Grid Computing Research---Road to economic and scientific progress for Pakistan

Arshad Ali
National University of Sciences & Technology, Rawalpindi, Pakistan

Abstract
We present a report on Grid activities in Pakistan over the last three years and conclude that there is a significant technical and economic activity due to participation in Grid research and development. We started collaboration with participation in CMS software development group at CERN and Caltech in 2001. This has led to the setup for CMS production and LCG Grid deployment in Pakistan. Our research group had been participating actively in the development work of PPDG and OSG, and now working in close collaboration with Caltech to create next generation infrastructure for data intensive science under the Interactive Grid Enabled Environment (IGAE) project with a broader context to Ultralight collaboration. The collaboration on Grid Monitoring and Digital divide activities with Caltech in MonaLIsa developments and with SLAC on Maggie has not only helped to train our manpower but it also helped to improve the infrastructure, bandwidth and computing capabilities in Pakistan. The dedicated team of researchers (faculty/students) are playing their role not only in international research activities but also supporting Industry and advising Govt to address the digital divide issues and internet/information access to its population. Discussions have already been started about establishment of National Grid in Pakistan and the means to mobilize resources in basic Sciences, Govt. departments and business community.

INTRODUCTION

Research and development have been important agents for the growth of any organization, community and country. Currently one of the latest and most innovative technologies is Grid Computing, which has become a high-end research platform not only in computing but also in all areas of sciences and technology.

The idea of Grid Computing has provided an economical solution for the applications and experiments that need high computation power and intensive resources. Different research communities in the world have adopted Grid Computing due to its flexibility, economy of scale and other attractive features. Grid Computing is playing a great role in sharing and globalization of information and technology. In this paper we will briefly focus on how Grid computing has boosted various diverse fields such as high-end scientific simulation to database systems. We will also discuss how Grid has helped to set up high speed and high performance cross continental internet and secure networks and how Grid Research could enhance cluster computing, Databases, bridging the Digital Divide, enhancing education and knowledge dissemination and Next Generation solutions for the complex problems.

In this paper, we will undertake to support our premise that Grid Computing [1] is a very suitable and practicable platform for the solution of a variety of problems challenging the Asian world today.

GRID AND DATABASES

Data are the most important asset of any organization or country. Storage of data into a secure repository that can easily be accessed by authentic users and can effortlessly be shared with other organizations and communities is the main goal of the database research community. The data should also be recoverable and available in case the data repository fails. The data production rate is much higher today as compared to the past rate and it will continue to increase with time. Examples of such high production sources are High Energy Physics (HEP) [2], Weather forecast, Earth Sciences etc. HEP experiment series at CERN will produce peta-bytes of data a year and storage of such a huge data is a challenge to the research community. To fulfil the above mentioned needs data should be stored and replicated in distributed repositories. A platform that can serve the very purpose is Grid Computing. PhedEX, a replica management system used by Compact Muon Solenoid experiment at CERN Geneva and Condor-STORK are some examples of projects focusing on replication and storage of large amount of data on Grid.

Database research communities have found the solution to most of their problems in an integrated framework of Grid computing. Grid computing in databases field has eased the job for replicating data, utilizing spare storage space, securing data and has made possible the sharing of data amongst geographically distant areas. Oracle Corporation has released its product Oracle 10g, the first database system designed for grid computing. Another notable project is Pool Of persistent Objects for LHC (POOL), which is a persistency framework for physics applications at LHC. It is a part of Large Computing Grid (LCG), a computing infrastructure for LHC. Main objective of this project is to allow the multi-petabyte of experiment data and associated metadata to be stored in distributed and grid-enabled fashion. Grid Database Integration and warehousing is another project followed by NUST Institute of Information Technology in collaboration with Caltech USA, aiming at replication of data on heterogeneous Grid nodes to store large amount of data produced by Compact Muon Solenoid (CMS) at CERN , Switzerland.

