Cost/Performance Tradeoffs in Network Interconnects for Clusters of Commodity PCs

Christian Kurmann Felix Rauch Thomas M. Stricker
Laboratory for Computer Systems
ETH - Swiss Institute of Technology
CH-8092 Zürich, Switzerland
{kurmann,rauch,tomstr}@inf.ethz.ch

Abstract

The definition of a commodity component is quite obvious when it comes to the PC as a basic compute engine and building block for clusters of PCs. Looking at the options for a more or less performant interconnect between those compute nodes it is much less obvious which interconnect still qualifies as commodity and which not. We are trying to answer this question based on an in-depth analysis of a few common more or less expensive interconnects on the market. Our measurements and observations are based on the experience of architecting, procuring and installing Xibalba, a 128 node - 192 processor versatile cluster for a variety of research applications in the CS department of ETH Zurich.

We define our unique way to measure the performance of an interconnect and use our performance characterization to find the best cost performance point for networks in PC clusters. Since our work is tied to the purchase of a machine at fair market value we can also reliably comment on cost performance of the four types of interconnects we considered. We analyze the reason for performance and non-performance for different Fast Ethernet architectures with a set of micro-benchmarks and conclude our study with performance numbers of some applications. Thus, the reader gets an idea about the impact of the interconnect on the overall application performance in commodity PC clusters.

Keywords: Clusters of commodity PCs, Ethernet, Myrinet, switch performance, application performance, full bisection bandwidth, all-to-all communication.

1 Introduction

Several authors have pointed out the architectural principle for constructing high performance systems out of widely available commodity components about a decade ago. The literature on the Beowulf [1], the Hyglac and Loki parallel workstations projects provide a good overview on the topic and an almost complete list of credits to these early projects is given in [17].

Microprocessor based computer systems leverage from a high volume to be competitive in computational speed and price. Such volumes can only be sustained if the node architectures are similar to the architecture of PCs and workstations. It remains an open question of whether this law of commoditization also holds for cluster interconnects. So far the prediction of a commodity one-for-all-needs network has not quite materialized and the market is still split between regular Ethernet and a few dedicated high-performance interconnects. Initially the first networking technology expected to become the universal standard was ATM/Sonet, but at this time Ethernets using the TCP/IP protocols seem to be a more viable candidate for the role of the universal interconnect.

PC Clusters are the successors to massively parallel computers. Networks for massively parallel computers are a well researched topic. It would be well beyond the scope of this paper to give a complete survey, but the two fundamental approaches can be mentioned easily:

- Networks for parallel computers should be scalable to a large number of nodes and should provide full bisection bandwidth across any arbitrary bisection of the parallel machine. As a tradeoff the performance of a single link in such a network could be a secondary concern. The best example of such a network is the fat tree used in the Thinking Machines CM-5 [11].

- Networks for parallel computers should be designed around a sophisticated tradeoff of technology factors (i.e. best possible pin counts, clock speeds) and the links should be as fast as possible, allowing only simple networks like tori or hierarchical rings. Representatives of this line of research are the Cosmic cube project [5] or the Hector project [19].

After many research papers on this topic, it appears that the essential question is still open and needs to be readdressed in the light of commodity clusters incorporating the technology factors of commodity cluster interconnects.

In related studies, [7] compares different networking technologies for parallel computing, focusing on system software aspects that balance the network load on different local area networks used in parallel. The requirements for a high performance compute cluster for a successful integration into a larger scale computational grid is nicely described in [15]. Some interconnects and protocols are analyzed, but the work focuses more on the communication of two single nodes within and outside the cluster, rather than traffic patterns requiring full bisection bandwidth. A communication cost model is presented in [12], characterizing the key communication resources for parallel applications in high performance networks of workstations. After a close examination of our networking architecture and our network model the performance results of this study could lead directly to the determination of their “gap” and “bulk gap” parameters for performance predictions of algorithms, whose communication system response can be defined as LogP model parameters.

In addition to the popular PC clusters there are several new platforms for wide area distributed computing. Those platforms use the regular Internet as an interconnect between the compute nodes and are therefore rather limited to embarrassingly parallel tasks at this time.

1.1 The Xibalba Cluster: Concept, Design and Implementation

During the past five years many research groups of the department of computer science at ETH Zurich have related some of their research to the cluster of PCs platform by working on the software technologies and the design of such systems, by parallelizing their database systems to run on such clusters, by investigating the scheduling of tasks and work flows in scientific computation on clusters or simply by bringing the important application of large scale car traffic simulation to clusters of PCs.

