Proposal to the SSC Laboratory for Research and Development for a Parallel Computing Farm

L.D. Gladney, N.S. Lockyer, and K.J. Ragan
University of Pennsylvania, Philadelphia, PA 19104

J.G. Heinrich and K.T. McDonald
Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544

Justin Rattner
Intel Scientific Computers, Beaverton, Oregon, 97006

(September 29, 1989)

Abstract

A study of large, concurrent microprocessor farms for use in SSC detector data acquisition systems is proposed. The feasibility of a teraOPS parallel supercomputer connected with an inter-node mesh network with a bandwidth of $100 - 2000$ gigaBytes/sec is addressed. A pilot farm of eight ultra-high-speed processors would be used to develop expertise and serve as a focal point for collaborative efforts between industrial and university researchers. This proposal will allow us to take advantage of an ambitious multi-year program of collaboration already underway between Intel and DARPA at a time when the high energy physics community can influence the directions of this rapidly evolving industry.

1 Executive Summary

The online and offline computing needs of SSC experiments will exceed several trillion operations per second (teraOPS). It is estimated that computing needs during proposal preparation exceeds that available at Fermilab even if all resources were dedicated to a single experiment. Ultra-fast, highly parallel processing farms, presently being developed and soon to be delivered by industry, represent an ideal start to the solution of this problem. The joint Intel/DARPA iPSC/2 Touchstone Project is a $27.5M program which will deliver multi-computer systems with combined performance of $\sim$ 100gigaFLOPS by 1991. The high bandwidth 2-dimensional mesh networks designed by Intel will have a major impact on applications. For example, it may obviate the need for a conventional event builder. Another Intel/DARPA program is developing specialized processors to use in hybrid processor farms. Equally important is the software, and Intel is investing heavily in exploring new and more efficient programming tools for large-scale parallel systems.

We propose the following goals for a research and development program:

- begin a collaboration with Intel Scientific Computers Corporation aimed at refocussing the resources of industry on problems specific to SSC triggering and data acquisition.
- initiate a pilot farm project which can study trigger algorithms and concurrent processor code development, Monte Carlo generation, and detector simulation.
- study issues relevant to setting minimum requirements for inter-processor communications.
- initiate cooperative efforts with university computer-science departments on developing software tools for managing parallel-processor systems.

2 Introduction

2.1 General Comments

Experience with both the hardware and software required to operate a large-scale microprocessor farm are necessary in order to refine the conceptual design of a farm that will fit into a high-rate data acquisition system. There is little practical experience in the high energy physics community with concurrent-processing techniques. A program to convert physics algorithms into code that can run "in parallel" on many separate processors and to debug such codes while in operation on the nodes of a farm will be crucial in refining the conceptual design of the farm architecture for several reasons:

- it will aid in developing expertise with technology for high-rate data transfer between the detector and farm.
- it will set requirements for communications between farm processors.
- it will define the amount of CPU power needed to achieve a rejection in the trigger from a particular algorithm.
- it will specify requirements for software packages needed to monitor, control and debug the farm.
- it will specify those data structures that will be most efficient for transfer of data to offline mass storage systems for subsequent offline processing.

2.2 Rationale for Industry, University Collaboration with SSC

The Defense Advanced Research Projects Agency (DARPA) is funding parallel computer projects at levels well beyond the means of the SSC laboratory. In particular Intel Scientific Computers Inc. has been funded to develop a prototype massive parallel computer that will be ~ 100 times faster than any existing computer. The computer can have up to 2048 nodes and uses the Intel i860 processor chip. The i860

has been generally recognized as the leading RISC (Reduced Instruction-Set Computer) device on the market today. It's computer power is roughly the equivalent to the Cray Research Inc. Cray-1 supercomputer. The system will provide about 100 gigaOPS of sustained computing, roughly 100 times that of a Cray Y-MP, the worlds present fastest supercomputer.

3 Overview of Processing Requirements at SSC

3.1 Introduction

As an example of the need for large scale computing at the SSC, we consider a B-physics experiment[?]. This experiment must exploit the high luminosity of the SSC to accumulate the large event samples needed to study CP violation. The raw $\bar{b}b$ rate is about 50 KhZ at 10^{32} cm^{-2} sec^{-1}. For rates this high, a teraOPS computer farm will be required to select the interesting B events in real time. The experiment cannot succeed without this filtering engine.

