the enhance the quality assessment, identification and object recognitions in real time, training of AI models, analyze massive amounts of data to extract relevant information for the decision making using parallel platforms in function to obtain most accurate decisions. It also can be used to create simulations, AI-enabled Telecom. AI-on-5G, extending AI to edge computing, Telecom edge services, AI using machine learning, deep learning solutions inside of the organizations [55].

The infrastructure can support data analysis tools to cover risk assessment, cybersecurity and detection algorithms from determination of processing units for AI is required for the organizations base on the hardware infrastructure available, such as hardware accelerators and proper storage. Hybrid computing HPC systems can benefit AI the telecommunications projects from using GPUs processors to more effectively process AI-related algorithms. Parallelism and co-processing speed up computations offer the conditions to process large data sets and run massive-scale experiments in the least time including the AI models training process. It is necessary to use checkpointing for long jobs and redundant components.

3. RESULTS AND DISCUSSION

The HPC architecture propose a technological ecosystem interconnected to help to engineers, data scientists, designers, architects, researchers and developers to deploy the HPC infrastructures of their own organizations from the integration of hardware and software components in function to resolve large and complex problems in far less time and at less cost than traditional computing. The small and medium-size telecommunications organizations put up high-performance, low-latency high-speed data transmission, high-capacity data storage, high-bandwidth to process dissimilar heterogeneous data streams included in real-time. The architecture contributes to horizontal and vertical scaling. The reducing the economic cost of this type of projects is also a key aspect based on their components.

The multilayer HPC architecture contribute to manage organically the components hardware and software from the designing stage to deploy HPC infrastructure as technological platform in small and medium size organizations of the telecommunication sectors. The main objective is to dispose an approach in what using logical reasoning will can obtain more outstanding performance than a single computer and the integration of the potentialities continually from add emergent components on the configuration of multiple computing nodes networked together and software in every layer of the architecture.

The building of the HPC infrastructures on premise usually require the availability of the financial resources to acquire the mainly the hardware components because the software is open source. The knowledge for the management of the software infrastructure and the capacity to design the technological solutions considering the specifics technical problems of the organizations. The small and medium-size organizations can be of the academic, scientific and business of the telecommunication sector. These need to impulse the technological adaptation and strategies in the context of digital transformation. In 2025 is routing complex landscape HPC platforms using hybrid environment: CPU multicore, GPUs and other accelerators considering the scalability demands and the integration of emerging technologies like AI, Edge computing, IoT and among others applicable in the telecommunication field.

In all cases mentioned before the processes must be accelerated. These challenges require an integral balance between innovation, standardizations, and usability improvements, the type of organizations object of study usually have limited budgets but the complexity of the intrinsic problems are increasing. The designers and developers of the own organization following the guidelines presented by this architecture can to assume the tendencies of HPC international community in function of their own organizations. The basics aspects conceived in the architecture contribute with the organizations to increase their computational potentialities to accelerate their intrinsic processes and also can share these with others remote clients on demands.

The idea behind the architecture is to propitiate the building of HPC infrastructures based on cloud reusing VMs, containers previously built and gradually maturated. The architecture can be reused in many different organizations of the telecommunications sector and beyond. It is a reference model to deploy all components of the infrastructures and execute complex applications from a common configuration in their own organizations. The architecture has been widely validated in small and medium-size organizations. The theoretical aspects discussed have been shared in every one of the designing and deployment processes. This approach is applicable to the HPC infrastructures regardless the hardware characteristics in organizations with different dimensions. In that sense, the theoretical maximum performance (Rpeak) and maximal achieved performance (Rmax) are irrelevant for the architecture.

The architecture facilitates the possibility to use hybrid models that combine bare metal and virtual nodes for performance-critical tasks using virtual machines and containers for flexible workloads and on demand. It is viable the migration of HPC services from an organization to another using virtual machines and containers that have an elevate maturity grade and it can be reused efficiently, ensuring the HPC configurations, base applications and data that aren't confined to a specific organization data center and also can be reused of the HPC solutions for thirds. This architecture has permitted to deploy different HPC environments on cloud using open source management tools. It contains levels used as reference to manage all HPC environments on cloud from the master node configuration. This contain all the tools necessary for the deployments.

The virtualization contributes to the reproduce and the scaling of the infrastructures most easily taking as reference an ecosystem to deploy HPC infrastructures and their relations mainly in telecommunications small and medium size organizations including knowledge areas of these field. HPC facilities are heavily used for solving computational problems that can't be done using conventional computers due these demands to the huge amount of CPU power, memory, network and disk space requirements. In order to facilitate intensive computations, which include diverse and emerging computing resources, aims to support both advanced computing and data-intensive research.