The core of the Xibalba cluster is made of 128 dual 1 GHz Pentium III compute nodes. Since the database users can not make use of an additional processor, only half of the nodes are equipped with dual processors and the memory is kept at 512 MByte per processor in all the nodes at this time. Still for the node architecture a powerful Intel STL2 dual-processing server board with Serverworks Serverset III LE chipset was chosen to provide a memory system with excellent characteristics using cost effective standard PC133 SDRAM memory. A 64 bit/66 MHz PCI bus provides maximal I/O throughput for existing and future high speed communication with Gigabit Ethernet or Myrinet PCI adapters. Each node is equipped with two Intel PRO/100+ Fast Ethernet controllers to attach to two separate networks for data and control traffic.

More details about Xibalbas hardware and software concepts can be found in an extended version of this paper [10].

The rest of our workshop contribution is organized as follows: In Section 2 we show how to build a full bisection cluster network with Fast Ethernet using commodity networking equipment and show why this is difficult. Section 3 explains our evaluation principle and discusses how to read the performance results. In Section 4 we attempt to characterize a fairly expensive central switch that was said to provide full bisection bandwidth but did hopelessly fall short of our expectations. After a presentation of the benchmarking results the vendor replaced the switch against a model with higher performance. The performance comparison in Section 5 discusses the performance and the cost/performance ratios of the different networks by using an all-to-all personalized communication micro-benchmark. Section 6 finally describes the applications we regularly use in our cluster and discusses the relevance or the irrelevance of a full bisection network to real applications. The quite surprising results and experiences provided by the design process, the installation and the evaluation by the micro-benchmarks are presented for a conclusion in Section 7.

2 Xibalba Network Options

2.1 Networks for Clusters

For the optimal cost/performance tradeoff the inter-processor communication facility is the most critical part of a cluster. The networks of Xibalba are based on commodity 100 MBit/s Fast Ethernet, like in most Beowulf class systems. Before inexpensive single backplane networking switches became readily available several different topologies were proposed for Beowulf clusters [16]. As a major difference to most other Beowulf clusters, Xibalba has two Fast Ethernet networks as specified below.

For dedicated networks in parallel computing several high speed interconnect technologies were developed, e.g. with Myrinet. A comparison between two such technologies and a traditional supercomputer network is given in [8]. Myrinet with its very low latency and high bandwidth was considered for Xibalba but initially rejected due to its high cost. The expense was not justifiable to the database experts as their database management middleware was not instrumented for high speed communication at all and therefore a high performance network appeared useless to them. In the mean time the traffic simulation group solicited funding to equip a 32 node sub-cluster with Myrinet 2000.

For a brief introduction we give a broad overview of the networking technologies considered and implemented in Xibalba at this point in the evolution of cluster technology (i.e. in the year 2002). The fractions of the network cost relative to the total cost of the cluster are as show in Table 1.

Cluster Network Technology	Cost Ratio Nodes : Network
High Perf. Myrinet	65% : 35%
High Perf. Shared Myrinet	70% : 30%
Full Bisection ER16	80% : 20%
Reduced Bisection E7	87% : 13%
Maintenance Ethernet	96% : 4%

Table 1: Cost ratios (nodes versus network) for different cluster networks in Xibalba.

Maintenance Network For the purpose of separating maintenance and operating system traffic from application traffic we designed a cheap secondary network for the Xibalba cluster in order to supplement the primary network. This network uses 100BaseT technology but is most reduced in its topology and the performance of the components used. The topology follows the physical design. At the price-performance point of 128 nodes, the cost of this network is only 4% of the cluster. This type of network is installed in addition to the primary data network. It is implemented by eight 24-port Enterasys Vertical Horizon VH-2402S Fast Ethernet switches which are interconnected further by a Fast Ethernet switch of the central communication facilities at ETH Zurich.

Full Bisection Ethernet The primary data network targeted at in our 128 node Xibalba cluster design is specified to sustain full-speed non-blocking, full-duplex communication on all ports simultaneously. Several networking product vendors offered their switches which shall comply to this specification. This network was first implemented by a large central Enterasys Matrix E7 network switch, including four 6H302-48 line-cards providing 48 Fast Ethernet ports each. We will explain the problem with this equipment and the reason for providing more ports than what seemed required in Section 4. Due to the many limitations

of the Matrix E7, the switch was upgraded to an Enterasys X-Pedition ER16 Switch Router with seven ER16-TX-24 switching modules (24 port 100Base-TX) and an ER16-8 Gigabit uplink switching module (8 port 1000Base-SX). At the price-performance point of 128 nodes, the cost of this network is 20% of the cluster while the E7 solution is 13%.

High Performance Network As a representative of the expensive networks we are considering Myrinet 2000. A technical introduction is given in [3]. The important difference to the previous networking concepts is the emphasis on expensive network interfaces and low-cost high-speed switches. At the price-performance point of 128 nodes the overall cost of Myrinet is about 35% of the total cost of the cluster. In our cluster a 32 dual processor node subsection is equipped with Myrinet allowing to use the network either solely with just one processor or in a shared dual configuration. The second option leads to a network cost *per processor* ratio of 30% (see Shared Myrinet in Tables 1 and 2).

