Fermi National Accelerator Laboratory

FERMILAB-Conf-00/053-E

Computing Farms

G.P. Yeh

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

April 2000

Presented at and to Appear in the Published Proceedings of the *7th International Conference on Instrumentation for Colliding Beam Physics (INSTR99)*, Hamamatsu, Japan, November 15-19, 1999

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy

Computing Farms

G. P. Yeh

Fermi National Accelerator Laboratory, Batavia, Illinois
Academia Sinica, Taipei, Taiwan

High energy physics, nuclear physics, space sciences, and many other fields have large challenges in computing. In recent years, PCs have achieved performance comparable to the high-end UNIX workstations, at a small fraction of the price. We review the development and broad applications of commodity PCs as the solution to CPU needs, and look forward to the important and exciting future of large scale PC computing.

Key words: Computing; PC; Farm; Linux; Cluster; Beowulf

1 Introduction

High energy physics (HEP) and nuclear physics have large challenges in computing. To meet the CPU requirement, Fermilab has built custom board based ACP and ACPMAPS [1] computers since the mid-1980s. Fermilab also pioneered 10 years ago loosely-coupled parallel computing "Farms" as an alternative to the costly mainframe computers. The Fermilab UNIX Farms at their peak in 1998 had a total of about 500 UNIX workstations.

With the extreme CPU needs in the future, it became natural to evaluate using PCs as the most cost-effective solution. Since the introduction of the Intel Pentium Pro processors in 1995, PCs have provided high performance at low cost, relative to RISC processors and UNIX workstations. The advances in CPU chips are accompanied by similar great advances in commodity network and memory. The PC Farms and general PC computing also importantly provide an upgrade path.

By April '97, a dozen groups from NASA, DESY, Fermilab, Sandia, NIH, other labs and universities reported [2] [3] their performance benchmarks for various applications on PCs, some with plans to build clusters with thousands of processors. By 1999, Linux [4] PCs have become the dominant choice for computing in high energy and nuclear physics. This success has also had an important role [5] [6] in establishing the acceptance of Linux by other scientific fields and the general public.

1.1 Large CPU needs in the future

Many major experiments in HEP and nuclear physics will have large, challenging computing needs. The three large experiments at three B Factories, BaBar (SLAC), Belle (KEK), HERA-B (DESY) start in '99 to collect large amounts of data for CP violation studies. The BRAHMS, PHENIX, PHOBOS, and STAR experiments at RHIC/BNL plan to start heavy ion collisions in 2000, with one of the main physics goals being to study quark-gluon plasma.

At Fermilab, CDF and D0 experiments will have data taking period Run IIa in 2001 – 2003, to study the newly discovered Top Quark, and search for B CP violation, Higgs, SUSY. The amount of data collected in Run IIa is expected to be 20 times that of Run I. The CPU requirement will be $N \times 10^5$ MIPS or 10^4 SPECint95 (Fermilab defines CPU equivalent to a VAX-11/780 to be 1 MIPS, and SPECint95 = 40 SPECint92 = 13 MIPS = 10 CERN_Units). The amount of data collected in Run IIb in 2004 – 2006, is expected to be 15 times that of Run IIa.

For the Large Hadron Collider starting in 2005, the Atlas and CMS experiments have estimated CPU needs of $N \times 10^7$ MIPS or 10^6 SPECint95. LHC-B has been approved to start when LHC starts, and BTeV is being proposed at Fermilab. Each of these new B-Factories will require thousands of Linux PCs for online trigger and offline computing. There are strong, active Linear Collider study groups in the 3 regions of Asia, Europe, and North America. Their hope is to complete by 2008 a Linear Collider with (Phase I) center-of-mass energy of 500 GeV, which can later be upgraded up to 1,500 GeV. There are also study groups for the longer term Muon Collider and the Very Large Hadron Collider.

