Classification of Weather Indices on Secured Computational Grid

Madhuri Bhavsar[1], Anupam Singh[2], Shrikant Pradhan[1]

[1] *Department of CSE, Institute of Technology, Nirma University*
[2] *Department of Civil Engineering, Institute of Technology,Nirma University*
madhuri.bhavsar@nirmauni.ac.in, anupam.singh@nirmauni.ac.in,
snpradhan@nirmauni.ac.in

Abstract

Grid promises significant advances in support of e-science application, whilst much of the predicted infrastructure is still under development. Deployment of these applications across the grid continues requiring a high level of expertise and a significant amount of efforts, mainly due to overall complexity of the grid. Large scale parallel scientific applications often require computational and data grids to obtain large compute and data resources necessary for execution regardless of heterogeneity, occurrences of faults and complexity.

This research paper reveals the grid enabled computing capabilities for climate impact modeling and change detection facilitating for rainfall prediction. The large dataset utilized in computations are Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite (HOAPS) data. It presents the basic field of air-sea interaction parameters and covers the period July 1987 to December 2005. In order to assess the valuable user friendliness of grid computing to a large community of researchers who need to perform large scale computation, we deployed an application developed for rainfall prediction on NirmaGrid Testbed.

Keyword: *Application on Grid, Computational Grid, , HOAPS Data Analysis, NirmaGrid,*

1. Introduction

Climate or weather is global phenomenon and its study, like developing model or prediction requires observations about various physical quantities spread across the globe. One of the parameter that climatologist uses is rainfall. The rainfall data obtained by surface based rain gauges and radar sites is simply inadequate to build any model. As a result rainfall estimation based on satellite observations of surface temperature, cloud temperature, infrared and microwave radiations have been used to provide the global coverage of rainfall. The global coverage of data leads to vast amount of data to be processed for making meaningful results. Such large amount of data is made available in public domain by Max plank institute under Hamburg Ocean Atmosphere Parameters and fluxex from satellite data (HOAPS)[13][14].

Grid computing enables the virtualization of distributed computing over a network of heterogeneous resources such as processing, network bandwidth and offers a single, unified resource for solving large scale computation and data intensive applications, such as high energy physics, Monte-Carlo applications, Brain activity analysis, Molecular modeling for drug design, communication intensive application (OptIPuter) and many more [1]. Thus grid

computing is not only interesting to business world, but offers less expensive computing power for complex computational models of various engineering and natural science disciplines. The computational power needed for the computation in our application is aggregated by harnessing the unused CPU cycles of desktop computers available on University campus which are partially controlled environment and a high speed LAN unlike internet which is totally autonomous.

There are many e-science applications that can benefit from the grid infrastructure. Large scale Complex calculations if needs faster processing, can be distributed on computational grid. DAME and MyGrid [8] are popular e-science projects developed at UK. DAME (Distributed Aircraft Maintenance Environment) is a System to access and analyze maintenance data from aircraft engines in flight [13]. These diagnostics provide huge amounts of data, which are modeled and analyzed using Grid-based programs. Computational problem defined in this paper processes for grid rainfall anomalies for 19-years from 1987-2005 on HOAPS-3 climate models.

1.1 Computational Grid

The continuous growth in the speed and capacity of workstation with high speed networks in the last few years has led to the emergence of High Performance Computing (HPC) systems. Keeping a low cost approach, high performance and interconnects optimized specifically to enhance both the parallel performance and high availability aspects. The availability of many low-cost and high-performance commodity clusters/Grids within many organizations has provoked the exploration of aggregating distributed resources for solving large scale problems pertaining to various disciplines of engineering and science. This has lead to the emergence of computational Grids for sharing distributed resources. The Grid community is generally focused on aggregation of distributed high-end machines. The LHC (Large Hadron Collider) is an international research project based at CERN which is world's largest highest-energy particle accelerator, responsible for the analysis and management of more than 15 million Gigabytes of data flowing from the LHC every year[2] ,whereas the P2P community (e.g., SETI@Home [3]) is looking into sharing low-end systems, such as PCs connected to the Internet and along with their contents (e.g., exchange music files via Napster and Gnutella networks). Given the number of projects and forums started all over the world in early 2000, it is clear that interest in the research, development, and deployment of Grid and P2P computing technologies, tools, and applications is rapidly growing[4][5]. Already application domains like Monte Carlo simulations and parameter sweep applications (drug design [6], Biomedical, operations research, electronic CAD, and ecological modeling, exploited large computational power provided grid. Major challenges solved were processing of large data which could be distributed on the grid and solved independently, and thus availing great advantage of Grid computing.[7]

High performance computers are required to process such large amount of data. Grid computing is one such viable solution available to majority of researchers intended in carrying out such studies using huge amount data.