3.2 Intermediate P_t Physics-BCD

The SSC is a prolific B factory and represents the ultimate machine for the study of CP violation in the B system. The interesting branching ratios are $< 10^{-5}$ and a very high rate experiment is needed in order to collect a sample of 1000 B's in several modes. This physics often requires that both the $\bar{b}$ and b are fully reconstructed, further emphasizing the need for high rates since reconstruction efficiencies will be of order 1%. CP violation has great importance within and beyond the standard model and is a subject of much interest for theorists and experimentalists.

- Cosmological models that try to explain the matter anti-matter asymmetry in the universe usually invoke CP violation and some Grand Unified Theory. Thus the existence of the universe is thought to be related to CP in some way.

- Multiple Higgs bosons can lead to relative complex phases that in turn have CP violating effects.
- CP is important to the generation puzzle. Of the 21 free parameters in the standard model, 18 are related to the fact that we have 3 generations of families. If there were two families, no complex phase would exist in the standard model and thus there would be no simple explanantion of CP violation. With 3 families, there is one complex phase, and this is consistent with present data. If there are 4 families, then there are 3 complex phases and new CP phenomena are expected.
- Measurements of CP violation in the B system determine C-K-M angles with little strong-interaction uncertainty. In the study of several modes, the C-K-M system can be overconstrained.
- Left-right symmetric models predict smaller CP violating effects in the B system than does standard model.

3.3 Hi-P_t Physics

Several SSC workshops have undertaken studies of the "keynote" physics topics which will be of interest at the SSC. These include searches for

- new quarks and leptons
- the Higgs boson
- new W's and Z's
- exotic new particles, including supersymmetric particles.

All these physics topics can be addressed by detecting the basic building blocks of nature as we currently know them, i.e. detection of quarks, leptons, and bosons. There is also a strong belief that any new particles or interactions which have not been considered will also couple to these "standard" particles. Thus, the principle

goal for a Hi-P_t detector will be to detect, and identify as well as possible, quark and gluon jets, electrons, muons, neutrinos (through the detection of missing transverse energy), and W and Z bosons.

Typical Hi-P_t detector designs emphasize detection of quark and gluon jets, and muons over a pseudo-rapidity range of ± 5, and detection of electrons over a rapidity range of ± 3. To achieve the position and momentum resolution deemed necessary for the physics goals over such a large region of phase space will probably demand detectors with at least 1 million channels of electronics and event sizes of order 1 megabyte. This fact, when coupled with the very general nature of the physics goals envisioned for such a detector, means that the event-building and farm-computation requirements for the Hi-P_t detector will likely be no less than that of the intermediate-P_t detectors. Thus, research on the highest level of the data acquisition system, namely data input into and data processing by the microprocessor farm, will be applicable to both the intermediate-and Hi-P_t -detector research and development programs. Rates for data throughput, event sizes, complexity of physics algorithms, etc., are all far beyond those within the current experience of the high energy physics community.

3.4 Triggering and Data Acquisition Overview

The trigger systems for the SSC detectors will have multi-levels and will involve massive parallel computing. There were two triggering approaches presented at the SSC Triggering and Data Acquisition Workshop held in Toronto, January, 1989. One scheme for a HiP_t experiment like the SDE, and another for an intermediate P_t experiment like the BCD.

The HiP_t experiment will run at the highest luminosity. The interaction rate at a luminosity of 10^{33} cm^{-2} sec^{-1} is 10^8 events per second. The Hi-P_t experiment will have a rejection factor of about 100 at the first stage (level 1) of triggering. SDE will require about 1μsec to prepare for the level 2 trigger at which point an

asynchronous rejection of about 10^2 events per second will begin. The event rate into a level 3 processor farm would be about 10^4 events per second. The Hi-P_t experiments will emphasize greater rejection at the hardware trigger level than is possible by the BCD.