HPC architecture play a critical role in the complex telecommunications systems, providing a valuable insight in this field and consolidate the strategic decisions related with HPC system management and sustainability. The migration of HPC to the cloud requires to focus the organizations jobs on the efficient based on HPC systemic design approaches from organizational point of view to use more efficiently all computational capacity installed, regardless of the physical location of the hardware resources with the purpose of minimizing communication latencies, improving the fault tolerance and data replication. The potentialities of the organizations HPC infrastructures constitute a platform to solve complex calculations belonging to their daily processes to obtain the efficiency of their own organizational jobs [56].

Organizations that succeed in building their own HPC computational infrastructures can solve intrinsic processes using their own computational capabilities installed and sharing HPC services to thirds at the edge. In this case, cybersecurity is strengthened: security, privacy and mitigation of the dangers of theft of information that may be confidential for the organization [57]. There is a noticeable improvement in performance and reduction of communication latency, using network with less time consuming data transfer. Both factors have a direct impact on the ROI. It contributes to obtaining the final results associated with each one organization’s service more rapidly or the more efficient reuse the available computational capacity. That is an alternative that whatever can help to accelerate organization’s process. The configuration of HPC infrastructures based on cloud computing offers better cost-effectiveness and flexibility for the execution of workloads in the organizations. The lightweight design provided by container technologies has made it possible to make adjustments to their capacity with minimal overhead, thus improving performance with a significant reduction in costs, while meeting other requirements such as reliability, automation and security.

The evolution of HPC to cloud operating model from the organizations point of view provides a common structured for use in any geographic location [58]. This strategy ensures that organizations or users can access to the services and also offer to the systems’ client the possibility to pay to access a virtual pool of and also provide shared resources: compute, storage and networking services, located on the configured HPC infrastructure. The architecture proposed has been improved gradually based on the experiences of successive process of design and deployment in a time period greater than 25 years in which it has occurred the migration from physical to virtual environment. This technological evolution has led to solve problems every time more complex on the part of the organizations on premise.

The Figure 3 contains four of the clusters HPC designed and built by authors of this paper in different stages specifically in small and medium-size universities organizations of the telecommunication sector. These designing and deployment processes generate a systematic feedback to the architectonic design, considering that it is a dynamic process with continue changes in correspondence with the evolution of hardware and software industries. The technical characteristics of the clusters HPC built and configured are briefly described below, considering the technological resources available and the cost at the date indicated.

The Figure 3(a) was building in 1997-2003 based in one master node and 128 computing nodes on motherboards with double processor in 16 nodes Intel Xeon P3 (32 CPUs) and 112 nodes Intel Xeon P4 (224 CPUs). All nodes were chassis rackmount 1U. Interconnection devices: switch and NIC cards Gigabit/Ethernet. Cabins 41u and 42u. The Rpeak was 1 Teraflops. The cost of this HPC infrastructure was 192 636.99 USD.

The Figure 3(b) show a tier pack cluster built in 2010 based in one master node and 8 computing nodes with motherboard supporting Intel core I7 920 LGA 1366/Socket a CPU Quad Core with 16Gb of RAM, 2 GPU cards Geforce GTX 260 per node, Rpeak 4.3 Teraflops. The cost of the hardware infrastructure was 10000 USD. The Figure 3(c) was building in 2015 based in a master node and 4 computing nodes tier pack with motherboard supporting Intel core I5 quad core 8 Gb of RAM per node, the cost of the hardware was 6000 USD. The Figure 3(d) was building in 2023 reusing a hardware infrastructure donated from the business sector to the Technological University of Havana. The HPC infrastructure developed has 2 enclosure PowerEdge m1000e dell. Each enclosure contains 16 blade servers PowerEdge M610SMP with Intel Xeon CPU L5520-2.27 Ghz, core speed 2.26 Ghz, bus speed 5.86 GT/s cache L2: 4×256 kb, cache L3: 8 MB 4 cores/CPU, 32 blade servers in total.

The master node has 96 GB of RAM memory ECC DDR3 1067 MHz. Each computing node (blade server) has 2 NICs Gbe; the transfer rate is 1 gbe. The computes nodes of the cluster have similar processors with 64 GB of RAM as average and the same NICs. Have been incorporated to the cluster others disperses nodes of the organizations with processors Intel Core I5, with 8 GB and 4 GB of RAM memory. For the HPC infrastructure deployment on cloud the master node must have if it is possible a high-end processor, a minimum of 96 Gb of RAM and interconnection devices that guarantee a high data transfer rate between nodes, at least of 1 Gbps.