2. Configuration of Nirma Grid – Testbed for Prototype Development

In most university colleges' large numbers of computers (PCs) are available using which one can generate a campus grid of the size no less than few hundred machines. The authors have made use of such an environment to generate a grid.

To compose a grid, the machines which we have coordinated and aggregated are heterogeneous in architecture and Operating systems. For experimentation total 162 machines connected in grid and has provided total aggregated computational power 165 GFlops [9]. The climate model is executed on the grid which is developed using globus toolkit and third party schedulers like OpenPBS and Condor. Globus Toolkit 4 is an open source toolkit organized as a collection of loosely coupled components [10]. These components consist of services, programming libraries and development tools designed for building grid-based applications.

Our system facilitates the user for acquiring flexibility in computation by either Grid-proxy-init or Globusrun as shown in figure 1. If the user opts for Grid-proxy-init and wants single-sign-on facility user will be prompted for lifetime of the proxy certificate, the strength of the key and password to access the private key of the user. The more the strength, the harder would be the effort to guess the key. Pass phrase is required which is used to protect the private key for proxy-certificate.

This paper describes deployment pf grid to analyze large amount large amount weather related data.

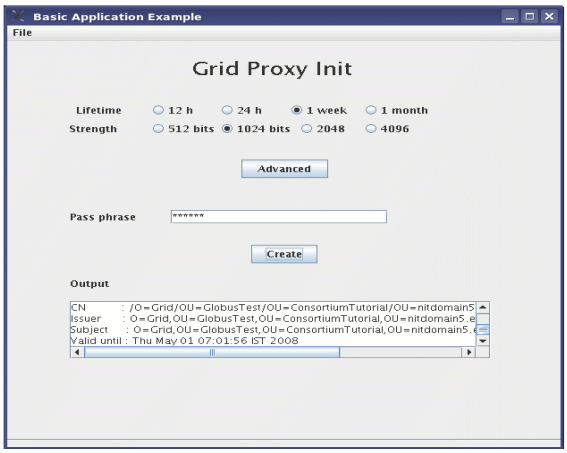


Figure 1. Grid Proxy Generated for Creating the Proxy of Head Node in Option with Globusrun in Interface

Proxy generated for creation and submission of jobs as shown in figure 1, provides authenticated and secured access to the grid. Globusrun allows user to submit a job on the grid. The window shown in the figure facilitates the user for Grid-proxy or Globusrun.

3. Analysis of HOAPS Data

The HOAPS (Hamburg Ocean Atmostphere Parameters and Fluxex from Satellite Data) data set is a completely satellite based climatology of precipitation, turbulent heat fluxes and freshwater budget (evaporation minus precipitation) as well as related atmospheric state variables over the global ice free oceans[11]. The HOAPS-3 grid rainfall data collected twice daily at 0 UTC and 12 UTC has been considered for calculation of anomalies.

The grid rainfall anomalies in computational grid model are calculated for 19-years from 1987-2005 on HOAPS-3 climate models. The rainfall anomalies are departure from mean normal rainfall condition for grided rainfall data collected six hourly over a window for

India. The results are plotted using formula as discussed under sub-section V for twice-daily rainfall anomalies. Later, the cross-correlation coefficients for yearly rainfall anomalies data series were calculated to understand the inter-annual climate relationship. As the climate change does not occur uniformly over various years therefore the study of rainfall variability in terms of rainfall anomalies become important.

