EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN/ECP 95–8
30 March 1995

THE APPLICATION OF THE T9000 TRANSPUTER TO THE CPLEAR EXPERIMENT AT CERN

R. Heeley, M.P. Ward and S. Fisher
CERN and The University of Liverpool, Liverpool, England

R.W. Dobinson and W. Lu
CERN, Geneva, Switzerland

D.J. Francis
Physics Department, Boston University, Boston, Mass., USA

Abstract

As part of the ESPRIT GP-MIMD project a network of 54 prototype T9000 Transputers has been integrated into the data acquisition system of the CPLEAR experiment at CERN to perform on-line event filtering. Initial experience with T9000s and their interconnection using C104 packet routing chips is presented together with performance measurements and prospects for the future.

(Submitted to Nuclear Instruments and Methods in Physics Research)

1 INTRODUCTION

Transputer systems have been successfully used in real-time High-Energy Physics (HEP) systems for some time, for example in the ZEUS, UA6, and OPAL experiments [1]–[6].

Data acquisition and triggering systems for experiments at the Large Hadron Collider (LHC) and other demanding applications will require large-scale parallel systems [7]–[9]. In these applications, systems will primarily be based on high-speed point-to-point serial links and switches rather than shared buses. The INMOS T9000 Transputer [10] and the associated C104 packet routing chip [11] are commercially-available integrated circuits which can be used to build large networks of the type required for future experiments.

The main objective of the ESPRIT project GP-MIMD is the design and construction of large-scalable parallel computers using the T9000 and C104. As part of this project a 54-node T9000 network, using C104 packet routing switches, has been integrated into an existing 32-node T805 real-time data acquisition system in the CPLEAR experiment [12]–[14]. The T805 system was also developed as part of the GP-MIMD project.

The T9000 network operated as a processor farm running the standard CPLEAR off-line event reconstruction program. Initial experience with this prototype system are presented, together with computational and communications performance measurements.

2 THE T9000 TRANSPUTER

The T9000 is the latest generation of Transputers from INMOS (see Fig. 1). It has a 32-bit pipelined processor with a 64-bit FPU and 16 Kbytes of cache. There are four bi-directional serial data links and a Virtual Channel Processor (VCP) allowing efficient T9000-to-T9000 communications. These components are combined onto a single integrated circuit.

The T9000 has several improvements over previous generations of Transputers in both performance and functionality. Improved performance has been gained through an increase in clock speed (50 MHz design), the implementation of an on-chip cache, and a pipelined superscalar architecture.

Improved communication is provided by the new Data/Strobe (DS) link technology which currently operates at 100 Mbits/s. Messages are divided into a sequence of packets, each of which has the structure shown in Fig. 2. All routing information is contained in the packet header. The T9000 implements a maximum packet body of 32 bytes. Any device receiving a data packet replies to the sender with an acknowledge packet. No further packets are transmitted by the sender until the corresponding acknowledge has been received by the sender.

Communication between T9000 processes is performed via virtual links. A virtual link is a single logical communication connection between two processes mapped onto a physical processor link. The VCP of the T9000 is a hardware communications processor which multiplexes the virtual links onto a specified physical processor link. Packets from separate virtual links are interleaved onto the physical link, allowing separate processes to communicate simultaneously. The virtual link to which the packet is being sent is contained in the packet header.

T9000 processors can be directly connected using their DS links or connected to a network of C104 packet routing chips, thus allowing the construction of large networks with scalable communication bandwidth between nodes. In the latter case, additional packet headers are required to route the packet through the switching network.

Two separate control links allow the T9000 to be controlled (processor initialization and loading) and monitored for errors, even when there are faults in the normal communications network. These control links may be daisy chained and/or connected via C104 packet routers.

3 THE C104 PACKET ROUTER

The C104, developed by INMOS, is an asynchronous 32-way dynamic packet routing switch with DS links operating at 100 Mbits/s. It can interconnect up to 32 devices (for instance T9000s), and may be cascaded to form large switching networks.

The technique used to route packets through a C104, to select the required output link, is that of interval labelling. In this technique each link of a C104 is assigned a range of device labels (a device interval) which indicates the physical devices that are accessible via that link. Each physical device has a unique label associated to it. When a packet enters a C104, the device label contained in the packet header is compared to the device intervals. The output link whose device interval contains the required device label is selected to route the packet out of the C104.

The C104 uses wormhole routing (see Fig. 3). A routing decision is made as soon as a packet header enters the C104. This routing decision leads to the creation of a temporary circuit through the C104 which vanishes as the packet terminator passes through. As a consequence of wormhole routing a single packet may pass through multiple C104s at any one time and the header may be received at the destination before the whole packet has been transmitted, thus minimizing communication latency.

In switching networks there will often be many possible routes that a packet may take to reach a specified destination. Should one of these links be in use or in error then it is desirable that an alternative link be chosen. To fulfil this requirement the C104 supports grouped adaptive routing. Output links can be grouped so that packets routed to the first link of a group can be routed to the other links of that group should the first link not be available.

The inter-connectivity offered by the C104 in combination with the VCP of the T9000 removes any requirement for through-routing software and virtual channel multiplexing, which was necessary with previous generations of the Transputer.

4 THE T805 SYSTEM

The T9000 processor farm has been integrated into the CPLEAR experiment via an existing 32-node T805 data acquisition system (see Fig. 4) which was also developed within the GP-MIMD project. The T805 system, which runs the Chorus distributed operating system [15], is interfaced to the experiment via a B016 module, a Transputer-to-VME interface [16]. Raw data is acquired by the T805 network via the B016 and passed to the T9000 farm, which performs event reconstruction. The reconstructed event data and raw data are transferred back to the T805 system for recording on one of six Exabyte drives. Different Exabyte drives are used for different event types.

Communication within the T805 system is through 20 Mbit/s serial-over-sampled (OS) links. Messages are transmitted between processors as a sequence of single bytes.

5 T805 TO T9000 INTERFACE

The T805 system is connected to the T9000 farm by the C100, an OS to DS link converter chip [10]. There is only minimal support available for the integration of T805 and T9000 systems, in the form of the low-level link protocol conversion offered by the C100. All communication protocols above this physical level had to be implemented in software.

The protocol used by T9000 DS links is based on the exchange of packets, whilst the OS link protocol of the T805 is based on the exchange of bytes. The DS link packet protocol has been emulated in software on the T805 in combination with the low-level DS to OS link conversion performed by the C100.

Each of the T9000 farm workers has seven virtual links to the T805 system. One virtual link corresponds to the raw data source and six to the individual Exabyte drives. Figure 5 shows how these virtual links are mapped onto the corresponding virtual links of the T805s. For each T9000, three virtual links map across to the T805 system. This is less than the seven used in the actual implementation, but demonstrates the principles involved. The tables for each T9000, in Fig. 5, show two pieces of information: the virtual link number (VL #), and the description of the use of that link. In this example virtual link 0 of each T9000 will be used to request raw data from the T805 network, and two other virtual links will be used to send back reconstructed event data. The virtual links 'Send A' and 'Send B' correspond to different destination Exabytes.