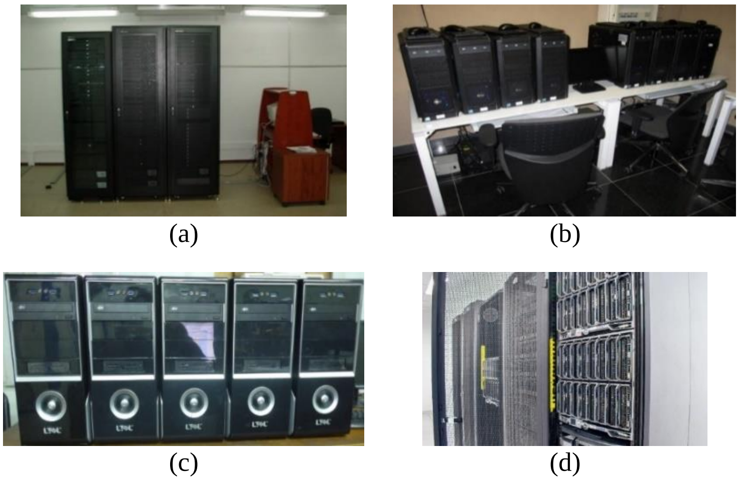


Figure 3. HPC physical infrastructures built with the participation of authors of this paper; HPC cluster belonging to: (a) Parallel Massive Research Group, (b) Integrated Technologies Research Center (CUJAE), (c) Telecommunications and Telematics Department (CUJAE), and (d) HPC Cluster Telecommunications HPC Laboratory in Center for Telecommunications and Informatics Studies (CUJAE)

In every case showed in the previously the organizations performance demands have been satisfied with the Rpeak achieved. The HPC infrastructures design contain the most important components involved: computer architectures, management tools, algorithm design, development tools for the designing, development and deployment of infrastructures and services using these HPC infrastructures to achieve landscape solutions for the accelerated execution. This type of design is very useful for the resolution of heterogeneous problems correspondent to each one of the organizations. The migration of high-performance computing from bare metal to virtual environments constitute a substantial benefit for the telecommunications organizations because it permits a better structuring of the potentiality and modularity of their infrastructures to accelerate heterogeneous complex process of this field reusing the precedent experiences.

The Table 1 contains 4 of the HPC infrastructures designed and deployment by work teams led or in which the authors have participated. The performance achieved has been correlated with the financial resources available in the organizations listed. The architecture presented in the research has been the reference model for the design, implementation, deployment and validation during a large time period. It has been enhanced gradually, from the integration of emerging hardware and software components. The cost of building a teraflop has decreased significantly since the 1980s transforming HPC to a tool accessible to small and medium organizations as startups, researchers and enterprises. This trend is expected to continue, further enabling innovations in AI, scientific research, and data analytics. For detailed cost breakdowns of specific HPC configurations.

Table 1. Cluster HPC designing and developed for authors

Cluster HPC	Nodes	Procesors	Devices Data transfer	Rpeak	Date	Cost
Figure 3(a)	128	32 Intel P3 Xeon/ 224 Intel P4 Xeon	1gb/E	1 Teraflop	1998-2003	192636.99 USD
Figure 3(b)	8	8 Cpu Quad Core Intel Core I7 920 Lga 16 Gpu Gtx 260 (2 per node)	1gb/E	4.3 Teraflops	2010	10000 USD
Figure 3(c)	5	5 Cpu Intel Core I5-9600k @37.73 Gflops	1gb/E	168.65 Gglops	2015	6000 USD
Figure 3(d)	32	Intel Xeon Cpu L5520-2.27 Ghz	1gb/E	2.7 Teraflop	2022	Donation to the CUJAE from Telecomm Business Organization

The architecture of the infrastructures is independent of its dimensions and the demands of the organizations. Other clusters have been built for these research team in telecommunication business organizations, but these don’t have been showed because is confidential information of the organizations. The authors consider these computing devices are enough to show the viability of the architecture in the described conditions. It is extensible from dissimilar organizations that need to increase continually (scaling) their performance from the emergent hardware components available in the market in correspondence with the organizations necessities.

Between the middleware configured and adjusted systematically are Platform Cluster Manager-Dell Edition, HPC Intel Orchestrator, xCAT Extreme Cloud Administration Toolkit, Warewulf, Openstack HPC, Apache CloudStack, Oscar, OpenNebula, among others as was described in the monography cluster HPC management tools: configuration, deployment, job schedulers, monitoring and supervision, wrote for authors of this paper [59]. In this level Parallel applications development is used mainly compositional models that are based on explicitly ways in which programs can be composed. Nichols *et al.* [60] the same authors of this paper made an analysis of the solutions methods and required knowledge to designing and implementing of parallel applications for paradigms of share and distributes memory.

In this context the most pertinent programming language is C++ because in the telecommunications field is necessary the programing at hardware level, system level and the application level to achieve the acceleration as rapidly as possible. Duany *et al.* [61] the same authors of this paper designed and implemented a module of capturing multiples IP video streams using the parallel technics. This process is very important to accelerate the video streams processing on HPC infrastructure as platform to diverse applications of video quality assessment and enhancement of video quality using deep learning whose knowledge have been also developed for this authors. The compositional way is adequate to develop parallel applications using parallel libraries based on the shared memory paradigms of OpenMP and distributed memory paradigms using message passing interface (MPI).

