Designing a PC Farm to Simultaneously Process Separate Computations Through Different Network Topologies

P. Dreher

Laboratory for Nuclear Science
Massachusetts Institute of Technology
Cambridge, MA 02139
U.S.A.

Abstract

This is a short report on a demonstration project at the MIT Laboratory for Nuclear Science to develop multiple types of network hardware interconnections among the individual PC processors in a single PC farm. The goal of the project was to develop a flexible system that would only require the hardware from one PC farm to provide the capabilities of handling specialized computations optimized for different types of PC farm hardware configurations. High energy and nuclear theorists frequently design the interconnections among the processors on a PC farm to be either a 2, 3, or 4 dimensional torus network topology. Experimentalists frequently prefer the PC farm processors to be simply connected to allow many processors to work independently on data analysis. This report will describe the network hardware modifications and system software needed to allow one PC farm to simultaneously handle both the lattice gauge theory calculations with an embedded torus network topology and the single independent processor PC farm to analyze experimental data.

Keywords: COTS, PC farm, torus network, LQCD

1 Introduction

In the last few years, the PC farm as become a standard hardware component for high energy and nuclear physics research projects [1] [2]. These systems can deliver equivalent computational throughput compared to the large commercial supercomputers at a fraction of their total hardware cost. As a result, research computing facilities at universities and national laboratories have adopted these types of computer hardware designs and have constructed PC farms use in their research projects.

Unfortunately, the there is no one PC farm hardware design in general that is optimal for both experimental and theoretical high energy and nuclear physics computations. One solution to this problem is to build dedicated PC farms for experimentalists and other high performance computational farms designed for large scale theoretical simulations. Although PC farms are quite reasonably priced compared to the cost of commercial machines, they are still can be a substantial expense within the overall budget of a research project. Furthermore, if the research center services a mixed environment of both experimentalists and theorists, it can become quite difficult to balance the development and production needs for both groups of physicists. The goal of this project was to build one low cost, yet sufficiently flexible PC farm to provide a platform that could be used simultaneously by both experimentalists and theorists for code development and testing as well as for small to mid-size production jobs.

2 PC Farm Design Requirements for Experimentalists and Theorists

The development of the PC farm has allowed both experimental and theoretical nuclear and high energy researchers to plan more complex and detailed numerical investigations that had ever been

attempted in the past. However, such complex numerical work requires some in depth understanding of how the PC farm architecture is constructed and how to organize the numerical work in order to maximize the efficiency and overall production throughput on the machine.

Experimentalists are usually interested in processing large quantities of statistically independent data. Examples of such problems include analysis of raw data from high energy and nuclear physics detectors in the experimental halls. Numerical work of this type can include event by event particle track reconstruction and analysis to search for new physics. Such problems are ideal for PC farms which consist of many independent CPUs that are capable of processing individual and statistically independent events simultaneously.

Problems of numerical interest in theoretical physics tend to be grouped around the technique of building large lattice structures within the computer and embedding the physics both on these lattice sites and on the links that connect this grid mesh. Theory simulations tend to be CPU intensive and require large amounts of memory to store the lattice constructs and variables. PC farms are ideal for harnessing the power of many PCs to focus on one production job distributed over all of the CPUs in the farm. However, such a design requires excellent network communications among processors in different hardware boxes. This can be achieved by either installing state of the art network switching equipment or designing the network topology among the individual PCs to have a 2, 3 or 4 dimension torus configuration. The budget for this project limited the hardware selection to COTS PC type equipment without state of the art network switching equipment and so the enhanced communication was achieved by configuring the PCs in a 2 dimensional torus network topology using fast ethernet technology.

In summary, the PC farm design requirements included:

- Use only COTS PCs without expensive or state of the art network switching equipment to keep the overall project within a minimum budget
- Use only freely available operating system software and utilities to operate the farm
- PC farm must be capable of simultaneously running
  - Analysis and reconstruction jobs for experimentalists
  - Small theoretical simulations and code development in a 2-d torus network

2.1 Hardware Selection

The PC farm consists of 20 dual Pentium II 400 Mhz PCs, each with 384 Mbytes of RAM, 13 Gbytes of disk space and 100 Mbit fast ethernet cards. Each PC in the farm is connected to Kingston EtheRx 100BaseTx Fast Ethernet Stackable Hubs. The PC farm has a front end with two fast ethernet cards. One card interfaces the front end to the Laboratory's LAN and the other card connects to the hub (Figure 1). The front end serves as the control unit for administrative operation of the PC farm and for controlling the jobs running on the farm. The PC farm had four of the PCs modified from the original configuration to include an additional 4 Port Adaptec Network card (figure 2).

2.2 Software Selection

The operating system that was chosen was Redhat Linux V5.2 for x86 machines built with an SMP kernel. In order to manage the production component and distribute the jobs throughout the PC farm as well as to partition groups of nodes, the Network Queueing System software was selected. This software met the need for a good UNIX-type batch and device queueing facility capable of supporting requests in a networked environment and partitioning the farm into subgroups of nodes. This software was freely available URL (`http://www.gnqs.org/`). Additional software to handle message passing among CPUs was installed on the 4 PC nodes with the 2-d torus mesh. The