To analyze the huge acquired data which was in the netCDF format, is transferred to an ASCII format and supplied to the grid for further computations which are requisite parameters for rainfall prediction

3.1 Dataset Description

This dataset contains 1 degree, daily twice, globally gridded multi-satellite rain values falling in time series of years from 1987 to 2005, providing high temporal resolution. The fields are stored for 0-12 and 12-24 UTC. Timestamps in the data files are at 0 UTC (0-12 UTC overpasses) and 12 UTC (12-24 UTC overpasses). Each grid-cell contains the average of data from the satellite that passed this gridbox closest to 12 and 24 UTC, respectively. The dataset

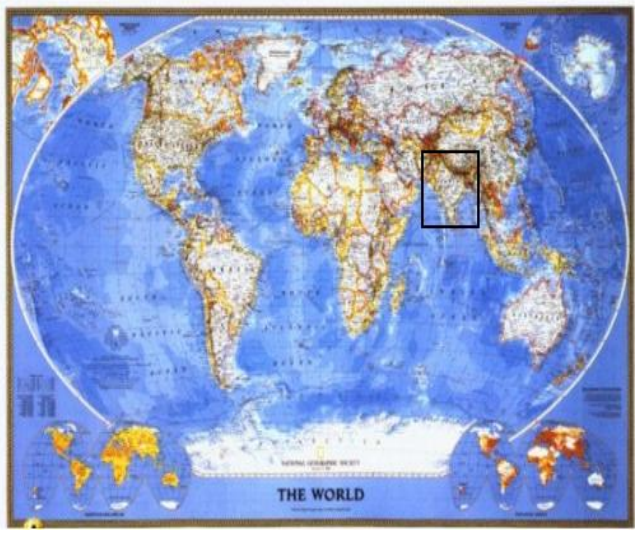


Figure 2. Geographical Location of Data Window Used in this Study Bounded by (5°N, 40°E; 35°N, 95°E)

Was processed to extract rain values for a predefined window out of the global grid.This window corresponds to the area covering India in the global grid shown in figure 2 with a rectangle.

The extraction of data was done as follows:

The netCDF dataset contained 228 files, each corresponding to a month from January, 1987 to December, 2005.

Each file was converted into readable ASCII format.

Data in the ASCII files was in following structure:

Header providing information about data

Rain values separated by ‘,’ with time steps mentioned above.

All the files of totalling size of 2.7 GB were read word by word for data extraction.

The rain values whose coordinates accommodated in the window were extracted on 12 hourly bases i,e, for every period of 12 hours from January 1, 1987 to December 31, 2005; one window was extracted and stored in a two dimensional tabular format in a central database system.

The format of each table was - "rain_<year>_<month>_<day_of_month>_<binary_value_representing_half_of_the_day>".

More than 13780 such tables were created each having 1736 records.Each table had three scaled fields corresponding to latitude, longitude and rain value respectively.

3.2 Design

The grid architecture which refers to pool of resources is modeled with Globus, PBS, Sun SGE, Condor and forms heterogeneous environment with different platforms. Figure 3 shows functional prototype of climate model on this heterogeneous grid platform.

This shows the flow of the data in the grid infrastructure after the jobs are submitted to the container. Out of created designs, major classes formed are:

DayMonthMean:	Used to generate RSL file for PBS and Condor and computes mean for a day or a month.
Main:	Abstract class used to call multiple threads of different classes.
Anomaly10yearlyexec:	Finds out anomaly for the period of 10 years.

4. Implementation and Results

Upon completion of grid programming using the globus programming components [12] and deploying climate model, many valuable results are obtained. Salient feature of this prototype is facility provided to the user. Instead of console based executions, flexible user interfaces allows user to interact with grid system more effortlessly.

The following Screenshot represents options provided to the user for selection of parameter and the frame for finding out the Standard Deviation and anomaly of the rain for a given day/month or year respectively. This also facilitates in turn, the computation for the required number of years.