Cluster Network Technology	Cost Ratio Switch : Cable : Interf.
High Perf. Myrinet	24% : 9% : 67%
High Perf. Shared Myrinet	24% : 9% : 67%
Full Bisection ER16	92% : 2% : 5%
Red. Bisection E7	87% : 4% : 9%
Maintenance Ethernet	65% : 5% : 30%

Table 2: Cost ratios (Switch versus Cabling versus Interface) for different cluster networks in Xibalba.

The relative costs of the network components are shown in Table 2. The main difference between a dedicated high performance network and an Ethernet lies in the cost ratio of the switches vs. the interface cards. While interface cards for Ethernet are nearly for free because already built on the main boards the cost lies in the switches. For Myrinet it is the other way around. The switches can be built very simple as the intelligence is in the interface hardware which is therefore much more expensive than Ethernet adapters.

All our cost calculations are for cost/performance evaluation only and are given relative to the total cost of the 128 node Xibalba cluster which amounts to about US$ 500'000. As we work in a country with exceptionally high wages and high cost of graduate students labor, the design of the Xibalba cluster was advertised in a public bidding process. The winning bid was by DALCO Inc., a local contractor that happily assumed the responsibility for systems integration and installation in the machine room of our university.

2.2 Full Bisection Bandwidth

Interconnect networks of most regular computing structures are characterized by their bisection bandwidth. In the discussion of bisection bandwidth the worst case performance critical bisection of the network is determined (according to the topology) and addressed. For the measurement the nodes are paired in such a way that all the communication must cross the links on the most critical bisection cut. If the network access provides full duplex links the communication between the pair of nodes must also be addressed as full duplex, i.e. must go simultaneously in both directions.

A network is said to have a full or—in somewhat more precise terms—a fully scalable bisection bandwidth, if it can sustain the full network access bandwidth of every node across the most critical bisection while all nodes communicate simultaneously. For a 100BaseT network this means that every node must send and receive data at the same time with 100 MBit/s. Network topologies with full or scalable bisection include the full fat tree, the hypercube and the full crossbar central switches. The mesh, the torus and the plain/skimmed tree network configurations do not offer scalable bisection bandwidth in general, but for some cases some full bisection communication might be achievable for machines up to a certain fixed size.

2.3 Cluster Networks with Full Bisection Bandwidth

In switch-based high-performance networks like Myrinet the Clos network used for their switches can readily sustain scalable bisection bandwidth up to 128 nodes at almost linear cost per port. After that scaling beyond 128 nodes will face some growing switch costs per port as multiple switches have to be cascaded into a larger network. Still full bisection bandwidth is doable for high performance networking in larger machines.

Ethernet based networks with full bisection can be constructed from either single backplane switching solutions or fat trees using small 8-way switches with up-links to a central backplane that are 8–10 times the speed of the basic links. Both kind of networks were considered for the primary data network of the Xibalba cluster, but finally the single backplane solution was given preference. The single switch solution can scale up to about 512 nodes for basic 100BaseT connections. Fat trees can scale up to somewhat larger configurations, even without any exploding costs in practice when multiple up-links and a moderate number of duplicated backbone switches are used.

3 Evaluation Principles and Micro-benchmarks

We intend to design and evaluate the performance of an interconnect for specific communication patterns, that can be represented as micro-benchmarks. The primary goal of looking at isolated communication primitives is to gain architectural insights into the bottlenecks. Despite the primitive function we can relate these benchmarks to some common communication patterns in real application code.

3.1 Bandwidth vs. Latency

There is a lot of emphasis on communication bandwidth in this study and the aspect of latency is not considered much. The commodity system architect can not do much about certain latency components in the system e.g. the PCI bus arbitration latency. A common wisdom says that additional bandwidth can be purchased easily while latency is given by the laws of nature (or by the boundary conditions of systems engineering). In the light of this background there are many more interesting cost/performance tradeoffs for additional bandwidth than there are for lower latency.

We understand the several order of magnitude difference of latency between a high-performance interconnect, using a flit level worm-hole routing scheme in the switches and a communication co-processor at the endpoints vs. the commodity Ethernet that uses store-and-forward routing and a simple host interface using delayed interrupt processing due to coalescing of interrupts. Still many applications are affected by the granularity of communication instead of pure latency and can therefore be reprogrammed accordingly to communicate in large blocks or use mechanisms of latency tolerance.

3.2 Communication Patterns Requiring Full Bisection Bandwidth

Communication patterns requiring full speed communication across the critical bisections are relatively rare and can be avoided in many cases by clever parallel programming or with probabilistic algorithms for large data sets (e.g. with sample sort) [2]. The most important parallel algorithm requiring full bisections are computations in a bitonic sorting network or an FFT butterfly network. The most common communication pattern limited by critical bisection is all-to-all personalized communication.