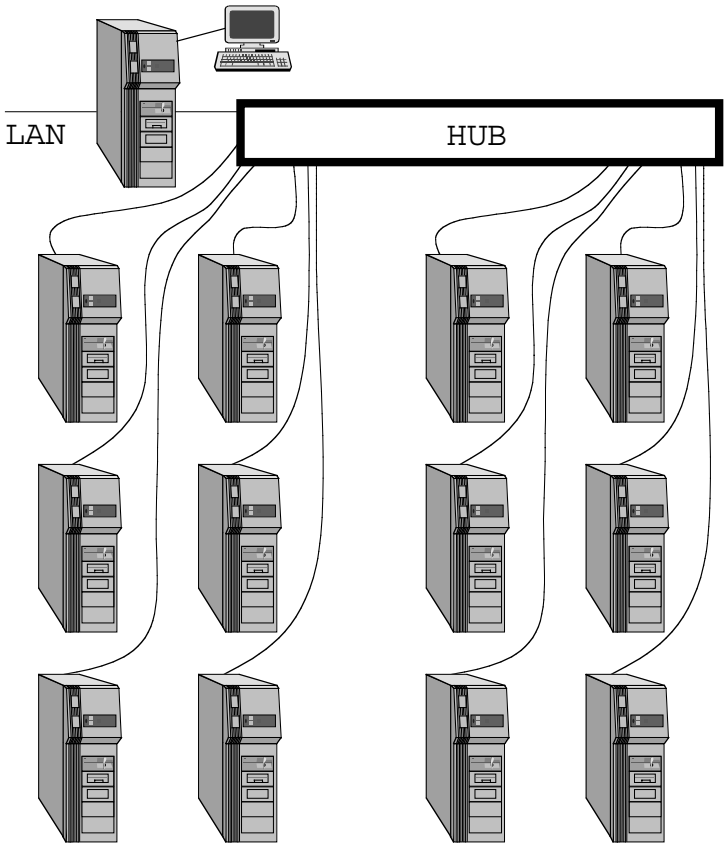

Figure 1: Figure illustrating the PC farm and front end connected to both the LAN and hub

MPI software is available at (URL: `http://www-unix.mcs.anl.gov/mpi/mpich/`). Modifications to the routing tables on the 4 PCs to identify the 2-d network topology among the 4 boxes as well as the ethernet card connecting the PCs to the hub were handled through a Perl script.

3 Operation of the PC farm

From the perspective of most users, the operation of the PC farm was controlled from the front end. The users interacted with the NQS batch system which provided the menu of batch queues and controlled the production aspects of the farm. There were two complete installations of NQS, although only one configuration of NQS could be operational at any given time. The NQS software operated in either a 20 dual pentium configuration (figure 1) or a 16 dual pentium configuration (figure 2).

When the PC farm was configured as in figure 2, 16 dual pentiums could be used by experimentalists for production type operations while the other 4 machines were dedicated to theorists for code development and testing. Access to the four machine system from the front end was handled by establishing an ssh connection from the front end to a node on of the 4 PCs. Theory simulations and code development work using the 2-d torus were launched directly from within,

and limited to the 4 PCs with the additional 4 port ethernet cards in each of them.

Given the budget constraints to purchase network hardware, this system worked quite well using the fast ethernet technology. The NQS system allowed a large experiment to always have at least 16 dual pentium nodes available for production while also allowing the theorists to simultaneously have access to a 4 node 2 dimensional torus network architecture within the PC farm for their code development work. During times when the 4 nodes for the theorists were not being used, the farm was re-configured under NQS with 20 nodes available to the experimentalists for production work.

This work is supported in part through funds provided by the U.S. Department of Energy under Contract Number DE-FC02-94ER40818.

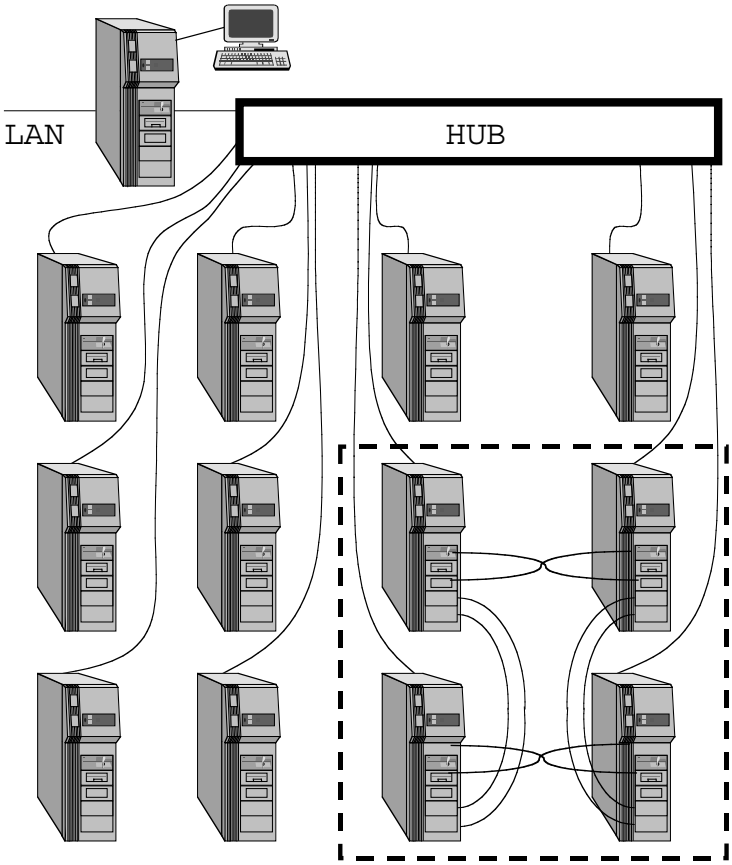

Figure 2: Figure showing the 2-d network torus interconnecting 4 nodes within the PC farm

References

1 O. Schwarzer, “Experience of using a LINUX PC farm for Physics Analysis and Reconstruction at the ZEUS Experiment”,CHEP’98, Chicago, Autumn 1998.

2 I. Bird, A. Kowalski, B. Lukens, “Experience With a PC-based Reconstruction Facility”, CHEP’98, Chicago, Autumn 1998.