1.2 Loosely-coupled parallel computing

In HEP, by its nature of independent collisions, the event reconstruction from data and the Monte Carlo simulation are excellent tasks for loosely coupled parallel computing. At Fermilab, to fulfill the CPU needs at an affordable price, the ACP systems were used in '87-'89 for CDF Level 3 and Offline, and also for other experiments. These systems were based on commercial (Motorola 68020) CPUs, and hundreds of custom boards. The ACPMAPS [1] system was developed for more tightly coupled lattice QCD calculations. The system had

5 Gflops with 256 Weitek XL-8032 processors in '88, and 50 Gflops since '92 with 612 i860 chips based on dual CPU boards.

For parallel computing on loosely-coupled processors, Fermilab developed CPS (Cooperative Processes Software). Other inter-processor communication software such as PVM [7] and MPI [8] have been developed, supported, and widely used by other labs.

UNIX workstation farms provided the main source of CPU at Fermilab and other labs since '91. The use of commercial UNIX workstations and Operating Systems instead of the custom board based systems such as the ACP provided the ability to upgrade the systems easily. To satisfy the large increase in CPU requirements in the future, however, a change in hardware platform is necessary.

1.3 PC in HEP computing since '95

In CHEP '95 (Rio de Janeiro, Sept. '95, www.hep.net/conferences/chep95/), there were the presentations "PC as Physics Computer for LHC ?" [S. Jarp], and NICE, desk top cluster [D. Foster]. In CHEP '97 (Berlin, April '97, www.ifh.de/CHEP97/), there were several talks on PC support and desktop applications for WNT or Linux, and two PC farms talks: Hermes Linux PC Farm [W. Wander], FNAL Linux PC Farm [S. Wolbers].

The First Pentium-Pro Cluster Workshop, April '97 (www.scl.ameslab.gov/ workshops/) was held at the same time as CHEP '97. Representatives from NASA, Caltech, Fermilab, NIH, and other institutions shared successes with Pentium-Pro cluster computing. The Linux/NT ratio was about 3:1.

The Super Computing '97 Conference (www.supercomp.org/sc97/) was in San Jose, in November '97. Several groups from labs and universities exhibited "Beowulf projects" or PC Farms. Almost all of these groups used Linux as Operating System. Fermilab and CDF demonstrated using PC Farms for Online Level 3 trigger and for Offline Reconstruction.

2 Non-HEP systems

2.1 Beowulf Projects

Among HEP and DOE labs, a loosely-coupled computer system is known as a "Farm". Following the UNIX farms, PC based systems are "PC Farms". The

NASA Beowulf Project started in Summer '94 at CESDIS to produce software for off-the-shelf clustered commodity PC-class hardware, using a high-bandwidth, low-latency network and open source Linux operating systems and message-passing. The system should provide super computer performance at a fraction of the cost ("GigaFlops at $50K"). An important contribution to the PC based parallel computing is the Linux device drivers for Ethernet by Donald Becker at CESDIS. There have been dozens of Beowulf systems worldwide [2], with applications ranging from space sciences, molecular biology, genetic programming, to weather, at NASA centers, NIH, NOAA, universities and labs. There are also Pentium-Pro Cluster "Cookbooks" for beginners.

2.2 Linux Alpha

In the movie Titanic, Digital Domain did the visual effects simulation on 100 Alpha nodes running Linux. The Avalon Linux Alpha cluster at Los Alamos has 70 nodes. The University of Wisconsin - Milwaukee used 48 Linux Alpha nodes to process data for the LIGO Experiment.

2.3 ASCI Teraflops Supercomputers

For large scale parallel computing, Teraflops supercomputers ASCI Red at Sandia has 9000 Intel P-Pro-200 processors, while ASCI Blue Mountain at LANL is based on SGI Origin 2000 clusters, and ASCI Blue Pacific at LLNL is based on IBM SP-2. Development and progress are being made from Teraflops supercomputers towards Petaflops super-supercomputers.

3 Linux Operating System

Linux is the Operating System of choice for both the non-HEP Beowulf projects and the HEP PC Farms. In recent years, there has been a very large UNIX base in high energy physics and other scientific fields. Linux is "generic UNIX", and enables easy, quick migration from UNIX workstations to PC computing. Http://hepwww.ph.qmw.ac.uk/HEPpc/ is a useful reference with many links.

