ANUPAM-Adhya Supercomputer

K. Rajesh, K. Bhatt, D.D. Sonvane, K. Vaibhav,
V. Duggal, U. Karnani, N. Chandorkar, K.R. Koli,
R.S. Mundada and A.G. Apte
Computer Division

Abstract

Modern day scientific research increasingly relies on high speed computers in some stage or other. This is true of BARC too, where a large number of scientists and engineers are engaged in research in advanced fields of science and engineering. Computer Division, BARC has developed the ANUPAM series of supercomputers to cater to this ever-increasing demand for computing power. The latest in the series of ANUPAM systems is the ANUPAM-Adhya supercomputer, developed in 2010-11 with a sustained performance of 47 Teraflops. This system is released to users and is being used extensively. This article describes the new supercomputer, its architecture, subsystems and some applications that run on the system.

Introduction

BARC is a premiere research organization working on the development, demonstration and deployment of technologies related to nuclear reactors, nuclear fuel cycle, isotopes and radiation applications. It carries out inter-disciplinary and multi-disciplinary R&D activities covering a wide range of disciplines in physical sciences, chemical sciences, biological sciences and engineering. Expertise at BARC covers the entire spectrum of science and technology. More than 4000 scientists and engineers working on various advanced R&D at BARC are extensively using computers for meeting their requirements of supercomputing, general scientific computing, scientific visualization, information processing and information exchange.

Computer Division, BARC has a mandate of providing centralized computing facilities for the scientists and engineers of BARC, a significant number of which work in fields that require access to high speed computers. The complex problems that these users attempt to tackle are such that they cannot be solved on conventional desktops or servers in a reasonable amount of time. The ANUPAM Supercomputer project, undertaken by Computer Division has been fulfilling this ever-increasing demand for number crunching power for the last two decades.

High Performance Computing is a branch of computer science that deals with the design, development and use of computer systems that have performances much exceeding those in normal everyday use. These machines are characterized by large processor performances, large memory sizes and large storages. Supercomputers are used for solving compute intensive problems in areas such as nuclear physics, weather forecasting, climate research, molecular dynamics, computational fluid dynamics, structural analysis and other problems commonly called Grand Challenge problems. The philosophy behind supercomputing is to divide such big tasks across multiple processors available in a supercomputer and get the job done in parallel and thus in a reasonable amount of time. Many of the scientific applications fall under HPC category and require supercomputers to solve the problems efficiently.

ANUPAM Series of Supercomputers

Computer Division, BARC has started development of supercomputers under the ANUPAM project in

1991 and till date, has developed more than 20 different computer systems. All ANUPAM systems have employed parallel processing as the underlying philosophy and MIMD (Multiple Instruction Multiple Data) as the core architecture. BARC, being a multidisciplinary research organization, has a large pool of Scientists and Engineers, working in various aspects of Nuclear Science and Technology and thus are involved in doing diverse nature of computations. To cater to the computational needs of this diverse set of users, the ANUPAM supercomputers have been developed as general-purpose parallel computers. To keep the gestation period short, the parallel computers were built with commercially available off-the-shelf components, with our major contribution being in the areas of system integration, system engineering, system software development, application software development, fine tuning of the system and support to a diverse set of users. The graph in Fig. 1 shows the road-map of ANUPAM Series of Supercomputers, with the year of inception on X-axis and the performance of the system on Y-axis.

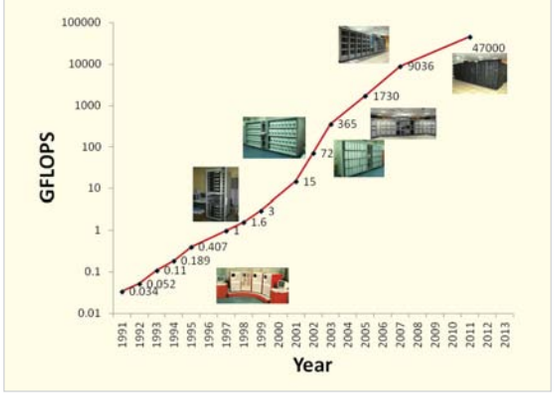

Fig. 1: ANUPAM Performance over the years

The series started with a small 4 processor system in 1991 with a sustained performance of 34 MFLOPS. Keeping in mind the ever increasing demands from the users, new systems have been built regularly with increasing computational power. The latest in the series of supercomputers is the 4608 core ANUPAM-Adhya system developed in 2010-11, with a sustained performance of 47 TeraFLOPS on the standard High Performance Linpack (HPL) benchmark. The system is in production mode and released to users. The detailed architecture of ANUPAM-Adhya system is described in the following sections. Fig. 2 is a photograph of ANUPAM-Adhya.