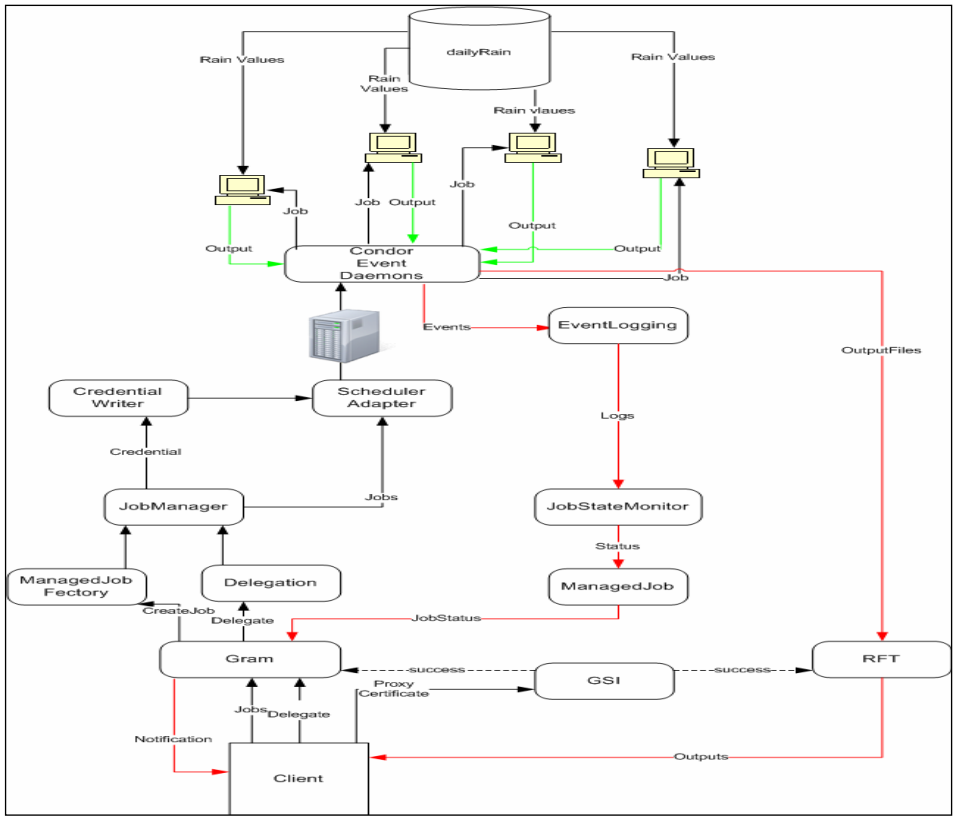


Figure 3. Functional Prototype of Climate Model

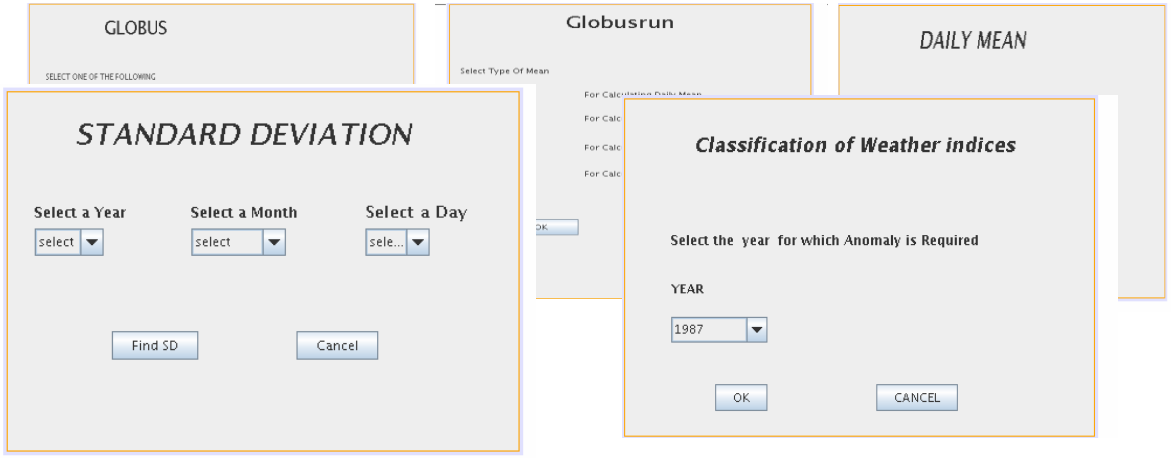


Figure 4 Interface Provides Computation Facility on the Grid for Year Wise Classification of Weather Indices

5. Performance Analysis

Grid exploits an idle computational cycles of the available resources and contributes in compute intensive applications. The application when developed and installed on the grid, the results obtained are shown graphs. Average time for finding mean on the grid is 92 seconds whereas sequential execution consumes 255 seconds. Standard deviation computation on grid acquires 175 seconds whereas sequential execution for the same could consume 415 seconds. Computation of these parameters as shown below in the expressions, leads to the calculation of anomalies for the required time period. Result shows the gain in execution time achieved on the grid platform. Figure 5-7 shows results obtained for the computation of anomalies with respect to time.