Packets leaving the T805 must have headers to guide them through the C104 network to the correct virtual link on the destination T9000. These headers must be hard-coded into the T805 receive and send process, and need to be pre-calculated for each T9000 network configuration.

When a T805 link receives a packet the header allows the identification of the virtual link on which the packet was received. This virtual link uniquely identifies which T9000 has sent the packet and which one of the three T9000 virtual links the packet was sent from. The receiving T805 process can then acknowledge and provide raw data to the correct virtual link on the appropriate T9000.

6 THE T9000 FARM

6.1 Development history of the T9000

The pre-production prototypes of the T9000 (gamma D02 versions) capable of running large programs were available in small quantities at the start of 1994. A small Parsys SN9400 system, consisting of six T9000s and one C104 packet routing chip, was installed and successfully run in the CPLEAR experiment in June 1994. In September 1994, four more SN9400 systems were added to the original CPLEAR set-up making a total of thirty T9000 Transputers (see Section 6.4).

During the summer of 1994, components of the GPMIMD machine became available. A subset of the machine comprising 24 pre-production prototypes of the T9000 nodes (gamma D02 versions) was installed, independently of the SN9400 systems, in October of 1994 (see Section 6.5).

6.2 Porting of CPREAD to the T9000

The CPLEAR event reconstruction code (CPREAD) has been ported onto the T9000. The whole package contains 230,000 lines of Fortran'77 code. The AT&T Fortran-to-C (f2c) compiler was used to translate all Fortran into C due to the present lack of a Fortran compiler for the T9000. No optimization of CPREAD was performed during this step and only minor changes were imposed due to the use of f2c.

In addition, the CERN programming libraries, used by CPREAD, have also been ported to the T9000. These include 170,000 lines of Fortran and additional C code. These libraries include data structure, histogram and mathematics packages. This work demonstrates that other large physics Fortran applications may be ported onto the T9000.

A native Fortran compiler is currently being developed by ACE for the T9000, which should improve the performance of applications written in Fortran.

6.3 Hosting T9000 systems

A T805 Transputer initializes and loads applications onto the T9000 network. It is connected to a SUN workstation via an Ethernet-to-OS link converter (B300). A single link of the T805 is used to configure the T9000 system via the control link. Another link of this T805 is used to load applications via a T9000 data link (see Fig. 6). The protocol conversion between the T805 and T9000 links is achieved as described in Section 5. Access to file systems and host system services by the T9000s is provided by the SUN workstation via the T805.

6.4 The SN9400 configuration

Thirty prototype D02 T9000 processors running at 20 MHz with five C104s operated as a real-time processing farm, performing standard CPLEAR event reconstruction and filtering. The C104s provided the exclusive method of communication and control between the T9000s. The T9000s were housed in five SN9400 units (see Fig. 7), each containing one C104 and six T9000s. The SN9400s were daisy chained with a single DS link cable. This did not affect the performance of the farm due to the relatively low bandwidth requirement of the application.

During the three-week run the system processed twenty million events in real-time. In this period of running a stable platform was achieved.

6.5 The GPMIMD machine

The GPMIMD machine is being developed as part of the ESPRIT program by Parsys (UK) and Telmat (F). It will consist of 64 T9000 processors with 56 C104s providing full inter-connectivity and is currently being assembled at CERN. Eight motherboards (see Fig. 8) will each carry eight T9000s and five C104s. Four switch cards, each with four C104s, will provide the connectivity between the motherboards. This architecture gives four independent networks, each network being associated with one of the four T9000 data links (see Fig. 9).

During a three-week period in September/October 1994, a subset of the machine processed events in real-time from the CPLEAR experiment. This machine had 24 processing nodes, on three motherboards, each running at 10 MHz.

This system was combined with the system described in Section 6.4, resulting in a processor farm comprised of 54 processing nodes and 20 C104s.

6.6 On-line monitoring

The event reconstruction code (CPREAD) produces a standard set of histograms for monitoring purposes. A process on a T9000 (a separate process for each network) was used to collect these histograms from each worker and communicate them to the SUN host, where the PAW (Physics Analysis Workstation) package [17] was used to view them.

As a method of checking the behaviour of the T9000 farm, the acceptance of CPREAD as seen by each processing node was monitored. This acceptance is defined as the fraction of events which passed the selection criteria of CPREAD. The acceptance as a function of the Worker Identifier is shown in Fig. 10. This approach provided information on the state of an individual worker and an indication of the quality of the data. More detailed monitoring was performed by using the standard set of histograms produced by each worker.

6.7 Communication requirements

Each farm worker had the following communication requirements:

- Host I/O. All workers must be able to communicate status information to the host SUN workstation
- Connection to the histogram collector
- Access to experimental data
- Access to all six output Exabytes
- Access to Calibration Data via a Calibration Server.

All these requirements were satisfied by using the C104 network in combination with the T805/C100 interfaces as shown in Fig. 11.

The host link to a SUN workstation is slow and a probable bottleneck, hence it was not feasible to have all of the workers read the calibration files over the host link. Therefore, an up-to-date copy of the calibration files was stored on a single T9000 (the Calibration Server) which was then accessed by all other workers via the C104 network. Thus, only one T9000 was required to load the calibration files over the host link.

All workers required access to the host to allow status information to be displayed. To facilitate this an I/O multiplexer was implemented. It took as input all I/O requests from the workers and multiplexed them onto the single link to the host.

7 COMMUNICATION MEASUREMENTS

The time between the sending of a single packet and the reception of an acknowledge packet between two directly connected T9000s was measured to be 7.5 µsecs. The additional delay incurred by connecting two T9000s via a C104 was measured to be 1 µsec (see Fig. 12). The curve intercepts the y-axis at 2 µsec which represents the time between the sending of a data packet and the reception of the acknowledge packet, i.e. the delay in transmitting two packets.

In Fig. 13 the dependency of the bandwidth on the number of virtual links used is demonstrated. The figure shows the bandwidth as a function of message size for one to five virtual links mapped onto a single physical link. The bandwidth represents the usable amount of data exchanged between two T9000s running at 20 MHz. The increase in bandwidth can be accounted for by the increased packet inter-leaving performed by the VCP and more efficient use of its pipelined architecture. When multiple virtual links are used, packets for different virtual links may be transmitted independently of the reception of acknowledge packets on other virtual links.

