ANUPAM – Ameya: A Teraflop Class Supercomputer

Rajesh Kalmady, Vaibhav Kumar, Digamber Sonvane, Kislay Bhatt, B.S.Jagadeesh, R.S.Mundada, A.G.Apte, P.S.Dhekne

Computer Division, Bhabha Atomic Research Centre,
Trombay, Mumbai, India

***Abstract*- High Performance Computers are essential tools of trade in modern day science. The current research problems scientists and engineers are attempting to solve are becoming increasingly complex and are beyond the reach of traditional laboratory style experimental methodologies. Access to supercomputers makes a great deal of difference as they help in dealing with very high resolution models which result in accurate simulations of the phenomena that one is trying to investigate, in reasonable timeframes. This, in turn helps in cutting down the number of experiments that one has to conduct and also in arriving at accurate inferences at a faster rate. The Computer Division of the Bhabha Atomic Research Centre (BARC) has been developing high performance parallel computers for its scientists and engineers for over a decade. The latest in the ANUPAM series of parallel supercomputers is called 'Anupam-Ameya', a 512-processor cluster running Linux. This paper describes the design goals, architecture, components and benchmarking figures of the Anupam-Ameya cluster.**

I. Introduction

"Out-compute to Out-compete" is the dictum of modern science. In order to carry out frontline research and development and remain internationally competitive, it is extremely important that scientists have access to high performance computers. BARC is a premier research organization working on the development of technologies related to atomic energy and its applications in a wide range of areas. BARC scientists are engaged in research in various fields of science such as physical sciences, chemical sciences, nuclear and atomic sciences, biology and engineering, using computers extensively to meet their requirements of high performance computing, scientific computing, visualization, information processing and information exchange. It is extremely important that the computing and communication facilities available at BARC are not only enough to meet the growing requirements but also comparable with the best.

We, at Computer Division, BARC have been striving towards meeting these requirements of users by providing advanced computing facilities on par with the rest of the world. As part of this programme, the ANUPAM series of parallel computers developed by BARC have been the computational workhorses for BARC users for over a decade. Our major efforts have essentially been in developing systems, software and applications based on open source platforms and are based on off-the-shelf commodity components.

So far, BARC has developed sixteen different models of the ANUPAM series using a variety of processors as compute nodes and various interconnection technologies. We started initially with parallel processing systems based on Intel i860 processors and Multibus-II interconnect, progressing finally to Linux clusters using Intel x86 processors and Gigabit Ethernet interconnect. The performance figures of these machines have gone up from a modest 30 MFLOPS in 1991 to 365 GFLOPS in 2003 and finally to 1.73 TFLOPS in 2005. The number of processors has also increased from 4 to 512 over the years.

The latest in the ANUPAM series and the first teraflop class machine is called Ameya. This is a Linux based cluster consisting of 512 processors, connected over a Gigabit Ethernet network.

II. Design Goals

Apart from providing raw computing power, the design goals of the ANUPAM series of parallel computers have also been

- To maximize the use of open standards and technologies
- To harness commodity components
- To establish that a general purpose architecture is able to cater to a wide variety of problems

These goals ensure that the time required to build a new machine is as short as possible and also keep the cost down. BARC being a multi-disciplinary research centre, it would not be feasible to cater to a specific subset of problems and build a system optimized for that class. Hence, the machine should be able to support all kinds of compute models and messaging patterns equally well.

This was the philosophy on which all ANUPAM systems were based. Starting with a bus based architecture with MultiBus-II and Intel i860 processors, we later on adopted the cluster of workstations architecture, which rose in popularity during the mid-nineties with the advent of high speed networking technologies. The rise in performance of Intel processors and emergence of Linux as a robust and stable

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 409–414.

operating system made this architecture the platform of choice for high performance computers.

Building Linux clusters for HPC applications is now a well-understood concept and plenty of expertise and literature is available for doing this[1]. However, building a large cluster consisting of hundreds of nodes is a different ballgame altogether. Along with issues of performance and scalability, there are issues like stability, availability, cluster monitoring and management to worry about. As a large cluster, the system should be able to handle a large number of jobs and users and thus should be designed to be efficiently managed. Large clusters also have issues such as optimization of space, layout, power and cooling, each of which has to be addressed properly.

III. Architecture And Design Rationale

The ANUPAM Ameya system is based on the concept of Cluster of Workstations. It is a compact, centralized, homogenous Linux cluster of 256 dual processor nodes interconnected by Gigabit Ethernet network. A logical view of ANUPAM Ameya architecture is described in figure 1 below.

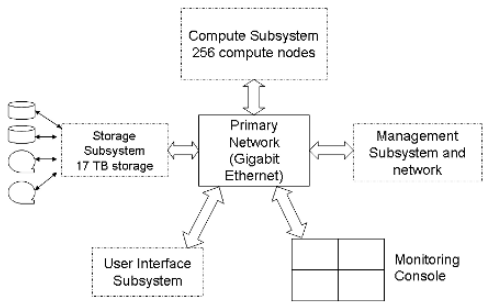

Fig. 1: Block Diagram of ANUPAM-Ameya Architecture

The major subsystems, key components and their design rationale are explained below:

A. Compute Subsystem

The Compute Subsystem consists of compute nodes and the software providing the parallel programming environment.

Compute Nodes**:** There are 256 compute nodes, which are rack mount servers of 1U size. Each of the compute nodes consists of two Intel Xeon EM64T processors @ 3.6GHz with 2MB cache each, shared 4GB memory and a SATA disk of 80GB capacity. In addition to the above-mentioned compute nodes, there are nine spare nodes in the system in order to increase the availability.

The choice of compute node was made after extensive evaluation of various models. The selection criteria for compute node models were performance, stability, acoustic noise, cooling and management features. The performance of various models was evaluated using a set of benchmark programs that evaluated the machine in different areas (such as network, memory, CPU and so on). Stability of the machine was tested by keeping the machine under the test-procedure for 3 days in continuous operation and checking the performance. The machine was deemed to have failed if the performance degraded with time or if the test failed. Some of the features like BIOS support for remote installation and console redirection to the serial port required for cluster installation and management were also tested.

Software**:** The Operating System on all the nodes is Scientific Linux 4.1. Parallel programming environment is provided by MPICH, LAM MPI, PVM and ANULIB libraries. Compilers for FORTRAN, C, C++ and various scientific libraries such as BLAS, LAPACK, SCALAPACK, etc. are also available.

B. Storage Subsystem

The storage subsystem consists of file servers, backup servers and tape libraries.

File Servers**:** There are 12 file servers, which constitute 17 terabytes of storage space. Each server is a 2U rack-mount server, equipped with dual Xeon processors.

Backup Servers**:** There are two backup servers for taking backup of users' data. Each server is a 5U rack-mount server equipped with dual Xeon processors and backup devices such as DVD writer, DAT and DLT drives. In addition, each server is connected to a tape library. The backup servers offload the backup load from the main file servers.

Tape Libraries**:** There are two tape libraries with a storage capacity of 3.2 terabytes each. The backup servers are programmed to take periodic backup of file servers to their local disks and then copy it onto the tape libraries.

Design challenges for the storage subsystem were performance, reliability and availability. Necessary redundancy is provided to reduce failures and downtime. Moreover, the system design ensures minimal effect of a server failure. RAID is configured on each server to overcome single disk failures. Each server has three Gigabit Ethernet network ports, which are link-aggregated to increase the availability and throughput by three fold. The users are distributed across the 12 file servers so that a single server failure affects only a fraction of users while others can still use the system.

C. Primary Network

The primary network in our cluster is Gigabit Ethernet The same network is used for inter-processor communication as