Fig. 2: ANUPAM-Adhya

ANUPAM-Adhya Architecture

The ANUPAM-Adhya supercomputer consists of the following subsystems:

Compute subsystem

The Compute Subsystem is the computational workhorse of the system, the place where user jobs are run and their problems get solved. This subsystem also determines the overall performance of the supercomputer. The compute subsystem of ANUPAM-Adhya is made up of 576 compute nodes, with each node having two Quad Core 3.0GHz processors and 32GB 800 MHz DDR2 FBDIMM memory. Thus, there are a total of 4608 processing cores, each with a peak performance of 12 GFLOPS. Scientific Linux 5.5 is used as the operating system along with OpenMPI, MVAPICH and MVAPICH2 libraries providing parallel environment. Since the number of computing cores in ANUPAM-Adhya is finite, the system should ensure that each user gets her fair share of the available resources. This is done by a resource management system which maintains

queues of user jobs and schedules jobs in the system using a fair share policy. Different queues have been implemented to cater to the needs of a variety of job types such as sequential jobs, long jobs, short jobs and so on.

Infiniband Interconnection Network

ANUPAM-Adhya consists of two independent interconnection networks – a primary network using Infiniband and a secondary network using Gigabit Ethernet. The Infiniband network is used for inter-process communication by jobs and Network File System I/O. The Gigabit Ethernet network is used for installation and management tasks.

Anupam-Adhya uses 4x DDR Infiniband network as a primary interconnect. 4x DDR Infiniband provides low latency (4 microseconds) and high bandwidth (20 Gbps) network for inter-process communication. Ideally, in order to connect 576 nodes in an Infiniband network, we need a network switch that has 576 ports in it. Since the largest infiniband switch commercially available at the time was of 288 ports only, the required 576 node network had to be realized using a multistage network using many switches.

In order to achieve best performance figures from an infiniband network, it is necessary to build them with 100% non-blocking factor, also called fully non-blocking. To construct a 576 port fully non-blocking network using 288 port switches, we need 6 numbers of 288 port switches and 1152 infiniband cables. Managing these many switches and cables was always going to be difficult; hence other less complex switch configurations were investigated. It was found out using experiments that even a 50% non-blocking network resulted only in insignificant drops in job performance but with much less switch and cabling complexities. Hence it was decided to build a 50% non-blocking network using one 288 port core switch and 36 numbers of 24 port edge switches. Fig. 3 shows the diagram of the network.

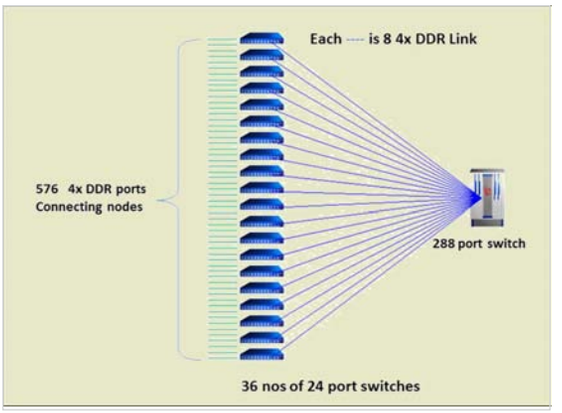

Fig. 3: ANUPAM-Adhya Infiniband network

Switching fabric of Infiniband network is managed by a subnet manager running inside the central switch. It sweeps the network regularly and generates its topology to determine the optimal route between the nodes. Since system has a 50% non-blocking network, the subnet manager uses a "top-down" routing algorithm to determine the routing tables instead of default balanced algorithm used for full non-blocking network.

Storage Subsystem

Storage is one of the critical components of a high performance computing system, which affects the usability of the system by the users. Traditional storage subsystems such as direct access storage and network access storage suffer from the lack of scalability and manageability when used for building large storage subsystem of the order of hundreds of terabytes. Storage Area Network (SAN) based systems separate the actual storage part from the servers and thus provide better scalability and reliability. Moreover, SAN based storage systems provide advanced features such as access control, volume management, snapshots (point-in-time copies), synchronous and asynchronous replication are critical to the design of a full-fledged storage system and play an important role in data security and disaster recovery.