The measured bandwidths as a function of message size when using 20 and 25 MHz processors are shown in Figs. 14 and 15. In these measurements the links are routed through different C104s, i.e. the curve for four links uses four independent C104s. This configuration is dictated by the architecture of the GPMIMD motherboards. The theoretical limit for the bandwidth is 9.5 Mbytes/s on each physical link. The bandwidths measured at 20 and 25 MHz fall short of this, but there is a clear improvement from the 20-to-25 MHz processors. The problem is believed to be due to the inability of the VCP, at 20 MHz, to fully exploit the capacity of the links. If the T9000 were running at 30 MHz the VCP should be able to reach the theoretical limits for the link bandwidth.

8 RESULTS

The results reported are based on a three-week run in September/October 1994. During this period two T9000 processor farms were combined into a 50-node processing farm which processed events at a rate of 64 Hz. For the last 128 hours of the run a stable platform was maintained and no failures of the network were observed. During a total running time of 285 hours, 26 million events were processed. The initial and main source of problems was the susceptibility of the DS links to noise. This problem was resolved with suitable hardware modifications. Samples of the histograms produced by CPREAD were recorded and all output from the farm was written to six Exabyte devices. The results have been verified against the standard CPLEAR off-line program.

The feasibility of a T9000 farm processing the full CPLEAR event rate (approx. several hundred Hz) was based on measurements made on the T805 system implemented in 1993. The 50 MHz T9000 was expected to be at least a factor of ten more performative than a 20 MHz T805 and a 64-node 50 MHz T9000 farm should have processed events at about 400 Hz. However, the current prototype T9000 performance falls short of its design goals and only a factor of four improvement over the T805 instead of the expected factor of ten has been measured.

9 PROJECTIONS AND SUMMARY

Despite the prototype nature of the T9000 and the known hardware bugs, some of them severe, we have been able to install and operate a network of 54 T9000s and 20 C104s. The system has been operated reliably at the CPLEAR experiment at CERN, running a very large event reconstruction program in real-time.

Currently the computational performance of the T9000 is below its design goals. In particular:

- The clock speed is 20 MHz and not 50 MHz
- Many floating point operations take three cycles instead of two
- Some trigonometric functions run three times slower than expected
- The lack of a native Fortran compiler necessitates the use of f2c, which on a T805 and a SUN Sparc station gives a performance penalty equal to a factor of 1.5.

However, if these issues are promptly addressed the T9000 might still prove to be a cost-effective solution for embedded system applications due to its high level of integration and functionality, ease of scalability, and low cost.

The new communication system using DS links and C104 switches shows considerable promise. The C104 in particular offers high density cost-effective commodity communications, which can be used to build very large switching networks. DS link technology is going through a standardization process (IEEE P1355), and the link speed is likely to increase by a factor of two within the next year.

A 64 T9000 GPMIMD machine using 56 C104s will be completed early in 1995 and will be upgraded with 40 MHz T9000s as soon as they are available. The performance of the machine within the CPLEAR application could also be improved by using a native Fortran compiler and optimization of the CPREAD code.

The GPMIMD machine will be used as a test bed for the type of communication architectures required for triggering and event building in future generation experiments. In addition, hybrid nodes consisting of a T9000 as a communications processor and a high-performance RISC chip to provide computational power will be tested.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support provided by the European Union via the GPMIMD project 5404. The collaboration with our industrial partners (INMOS, Parsys and Telmat) in this project has been close and fruitful.

We would like to thank B. Martin, X. Liu, and M. Zhu who developed the C100 converter board used to interconnect the T805 and T9000 systems.

The work presented in this paper would not have been possible without the close co-operation of the CPLEAR experiment.

REFERENCES

[1] H. van der Lugt, The Data-Acquisition and Second Level Trigger System for the ZEUS Calorimeter, Academisch Proefschrift, April 1993.

[2] R.W. Dobinson, J.L. Pages and J.C. Vermeulen, Transputers in Particle Physics Experiments, Particle Physics, Vol. 2, No. 2, 1991.

[3] J.L. Pages et al., Parallel readout of the CERN RMH system using Transputers, North-Holland Physics Publishing, May 1992.

[4] L.W. Wiggers, J.C. Vermeulen, The use of Transputers in the ZEUS on-line system, Comput. Phys. Commun. **57** (1989) 316.

[5] H. Boterenbrood et al., A two-Transputer VME module for data acquisition and on-line event selection in ZEUS, Nucl. Instrum. Methods Phys. Research Section **A332** (1993) 263–268.

[6] Ein Transputernetzwerk zur Datenerfassung in einem Vielkanal FADC-System.
M. Feuerstack, Doctoral Thesis, University of Heidelberg, 1994.

[7] E. Barsotti, Data Acquisition System Architectures, AIP Conf. Proc. **209** (1990) 82–102.

[8] The ATLAS Technical Proposal, CERN/LHCC/94–43, LHCC/P2, ISBN: 92–9083–067–0.

[9] The CMS Technical Proposal, CERN/LHCC/94–38, LHCC/P1.

[10] The T9000 Transputer Hardware Reference Manual. Inmos Ltd., Inmos document number 72 TRN 238 01.

[11] Networks, Routers and Transputers: Function, Performance and Application. Edited by: M.D. May, P.W. Thompson, P.H. Welch. IOS Press 1993, ISBN: 90–5199–129–0, ISSN: 0925–4986.

[12] M.P. Ward, A Transputer Based Scalable Data Acquisition System, Ph.D. thesis submitted to the University of Liverpool, 1995.

[13] R.W. Dobinson et al., A Scalable Data Acquisition System Based on Transputers and the Chorus Operating System, submitted to Scientific Programming, 1995.

[14] L. Adiels et al., Proposals for the experiment PS195. CERN/PSCC/85–6/P82, PSCC/86–34/M263, PSCC/87–14/M272.

[15] Chorus Systems, Chorus V3 Programmers reference manual. CS/TR–90–25.1 1991.

[16] The B016 VMEbus Master Board. Inmos Ltd., Inmos document 72 OEK 260 00.

[17] Rene Brun et al., PAW - Physics Analysis Workstation. CERN Program Library Entry Q121, October 1989.