Linux [4] is an open source, well supported, widely used Operating System with multitasking capability. Most of the Linux system infrastructure (libraries, compilers, utilities) outside of the kernel comes from the Free Software Foundation (GNU). Codes written in the C language can be compiled using GNU

CC (gcc), and Fortran codes are often compiled using g77. Other compilers are also available (e.g. Microway, Absoft, f2c converter, g++, egcs, and KAI). Linux runs on many platforms, including Intel 386, 486, Pentium, P-Pro, P-II, DEC Alpha, MIPS R4x00 and SPARC Fujitsu(AP+).

Many of the software packages that physicists use as part of everyday work are available for Linux, including TEX/LATEX, editors (Nedit, EMACS), TCP/IP, X11R6, Netscape, HTTP, FTP, Telnet, SLIP, etc. Versions of Cernlib with different compilers (v95b (f2c), v96a(g77), v96a_ absoft, v97a, v98, v99) are available from CERN (ftp://asisftp.cern.ch/). Various pieces of the Fermilab UNIX environment have been ported as well, including UPS (a configuration and product management system), UPD (a product distribution system), and CPS (Cooperative Processes Software). Porting of software to Linux has in general been straightforward. Linux is now used in many experiments. In most cases, the porting takes only a "few weeks".

3.1 Physics applications

Data analysis in high energy physics often require lots of Monte Carlo studies, which need large samples of generated events with detector simulations. Currently PYTHIA, ISAJET, GEANT, VECBOS, NJETS, JETRAD, DYRAD and other programs can all be run on PCs under Linux. PC Farms can be an excellent and cost-effective platform for Monte Carlo studies.

In CDF, for example, most of the offline codes, enough for Run I Production (UNIX in Run I), event display, and Run II codes have been ported to Linux. Most modifications are on OPEN() statement and system calls. The byte-swapping from UNIX is similar to reading/writing VAX format. We used the # derivatives of C Preprocessor (cpp). The g77 Fortran compiler automatically calls cpp to preprocess *.F before compilation. Linux and PCs are now widely used for the development of Run II CDF codes, which are mostly in C^{++}. The collaborators will be able to do analysis from their PCs at home or in offices.

We also use PC Farms for online level 3 triggers. The open source nature of Linux has been particularly useful in testing, understanding and improving the Level 3 PC Farm.

By Autumn '99, Fermilab has built the test cluster PCQCD for the more tightly coupled lattice QCD calculations. The goal is to replace/upgrade the ACPMAPS [1] system by at least an order of magnitude, with a cluster of thousands of PCs.

4 Price/Performance

The main reason for PC computing is its high performance at low price. In CHEP '97 and the Pentium-Pro Cluster Workshop, the PCs had P-Pro 200 MHz and DRAM (66 MHz) at $4 /MB (vs. $40 /MB for UNIX Workstations). In CHEP '98, the PCs had 400 MHz P-II and SDRAM (100 MHz) at $1.5 /MB. The price/performance came down by a factor of 2 between CHEP'97 and CHEP'98. By Autumn '99, the price is less than $4 per MIPS (CPU equivalent to a VAX-11/780, which used to cost about $50,000 in 1985). The Price/Performance is decreasing at the rate of about 1% per week.

Our tests, as well as tests done by others, show that the performance of the newer P-II and P-III chips scale with the clock speed. The 400 MHz P-II is twice as fast as the 200 MHz Pentium-Pro. The doubling in speed every 12-18 months is expected to continue. The 700 MHz P-III chips are available now at the end of 1999, with GHz chips planned by Intel for the year 2000.

4.1 Dual Processor PC

A dual processor PC running Linux, which has Symmetric Multi-Processing (SMP) capability, is twice as fast as single processor PC, for HEP Monte Carlo or CDF Data. The 200 MHz Pentium-Pro chip has a processing power of 113 MIPS as indicated by timing studies using our benchmark programs Pythia and Geant, "Tiny", and the CDF offline reconstruction code. For comparison, the processing power of a R4400 node on SGI Challenge was 83 MIPS. The dual Pentium-Pro 200 MHz PC had the same combined speed as SGI R10000 (which has about 2.5 times the speed of R4400). By February '97, the Price/Performance of dual Pentium-Pro PC was more than a factor of 3 better than a R10000 workstation.