The $\bar{b}b$ events are produced with a P_t scale set by the b mass. The B events are similar to minimum bias events, making the trigger selection very difficult. Consequently the trigger rejection at level 1 is estimated to be about 50. After this the data are transferred from the detector to the processor farm. If a luminosity of 10^{32} cm^{-2} sec^{-1} is delivered to the BCD, the rate into the farm would be 200,000 events per second. If the rate to permanent storage is a maximum of 1000 events per second, a rejection of 200 is necessary from the farm algorithms.

3.5 BCD Data Acquisition Requirements

The BCD online and offline computer is considered the same machine. We present two estimates of the scale of this problem. We estimate it is necessary to generate, simulate and analyze 10^6 events per week to study B-decay modes with branching fractions of 10^{-5} or smaller. Essentially all CP-violating modes have branching ratios of this order. Consider first an estimate of the amount of computing necessary to prepare an SSC detector proposal.

1. A simulation of $B \to \pi^+\pi^-$ using only the silicon vertex detector and a beam pipe required 24 CPU hours of a 30-mips machine to simulate 10^5 events. Eight physical days were spent waiting in queues and running the jobs. The BCD has at least 7 major subsystems to simulate. The simulation complexity grows as some power of the number umber of systems involved, probably between a power of 1 and 2. Taking the power as 1.5, we would need about 20 times as much cpu power per event for the full detector simulation. That is, 20 days of a 30 mips machine to simulate 10^5 events, or 200 days to simulate 10^6. We want to do the 10^6 events in 1 week, which implies a need for 30 times as much

cpu power, or 1000 mips.

2. This estimate is based on tracking times in CDF and MarkII. A tracking device with ~ 100 samples requires ires about 300 msec on a 1-mip machine to analyze each track. As we must provide for 100 tracks per event, this implies 30 sec per event on a 1-mip machine for tracking analysis only. Our experience with GEANT indicates that the time to simulate an event is about 4 times that to analyze that event. Then we would need about 150 sec per event to simulate and analyze the tracks on a 1-mip machine. If we also suppose that the tracking is the most complicated feature of the simulation, we estimate that the 'full' simulation requires twice as much time, or 300 sec per event on a 1-mip machine. Then 10^6 events would require about 3×10^8 seconds on a 1-mip machine. A week is 6×10^5 sec, so to accomplish the simulation of 10^6 event in 1 week, we would need 500 mips.

Thus we arrive at preliminary estimates of the cpu power needed for BCD detector simulation in the range of 500-1000 mips. We estimate the online cpu power needed as 10^6 mips. s. The figure of 10^6 mips (= 1 tips or tOPS) is based on our ambition to process 10^5 events/sec online as our second-level trigger. This gives us 10^7 instructions per event for the online processing (= 10 sec on a VAX 780). t this

4 Technical Discussion of the Farm

4.1 Overview of Intel/DARPA-Touchstone Project

The Intel Parallel Supercomputer is a collection of high-speed processors interconnected by a large-bandwidth network. The system offers the capability of accepting data in any processor node, performing local processing, and sending data directly to any other node in the system, including those for mass storage. The development of special-purpose interfaces to these nodes, improvements to the processors, I/O systems, and inter-processor network is underway at Intel and has been

funded at $27.5 as part of the DARPA Touchstone project.

The problem in assembling a useful, powerful, large parallel-processor supercomputer is not the processors themselves, but rather the interconnect hardware and system software. The Touchstone project brings together industry and university resources to solve these problems. MACH, a U.C. at Berkeley Unix based operating system, developed at Carnegie-Mellon University, will be the starting point for the i860 distributed-memory architecture. Princeton University has developed shared-memory technology for farms and this will be employed by Intel. The University of Illinois is developing the tools to monitor the performance of this large parallel system. The node-to-node communications is funded separately by Intel-DARPA to Massachusetts Institute of Technology and Caltech. The goal is to have inter-node communication at the 100-200 MegaBytes/sec rate. This 2-D mesh net architecture is a result of the thesis of W.J.Dally[1] of CalTech, now at M.I.T. and his advisor, Professor C. Seitz. This new network architecture has now replaced the Hypercube architecture (3-D) presently used by many parallel-processor supercomputer companies. We have contacted Dally and Seitz and there is interest by them in involving a thesis student, M. Pertel, in studies of network architectures tailored to the needs of high-energy physics. a very interesting problem. We intend to pursue this collaboration.