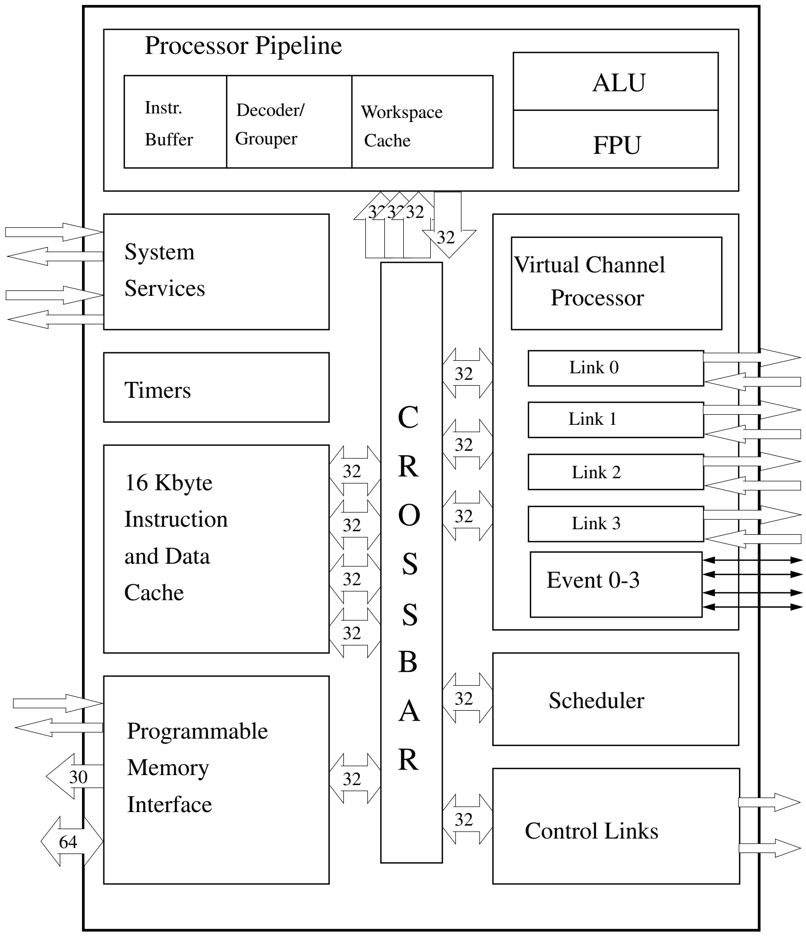

Fig. 1 The T9000 Transputer

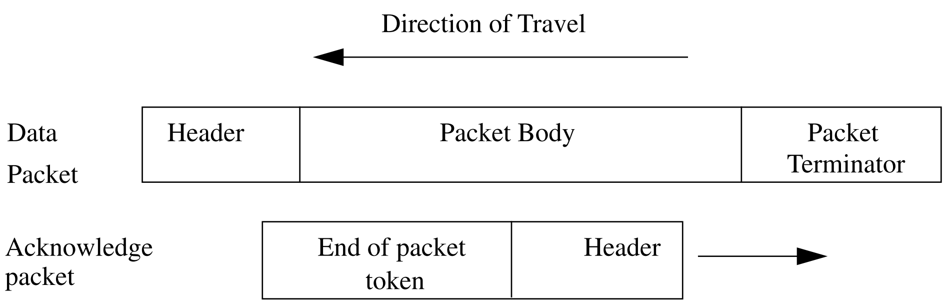

Fig. 2 Packets on a DS Link

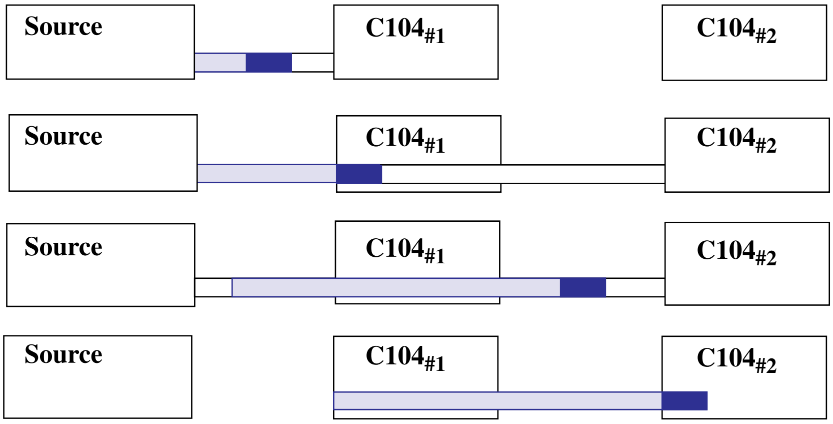

Fig. 3 Wormhole Routing

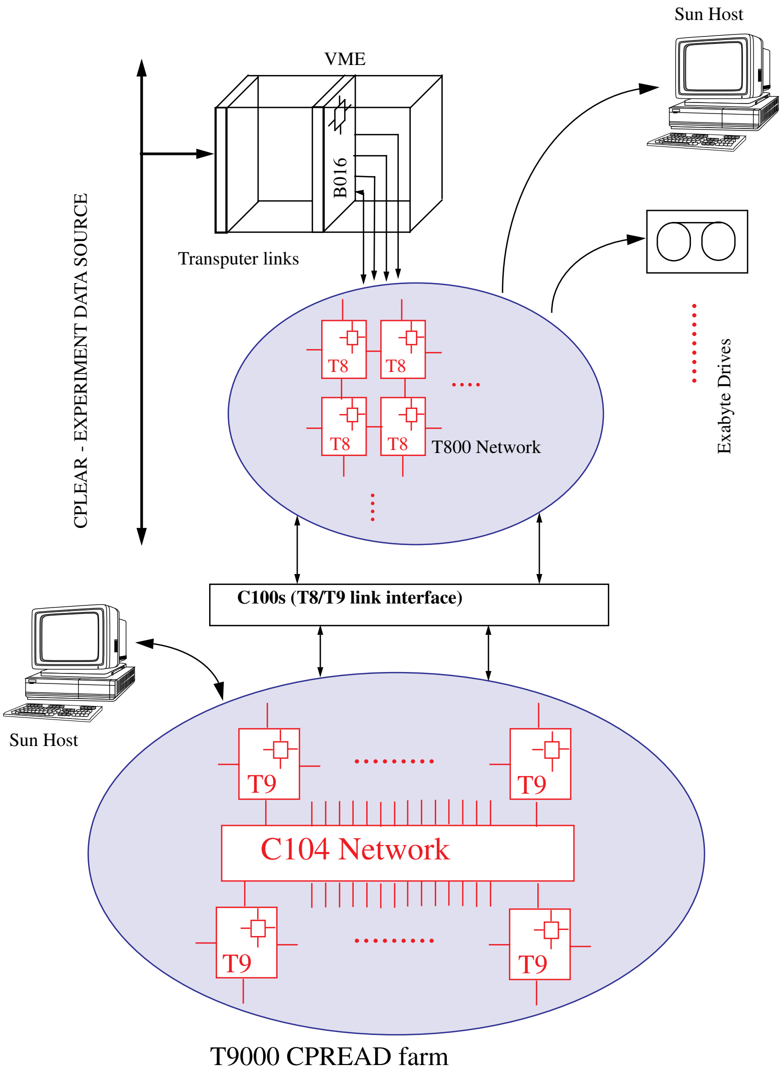