There are three major categories of AI algorithms: supervised learning, unsupervised learning, reinforce learning. In all cases the AI model contains a set of selected algorithms and the data used for their training to obtain most accurate predictions. These algorithms are by nature more complex, because the process of model training is very dense and it is usually being necessary to every learning process. The data are acquired and labeled to be used by the AI algorithms and get output data. Vlaović *et al.* [62] the same authors of this paper made a comparative analysis of the objective methods for video quality assessment using deep learning. The main ideas of this methods are based on the deep learning, as parallelizable solution to realize the evaluation in real time of the video quality. This is a very useful solution in the previous stage of the processing and enhancement the content of the video stream.

Leszczuk *et al.* [63] some of same authors of this paper development the paper “object recognition from video sequences in real time using deep learning technics”, In this research was proposed an implementation based on Yolo for object recognition in video sequences, which uses the library for computer vision OpenCV. The experiments were realized in two computing node and the application can be extended to many parallel nodes.

The algorithms usually need massive amounts of data that capture the full range of incoming data. To train models that “learn faster” should implement a HPC infrastructure scalable for the solutions of heterogeneous problems of telecommunications field. Therefore, the data scientists in this field always need a platform that are not only powerful processing huge amounts of “raw” data, but need to not have worry too much about the programming language these are using offering an integrated experience.

The architecture has been conceived as result of successive analysis processes and their practical validations in different research projects during a long period of time. The main motivation has been the building of heterogeneous systems integrating CPUs, GPUs, and specialized accelerators as FPGAs, to optimize the performance of the workloads. This heterogeneity addresses energy efficiency and computational demands, particularity in all systems to deploy. The architecture has been validated mainly in the telecommunication sector, but it isn’t exclusive of this sector.

4. CONCLUSION

The architecture presented have been very useful as reference model to build successively HPC infrastructures, in small and medium size organization of the telecommunications sector. The approach is continually actualized from the practical experience and emergent hardware/software components. This type of infrastructures contributes to solve complex computationally problems on own HPC infrastructure of the organizations. In this work have been discussed relevant aspects that must take in consideration from the design stage to the full deployments in the organizations. The architecture is key to achieve the acceleration of the organizations improving their efficiency and efficacy, therefore the productive and business indicators. The architecture describes step by step the ecosystem with six linked layer that contribute to integrate knowledge areas that have a rapid and concurrent development and add a wide diversity of capability to the infrastructures. Each layer is responsible for providing specific functionalities for the design and deployment of the components step by step in the route to deploy the well-organized HPC infrastructure. This reference model combines the most recent advances of hardware and software industries, virtualization, management tools, design and implementation of parallel applications technics. All these oriented to develop of own solutions based on advanced research like the AI, big data, IoT in function of specific applications in telecommunications field that need to be accelerated for their functioning 24/7/365. The architecture also includes, from the starting point, the analysis of the cost/performance relation to acquire the hardware components in the market with the purpose to achieve the optimization of their budgets and the impact of ROI for the organization. The hardware infrastructure design also demands a previous analysis of the computational complexity of the problems to estimate the financial resources required to acquire the hardware components commercialized in the international market for the infrastructure building. The consideration of these key elements are important to evaluate the period required for the amortization of the investment, the utilities generation and the achievement of the organizational results. The utilities can be directly linked to the process of actualization and systematically scaling of the infrastructure performances in the proportion that it is necessary. To achieve optimization of the problems is essential the compatibility between the demands of the high complexity problems in correspondence with the potentialities of the available hardware components available in the market. The adoption of HPC on cloud computing by the organizations should contribute to reproduce or to scale most easily the HPC to relative low cost and reducing the time required for the deployment to solve every time more complex challenges from point of view of computational complexity. The prioritized problems of the organizations must be considered for the significant reduction of their execution times based on the systematic optimizing of HPC infrastructure utilization. The business organizations could monetize offering HPC infrastructure as service on demand to thirds and also configuring new and extended different services of designing and deployment the HPC infrastructures to other organizations that have understood the importance of these infrastructures for improve the functioning of their organizations. The use of virtual HPC infrastructure plays an important role in supporting the computational demands of the telecommunications organizations because it can contribute to accelerate dissimilar process. It permits to guarantee the migration considering on different infrastructures and their benefits and architecture facilitating reasonably the compatibility with different environments with relatively low cost and adding new functionalities on demand to the organizations, considering the possibility of vertical and horizontal scaling to bring HPC as service.

ACKNOWLEDGMENTS

The authors would like to thank the Technological University of Habana and its High Performance Computing Laboratory for Telecommunications (TeleHPC) belonging to the Telematics Research Group in the Telecommunications and Informatics Studies Center (CETI) of the Telecommunications and Electronics Schools in collaboration, through a research grant or contract, with the Telecommunications and Informatics Enterprise (ETI) in BIOCUBAFARMA, and the Centro de Investigación en Computación, Instituto Politécnico Nacional, México.