$$\mu = \sum_{i=i}^{N} x_i \times 1/N$$

$$\sigma = \sqrt{\sum_{i=1}^{N} (x_i - \mu)^2 \times 1/N}$$

$$\{x \mid x = (x_i - \mu)/\sigma\}$$

Since grid aggregates and exploits only idle computational power, we have carried out analysis with different time slots when the network traffic was more, less and medium with the usages of grid nodes by other users. And the impact of traffic gives varying grid cycles as shown in figure8.Since Communication and Processor Technology is on loosely coupled distributed system - issues involved are -

- Changing loads on the nodes of Network,
- Changing node availability on the Network
- Differences in Processor Speed and Network Speed
- Heterogeneity in Architecture and Operating System

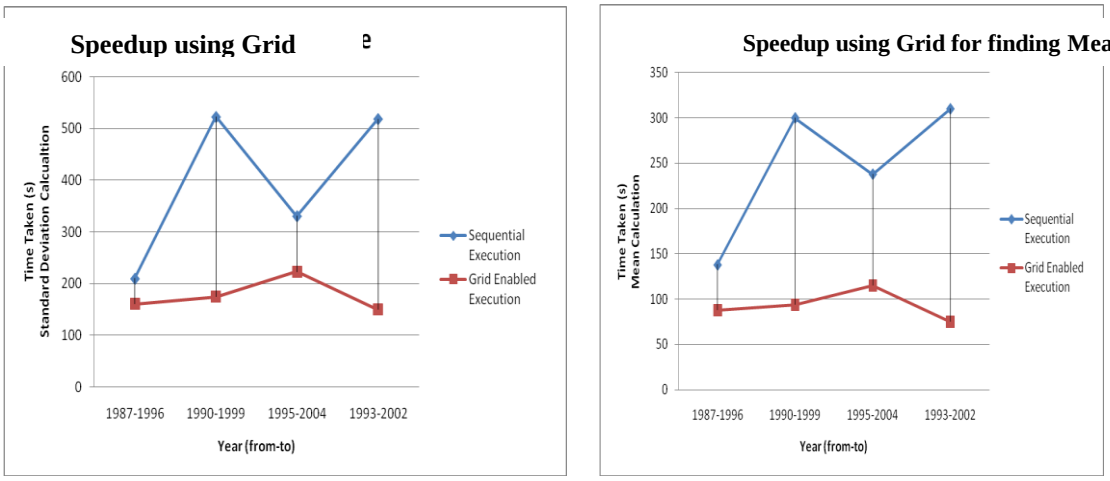


Figure 5 Variation in Performance of Both Types of Execution

5.1 Analysis of Anomalies

The HOAPS-3 grid rainfall data consists of total rainfall in the grid area 12 hours around twice daily at 0 UTC and 12 UTC which has analyzed to determine anomalies. The grid rainfall anomalies in computational grid model are calculated for 19-years from 1987-2005 on HOAPS-3 climate models, shown in 9.1. The rainfall anomalies are departure from mean normal rainfall condition for grided rainfall data collected six hourly over a window for India. The results are plotted using formula as discussed under sub-section V for twice-daily rainfall anomalies. Later, the cross-correlation coefficients for yearly rainfall anomalies data series were calculated to understand the inter-annual climate relationship. As the climate change does not occur uniformly over various years therefore the study of rainfall variability in terms of rainfall anomalies become important.

During seasonal analysis, it is observed that rainfall anomalies have been found large during Jan-Feb-March (JFM) and Oct-Nov-Dec (OND) months. While the period between Apr-May-June (AMJ) and July-Aug-Sept (JAS) have been recorded with low anomalies. The annual rainfall anomalies exhibit striking degree of variability pattern from January to December. This variability is due to the fact that rainfall in Indian peninsular spells during June-October (JJASO) months. The plot of mean rainfall anomalies shows large variability in twice-daily data. The rainfall anomalies found to be as high as 0.75 and as low as -0.20 for the early cycle during various years. In general, the mean rainfall anomalies for a normal year found to be 0.2 for months of January-May; 0.0 for June-October, and 0.1 for November-December. Furthermore, it can be stated that during drought years such as 1989, 1995, 1996, 1999, 2000, 2001 and 2002 showed very good cross correlation which has been found to be 0.40 to 0.50 The years such as 1990, 1992 and 1998 were classified as wet years found good correlation coefficient of 0.40. The other years were normal years and their correlation coefficient found to vary from 0.2 to 0.30.