It would be great if PCs with higher number of processors per node could be competitive in Price /Performance, since multi-processor nodes could reduce floor-space and system maintenance, and even improve performance. For now, the dual processor PCs offer the best Price/Performance.

5 PC Farms in HEP

Combining a large number of PCs provides the computing power of super computers at a small fraction of the cost. Similarly to the non-HEP Beowulf systems, the scale of the PC Farms and PC computing in the high HEP and nuclear physics community have increased significantly since '97.

5.1 DESY

The HERMES experiment has been successfully running a Linux Production Farm since Sept '96, with 10 Dual P-Pro PCs. This was one of the earliest PC Farms. The ZEUS experiment has 20 Linux PCs, which processed 30% of ZEUS data (10 Million events) in '97. The system reprocessed data during the '97-'98 HERA shutdown, and has also been used for data analysis. The HERA-B experiment has 100 P-II Linux nodes installed for Second Level and Third Level Triggers. The Fourth Level Trigger currently has 10 Linux nodes, with a goal of 100-200 nodes.

The Linux-PC Farms at DESY Zeuthen have been running since February '98. The three farms from AMANDA, H1, and HERA-B have 40 PCs, which doubled the CPU for the experimental groups. The installation of Linux and Post-Installation for setting up a farm PC (e.g. batch system, file server access) and for setting up DESY-Zeuthen environment (e.g. AFS, user environment) are all via network.

5.2 RHIC at BNL

The 4 experiments BRAHMS, PHENIX, PHOBOS, STAR starting in 2000 will record 1.5 PB of raw data per year, requiring 10,000 SPECInt95 (500 P-II 400 MHz) of CPU. The Mock Data Challenge started in Sept. '98, using Linux PC Farm with 40 dual Pentium II is a 10% prototype for the full system. The Linux PC farms will be based on either the commodity PCs, or custom 8-processor nodes being evaluated. The in-house control software is written in perl. Remote Management of individual PCs and a console multiplexing system, as well as a batch processing system are available. RHIC Computing Facility (RCF) is planning 3 additional farms: Central Reconstruction Server Farm, (part of) Data Mining Farm, and (part of) Central Analysis Server Farm.

5.3 CERN

Most experiments at CERN still rely on UNIX workstations, since LEP experiments have been running for many years, and plan to stop running in 2000. The PC based Simulation Facility running Windows NT (PCSF) began in Summer '97, with 35 dual processors in spring '98. Some tools have been developed to deal with this OS. Scalability issues are being studied. There are also Linux PC Farm and Linux PC clusters, which are used in Chorus, NA48, NA49 and other experiments (wwwinfo.cern.ch/dci/linux/).

The CPU requirement for the Atlas experiment has been estimated (J. Knobloch in www.ifh.de/CHEP97/) for Event Reconstruction, Monte Carlo Production, and Physics Analysis to be 7×10^4, 5×10^4, and 15×10^4 SPECint95, respectively. The Online triggers will have even higher CPU requirement. Similar numbers are expected for CMS. Online and offline computing for LHC-B will require thousands of Linux PCs. PC Farms and PC computing are clearly the leading solutions to such CPU challenges for LHC.

5.4 Other labs

At KEK, the Online groups have provided software for Linux (VMElib). The Belle experiment officially supports and has ported their software to Linux. K2K and other experiments use Linux for DAQ. The JLC software also runs on Linux PCs.

At SLAC, for BaBar, the officially supported Operating Systems have been AIX, Solaris, OSF, and HPUX. Both the BaBar and NLC software have been ported to Linux.

At Jefferson Lab, the 3 experimental areas record 1 TB of raw data per day. To process this large amount of data, the Linux PC Farm had 50 dual P-II nodes by the end of '98. The PC Farm has $\geq 85\%$ usage. There is also a small Linux PC interactive cluster.