4.2 Status of BCD-Intel Collaboration

Generic SSC R&D funds were provided to N. S. Lockyer, E. Barsotti and J. Morfin et al. We have been simulating benchmark programs supplied by Irwin Gaines of Fermilab. All three Fortran programs compiled the first time with the i860 Greenhills Fortran compiler. The codes have been run under the i860 simulator/debugger on a 80386 processor and we intend to report results soon. The actual i860 computer has a delivery date of Oct. 25. 1989.

4.3 Technical Overview

4.3.1 Possible Network Architectures

The SSC Farm System components are shown in figure 1. Initial BCD/SSC data-flow simulations, performed at Beaverton by Intel scientists, indicate that there are two system architectures best suited to the SSC environment. First, the Injector Model is shown in Figure 2. There are three types of nodes, injector, processor, and the I/O. The detector injects event records directly into the system network through a number of specially designed buffers or injector nodes. These buffers balance the burst speed of the detector against the average speed of the network, and route each injected message to the appropriate (predesignated) processor. The injector node would take the place of a processor in an an otherwise completely-connected inter-processor network. Injection of data at the network boundary is shown in Figure 2. The processor nodes are the i860 computers. The I/O node allows output to a permanent storage device or input for testing and debugging the network.

The second approach is called the Distributed Model and is shown in Figure 3. A special interface connects a selected number of processor nodes to the detector. This interface would allow standard processors to act as injectors as well as computation nodes. This would give a great deal of flexibility in setting up the optimal configuration for the network.

Initial simulations by Intel have begun that address whether these two network models can be used to build events at required 200kHz rate.

4.3.2 System Software

An important aspect of the system software is scalability. The Intel hardware and software changes will not affect application-software. The iPSC/2 system software maintains all the features currently available for farms using the i386 processor. It includes debugging facilities for processor nodes and the network, monitoring of message traffic, and gives special emphasis to minimum latency and maximum

throughput. The details of the network topology are normally hidden from the user. Any node can send messages to any other node with very small differences between sending to neighbors or more distant nodes. It is this feature that raised the issue of whether the farm network could replace the conventional event builder.

4.3.3 Proposed SSC 1995 Configuration - Farm Architecture

It is estimated that between 2,000 and 20,000 processing nodes will be connected using a 2-dimensional mesh network offering between 100 and 200 MBytes/sec in each direction. The specific number of nodes will depend on the peak event rate and the result of algorithm timing studies. The combined processing capacity, between 200 gigaFLOPS and 2 teraFLOPS, and the combined internal network capacity of between 300 gigaBytes/sec and 3 terabytes/sec should adequately accomodate the 200KhZ, 1 MByte/event, data rate into the farm. The File system described below can store roughly 1% of the event rate. The nodes will be arranged into a logical rectangle with a predetermined routing algorithm designed to evenly distribute the event traffic.

4.3.4 Event Builder Architecture

It is estimated that between 200 and 1000 processing nodes would be fitted with a special-purpose connection to the detector output streams. This connection would probably involve fiber-optic channels with the same capacity as the mesh network and terminate in an on-board buffer that feeds the local memory directly. A store-and-forward driver will route messages into the network if not targeted for the injector. These injector nodes would be spaced throughout the rectangular processor array in a pattern to be determined by the results of this proposal.

4.3.5 File Architecture

Roughly 200-500 I/O nodes will be incorporated into the processing array to manage the online storage and user interaction with the farm. These special-purpose I/O

nodes will drive SCSI-based storage devices (magnetic and optical) each offering a total storage bandwidth in excess of 2 gigaBytes/sec. The number of nodes would depend on the amount of data being recorded for each interesting event and the amount (and location) of subsequent processing. Intel expects this portion of the system to augmented in several ways.

1. Incorporation of special graphics nodes to assist in visualization of events.
2. Incorporation of high-speed LAN and GAN controllers to facilitate online removal and copying of event records to other computer systems and workstations.
3. Incorporation of additional processing nodes devoted to correlation of event records that match specific characteristics and transferring to permanent storage.

5 Statement of Work

5.1 Introduction

This proposal will answer the following questions based on simulations of the network and tests on the pilot parallel processing farm. First, we discuss the farm configuration.