Fig. 4 A Schematic of the T805/T9000 System

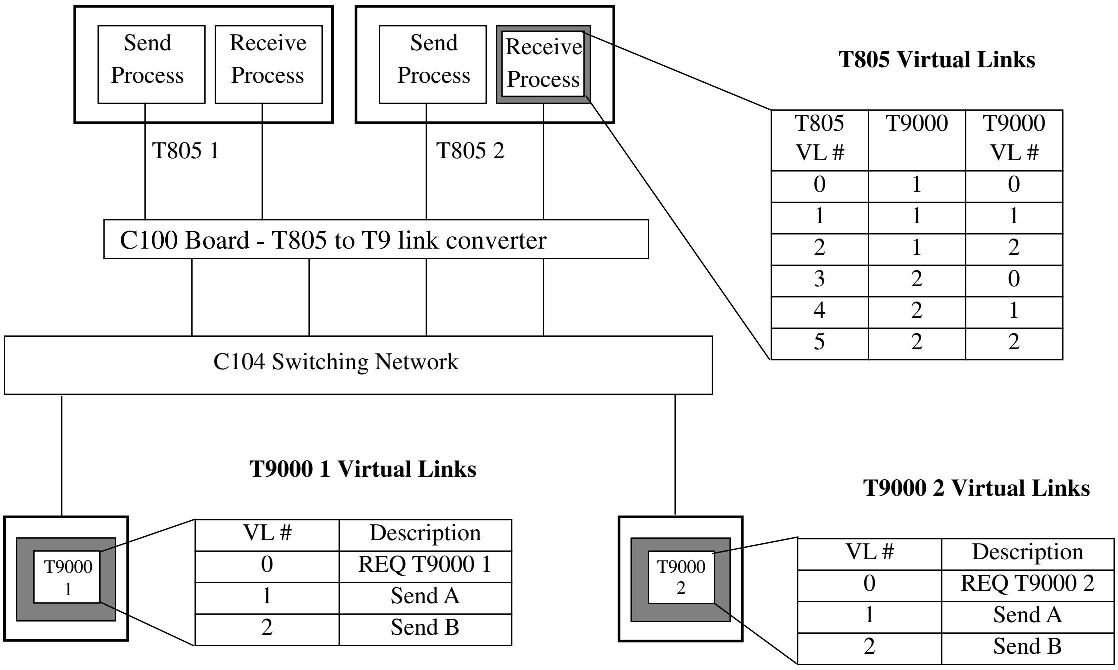

Fig. 5 Schematic of the T805 to T9000 Interface

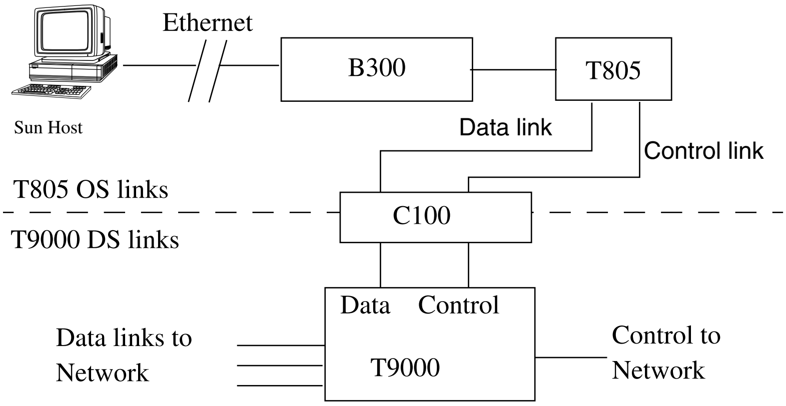

Fig. 6 Hosting T9000 Systems