ANUPAM-Adhya has a 100 Terabyte SAN based storage subsystem that is based on iSCSI Storage

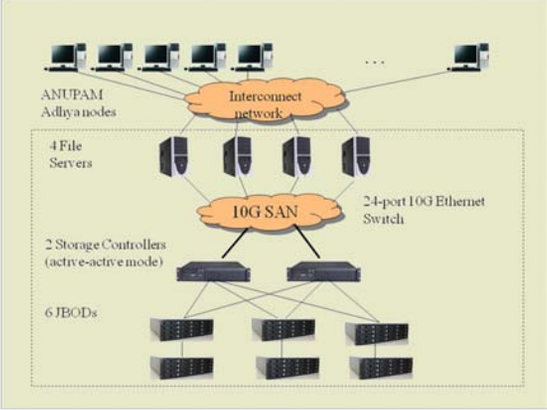

Fig. 4: ANUPAM-Adhya Storage subsystem

protocol. Fig. 4 shows the architectural diagram of the storage. The design of the system is modular, which means upgradation and enhancement in storage is easy. The 100 TB SAN storage system is housed in a 42U rack along with ANUPAM-Adhya that consists of the following components:

Storage Targets: These are two storage controllers connected in active-active mode for high availability and six SAS disk arrays (JBODS) each hosting 16 1-Terabyte SAS Disks.

File Servers / Initiators: Four file servers act as a frontend to the SAN storage. The compute nodes requests data to the file servers over NFS, the file servers in turn access the data from Storage Targets using iSCSI protocol over 10G Ethernet network.

Storage Network: A 24-port 10 G Ethernet switch provides the connectivity between the storage targets and the file servers/initiators.

Heat Removal Sub-system

ANUPAM-Adhya has 4608 computing cores, which are distributed in 576 servers and these servers are distributed across 12 standard 42U, 19 inch racks. Each rack dissipates 22KW of heat under full load and this heat needs to be removed efficiently to keep the system temperature under control. At 22 KW heat load per rack, each rack needs to be supplied with 72000 litres of cold air per minute (2540 cfm) to remove the heat with a Delta T of 15 degrees Celsius. Traditional heat removal methods employed in computer centres such as in BARC are not able to handle this kind of dense heat loads. To solve the problem of heat removal, an advanced chilled water based heat removal system was deployed. Water is a far better coolant than air because of its high specific heat (four times that of air) and high density (800 times that of air). Only 40 litres of water per minute are needed for a 22KW rack with 8 degrees C delta T. The 12 server racks were arranged in two rows, each row containing 6 server racks and 7 air-water heat exchangers. Cold air is blown through the server racks, where the heat dissipated by the servers is transferred to the air. The heat absorbed by air is immediately transferred to chilled de-mineralized water by the air-water heat exchangers. The requirement of 2500 cfm of air for cooling is fulfilled by large size blowers that circulate air from the server racks to the air-water heat exchangers. De-mineralized water is used as a primary coolant and this water is cooled by chilled water from the central AC plant by means of water-water heat exchangers. The reason behind using demineralized water in the primary circuit is that it is non-conductive, ultra-pure, contains no particulate matter and hence poses less risk to electronics in case of leakage. And, the purpose behind using two coolant circuits is to limit the flooding of water up to the amount of water in the

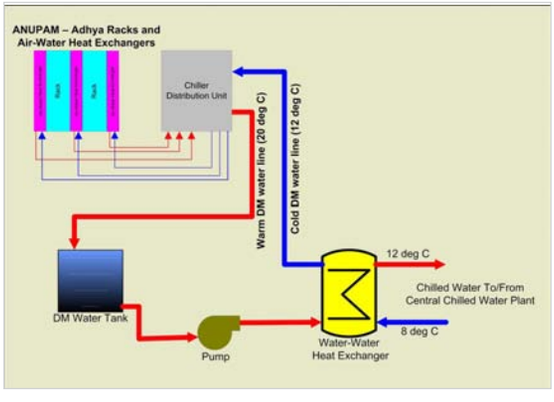

Fig. 5: Heat Removal Subsystem

primary circuit in case of failure in water pipe. The schematic of the heat removal sub system is shown Fig. 5.

Environment Monitoring and Protection System

The ANUPAM-Adhya supercomputer is arranged in 12 racks with each rack consisting of densely packed servers. Since this electronic hardware is sensitive to high temperature, it is necessary to continuously monitor the heat removal system for any variation in temperature. For this purpose, an integrated central environment monitoring system is developed to monitor environmental parameters, such as temperature and humidity of the system continuously. The monitoring system frequently reads environmental parameters from large number of sensors placed in the system racks and stores these values in a database. It also analyzes this data to detect abnormal changes and alerts the operator. The operator can track the current environment around the system using a web interface.