5.2 Initial Configuration of the Farm

In the first year of this proposal, we propose to purchase a pilot farm. In particular, we envisage the following initial configuration:

1. A Front-end system consisting of an 8 MByte, 80386 UNIX V.2 system with 380 MByte disk, streamer tape drive, NFS and TCP/IP Ethernet interface, 6-port terminal interface, three external modems, FORTRAN and C compilers, VAST 2 vectorizing FORTRAN preprocessor, and assembler.

2. An eight node i860 system with mesh architecture, with each node supporting up to 8 MBytes of local memory.

3. A Concurrent File system with two 80386 disk controllers, each with two 760 MByte disks and two 80386 8 MByte I/O nodes with VME interfaces to be used for input channels. This system can be housed in a 19-in. cabinet.

An additional front-end system would be housed at Princeton, and a high-speed line would connect them to the computer.

5.3 Issues

- Does a farm of 1,000-10,000 nodes with a computing power of a TeraOPS appear technically feasible by the year 1997?
- Will this farm be able to handle data rates of 300 gigaBytes/sec directly from a detector?
- Can the sophisticated new 2-D mesh network architectures with wormhole routing being pursued by the Intel Corporation and academia replace the function of the high-speed event builder? The challenge of combining perhaps 1000 parallel data streams into an event of 1 Mbyte each at rates of 100 kHz or higher is an exciting new application for farm architectures. We look forward to working closely with the Intel group on this project. Should design studies prove favorable, we foresee the need for a hardware simulator of the detector data streams to be fed to input nodes of the farm. We envisage building such a device in the second year of the R&D program, and do not request funds for this now.
- Can currently envisioned filter algorithms running on parallel processors provide sufficient rejection online to meet trigger demands?
- Can the graphics features of the i860 be used to provide high-rate, real-time graphics for use in SSC experiments' control rooms and for data analysis?

- To gain practical experience with an industrial farm and its software, we will generate, simulate, and analyse 50 million events to study detector design issues in the next year. Estimates based on our partial simulations indicate that tracking studies alone will require 250-500 mips to achieve a throughput of 10^6 events per week. As with event analysis, event simulation is well suited for a network architecture, and will require development of software organizational tools also needed for the eventual event analysis.

At the University of Pennsylvania, L. Gladney and K. Ragan, will investigate the algorithm issues. Initially, we would learn how to convert present FORTRAN source code so that it will run on a concurrent-processing system such as the Intel iPSC/2 system. We would begin by incorporating trigger algorithms that are already in use in dedicated trigger processors for the CDF experiment into the concurrent-processing system. One such algorithm is a simulation of the CDF fast track-finding system based on track segment location in CDF's large central drift chamber. This algorithm locates Hi-P_t track segments by searching for hits which follow the wire vs. time patterns expected for straight (i.e. large P_t) tracks which have crossed the supercells of the chamber.

The University of Pennsylvania physicists have built a specialized, hardware processor that locates Hi-P_t track segments by identifying such patterns, then links the segments to form full tracks for use by the Level 2 trigger of CDF during the 1991 Tevatron run. This fast track processor is implemented in ECL technology and runs with a clock speed of 50 MHz, taking approximately 5 microseconds to locate track segments with transverse momentum above 5 GeV/c in a single supercell. Full track reconstruction using segments from all superlayers takes roughly 10 microseconds, depending on the complexity of the event, and will achieve a momentum resolution of $\Delta P_t/P_t = 1\%$. The tracking information will be combined with calorimeter clusters and muon chamber hits found by the CDF Level 2 trigger in order to identify electrons and muons on a time-scale of less than 20 microseconds.

We will transfer a software version of our track processor algorithm to the iPSC/2 system to test its effectiveness and speed with the 40 MHz i860 general purpose processors. Two-dimensional track reconstruction with good momentum resolution can produce a factor of 1000 reduction in fake electron trigger rates when combined with calorimetric information (ref. [2]) for Tevatron event rates. Successful implementation of an algorithm that reconstructs charged tracks in < 10 μsec would impact heavily on Level 2 electron trigger designs for the SSC.