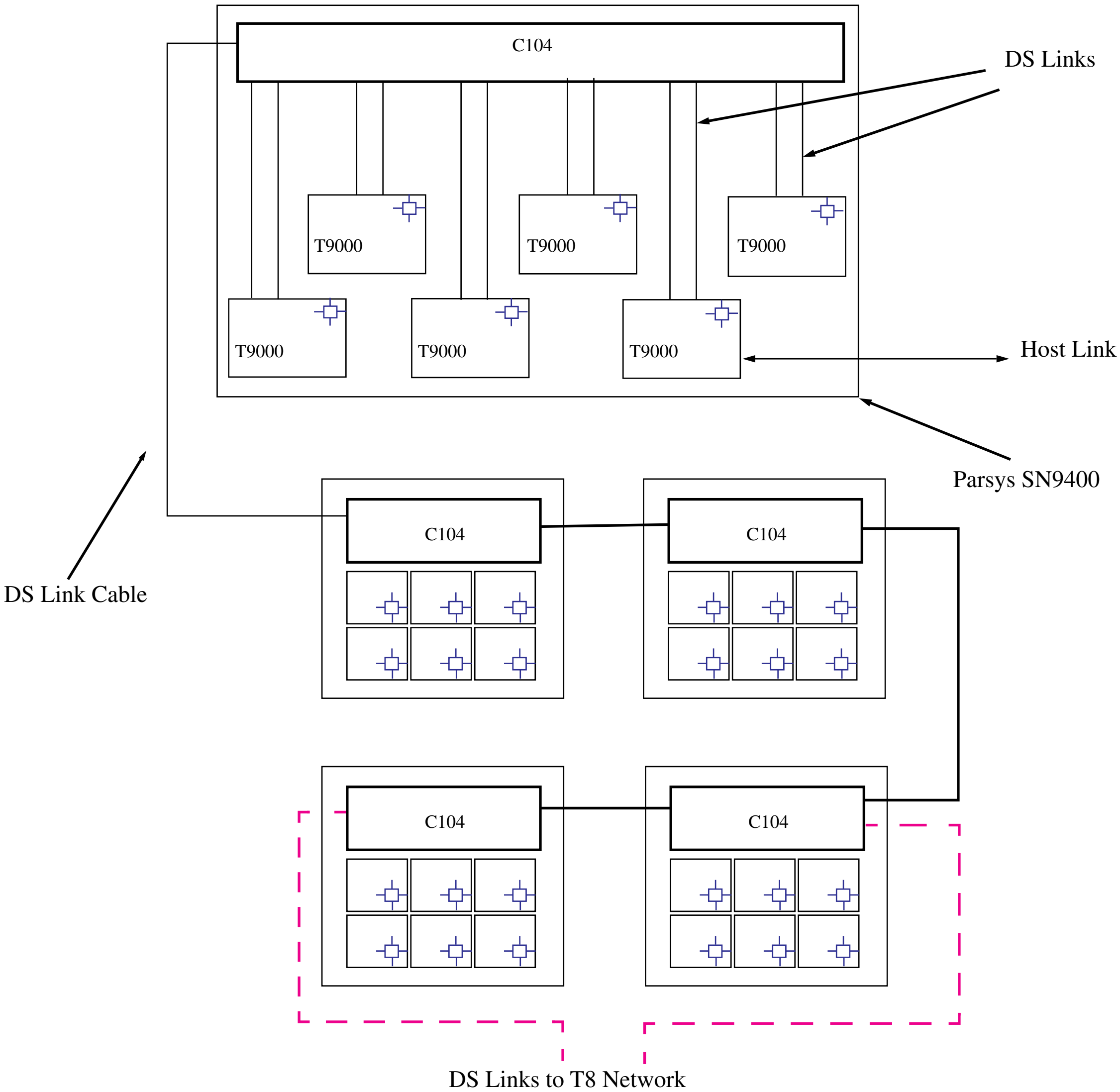

Fig. 7 The SN9400 Configuration

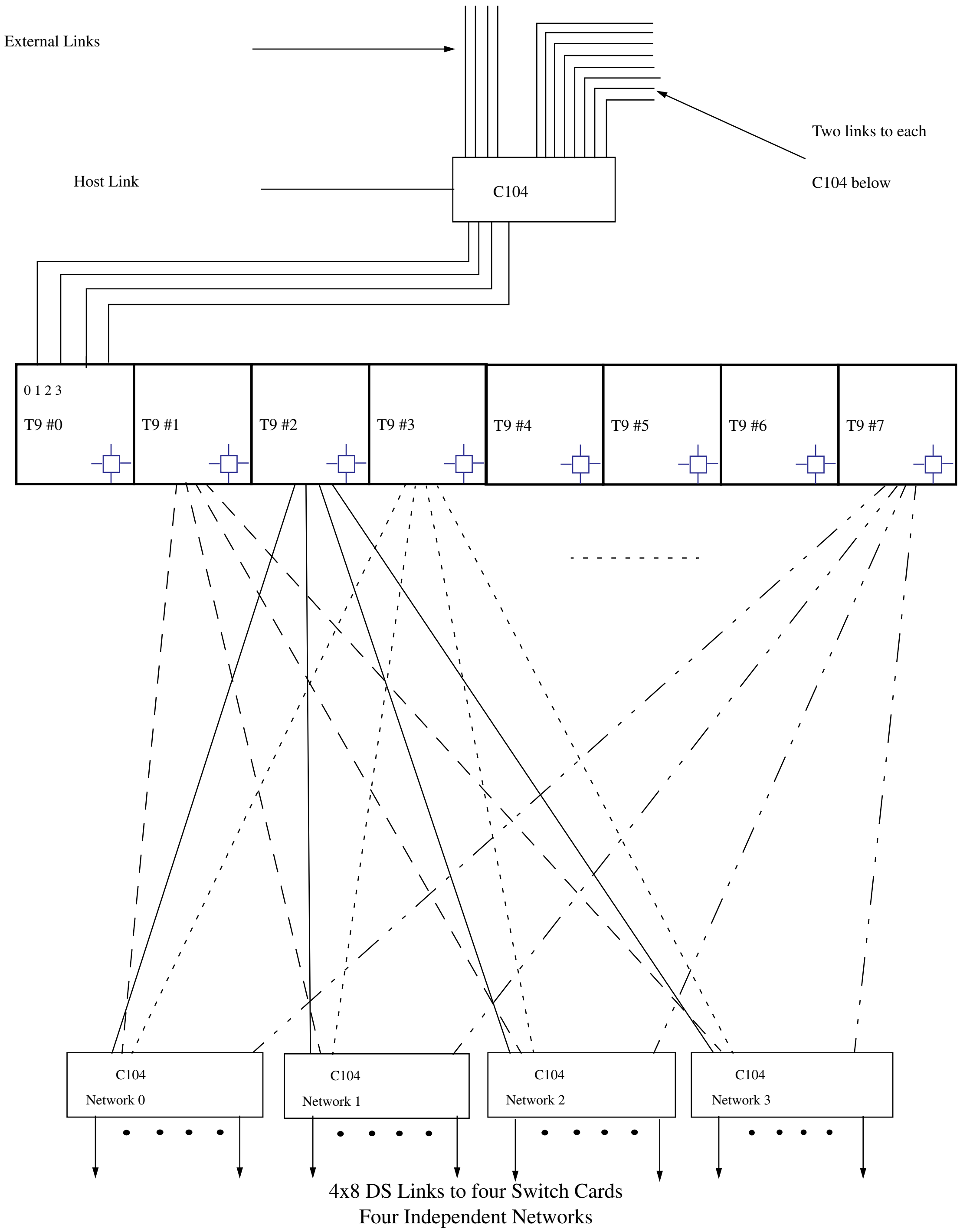

Fig. 8 A GPMIMD Motherboard

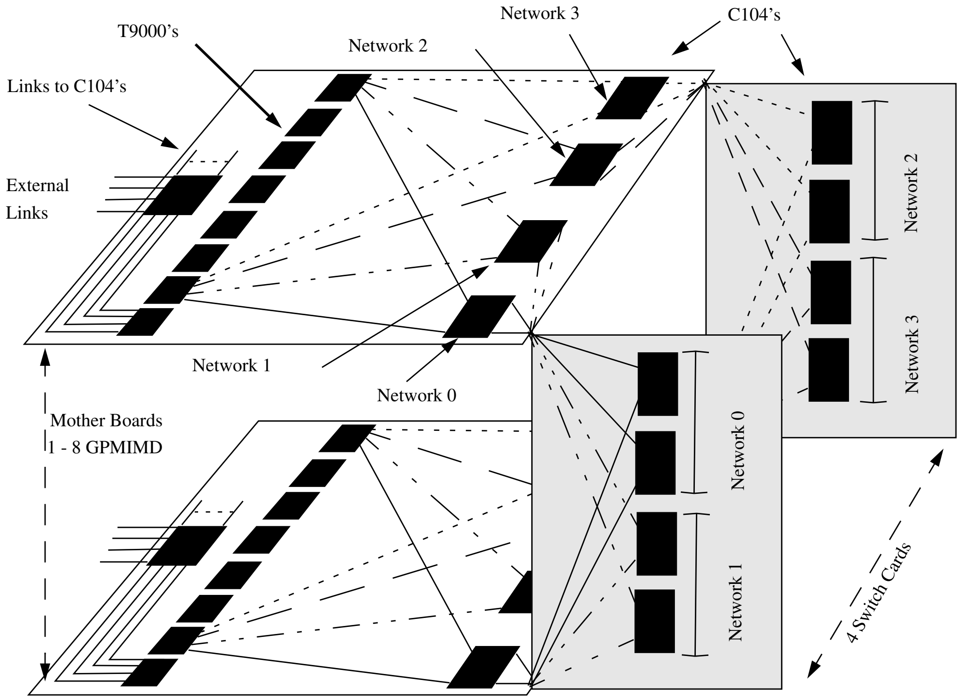

Fig. 9 A Schematic of the GPMIMD Machine

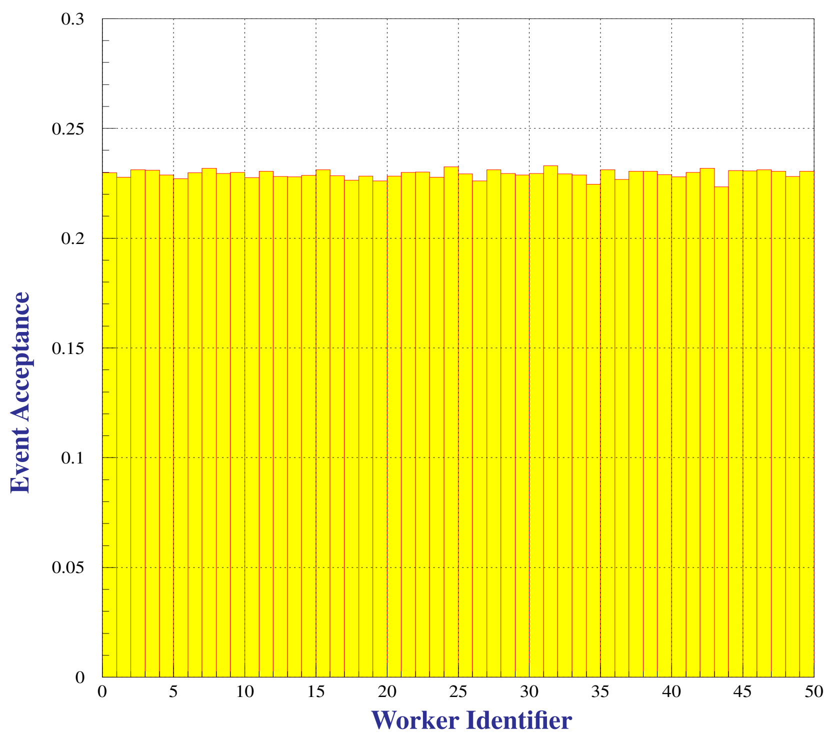