FUNDING INFORMATION

The funding for this research have been provided for the Telecommunications and Informatics Studies Center of the Telecommunications and Electronics School in Technological University of Havana, Havana, Cuba and the Information technologies Enterprise (ETI in Spanish) belonging to the BioCubaFarma Group.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Omar Antonio Hernández Duany	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Caridad Anías Calderón	✓	✓		✓	✓	✓	✓	✓		✓		✓		
Roberto Sepúlveda Lima		✓			✓	✓				✓		✓		
Fernando de la Nuez García		✓	✓	✓		✓		✓	✓		✓			
Cornelio Yáñez-Márquez	✓	✓			✓	✓				✓	✓	✓		

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest between the authors.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This study was approved following the organization's police by the Center for Telecommunications and Informatics Studies belonging to the Telecommunications and Electronic School in the Technological University of Havana "José Antonio Echeverría" (CUJAE) in cooperation with the information Technologies Enterprise (ETI) belonging to BioCubaFarma Group. The results obtained by the research have been validated in the telecommunications HPC Laboratory in HPC Cluster (CUJAE).

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study. This study is a transversal architecture as reference model to build HPC infrastructures in different organizations.

REFERENCES

[1] D. Reed, D. Gannon, and J. Dongarra, "Reinventing high performance computing: Challenges and opportunities," *Prepr. arXiv.2203.02544*, Mar. 2022.

[2] G. Li, J. Woo, and S. B. Lim, "HPC cloud architecture to reduce HPC workflow complexity in Containerized environments," *Applied Sciences*, vol. 11, no. 3, Jan. 2021, doi: 10.3390/app11030923.

[3] R. Hoerl, W. Jensen, and J. de Mast, "Understanding and addressing complexity in problem solving," *Quality Engineering*, vol. 33, no. 4, pp. 612–626, Oct. 2021, doi: 10.1080/08982112.2021.1952230.

[4] M. Merz and H. Sorgner, "Organizational complexity in big science: strategies and practices," *Synthese*, vol. 200, no. 3, May 2022, doi: 10.1007/s11229-022-03649-3.

[5] T. Gamblin and D. S. Katz, "Overcoming challenges to continuous integration in HPC," *Computing in Science & Engineering*, vol. 24, no. 6, pp. 54–59, Nov. 2022, doi: 10.1109/MCSE.2023.3263458.

[6] "Perspectives from the Global Telecom Outlook 2023–2027," *PwC*, 2023. https://www.pwc.com/gx/en/industries/tmt/assets/pwc-perspectives-from-the-global-telecom-outlook-2024-2028.pdf (accessed May 05, 2024).

[7] F. Nadeem, "Evaluating and ranking cloud IaaS, PaaS and SaaS models based on functional and non-functional key performance indicators," *IEEE Access*, vol. 10, pp. 63245–63257, 2022, doi: 10.1109/ACCESS.2022.3182688.

[8] "High performance computing market size, share & trends analysis report by component (server, storage, networking devices), by deployment (on-premise, cloud), by end-use, by region, and segment forecast, 2023-2030," *Market Analysis Report*, 2023. https://www.grandviewresearch.com/industry-analysis/high-performance-computing-market (accessed May 05, 2024).

[9] "Perspectives from the Global telecom Outlook 2024-2028," *PwC*, 2025. https://www.pwc.com/gx/en/industries/tmt/assets/pwc-perspectives-from-the-global-telecom-outlook-2024-2028.pdf (accessed May 05, 2024).

[10] U. Beyer and O. Ullrich, "Organizational complexity as a contributing factor to underperformance," *Businesses*, vol. 2, no. 1, pp. 82–96, Mar. 2022, doi: 10.3390/businesses2010005.

[11] "2024 telecommunications industry outlook," *deloitte*, 2024. https://www.deloitte.com/us/en/Industries/tmt/articles/telecommunications-industry-outlook.html (accessed May 11, 2024).

[12] J. Leverton, "Five key strategies for harnessing high performance computing," *Forbes*, 2023. https://www.forbes.com/councils/forbesbusinesscouncil/2023/10/05/five-key-strategies-for-harnessing-high-performance-computing (accessed May 05, 2024).

[13] H. Howard, "HPC revolutionizing telecom: Unveiling five transformative benefits," *FS*, 2024. https://www.fs.com/uk/blog/hpc-revolutionizing-telecom-unveiling-five-transformative-benefits-13604.html (accessed Apr. 10, 2024).

[14] A. H. L. Porto, M. Coelho, H. M. G. A. Rocha, C. Osthoff, K. Ocaña, and D. O. Cardoso, "Assuming the best: Towards a reliable protocol for resource usage prediction for high-performance computing based on machine learning," *Future Generation Computer Systems*, vol. 175, Feb. 2026, doi: 10.1016/j.future.2025.108070.