Since the ANUPAM-Adhya supercomputer dissipates about 240 KW of heat in a totally enclosed rack space, a failure of supply of chilled water can cause a temperature increase of tens of degrees with a space of a few minutes. Unless attended to immediately, this can have a catastrophic effect on the electronic components of the system. In order to cater to such circumstances, a protection system is put in place with three independent protection schemes.

- In case of break in water supply due to failure of a water pump, the air-water heat exchangers are configured to send traps. These traps are caught by listening software that immediately triggers a safe shutdown of the system.
- The BIOS of all the server machines are configured in "STAY OFF" mode so that whenever the power resumes after a failure, the power of server machines remain off and can be switched on once the cooling system is restarted after the power failure.
- In every server machine the processor temperature is monitored using the IPMI tools. If the temperature crosses the limit of 45 degrees, the server is automatically shutdown.

System Software for ANUPAM-Adhya

A large system such as ANUPAM-Adhya needs a wide gamut of software to keep it running and available to users. The software environment of the system is made up of components that are a mix of free and open source software along with several in-house developed tools and utilities. Scientific Linux 5.5 is the operating system that runs on all nodes of the system. The program development and runtime environment comprises MPI implementations such as MVAPICH and OpenMPI and numerical libraries such as BLAS.

Several in-house monitoring and management tools such as Anunetra (monitoring), AnuInstall (Rapid deployment), Anupam Accounting system (User Job accounting), AnuSakshi (Hardware Life Cycle management) and Load Sharing and Queuing system, which were used in earlier supercomputers are also used in the ANUPAM-Adhya. Some of the newer software development efforts are listed below

LDAPSync

Earlier ANUPAM systems made use of synchronized local files for storing user authentication information. Because of limitations of this approach, this technique was abandoned in favour of centralized LDAP based authentication in ANUPAM-Adhya. But the centralized authentication scheme was found to have scalability issues, especially during the boot process. During boot up, hundreds of nodes tried to communicate with the LDAP server simultaneously, thereby exhausting the resources of the LDAP server. To overcome these issues LDAP sync utility is developed, which regularly

synchronizes the user information from central LDAP server to local files. If the central server is down then the information is not refreshed but still the old information can be used for authentication. A random sleep is used to randomize the connection time of nodes to LDAP server, to provide load balancing and avoid resource scarcity and connection denial at server. Currently this system is being successfully used in ANUPAM-Adhya supercomputer.

Use of Virtualized Service nodes

Any ANUPAM system has several nodes that run critical services to keep the system running. Some of these critical services are the authentication service, resource management service, installation service, management and monitoring services and so on. Even though, it was possible to run multiple services on a single machine as in smaller systems, the approach was abandoned in larger machines in order to eliminate dependencies between services and failure of multiple services when a shared host failed. In earlier large ANUPAM systems, each service ran on a separate dedicated node. Nowadays with nodes having multiple cpu cores, running each service on a dedicated machine seemed a waste of computing power. Therefore, in ANUPAM-Adhya a new approach has been used for hosting these services. Each service now runs within a dedicated virtual machine. Multiple virtual machines are hosted within a single physical machine. This solves the problem of wastage of resources and also does not compromise service isolation. Since these virtual machine images are backed up, it is easy to re-deploy them in another host in case of hardware failures

User Applications

The ANUPAM-Adhya supercomputer is being used by scientists and engineers of BARC in solving computationally intensive problems in diverse fields ranging from physics, chemistry and engineering. The applications that run on ANUPAM-Adhya are a mix of in-house developed, open source and commercially purchased applications. The results obtained have enabled the scientists to publish their research work in prestigious international journals in physics and chemistry. Some of the work done is listed below:

- Study of new materials and their properties under extreme conditions
- Exploring interplay of structural, magnetic, optical and transport properties of a wide range of novel and emerging material systems
- Study of nano-materials and nano-catalysts
- Micro-mechanical analysis of PHT piping in reactors
- MACE telescope simulation studies
- Designing suitable materials for reversible hydrogen storage, fuel cell and water splitting
- Molecular Dynamics (MD) simulations of Bio-macromolecules in Explicit Solvent
- Study of Electric response properties of carbon nanostructures
- Study of properties of artificially synthesized elements
- Design of new drugs to combat radiation damage
- Study of radiation damage in materials
- Structural, electronic and magnetic properties of Bimetallic Nanowires
- Design and screening of ligand/solvent system for metal ion separation and isotope separation
- Applications in nuclear waste management

The following pictures depict the outputs of some of the applications that run on ANUPAM-Adhya.