3.2.1 All-to-all Personalized Communication

The all-to-all personalized communication (AAPC) step is frequently encountered in parallel programs. In an AAPC step, each processor sends a block of distinct data to every other processor. An AAPC contains many different communication patterns and a large number of network properties are exercised. Next neighbor patterns or some bitonic sorting exchange pattern are subsets of AAPC and their performance can be derived from the general AAPC performance data.

The AAPC step occurs frequently in multi-dimensional convolutions (e.g. FFTs) and in array transposes where only one dimension of the array is distributed [18]. Transforming a two-dimensional 4096×4096 HDTV video image to Fourier space and back for filtering at 30 frames per second would require 60 GFlops/s sustained performance and can certainly only be done with an entire PC cluster or a big array of dedicated DSPs. Also transforming a $128 \times 128 \times 128$ grid for a particle mesh Ewald force calculation in a molecular dynamics simulation code at 1000 time-steps per second costs 70 GFlops/s for the FFTs alone, not including any additional work for force calculations or total energy evaluations. In addition to the GFlops/s those applications require a GBytes/s communication performance.

3.2.2 Congestion Controlled AAPC as a Micro-benchmark

A phased AAPC algorithm as described below can be devised and can achieve optimal aggregate bandwidth once the different phases are carefully separated. Phase separation can be maintained by globally synchronizing the entire machine after each phase is completed. This strategy adds some overhead for synchronizations and might require additional communication resources and/or dedicated hardware mechanisms, but it makes sure that no communication resources are wasted due to inefficient scheduling and due to unnecessary congestion.

A simple algorithm for AAPC proceeds according the following algorithm:

parallel algorithm *all-to-all*
1 **for** $i = 1$ **to** $n - 1$ **do**
2 concurrently send data to node $n_{(self+i) \bmod n}$ and receive data from node $n_{(self-i) \bmod n}$
3 wait for barrier

For the optimization of the AAPC performance we try to use information about the network topology or the internal structure of the switches to minimize the congestion in each phase. For every phase each node has a fixed communication partner to send to and to receive from. In the simple algorithm given above the logical communication distance increases with each phase of the algorithm. The true physical distance between the communicating nodes also depends on the mapping of node numbers to communication ports. We use a simple linear mapping and therfore the next neighbor communication stays mostly within a switching module while some long range communications traverses the module boundaries and the backplanes.

In the common application scenario of a balanced AAPC the same amount of data is sent/received by each node in each phase. Therefore a slower phase is an indication of congestion due to a particular communication pattern. Since phases are synchronized across the entire machine the duration and the final throughput is determined by the slowest connection of a phase.

We show the detailed result of an AAPC benchmark in two different graphical representations. First we time the performance of the communication for each corresponding pair/route and look for slow routes indicating hot-spots. For an AAPC of 64 nodes we have 4096 distinct source-destination pairs. We graph a histogram across all 4096 communications (left part of Figure 1). Since the performance of each phase is determined by the slowest route we give a second histogram in the middle section of Figure 1 that captures the distribution of phases according to their throughput. The total execution time is truly cumulative and therefore the weighted average over all phase throughputs indicates the total throughput in the small bar on the right of Figure 1. For large data blocks the cost of the barrier synchronization itself can be neglected, for small data blocks congestion is not a limiting issue and the barriers are omitted.

As a second consideration we look for the reason for slower communication pairs and relate the route/phase performance to the communication distance. A single graph in Figure 3 of Section 5 shows how different networks react with congestion when a lot of routes travel for longer distances. A graph of congestion vs. logical distance exposes the limitations in inter-module communication in switched Ethernet or unexptected dependencies on some topological features in the network. In most cases the claim of the vendor that the delivered switch is a full crossbar does not hold and our Switchbench measurement tool shows clearly to what extent this is true or not. The source code can be downloaded from [9].

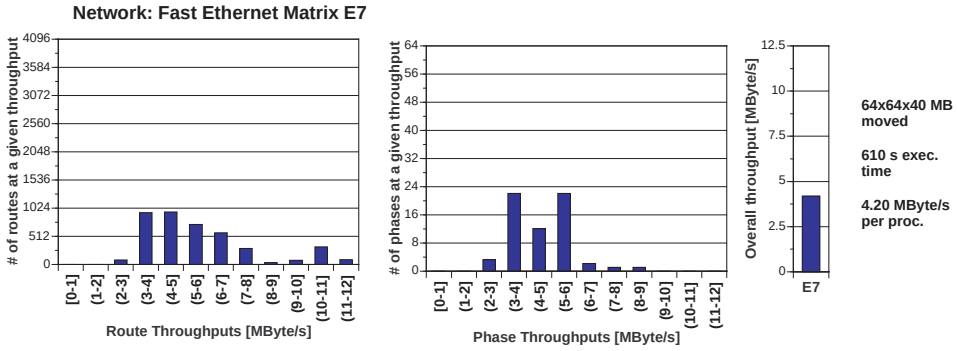

Figure 1: Histograms of transfer rates for the 64^2 routes (left), the transfer rates of the 64 phases (right) and the overall performance in a phased AAPC algorithm on a 64 processor cluster for the E7 network configuration.

