INTRODUCTION TO THE CDF PRODUCTION FARM A PARALLEL COMPUTING SYSTEM FOR HIGH ENERGY PHYSICS EXPERIMENT

YEN-CHU CHEN, TONY CHAN [a], PAOTI CHANG [b], PING YEH

Institute of Physics, Academia Sinica, Taiwan

MARK BREITUNG, TROY DAWSON, JIM FROMM, LISA GIACCHETTI, TANYA LEVSHINA, IGOR MANDRICHENKO, RAMON PASETES, MARILYN SCHWEITZER, KAREN SHEPELAK, DANE SKOW, STEPHEN WOLBERS, GONG PING YEH

Fermi National Accelerator Laboratory, Batavia, IL, USA

JAROSLAV ANTOS

Slovak Academy of Science, Slovak Republic

MIROSLAV SIKET

Comenius University, Slovak Republic

The CDF experiment at the Fermi National Accelerator Laboratory will record data in the year 2001 with a maximum throughput rate of 20 MB/s, twenty times the rate in the previous data taking period. To analyze data in quasi-real time, great computing power is needed. A production farm composed of up to 200 PC's is under construction. A prototype farm of 18 PC's was built and tested with Monte Carlo data for investigation. A production farm with 50 PC's was constructed and tested to understand the performance. The hardware and software used in these tests will be described. Results from these tests and future plans to construct the production farm are discussed.

1 Introduction

The Higgs particle is the missing part of the Standard Model. The existence of the Higgs particle can help us to understand the generation of mass and creation of the Universe. The search of the Higgs particles is one of the main goals of the CDF experiment [1]. Beside that the CDF experiment will also search for new physics beyond the Standard Model, such as SUSY particles [2]. To accomplish the goals the CDF experiment requires that data be analyzed as quickly as possible. To keep up with data taking, a large cluster of computers, called a production farm, is required. The CDF experiment and FNAL have decided to use as much commodity hardware as possible to build the farm. To have the best cost/performance ratio, the choice is to build a system composed of hundreds of off-the-shelf PC's running the Linux operating system. The main usage of the offline farm is for data production in quasi-real time, which includes reconstructing raw data, splitting the resulting output files into physics datasets and storing physics data on tape. It can also be used for data re-processing and Monte Carlo simulation. In this paper we describe the requirements, the design of the farm, tests of performance and the future plan. The goal is to have a powerful system, which is easy to use, maintain, and expand.

2 Requirements

The major requirements for the CDF production farm are the following:

- High throughput rate and IO capability.
- Ability to easily configure farm processing to control priorities.
- Excellent bookkeeping.
- Excellent error handling and error preventing.

The throughput is driven by the rate events are accepted and stored by the experiment. On average the farm is estimated to process 28 event/s, based on a peak output rate from the experiment of 75 event/s and factoring in all the known down-times. Given the input event size of 250 Kbytes, accordingly the peak input rate is 20 MB/s. The peak output rate is estimated to be 11 Mbytes/s.

The ease of use requirement includes the ability to control priorities on the farms. This is important to allow the farm to be reconfigured quickly to handle rapidly changing needs of the experiment. There may be new triggers, which need to be understood and tuned, physics results which require rapid processing, or other conditions under which the farm has to be reconfigured to

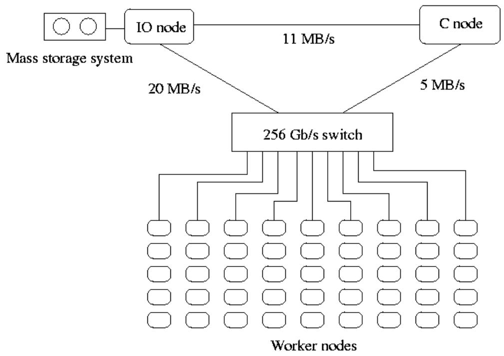

Figure 1: Hardware configuration of the CDF production farm.

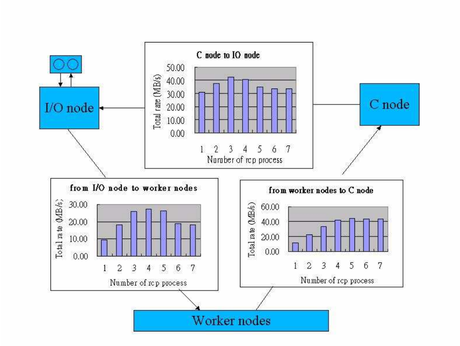

Figure 2: Hardware rates of the CDF production farm.

process specified sets of data using certain code and constants.

The bookkeeping that is required on the farm includes tracking the files being processed inside the farm, the versions of code, and the set of constants used to process each file. The number of events of each file is also recorded to handle the requests about the number of events run through each version of the reconstruction packages, etc. The exact information that must be kept will be driven by the needs of the people who run the farm, as well as the needs of the physicists who will analyze the data.

One additional requirement of such a system is to have excellent error handling. This makes such a system much easier to run and to maintain, and good attention to this will pay off over the long run. This effort is based on the experience of Run I farms and other related experiences as the system is designed.

3 System design

To reach the goal, great effort went into the hardware configuration, testing as well as software design [3].

3.1 Hardware configuration

The whole system is composed of an Input/Output (IO) node with mass storage system and RAID disks attached, a switch capable of 256 Gigabit/s throughput, a group of PC's as worker nodes and a Concatenation node with RAID disks. This is shown in the Fig. 1. To avoid bandwidth congestion Gigabit links are used in between the IO node and the switch, between the switch and the Concatenation node, also between the Concatenation node and IO node.