OPERATING SYSTEMS

Grid computing is a framework for efficient resource sharing. In this context, Grid computing plays somewhat a similar role to an operating system. Keeping in view the best mechanism for resources sharing, Grid computing is contributing significantly in the area of operating systems

research. Grid Operating Systems has been proposed at NIIT for taking advantage of services provided by Grid Computing. GLinux also Gridified Linux with Grid Virtual Machine (GVM) presents the idea of introducing GVM in the normal operating system as a layer. This layer will be responsible for connecting to Grid and becoming part of grid in order to use resources of the grid. Research is under progress in developing operating systems that have special compatibility with Grid and using services of Grid. Examples of such operating systems are CERN Linux, Fermi Linux, SuSe Linux, Fedora etc. CERN has a research group working on their own operating system known as CERN Linux; it has special libraries and development support for Grid Computing. They have recently released Scientific Linux-4.2 specialized for Grid computing and research community. SuSE Linux has also been developed with special support for Grid Computing. SuSE Linux has had special features of Sun Grid Engine since version 8.0. They claim that new SuSE Linux provides powerful grid computing solution to the Linux marketplace. Thus, Grid has given a new direction to Operating System research, by introducing innovative features of Grid Computing into operating systems. Similarly, Fermi Linux is another example of an operating system for Grid researchers. Such change of trend in the area of operating system focused on providing new features for efficient resource management, fast processing and large storage are all advantages of Grid computing.

HIGH-END ANALYSIS AND COMPUTING

High-end analysis frameworks are mostly resource intensive. They need high computational power, large and efficient storage system and integrated framework. Grid computing is serving the scientists by providing an efficient framework for analysis. Grid computing has made it possible to provide high processing speed by using and combining the processing power of different computers spread all around the world thereby creating an infrastructure that is accessible to scientists worldwide.

Researchers in Bioinformatics also need a framework that can solve problems, which they are facing such as sharing of knowledge, sources of data and sharing of large-scale computational resources etc. Different projects on bioinformatics that make use of grid computing have been started. “Neurogrid”, which focuses on distributed analysis of brain, is grid infrastructure in bioinformatics area.

TeraGrid is a working Grid infrastructure for High Performance Computing, which is focusing on experimentation in different areas of sciences. Particle Physics Data Grid is an infrastructure for providing physicists with a framework for physical data analysis. International Virtual Data Grid Laboratory (IVDGL) is another Grid infrastructure for scientists in physical sciences and astronomy.

Cluster Computing is also adopting Grid into its domain and is an idea for Grid of Specialized cluster interconnected by high performance LAN, Metro or Wide Area Network. Grid of Clusters can provide a good balance of flexibility and scalability as compared to loosely coupled Grids and at the same time gives a performance advantage of tightly coupled clusters. These clusters are taking the shape of a system that is highly flexible and has immense computational power.

Astronomy is another application area for Grid in which interesting work is under progress. Storing and sharing huge data generated by space research, processing it, and deducing results from this analysis can be made possible by the use of Grid computing power.

Chemistry has close links with 3D graphics and visualization. To compute the exact geometry of a molecule from raw data obtained from experiments is a highly computation-intensive task, and Grid computing is a very applicable field that is already proving its capabilities for these tasks. Grid computing has wide applications in Healthcare. Hospitals and healthcare centres can be connected through Grid to share the information and resources with one another. Using Grid Computing in this area will speed up the data sharing process and consultancy among hospitals, which in turn will keep all hospitals updated about diseases, medicines and new inventions.

MonALISA (MONitoring Agents using a Large Integrated Services Architecture) is a product of Caltech. This service is being deployed at various sites all over the world which includes CERN, Caltech, GATech, UPB, ASCCm NUST and many others. NUST has been involved in the development of End to End Performance monitoring tool for MonALISA and Automatic Discovery of the Network Topology.

GRID COMPUTING AND NETWORKS

Networks are the backbone for Grids. Evolution of Grid has also defined new dimensions for high-speed networks. Network research has new focus towards high performance network. The network research community is continuously improving the bandwidth, performance and efficiency of networks. Grid computing has caused networks to move in new areas of high performance networks, optical networks and development of networks that span over large geographical areas to serve as high performance backbones. These high-speed networks have eliminated the threat of under usage of a Grid.

The GEANT network, a collaborative project of 26 National Research and Educational Networks (NREN) representing 30 countries, is a running example of such infrastructure in Europe. The GEANT project aimed to develop gigabit speed network called “The GEANT Network”. Its main objectives include gigabit speeds, geographical expansion, global connectivity and a guaranteed service quality. GEANT2 is successor to GEANT. It is a seventh generation pan-European research and educational network. The project objectives are to plan and build multi-gigabit pan-European backbone research Network interconnecting Europe’s national research and educational networks to support advance

projects and users, which need advanced networking requirements.