4 Analyzing a Non-Performing Ethernet Switch

4.1 Different Network Configurations

As announced in Section 1 and described in Section 1.1 we have three networks installed in the Xibalba cluster: A full bisection primary data network implemented by a large central Ethernet switch, a secondary maintenance network implemented by 8 small switches mounted in the 8 racks including an up-link switch that interconnects these 8 switches by a Fast Ethernet link and a Myrinet 2000 network in a part of the cluster. For these tests TCP/IP and the socket interface was chosen as the software API.

The limited performance of the primary network as it was first installed gives us a good picture how a fully switched off-the-shelf backplane switch with reduced bisection bandwidth would operate with 128 nodes. In this Section we focus on this reduced bisection network and study the limitations introduced by its design. Still as we paid for a full bisection network this configuration was upgraded to full bisection.

We distinguish between some switch connectivity patterns as shown in Figure 2. Those patterns were instrumental to characterize the bottlenecks in the line modules of the Matrix E7 switch. A Matrix E7 switching module consists of two ASICs that provide 24 ports each. They communicate over an internal bus at full speed or with reduced speed to the E7 backplane. But as measured later in the benchmark results each ASIC provides barely enough bandwidth for 16 ports. Such bottlenecks are fairly typical for equipment that is optimized for LAN use.

Ethernet E7 (configuration 1) This switch configuration was the configuration we run on with the E7. Each ASIC was populated with 16 machines only to achieve a better bisection communication.

Ethernet E7 (configuration 2) This configuration uses all the ports provided by a module and uses only three modules with 48 ports each.

Ethernet E7 (configuration 3) To further test the communication between switching modules we setup a configuration where just one ASIC is used per module which results in 16 nodes per module.

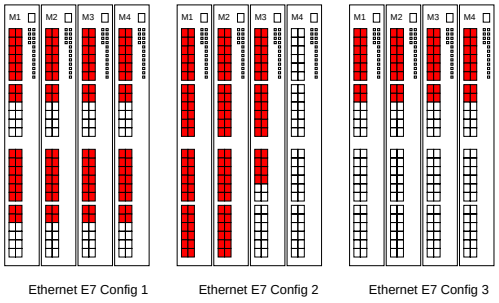

Figure 2: Different network setups and switch utilizations of the primary network for the benchmark tests.

4.2 Performance Measurements

Pairwise Traffic Tests

We first use a pairwise traffic generator to analyze all kinds of different bisections of the E7. In this micro-benchmark a set of machines communicates in pairs. All pairs send and receive a large amount of data between the two nodes, in parallel and at full duplex. The pairs in a set N of n machines are separated at an arbitrary distance of node ids, called a stride t, and with wrap around, so that the node pairs $(i, i + t \bmod n)$, where $0 \leq i < n$, $n \bmod 2 = 0$, communicate with each other. The stride parameter t allows to test different bisectional communication patterns, thereby varying the amount of data crossing a well defined bisection line.

The switch consists of a rack with a switching backplane as well as single switching modules providing 48 ports each. To consider this architecture we divide the study in two parts: intra-module and inter-module communication.

To measure the *intra-module* capability of a switch module we first generate pairwise traffic within the first internal ASIC. The upper part of Table 3 shows the achieved bandwidth that drops for the reduced bisection network as soon as more then 16 ports communicate. The aggregate bandwidth limitation of an E7 ASIC seems to be 2×1600 MBit/s. If the number of pairwise partners is increased to two ASICs (48 ports, middle part of Table 3) the performance is the same as in the intra-ASIC case, which means that the intra-module communication bandwidth is not reduced below the limitation of a single ASIC. A limited usage of the switching modules to no more than 16 nodes

per ASIC enables full bisection bandwidth within a module and was the reason for using the data network configuration 1 with the E7 switch in the Xibalba cluster.

The *inter-module* communication was measured by pairwise traffic between each port of the first module to each port of the second module with the reduced bisection E7 network. The lower part of Table 3 shows the results which reveal a drastic drop in bandwidth for more than 8 ports communicating to their partners. The aggregate bandwidth limitation for inter-module communication seems to be limited to 2×800 MBit/s. This is a severe limitation down to just $1/6$ of the specified bandwidth. Only very reduced bisection bandwidth is possible as soon as the number of nodes reaches 8 per module.