[15] M. Taufer *et al.*, "HPC and cloud convergence Beyond technical boundaries: Strategies for economic sustainability, standardization, and data accessibility," *Computer*, vol. 57, no. 6, pp. 128–136, Jun. 2024, doi: 10.1109/MC.2024.3387013.

[16] J. Chou and W.-C. Chung, "Cloud computing and high performance computing (HPC) advances for next generation internet," *Future Internet*, vol. 16, no. 12, Dec. 2024, doi: 10.3390/fi16120465.

[17] R. Purohit, K. R. Chowdhary, and S. D. Purohit, "On the design and analysis of parallel and distributed algorithms," *Prepr. arXiv.2311.05857*, Nov. 2023.

[18] E. Commission, “Transnational access programme for a pan-european network of hpc research infrastructures and laboratories for scientic computing,” 2022.

[19] Y. Hu, Y. Pan, M. Yu, and P. Chen, “Navigating digital transformation and knowledge structures: insights for small and medium-sized enterprises,” *Journal of the Knowledge Economy*, vol. 15, no. 4, pp. 16311–16344, Jan. 2024, doi: 10.1007/s13132-024-01754-x.

[20] “Improving the roi for enterprise HPC and AI with hpe end-to-end clusters,” *Hewlett Packard Enterprise*, 2025. https://www.cabotpartners.com/wp-content/uploads/2024/03/IMPROVING-THE-ROI-FOR-ENTERPRISE-HPC-AND-AI-WITH-HPE-END-TO-END-CLUSTERS.pdf (accessed Apr. 07, 2024).

[21] C. Schulz, “How enterprises can increase ROI in their HPC environment,” *Hewlett Packard Enterprise*, 2024. https://www.cio.com/article/189190/how-enterprises-can-increase-roi-in-their-hpc-environment.html (accessed Feb. 09, 2025).

[22] U.-U. Haus, S. Narasimhamurthy, M. S. Perez, D. Pleiter, and A. Wierse, “Federated HPC, cloud and data infrastructures,” *European Technology Platform for HPC*, 2022.

[23] Y. Wang, “Artificial-Intelligence integrated circuits: Comparison of GPU, FPGA and ASIC,” *Applied and Computational Engineering*, vol. 4, no. 1, pp. 99–104, May 2023, doi: 10.54254/2755-2721/4/20230358.

[24] M. Cengiz, M. Forshaw, A. Atapour-Abarghouei, and A. S. McGough, “Predicting the performance of a computing system with Deep networks,” in *Proceedings of the 2023 ACM/SPEC International Conference on Performance Engineering*, Apr. 2023, pp. 91–98, doi: 10.1145/3578244.3583731.

[25] P.-L. Bene, “Supercomputing market trends,” *Luxinnovation Market Intelligence*, 2023. https://luxinnovation.lu/resources/supercomputing-market-trends (accessed Apr. 15, 2024).

[26] N. Jouppi *et al.*, “TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,” in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, Jun. 2023, pp. 1–14, doi: 10.1145/3579371.3589350.

[27] T. Lehman *et al.*, “Data transfer and network services management for domain science workflows,” *Prepr. arXiv.2203.08280*, Mar. 2022.

[28] S. Denisov, K. Volovich, and A. Zatsarinny, “Providing high-speed data access for parallel computing in the HPC cluster,” in *INTELS’22*, Jul. 2023, p. 54, doi: 10.3390/engproc2023033054.

[29] G. Vitali, A. Scionti, P. Viviani, C. Vercellino, and O. Terzo, “Dynamic job allocation on federated Cloud-HPC environments,” in *Complex, Intelligent and Software Intensive Systems (CISIS 2022)*, 2022, pp. 71–82, doi: 10.1007/978-3-031-08812-4_8.

[30] “The list of T0P 500,” 2024. https://www.top500.org (accessed May 01, 2025).

[31] C. S. Ahmad, A. S. B. Mahomed, and H. Hashim, “Cloud computing adoption in SMEs: Exploring IaaS, PaaS and SaaS through a bibliometric study,” *International Journal of Academic Research in Business and Social Sciences*, vol. 15, no. 1, Jan. 2025, doi: 10.6007/IJARBSS/v15-i1/24452.

[32] A. Abdurahman, A. Hossain, K. A. Brown, K. Yoshii, and K. Ahmed, “Scalable HPC job scheduling and resource management in SST,” in *2024 Winter Simulation Conference (WSC)*, Dec. 2024, pp. 2226–2237, doi: 10.1109/WSC63780.2024.10838714.

[33] G. Schryen, “Speedup and efficiency of computational parallelization: A unifying approach and asymptotic analysis,” *Prepr. arXiv.2212.11223*, Nov. 2023.

[34] “2025 high performance computing market report: key trends, market size & future projections,” *The Bussines Research Company*, 2025. https://www.thebusinessresearchcompany.com/market-insights/high-performance-computing-market-insights-2025 (accessed Aug. 08, 2025).