Argonne announced in November '99, the "Chiba City", scalable Linux test bed dedicated to open source software development. This Linux cluster has 256 dual processor computational nodes from VA Linux Systems, and IBM Netfinity servers for cluster management, file storage, and visualization.

There are also small Linux PC farms/clusters at LBL and many universities and other labs.

6 Fermilab

At CHEP and the Pentium-Pro Cluster Workshop in April '97, Fermilab and CDF presented PC Farm benchmarks based on studies using Pythia, Geant, CDF Run I Production, and the CDF Event Display. At the Super Computing Conference in November '97, Fermilab and CDF exhibited a CDF Offline

Production PC Farm and a CDF Online Level 3 PC Farm. For the Online Farm, the initial discussion was in January, the hardware was ordered in June and arrived in August. The system with simplified Run I Level 3 code ran successfully by October '97. The Demo systems ran at San Jose Convention Center (trucked to San Jose and back) and ran in Fermilab HighRise lobby during Dec '97 - Feb '98. The systems are robust. By '97, the Fermilab Theory Group, and other experiments such as E815, E871, and SDSS also took advantage of PC computing using Linux.

In January '98, Fermilab's Computing Divison announced official support [9] for Linux. The Computing Divison selected /incorporated the RedHat distribution, and makes recommendations and guide lines for Fermilab users on major aspects of the Linux PC computing.

6.1 CDF Level 3 Trigger

With the 20 fold increase in luminosity and data for Run IIa in 2001 - 2003, the online Level 3 trigger will need to input, select, and output data at rates about 20 times Run I rates. The rates and data will be another factor of 15 higher for Run IIb in 2004 - 2006. The Level 3 trigger for Run I required 3,000 MIPS and was based on 4 SGI Challenges. PC Farms provide a cost-effective solution to online Level 3 challenges for Run II and beyond.

I/O performance may be a major concern for Level 3 PC Farm configurations. CDF has demonstrated and adopted the solution of using ATM switch distributing data to Linux PC Farms. The ATM switch has 16 inputs and 16 outputs. The event size is expected to be 200 KB. It has been measured and demonstrated that each output can perform at 20 MB/s (while the peak requirement is 3.8 MB/s), for a system total of 320 MB/s. The Level 3 Trigger using PC Farms can process 5 times the Run IIa TDR rate requirement. The output rate will be limited mainly by the offline storage, accessing and analysis of data. The system can easily be upgraded.

The open source nature of Linux has been valuable, especially for the Level 3 system, to test, understand, and improve the performance. The system has also been demonstrated to be stable. The system with 50 dual PCs had been installed by Summer '99 to further test the I/O, stability, and scalability of the system, and to exercise Level 3 trigger software. Scaling up the number of nodes and total CPU for Run IIa and IIb is expected to be straightforward.

6.2 D0 Level 3 Trigger PC Farm

For Run II, D0 Level 3 group plans for 50 dual or quad processor nodes, each node with 128-256 MB memory and 2-4 GB local disk, running Windows NT. This is one of the very few PC Farms that plans to use NT. The system should be able to process at an input rate of 1500-2000 Hz (designed for 10,000 Hz for Run IIb). The output rate is expected to be 50-70 Hz, with event size of 256KB.

6.3 Fermilab Plans for Offline Farms

Offline Production for experiments at Fermilab, including CDF and D0 Run II, will be accomplished by Linux PC Farms. In Summer '98, Fermilab started De - Commissioning UNIX workstation farms, and Commissioning Production PC Farm with 64 P-II 333 MHz processors for E871. The CDF / D0 / Theory Run II prototype farm started with 18 Dual P-II 400 MHz processors, each with 256 MB SDRAM, including an evaluation of Gigabit Ethernet. In the autumn of '99, 3 PC Farms each with 50 dual 500 MHz P-III nodes, for a total of 150 nodes and 150 GHz, were installed for CDF, D0, and other experiments. These PC Farms will be several times larger in Run IIa starting in 2001.