Performance Matrix E7		
Communication Partners (from, to)	Nr of Nodes	Transfer Rate [MByte/s]
Intra-Module comm.	7+7	11.2
(ASIC 1,ASIC 1)	8+8	10.5
Pairs: (1,2)(3,4)..	9+9	9.7
..(23,24)	12+12	**7.8**
Intra-Module comm.	14+14	11.3
(ASIC 1,ASIC 2)	15+15	10.7
Pairs: (1,25)(2,26)..	16+16	10.4
..(24,48)	24+24	**7.8**
Inter-Module comm.	7+7	11.3
(Module 1,Module 2)	8+8	10.2
Pairs: (1,49)(2,50)..	12+12	6.9
..(48,96)	48+48	**2.2**

Table 3: The results of the pairwise tests for different number of pairs with the reduced bisection network (Matrix E7 config. 2). We measure intra- and inter-module communication.

All-to-all Communication Tests

The all-to-all communication tests show the limitations of a reduced bisection network over a full bisection network with a slightly more realistic workload.

The reduced bisection network implemented by the E7 configuration 1 performs very well for 32 nodes by over 10 MByte/s for all communication steps and provides nearly bisection bandwidth. But going up to 64 nodes an inter-module communication limitation of the E7 switch reduces the resulting total bandwidth significantly.

For the E7 configuration 2 we have again the ASIC limitation inside a module resulting in a sustained bandwidth of 6 MByte/s for all patterns. We have less inter-module communication here, therefore this limitation does not carry weight.

The interesting test with the E7 configuration 3 shows the inter-module limitation quite clearly. The more communication paths cross the module boundary, the more the bandwidth drops. As soon as the inter-module communication limit is reached the bandwidth continuously stays at 2 MByte/s for some phases.

The overall execution times for all-to-all tests with 64 nodes and a message size of 40 MByte on the different network setups for the E7 matrix switch are 610 s for config 1, 711 s for config 2 and 468 s for config 3. As implied by the bandwidth results the Ethernet E7 configuration 2 needs roughly 16% more time than the configuration 1 where the underpopulated configuration 3 results in 23% better performance.

4.3 Vendor Promises vs. Reality

The outcome of this switch evaluation seems disappointing. It is well known that data sheets sometimes do not reflect the performance of the real hardware implementation. Confronted with our test results the representative of the vendor readily checked with engineering and admitted that there is an inter-module communication limitation in the line modules of the Matrix E7. The local representative also stated that marketing inflates the total bandwidth numbers to take into account that in a “normal” network setting not all the users on a switch will communicate with all other users on the switch and that we are the first customers that have a problem with this limitation. The system integrator relied on the data sheets of the vendor and was rather puzzled by those explanations of his network equipment supplier.

Still during the renegotiation of the acceptance criteria the vendor has offered to upgrade the network to full bisection bandwidth for all nodes by exchanging the Matrix E7 by the X-Pedition ER16 which is referred to as the full bisection network in this paper.

We repeated the previous micro-benchmarks with the ER16 switch. The satisfying results are presented in [10].

5 Performance of the AAPC Micro-benchmark

In this Chapter we compare the performance of the AAPC micro-benchmark as presented in Section 3.2.2 on the four principal networks presented in Chapter 2. For Ethernet TCP/IP and the socket interface was chosen as API, for Myrinet MPICH-GM was used instead. The all-to-all communication tests show the different performance figures of the networks with a slightly more realistic workload than the isolated pairwise test presented in Section 4.2.

Figure 3 shows the minimal bandwidth achieved by each single communication phase of an all-to-all communication for 60 nodes with the three Fast Ethernet based networks at the bottom. The upper part of the Figure also depicts the performance for the Myrinet networks on 30 nodes in single and dual node processor configurations respectively.

Looking at the numbers for the maintenance network in Figure 3, we see the expected sharp drop in performance where all nodes of the cluster attempt to communicate over the highly limited bisection of a single 100 MBit/s link. The bandwidth is slightly higher when a limited amount of intra-switch communication occurs in phases 1–15 and 44–59. The reduced bisection network of the E7 in configuration 1 shows the inter-module bandwidth bottleneck of the switch clearly when more than 8 nodes communicate to another switch module in phases 9–51. The ER16 full bisection network performs very well by over 10 MByte/s in average.

Looking at the results for the high performance Myrinet interconnect in the 30 nodes case in Figure 3 (note the different axis on the right side) we remark the very uniform performance of the highly symmetrical switch architecture.