In operation, data files will first be staged on the IO node then transferred to the worker where they will be processed, the resulted output files are transferred to the Concatenation node for further processing to produce useful physics datasets for further analysis. The final files are transferred back to the IO node and stored in the mass storage system.

To ensure that the hardware can match the requirement, a series of tests were made and effort was made to fine tune the system especially on the RAID disk to get the best performance [4]. The result is shown in Fig. 2.

3.2 Software design

The production farm system can be viewed as a 6-stage pipeline (input, dispatch, reconstruction, collect, split/concatenate, output), plus peripheral input/outputs. But it is not a usual straightforward pipeline due to the following reasons:

- The parallel processing of worker nodes breaks the sequence of files, especially when errors occur.
- The staging and spooling stage expect a collection of files as input while other stages could operate with one file at a time.

It is mandatory to have data buffers at the input end of the split/concatenation and spool stage to accumulate files until they can be processed. These buffers enable the most efficient operation of each step and allow the sorting of output files to be concatenated together. This is shown schematically in Fig. 3.

Each file process has an entry in the database and is updated step by step as it passes through the farm. This ensures that there is no file missing during the process. If the system breaks down due to a power outage, for example, this information can be used for automated recovering.

4 Performance

To test the integrated system of data taking and processing, CDF has exercises, called Mock Data Challenges

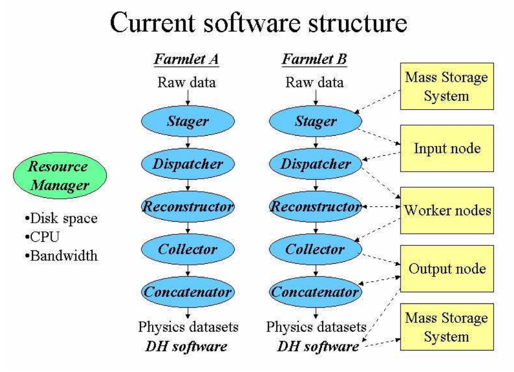

Figure 3: Current software structure of the CDF production farm.

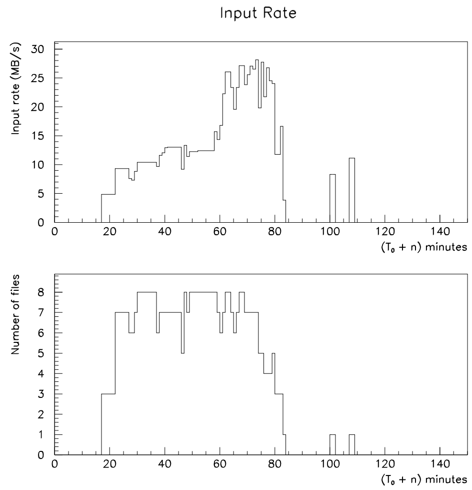

Figure 4: Input rate of the CDF production farm.

(MDC). In the latest MDC (MDC-2A), we measured the input rate and cpu processing rate in term of MB/s. These are shown in figure 4 and figure 5. In figure 4 the bottom plot shows the number of files being transferred at n minutes later after the farm was started. The upper plot shows the correspondent throughput at the same moment. The farm gradually reached 26 MB/s than dropped down because there was no more file for processing. The two spikes after the main peak correspond to additional file adding. The system successfully processed the new files automatically once they were added into the read-in area.

The two plots in figure 5 show the performance of the CPU's processing the data files. The bottom plot shows the number of files being processed inside of the farm. They stacked up as more and more files were transferred into the worker nodes. The upper plot shows the correspondent processing rate in terms of MB/s. The peak value was 26 MB/s or so, which is consistent with the input rate.

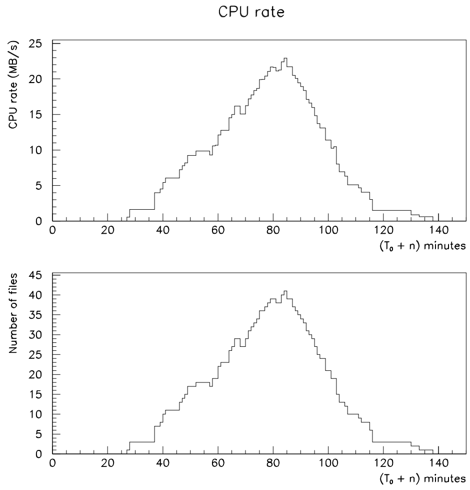

Figure 5: CPU rate of the CDF production farm.

5 Conclusion

The CDF production farm has reached 26 MB/s in input and cpu processing, which is more than the required throughput. More effort and tests are on-going on the parts of concatenation and error handling. Overall the farm is on a smooth path toward a fully integrated system, which will be capable of processing data at the required throughput.

Acknowledgments

This work is supported by the U.S. Department of Energy under contract No. DE-AC02-76CH03000, the National Science Council of the Republic of China under the project number NSC 89-2112-M-001-071, and the U.S.-Slovak Science and Technology Joint Fund under Project Number 00198.

References

1. T. Affolder *et al*, The CDF Collaboration, FERMILAB-PUB-00/238-E. Submitted to *Phys. Rev. Lett.Oct. 9, 2000.*
2. *T. Affolder et al, The CDF Collaboration,* Phys. Rev. Lett.84, 5704 (2000).
3. Stephen Wolbers *et al*, FERMILAB-Conf-00-095-E.
4. "Performance Tests of the CDF Run 2 Production Farm", CDF note 5467, Oct. 25, 2000.