[35] J. Tarraga-Moreno, J. Escudero-Sahuquillo, P. J. Garcia, and F. J. Quiles, “Understanding intra-node communication in HPC systems and Datacenters,” *Prepr. arXiv.2502.20965*, Feb. 2025.

[36] J. Baumgartner, C. Lillo, and S. Rumley, “Performance losses with virtualization: Comparing bare metal to VMs and containers,” in *High Performance Computing*, 2023, pp. 107–120, doi: 10.1007/978-3-031-40843-4_9.

[37] S. Surbhi, “Evolution of cloud computing: trends and future,” *Networks Kings*, 2023. https://web.nwkings.com/evolution-of-cloud-computing (accessed Feb. 01, 2024).

[38] T. Patki *et al.*, “Fluxion: A scalable graph-based resource model for HPC scheduling challenges,” in *Proceedings of the SC ’23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, Nov. 2023, pp. 2077–2088, doi: 10.1145/3624062.3624286.

[39] N. A. Al Etawi, “A comparison between cluster, grid, and cloud computing,” *International Journal of Computer Applications*, vol. 179, no. 32, pp. 37–42, 2018.

[40] P. (Guha) Ghosh, “Cloud computing trends in 2025,” *Dataversity*, 2025. https://www.dataversity.net/cloud-computing-trends-in-2025/ (accessed Aug. 08, 2025).

[41] P. Bonderson, “Transitioning from on-premise computing to cloud computing,” *UPPSALA UNIVERSITET*, 2023. https://uu.diva-portal.org/smash/get/diva2:1752929/FULLTEXT01.pdf

[42] M. K. Patra, B. Sahoo, and A. K. Turuk, “Containerization in cloud computing for OS-level virtualization,” in *Advances in Cyber Security and Intelligent Analytics*, Boca Raton: CRC Press, 2022, pp. 261–276, doi: 10.1201/9781003269144-16.

[43] V. P. Oleksiuk and O. R. Oleksiuk, “The practice of developing the academic cloud using the Proxmox VE platform,” *Educational Technology Quarterly*, vol. 2021, no. 4, pp. 605–616, Dec. 2021, doi: 10.55056/etq.36.

[44] J. Ford, D. Arnold, and J. Saniie, “Environment Provisioning and Management for Cybersecurity Education,” in *2023 IEEE International Conference on Electro Information Technology (eIT)*, May 2023, pp. 368–372, doi: 10.1109/eIT57321.2023.10187365.

[45] D. K. Panda, H. Subramoni, M. Abduljabbar, A. Shafi, N. Alnaasan, and S. Xu, “Designing converged middleware for HPC, AI, and big data: Challenges and opportunities,” in *Artificial Intelligence and High Performance Computing in the Cloud*, 2024, pp. 40–63, doi: 10.1007/978-3-031-78698-3_4.

[46] N. Brooks, C. Vance, and D. Ames, “Cloud computing: A review of evolution, challenges, and emerging trends,” *Journal of Computer Science and Software Applications*, vol. 5, no. 4, 2025.

[47] S. Santillan and C. L. Abad, “An analysis of HPC and edge architectures in the cloud,” *Prepr. arXiv.2508.01494*, Aug. 2025.

[48] P. H. B. Patel and P. N. Kansara, “Cloud computing deployment models: A comparative study,” *International Journal of Innovative Research in Computer Science & Technology*, vol. 9, no. 2, pp. 45–50, Mar. 2021, doi: 10.21276/ijircst.2021.9.2.8.

[49] A. Rajković and M. Antić, “An environment for orchestrating containers on a local ProxMox server,” in *2024 Zooming Innovation in Consumer Technologies Conference (ZINC)*, May 2024, pp. 179–181, doi: 10.1109/ZINC61849.2024.10579438.

[50] F. Almeida and E. Okon, “Assessing the impact of high-performance computing on digital transformation: benefits, challenges, and size-dependent differences,” *The Journal of Supercomputing*, vol. 81, no. 6, Apr. 2025, doi: 10.1007/s11227-025-07281-z.

[51] Z. Zhang, “Lower bound of computational complexity of knapsack problems,” *AIMS Mathematics*, vol. 10, no. 5, pp. 11918–11938,