Ultralight [3] is another effort towards high performance networks. It is a collaboration of physicists and network engineers to provide network services to enable petabyte-scale Grid analysis of globally distributed data. Ultralight aims to develop and deploy a prototype that will broaden existing Grids by promoting the networks as an activity-managed component and integrating current grid based physics production and analysis systems in ATLAS and CMS. UKLight is another project to connect several leading networks in the world to create an international test bed for optical networking.

Grid related network monitoring tools as well as other projects are also helping in bridging the digital divide between developed and developing countries. One such project is PingER, the Ping End-to-end Reporting project, has been measuring Internet connectivity around the world for over ten years, and now monitors over 600 Internet sites in 114 countries. It is a collaborative project of NUST Institute of Information Technology Pakistan, SLAC USA and Fermi National Accelerator Laboratory USA.

The technologies being used and developed by NUST faculty and students include JClarens which is a robust Grid Middleware for data intensive sciences and various other services which can facilitate the scientists for efficient manipulation and use of the data. Our research group had been participating actively in the development work of PPDG and OSG, and now working in close collaboration with Caltech to create next generation infrastructure for data intensive science under the Interactive Grid Enabled Environment (IGAE) project with a broader context to Ultralight collaboration. The collaboration on Grid Monitoring and Digital divide activities with Caltech in MonaLIsa developments and with SLAC on Maggie has helped to train our manpower but it also helped to improve the infrastructure, bandwidth and computing capabilities in Pakistan.

GRID AND DIGITAL DIVIDE

Digital divide can be defined as a social gap leading to differences in access to resources and communication tools among people from different geographical regions. These resources can be internet, computational tools, and communication data rate etc. Main issues that can be considered under the field of information technology are the concentration of supercomputing power, inequality in Internet use and the uneven growth of e-commerce. These issues constitute the main hurdles in the area of information and communication technology (ICT) for developing nations. All around the world, out of 500 best-equipped sites, 51% are in U.S; together with European Union and Japan, they concentrate more than 90% of the global supercomputing. Round about 20 nations are providing more than 90% of the world active internet users.

Grid computing can play a vital role in bridging the digital divide to some extent. Grid Computing is also an effort toward establishing a global village in which resources will be available to all users irrespective of their geographical location. Countries that cannot afford supercomputing can use services provided by grid for resource intensive jobs. It will provide easier access to data and computing intensive for smaller research groups, new sciences and developing countries. Grid will also provide an easier access to global market.

High Energy Physics (HEP) is being prepared for a series of experiments that will generate peta-bytes (10^{15}) of data. These are mega-budget experiments, which developing countries like Pakistan cannot afford. However, Pakistani scientists can still benefit from the experiments and benefit form the huge expenditure of Western countries in this field, through Grid Technologies.

NUST has been playing an active role in bridging the digital divide. NUST has been sending its research staff regularly to research centres around the world to let them gain experience and exposure.

GRID AND KNOWLEDGE SHARING

Grid provides resource sharing; these resources can be of any type for example hardware, software, services etc. Many of the projects have been completed for data sharing on Grid. These projects enable grid users to share data, replicate data on different sites, or do analysis on it etc; in fact, all becomes possible by sharing data. When this data is refined, it takes the shape of information. Information sharing is a very important element in establishing Global Village. Due to its inherent properties, Grid can play an important role in globalization of information. It is evident from an ordinary example of any organization that fast flow of information is the key to success in today's world. The organization that has best flow of information internally and externally performs best in its business. Grid can be used for sharing information between Asian world and Western Countries. Grid Computing can bridge the information gap between East and West. This information sharing will lead to large-scale knowledge. People will be able to get benefits from one another's knowledge and experience. In other words, this information flow will cause the flow of technology and expertise. Institutions all around world can be connected through a Grid infrastructure for high quality educational services. With such an infrastructure, scientists in Eastern world will be able to take good advantage of resources put forward in research by Western countries. We can also create a Grid to share lectures, information, business, news, forecasts and technology expertise.

Pakistan has initiated a project titled PERN (Pakistan Educational Research Network) which aims to connect educational grids and networks for updated knowledge and distant learning. Educational libraries can be connected to universities through grid computing.