Thus, the cross correlation coefficients for rainfall anomalies shown in figure 10 can be an indicator that whether a year will be dry, normal or wet. This will facilitate the use this information for climate and meteorological studies, namely in predicting the agricultural crop yield, assessment of surface water and flood control measures. The study of rainfall anomalies is not just important to understand the performance of grid computing but, also for bringing out climate variability and climate issue at regional or local scale.

Year (from-to)	Time Taken (sec) (Mean Calculation)		Speedup (s/g)
	Sequential Deployment (s)	Grid Enabled Execution (g)	
1987-1996	138	88	1.57
1990-1999	300	94	3.19
1995-2004	238	115	2.07
1993-2002	310	75	4.13
Year (from-to)	Time Taken (sec) (Standard Deviation Calculation)		Speedup (s/g)
	Sequential Deployment (s)	Grid Enabled Execution (g)	
1987-1996	210	161	1.30
1990-1999	523	175	2.99
1995-2004	331	223	1.48
1993-2002	519	150	3.46

Figure 6. Speedup Obtained

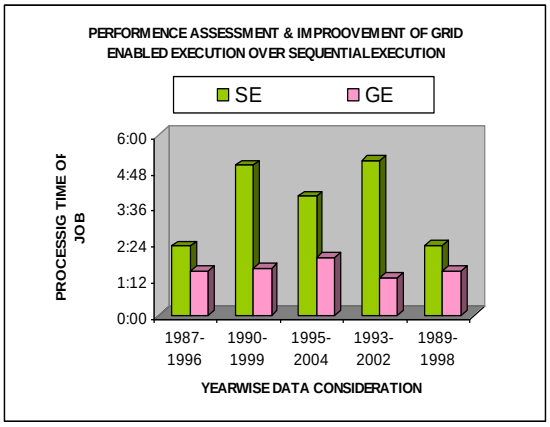


Figure 7. Performance Enhancement

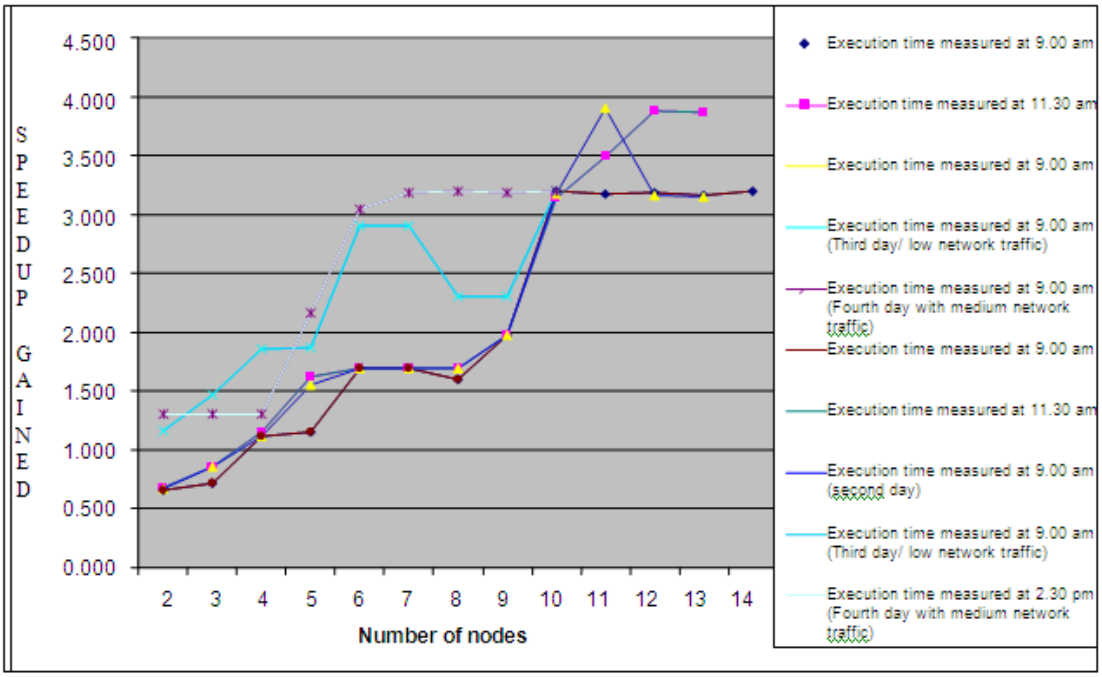
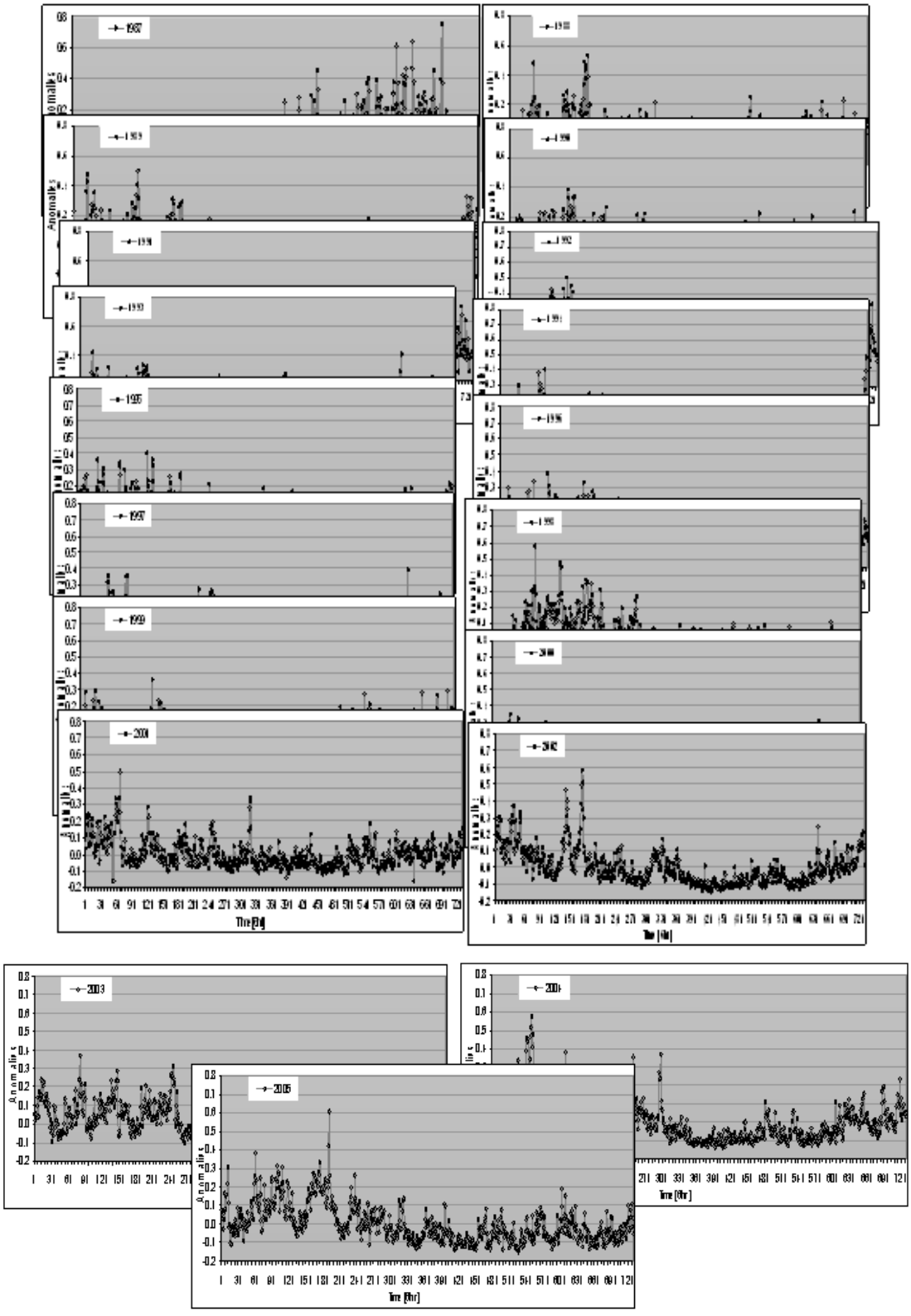


Figure 8. Grid cycles exploitation during network load

6. Conclusion

The heterogeneous grid setup on the campus with various software components such as middleware, schedulers and job submission components like Globus, PBS, SGE and Condor is exigent. The large computing power generated by exploiting idle computers, helped in achieving gain in the execution time required for the computation and classification of weather indices of climate model.