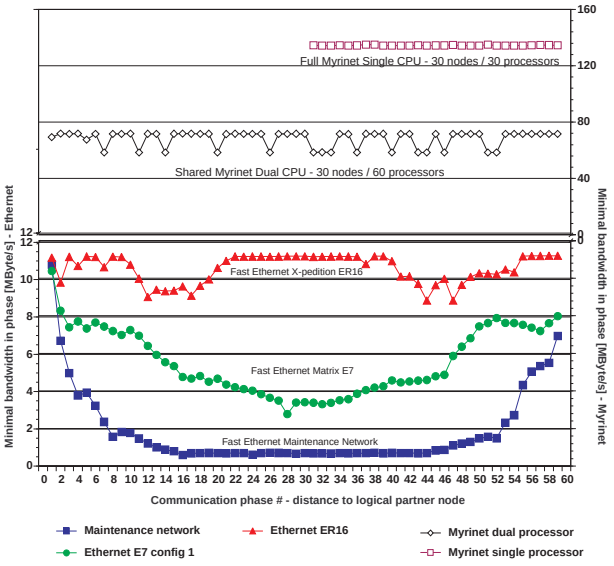

Figure 3: All-to-all minimal per-node bandwidth for different communication phases for all the discussed network architectures.

The performance of the 60 processor case (dual processor nodes) is roughly halved since the two processors of a node share a single network adapter and the bandwidth of a single network link. We attribute the irregular performance variations to a measurement uncertainty in an SMP environment with shared resources.

We instrumented a congestion controlled AAPC implementation to record the performance of each individual transfer (see Section 3.2.2). Figure 4 shows histograms of transfer rates over all 64×64 routes and all the communication phases of the algorithm on 64 processors. The Figure includes also overall throughput and execution time numbers. The histograms show that the maintenance network has an extremely limited performance on almost all routes and phases. On the reduced bisection network of the E7 reduced bisection bandwidth switch the performance for the routes varies considerably. The routes performance varies depending on intra-module or inter-module communication. The network of the ER16 offers very well performance on almost all routes. Since every single route with bad performance reduces the performance of a whole phase, there are slightly more phases with reduced performance than routes. Looking at the Myrinet interconnect shared between two processors we note a high percentage of routes with slightly reduced performance due to the resource sharing between the two processors. With single processors per node the performance of the Myrinet network achieves almost a perfect distribution of high route and phase bandwidths.

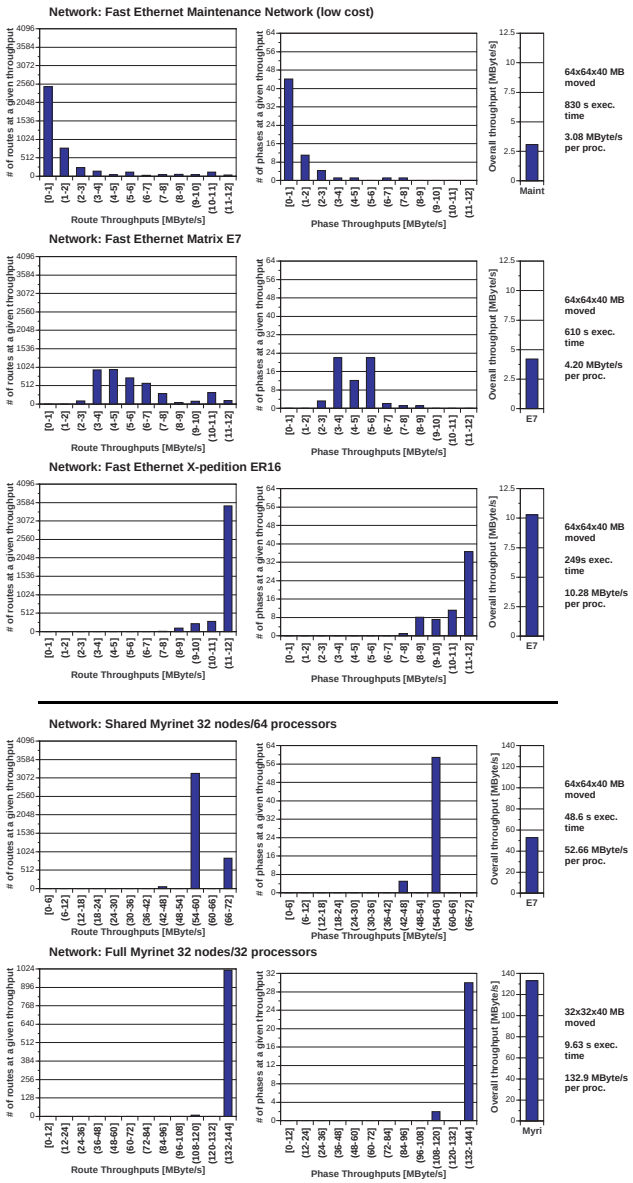

Figure 4: Comparison of the four networks with regard to transfer rates for the 64^2 routes (left), the 64 phases (center) and the overall throughput per processor (right).