Fig. 10 Worker Acceptance versus Worker ID

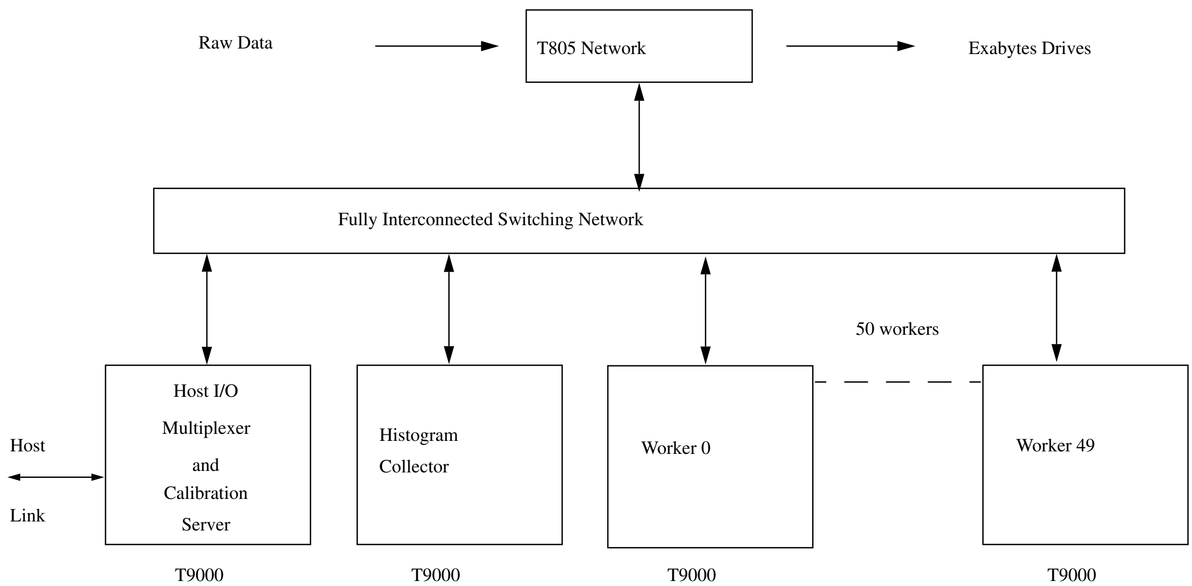

Fig. 11 Communication Requirements

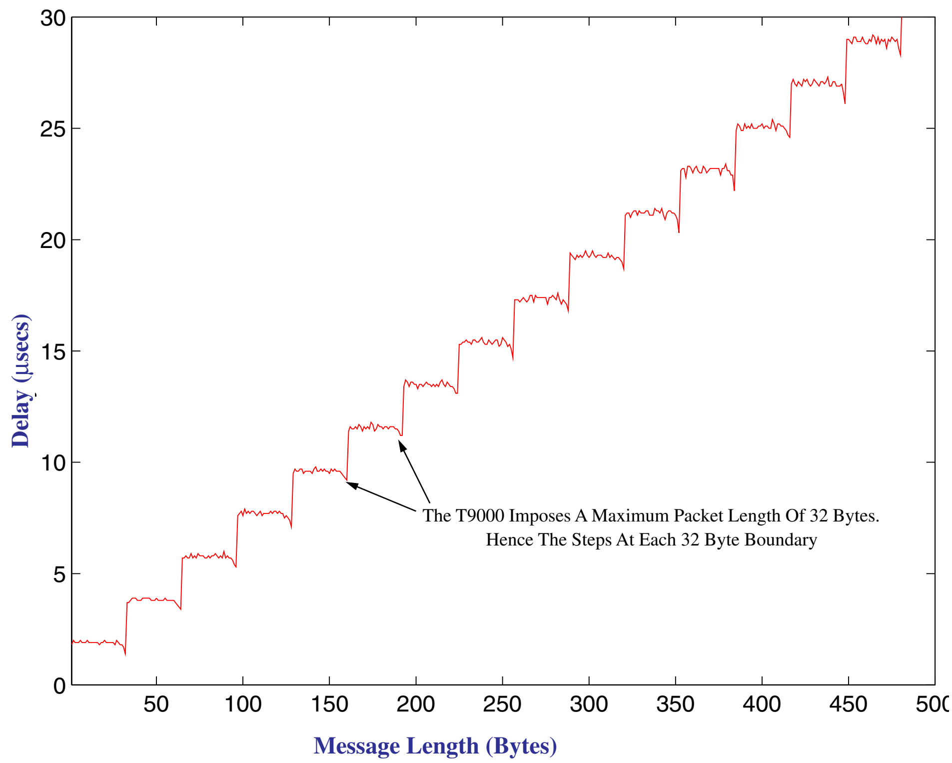

Fig. 12 C104 delay versus Message Length