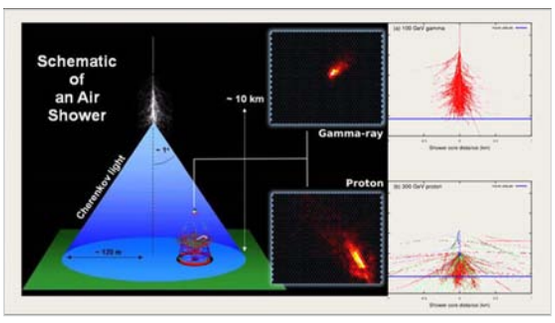

Fig. 6: MACE Telescope Simulations

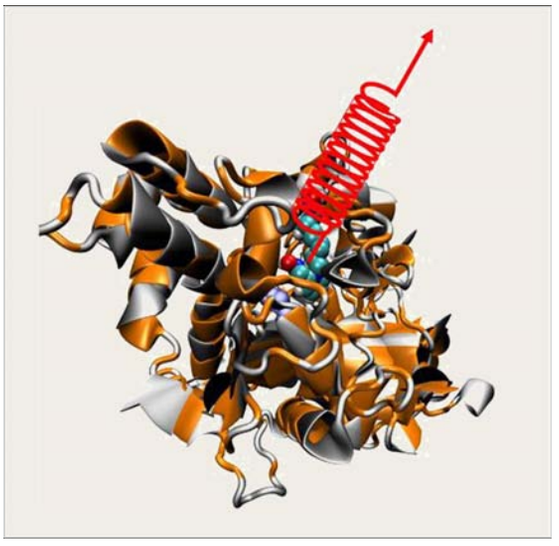

Fig. 7: Unbinding of drug molecule bound to the active-site of protein molecule

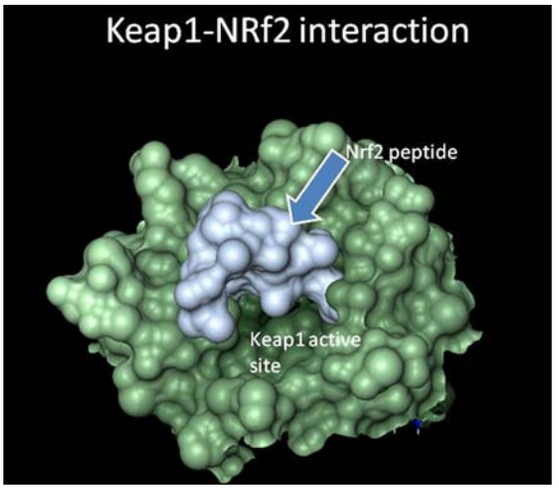

Fig. 8: Design of new drugs

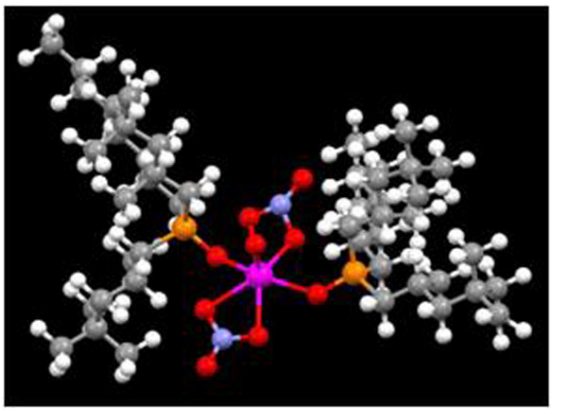

Fig. 9: Ligand/Solvent system for metal ion separation and isotope separation

Acknowledgements

Building and commissioning a large supercomputer such as ANUPAM-Adhya requires team effort from a large number of people. We gratefully acknowledge the efforts of the operations and administrative staff of Computer Division for full support during the work. The help of the erstwhile Infrastructure Projects Division (now part of TSD) in setting up of the heat removal system and the Technical Services Division for electrical support is also acknowledged. We thank Desalination Division for supplying the demineralized water for ANUPAM-Adhya's heat removal system. We also thank the users of the system for keeping the system fully utilized and their inputs for this article.