GRID PROJECTS AND THEIR ROLES

Different projects for Grid infrastructure are under development and some of them have been completed. We discuss some of the projects below.

The EGEE infrastructure was used by researchers at Institute de Physique du Globe de Paris (IPGP), France to analyze the large earthquake, which struck on 28 March 2005. The analysis was done within 30 hours of its occurrence. An educational Grid, GridPP is collaboration amongst particle physicists and computer scientists from the UK and CERN, who are building a computing Grid for particle physics. GridPP is funded by PPARC as part of its e-Science Program. GridPP has built a working prototype grid across 17 UK institutions. Over the next three years, this will be extended to the equivalent of 10,000 PCs. Particle Physics Data Grid (PPDG) is another effort towards deployment of production grid systems in High Energy Physics. It aims to integrate experiment specific applications, storage resources and provide services such as data management, Job Management, Production Grid Systems, Authentication, authorization etc. Grid3,, which focuses on deployment of a functional Grid, will provide services to LHC institutions as well as non-LHC institutions. It is expected to provide a throughput of 500-900 concurrent job execution with 75% efficiency.

China National Grid (CNGRID) [5] a key project by National High-Tech R&D program is a new generation testbed that integrates high performance computing and transaction capacity. Other projects such as Discovery Net and MyGrid of UK Sciences aim at biological research. Comb-e-Chem is a part of UK Science program focusing on combinatorial chemistry and SkyQuery, an infrastructure for integrating astronomical data are followed by different institutions to address different areas of sciences.

The collaboration on Grid Monitoring and Digital divide activities with Caltech in MonaLIsa developments and with SLAC on Maggie has opened further avenues for research collaboration in information technology with California Institute of Technology (Caltech) USA, University of the West of England (UWE) UK, University of South Brittany France, Beijing Institute of Technology (BIT) China and European Union. NUST students can now register in foreign universities while pursuing research under the joint supervision of NUST faculty and internationally acclaimed scientists. In this regard a pioneering case was initiated by registering an MPhil student at the University of the West of England, UK and research conducted at NUST Institute of Information Technology, Rawalpindi.

CMS Production

In the year 2007 a new particle accelerator, the Large Hadron Collider (LHC), is scheduled to be in operation at CERN. Four High Energy Physics (HEP) experiments will start to produce several Peta bytes of data per year over a life time of 15 to 20 years. Since this amount of data has never been produced before, special efforts concerning data management and data storage are required

NUST has joined CERN in this effort by providing data processing and data storages capabilities through a project named as CMS Production Center. The scope of this project is to build a cluster of computers for this experiment of CERN. This cluster, deployed at NIIT, serves as a Grid node to process the events generated by CMS Detector installed at CERN and is capable of simulating different aspects of CMS using various Production Cycle Steps such as Generation, Simulation, Digitization, Reconstruction and Analysis.

CONCLUSION

Faster and efficient communication among countries is the gateway to development. If we look at the world, we see that countries having strong communication infrastructure can boast higher development rates then those, which have a inadequate connection with outer world. Resource sharing is the only way for developing and underdeveloped countries to achieve Globalization. For Asian countries to develop and connect with the developed world, they have to have a strong Grid infrastructure connecting all of them together for a safe and progressive future for information, knowledge and expertise sharing. This will also bring tremendous benefits and uplift computer science and Information Technology research, which is presently in an extremely abysmal state in south Asian countries.

REFERENCES

[1] “The grid: blueprint for a new computing infrastructure”, I Foster, C Kesselman - San Francisco, 1999 - portal.acm.org

[2] Grid Computing in High-Energy Physics_Paul Avery University of Florida, Department of Physics,Gainesville, FL 32611-8440, U.S.A.February 2, 2003

[3] “JClarens: A Java Framework for Developing and Deploying Web Services for Grid Computing", M. Thomas,C. Steenberg, F. van Lingen, H. Newman, J. Bunn, A. Ali, R. McClatchey, A. Anjum, T. Azim,W. ur Rehman, F. Khan

[4] “Using the GridSim Toolkit for Enabling Grid Computing Education”, M Murshed, R Buyya

[5] “China’s E-Science Knowledge Grid Environment”, H Zhuge - Information and Management, 2003 - ieeexplore.ieee.org