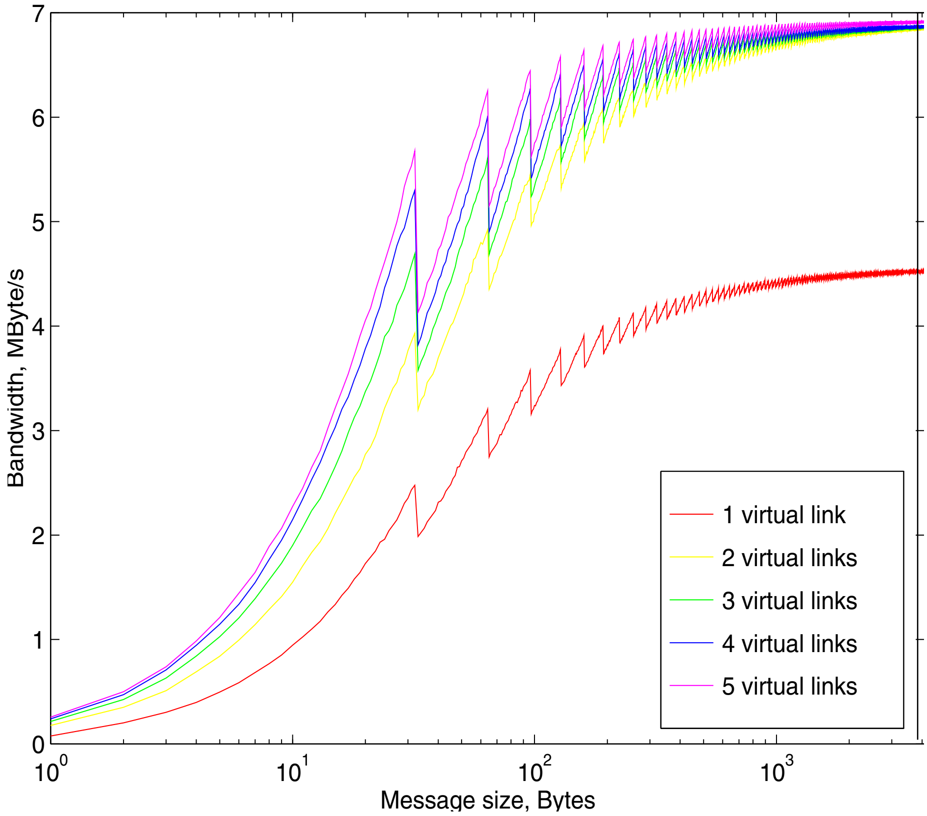

Fig. 13 Bandwidth, one uni-directional link

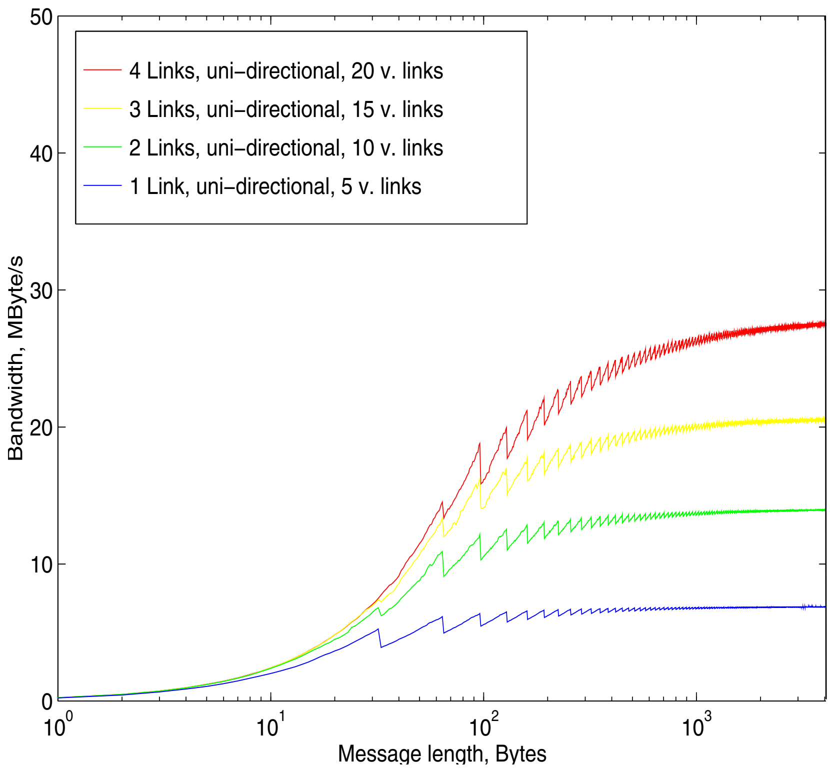

Fig. 14 Bandwidth, for a 20 MHz processor

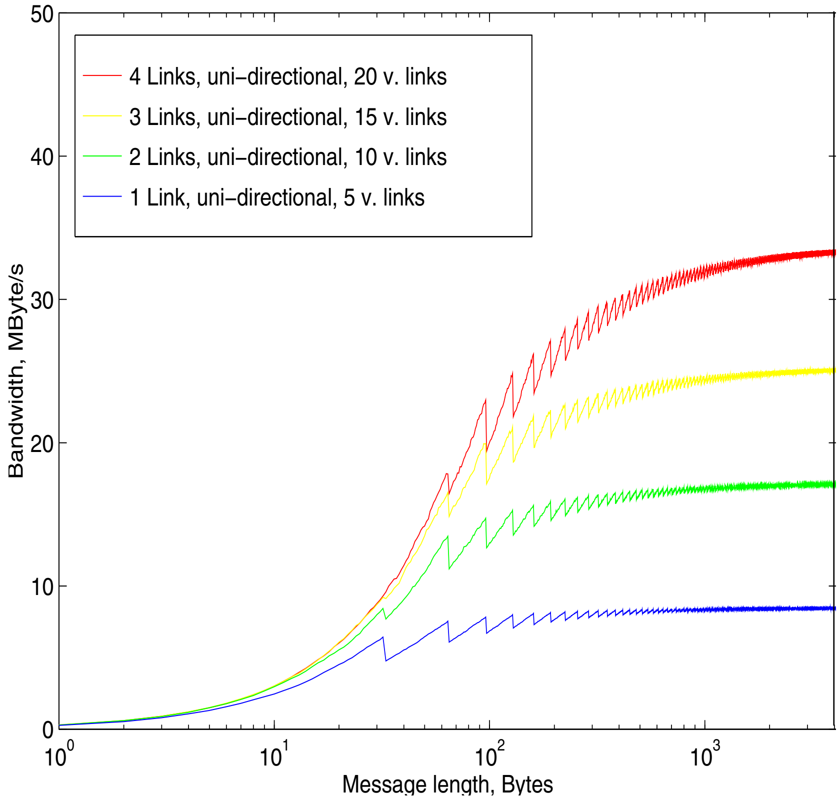

Fig. 15 Bandwidth for a 25 MHz processor