Several local "white box" companies have participated in the Fermilab purchasing/bidding process for the PC Farms. In the bidding and purchasing, Fermilab has specified hardware components such as the processor and motherboard. The PCs were delivered in 3 weeks, preinstalled with Fermi Linux, and other CERN, Fermilab, HEP software. The PCs were put together, including network, as a PC Farm within a few days of delivery.

6.4 Linux at Fermilab

The instructions for distribution and installation of Fermi Linux are available [9]. The software distributions are by free CD or download via network. CDF Run II software distribution is by similar methods. Fermilab users can also order PCs using the above web site, make hardware choices, and charge the order to a budget code. The PCs can arrive in about 3 weeks, optionally with Fermi Linux pre-installed. Vendor support is available (e.g. contract with Red Hat).

To help improve the usability of PC computing, the Computing Division offers Linux Administration (day long) classes. The monthly UNIX Users' Meeting has become the Linux Users' Meeting.

7 PC Computing

By the end of 1999, there are dozens of PC Farms in HEP and dozens of non-HEP Beowulf systems, each with tens to hundreds of Linux PCs. The PC Farms and the Beowulf systems worldwide continue to grow in size and applications.

The number of such systems will also dramatically increase. Software tools and manuals for clustered computing, both collected and developed by the NASA Beowulf Project, have been packaged as the Extreme Linux (www. extremelinux.org, and www.redhat.com/extreme). Extreme Linux allows a person/group to set up a computer cluster for parallel processing, with the OS and tools that are contained in a $29.95 CD-ROM. A person can have PCs for a "Personal Super Computer".

In addition to the PC Farms and Beowulf systems, desktop computing is also migrating from UNIX workstations to PCs. Many related issues are being solved, e.g. clustering, batch system, I/O, even via wireless communication. Physicists will soon be able to easily analyze data using his/her home and/or Notebook PC.

7.1 Recent port and support for Linux

Linux is officially supported by the leading companies providing database software, Oracle, Informix, Computer Associates, Sybase, IBM, Objectivity, and other companies such as Netscape, Corel, Interbase, Adaptec, Cygnus, Sun, SGI, DELL, Compaq. In September '98, Intel, Netscape, and other companies acquired equity positions in the Red Hat company. The strong support from such leading companies greatly strengthens future development of Linux. In August '99, Red Hat had a very successful IPO, followed by VA Linux Systems' IPO in December. Both Red Hat and VA Linux are now multi-Billion $ companies.

8 Conclusions

PC Farms are a cost-effective solution to future CPU needs. The commodity hardware has advanced to provide high performance at low price, and excellent upgrade path. Linux is an excellent Operating System for the PC hardware platform. PC Farms will be used for Monte Carlo studies, Offline data Production, and Online Trigger event selection. General PC computing will also benefit from, and contribute to the development of the PC Farms. These solutions are now well accepted by the high energy physics and nuclear physics community, and by many other scientific fields. The development of PC Farms and PC computing has been important and holds promise for a great future.

9 Acknowledgement

We thank the Fermilab PC Farms group, the CDF Taiwan group, and the CDF Level 3 group for their pioneering work on the PC farms. We thank Linus Torvalds, Linux and GNU developers worldwide, and the open source community for providing Linux and other software. Thanks also to the Fermilab Computing Division, DESY, RHIC, CERN, and other labs for strong support for Linux and PC computing.

References

[1] www-hppc.fnal.gov/acpmaps/acpmaps.html

[2] www-cdf.fnal.gov/cd/gp.html

[3] Super Computing '96, Pittsburgh; Computing in High Energy Physics CHEP '97, Berlin; Pentium-Pro Cluster Workshop, April '97 Iowa; www.hep.net/chep98 Chicago, Sept. '98

[4] www.linux.org and www.redhat.com

[5] Maddog Jon Hall, "a Geek in Paradise", Linux Journal, July, 1999.

[6] Bob Young and Wendy Goldman Rohm, "Under the Radar", 1999.

[7] Geist, Beguelin, Dongarra, Jiang, Manchek, Sunderam, "PVM: Parallel Virtual Machine", MIT Press, 1994.

[8] www-unix.mcs.anl.gov/mpi/index.html

[9] www.fnal.gov/cd/unix/linux