Fig 9.1 Shows Year Wise Computation of Anomalies Deployed on Grid.



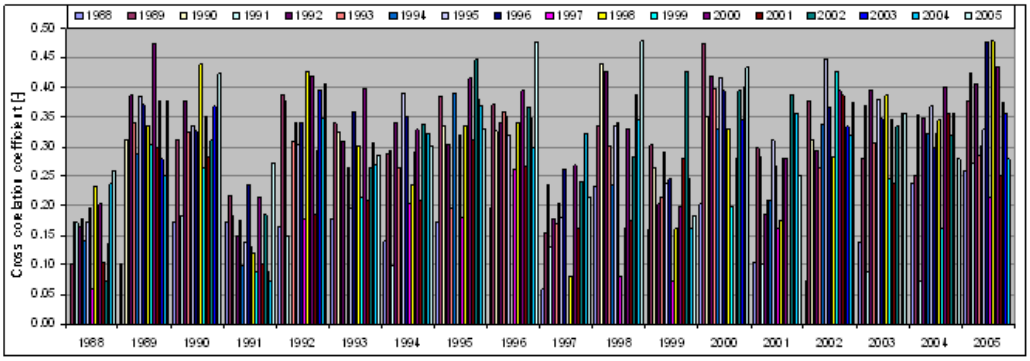


Fig 10 Cross Correlation Coefficients for Yearly Rainfall Anomalies

References

[1] Xingfu Wu, Valerie Tayler, Performance Analysis of parallel visualization Applications and Scientific Applications on an Optical Grid, , International Conference on Cyberworlds 2008, IEEE Computer Society , PP 1-2

[2] http://www.lhc.ac.uk

[3] http://setiathome.ssl.berkeley.edu/

[4] M. Baker, R. Buyya, and D. Laforenza, The Grid: International Efforts in Global Computing, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet , Rome, Italy, July 31 - August 6. 2000.

[5] http://www.GridComputing.com/

[6] R. Buyya, K. Branson, J. Giddy, and D. Abramson, The Virtual Laboratory: Enabling Molecular Modelling for Drug Design on the World Wide Grid, Technical Report, Monash-CSSE-2001-103, Monash University, Melbourne, Australia, December 2001.

[7] R Buyya, Ph D theis on "Economic-based Distributed Resource Management and Scheduling for Grid Computing", Monash University, April 2002

[8] http://www.parliament.uk/documents/post/postpn286.pdf

[9] M Bhavsar, S N Pradhan "Scavenging Idle CPU Cycles for Creation of Inexpensive Supercomputing Power:, International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009 ,PP 602-603

[10] I. Foster and C Kesselman, " The Grid2:Blueprint for a New Computing Infrastructure." Morgan Kaufmann publisher, 2004

[11] http:// http://cera-www.dkrz.de/WDCC/ui/

[12] http://www.globus.org/toolkit/docs/4.0/Exution/wsgram

[13] http://www.cs.york.ac.uk/dame/

[14] http://www.hoaps.org

Authors

Madhuri Bhavsar

She received M.E. From M S University Baroda, India in 2001. Currently she is pursuing Ph D in the field of Grid Computing from Nirma University, Ahmadabad, India. She has published total 15 papers in National and International Journals and conferences in the area of High Performance Computing. Madhuri Bhavsar is currently working as Senior Associate Professor in Computer Engineering department,

Institute of Technology, Nirma University, Ahmadabad. She is having total 16 years of teaching experience.

Dr Anupam K Singh

Dr.-Ing. Anupam K Singh is Professor in Civil Engineering Department, Nirma University Ahmedabad, India. He has more than 18 years of professional national and international experience viz. 6 years in academics, 8 years in research, and 4 years in consulting. Dr Singh has received Dr.-Ing. (PhD) from University of Karlsruhe (TH) in Germany. Dr Singh has published 8 book chapters, 12 papers in referred Journals, and presented 26 papers in International Conferences & 11 monographs.

Dr S N Pradhan

Dr S N Pradhan is PhD in Computer Science. He worked as a Scientist at PRL, Ahmedabad, India for 25 years. He is associated with academic field since 10 years. His areas of interest are Embedded System, Image Processing, Signal Processing & Networking. Currently he is working as a Professor and Coordinator of Post Graduate section, Institute of Technology, Nirma University, Ahmadabad, India.