5.4 Responsibilities and Commitments

5.4.1 Summary-Physicists

The overall system issues will be addressed by the Intel Corportion, K. T. McDonald and N. S. Lockyer. The system issues include the event-builder feasibility study using the new 2-D network. The study of algorithms will be done by L. D. Gladney and K. J. Ragan. The study of BCD simulations will be done by J. G. Heinrich, K.T. McDonald, and N. S. Lockyer. A study of issues related to data acquisition with a Pixel vertex detector is also planned. J. G. Heinrich, K. T. McDonald, and N. Lockyer will devote significant time, up to 40% of their research time to this project. L. D. Gladney and K. J. Ragan will devote 30% of their research time.

5.4.2 Commitment of the University of Pennsylvania

It is planned that the system will be located at the University of Pennsylvania. If this proposal is approved the university will provide:

- A one-half time system programmer.
- 600 square feet of floor space for the system.
- Fire protection and climate-controlled environment.
- Three Unix workstations.
- Dedicated T1 phone line to Princeton University.

- System hardware and software maintenance.
- Two thesis students from the Distributed Systems Laboratory, Department of Computer Science, University of Pennsylvania.
- The authors have in addition applied for matching funds from the university. The maximum amount is $50,000.

5.4.3 Commitment of Princeton University

Should it become appropriate that Princeton be the host institution for the farm hardware, we can offer several support facilities. The Princeton High Energy Physics group has a long tradition of support of in-house computing facilities, including an air-conditioned computer room, full-time computer maintainance engineer, and full-time system manager.

5.4.4 Commitment of the Intel Corporation

The Intel Corporation considers the development of large processing farms a major part of their program. Intel Scientific Computers will:

- Provide expert consulting and assistance in developing simulations and models of the processor farm.
- Make available a system expert during the installation stage at the University of Pennsylvania.
- Provide access to a very large system at Beaverton for testing of algorithms and network simulation tests.
- Make available to members of this proposal the latest developments from the Touchstone project, contingent upon approval by the DARPA program officer.
- Provide a lead scientist to co-author scientific papers that summarize the work done during this proposal.

- Provide a cost estimate of the 1000-10,000 node SSC system.
- Evaluate and propose a network technology with performance simulations and make results available to the SSC laboratory.

6 Budget - FY90 and beyond

The prices are a quotation from Intel. We list the items and the estimated costs. The i860 system with a hypercube architecture is available in early 1990. The mesh network technology will be developed by fall of 1990.

1. An 8-node Parallel Supercomputer with 8 MBytes per i860 node, Fortran, C, TCP/IP network interface and documentation. This price includes one Front End System for $\$25,000$. A second one would cost $\$20,000$ extra.
   Product PSC/2 D3/RX-8. The RX implies the product is not currently available.
   Cost: $\$225,700 + \$20,000 = 245,700$
2. 1.9 GBytes of disk, an 8mm Tape unit, 2 80386 I/O nodes, concurrent file system software, 8 MBytes of memory, and housing.
   Product MIO32T1
   Cost: $\$87,300$.
3. VME interface adapter and 8 MByte I/O node.
   Product PSC386VME-8
   Cost: $\$17,300$.
4. 380 MByte disk option for the iPSC/2.
   Product PSC386MGR380
   Cost: $\$3,500$
5. Network File System software for iPSC/2.
   Product PSC386NFS

Cost: $\$10,000$

6. 8-node license to Interwork II
   Product Interwork II
   Cost: $\$10,000$

7. VAST 2 Fortran compiler preprocessor to Vectorize Fortran
   Product PSC860VAST2
   Cost: $\$10,000$.

In addition we request travel funds. The University of Pennsylvania requests:
Three trips to Oregon for 2 people at \$700 per trip or $\$4,200$.
Princeton University requests:
Three trips to Oregon for 2 people or $\$4,200$.

The total requested funds minus the university discount of 10% on the first 4 items. is $\$349,820$.

References

[1] Dally,William J., *A VLSI Architecture for Concurrent Data Structures* Kluwer,Hingham,MA.

[2] Tarem, S. and Bensinger, **Level 2 Trigger Study using CTC Information**, *CDF note No. 269*. January 3, 1985.

List of Figures