2025, doi: 10.3934/math.2025538.
[52] L. Wu, "Research on the development and application of parallel programming technology in heterogeneous systems," *Journal of Physics: Conference Series*, vol. 2173, no. 1, Jan. 2022, doi: 10.1088/1742-6596/2173/1/012042.
[53] H. N. Alshareef, "Current development, challenges, and future trends in cloud computing: A survey," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140337.
[54] A. K. Kushwaha, "The evolution and impact of cloud computing on modern enterprises," *International Journal of Innovative Research in Technology*, vol. 11, no. 8, pp. 803–805, 2025.
[55] K. F. Pilz, J. Sanders, R. Rahman, and L. Heim, "Trends in AI supercomputers," *Prepr. arXiv.2504.16026*, Apr. 2025.
[56] J. Lange *et al.*, "Evaluating the cloud for capability class leadership workloads," Oak Ridge, TN (United States), Sep. 2023, doi: 10.2172/2000306.
[57] S. Nikolic, L. Filipovic, T. Ilijas, and M. Vukotic, "FIT4HPC?—Accelerating digital transformation by supercomputing opportunities," *The Journal of Supercomputing*, vol. 81, no. 9, Jun. 2025, doi: 10.1007/s11227-025-07559-2.
[58] S. Rangasamy, "Comparative analysis of cloud computing deployment models," *International Journal For Multidisciplinary Research*, vol. 7, no. 4, Jul. 2025, doi: 10.36948/ijfmr.2025.v07i04.51359.
[59] A. Elola and J. R. Wilson, "Cluster management and policy learning: the value of strategic intelligence," *European Planning Studies*, pp. 1–19, Jul. 2025, doi: 10.1080/09654313.2025.2530027.
[60] D. Nichols, A. Marathe, H. Menon, T. Gamblin, and A. Bhatele, "HPC-coder: Modeling parallel programs using large language models," in *ISC High Performance 2024 Research Paper Proceedings (39th International Conference)*, May 2024, pp. 1–12, doi: 10.23919/ISC.2024.10528929.
[61] Omar Antonio Hernández Duany, "Scalable parallel module for capturing multiple video streams in real time," *Journal of Information Systems Engineering and Management*, vol. 10, no. 38s, pp. 146–154, Apr. 2025, doi: 10.52783/jisem.v10i38s.6835.
[62] J. Vlaović, D. Žagar, S. Rimac-Drlje, and M. Vranješ, "Evaluation of objective video quality assessment methods on video sequences with different spatial and temporal activity encoded at different spatial resolutions," *International journal of electrical and computer engineering systems*, vol. 12, no. 1, pp. 1–9, Apr. 2021, doi: 10.32985/ijeces.12.1.1.
[63] M. Leszczuk, L. Janowski, J. Nawała, and A. Boev, "Objective video quality assessment method for object recognition tasks," *Electronics*, vol. 13, no. 9, May 2024, doi: 10.3390/electronics13091750.

BIOGRAPHIES OF AUTHORS

Omar Antonio Hernández Duany Associated Professor and Researcher. Master in Sciences. Bachelor in Communications Specials Systems, Bachelor in Computer Sciences from University of Havana, Leader of Telecommunication HPC Laboratory in the Center for Telecommunications and Informatics Studies in the Telecommunication and Electronics Engineering School at the Technological University of Habana, "José Antonio Echeverría" Research interest: HPC, AI, Parallel and distributed algorithms design, computer programmer, computational vision, digital video processing, and Big Data. He can be contacted at email: omar.hd@tele.cujae.edu.cu.

Caridad Anías Calderón Full Professor, Ph.D., Master in Sciences, Telecommunications Engineer from the Technological University of Havana "José Antonio Echeverría" CUJAE, Director of Center of Studies of Telecommunications and Informatics, Telecommunication and Electronics Engineering School of Technological University of Habana, "José Antonio Echeverría". Leader of the Telematics Research Group. President of the Nacional Commission of carrier of Telecommunications and Electronics Engineering, President of Informatics Event for more than a decade. Research interest: telematics, cloud computing, computing networks. She can be contacted at email: cacha@tesla.cujae.edu.cu.

Roberto Sepúlveda Lima Full Professor, Ph.D., Master in Sciences, Electricity Engineer from the Technological, National Commission of Scientific Grades, ex-Director of Studies of Systems Engineering Center, Dean of Computing Sciences Engineering School of Technological University of Habana, "José Antonio Echeverría". Research interest: informatics systems, AI, software engineering, data communications and cryptology. National commission of informatics engineering carrier, supervisor of master and Ph.D. thesis in computer sciences and affine. He can be contacted at email: rsepulvedalima@gmail.com.

Fernando de la Nuez García Telecommunications and Electronics Engineer from the Technological University of Havana "José Antonio Echeverría" CUJAE, Summa Cum laude, Research interest: HPC, AI, parallel and distributed design, and systems administration. He can be contacted at email: fernandela@tele.cujae.edu.cu.

Cornelio Yañez-Márquez Full Professor and Researcher Ph.D. Master in Sciences in Computing Engineneer. Bachelor degree in Physics and Mathematics. Chief of the Intelligent Computing Laboratory at CIC. IPN. Member of Researcher Nacional System. Intelligence Computating Laboratory. Prize "Lazaro Cárdenas". He is also a receipt of the National Researcher System (SNI) Research interests: associative memories, neural networks, mathematics morphology and software engineneering. He established the Alpha-Beta research group. Research interest: associative memories, neural networks, machine learning, and software engineering. He can be contacted at email: cyanez@cic.ipn.mx.