6 Performance of Application Benchmarks and Applications

In Section 5 we determined significant differences in performance for the different switched Ethernet and the Myrinet configurations. While these differences are quite interesting for the architects of a cluster they might not matter much for the typical application user of the Xibalba cluster.

We measured three different kinds of application benchmarks to find out if the differences in networking performance really matter. The codes represent a communication bound workload, a mostly compute bound workload and a mixed workload respectively. The application programs and the corresponding results are presented in the following subsections.

6.1 Dolly

Dolly is our partition-multicast utility programm for system administration on clusters as described in [14]. Dolly is a small program that distributes large amounts of data to many nodes in a cluster in a highly efficient way. It is

mostly used to install new operating systems in partitions on the hard-disk drives of clusters or replicating database images with maximal performance. In short, it sends the partition data from a master in a virtual multi-drop-chain over TCP/IP to the first participating node in a cluster, which writes the data to the local hard disk drive and forwards it concurrently to the next participating node and so on. With single or dual Fast Ethernet as networking technology and fast hard disk drives the nodes in the Xibalba cluster are capable of saturating their network interfaces for sending and receiving data concurrently. For the purpose of this benchmark Dolly does not access the local hard disk drives but sends dummy data through its multi-drop-chain.

Dolly is communication bound, but its communication pattern is limited to two high-speed connections to the nearest neighbors of each node. Thus, there is only very limited data traffic over any bisection for reasonable configurations.

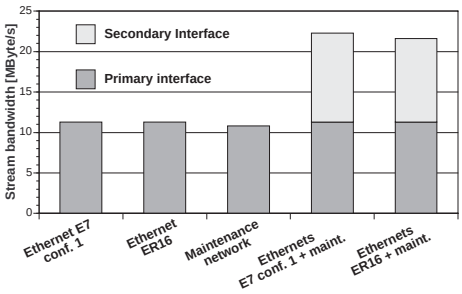

Figure 5: Average bandwidth of Dolly stream through 30 nodes with the the different networks alone and combined.

In Figure 5 we measure a dolly partition broadcast for a distribution to 60 nodes in parallel and examine the data distribution over the maintenance network, the reduced bisection switch E7 and the full bisection switch ER16. As expected from the results in Section 4.2, Dolly is able to use the full bandwidth on the E7 and ER16 switches. The maintenance network is able to handle nearly the same stream bandwidth since its minimal cost switches do not run into any bandwidth limitation for 16 connected nodes. Furthermore, we combine the maintenance network with the different data networks using Dollys capability to send data over both interfaces to double the throughput.

For the data distribution application Dolly, the cheap maintenance network offers roughly the same performance as the expensive full bisection Ethernet with its large central switch and therefore the maintenance network is a cost effective investment to double Dollys performance.

6.2 HPL

HPL (High Performance Linpack [6]) is a popular benchmark suite to evaluate the computational capabilities of supercomputers and clusters. The results of that benchmark are published semi-annually in the *Top500* list of the worlds most powerful computers [13]. The benchmark involves solving a system of linear equations.

The results of the benchmark depend only moderately on the performance of the underlying communication network and the tasks executing at the different nodes of the cluster are mostly compute bound. The communication pattern involves broadcasting panels of columns, which can be done by six different broadcasting algorithms. We used the algorithms “Increasing-ring” and “Increasing-2-ring(modified)” as they gave the best performance. The communication is mostly between near neighbor nodes in any time-step and does not seem to require a high bisection bandwidth.

We examine the results of the HPL benchmark which was run on 16, 24, 32 and 64 processors with the high performance Myrinet network (in shared mode for 64 processors), the full bisection ER16, the reduced bisection E7 configuration 1 and the maintenance network. The results of the benchmark are shown in Figure 6. The HPL benchmark was not tuned for maximal performance on the Xibalba cluster, as every node uses 50 MByte of memory during all the experiments. The results are fine to compare the different networking architectures against eachother, but should not be used to compare the performance of the Xibalba cluster with other clusters (a Top500 test with optimal parameters resulted in approximately 60 GFlops on Xibalba).

The results of the HPL benchmark on 16 nodes reveal no difference between the Ethernet architectures, as all the nodes are directly connected to the same switching module/switch in all cases and are thus practically identical. As more nodes are used the limited bisection bandwidth of the maintenance and the E7 network become more important factors. The same holds for the additional latency due to the stacked switches in the maintenance network. The latter low cost architecture shows clearly worse performance than the fully switched Ethernets networks with single central switches. The two Ethernet networks E7 and ER16 achieve about the same performance, the only difference being the slightly higher latency of the ER16 due to its more sophisticated higher level switching features. Myrinet with its much higher bandwidth and lower latency surpassed all the Ethernet network architectures by more than 50%.

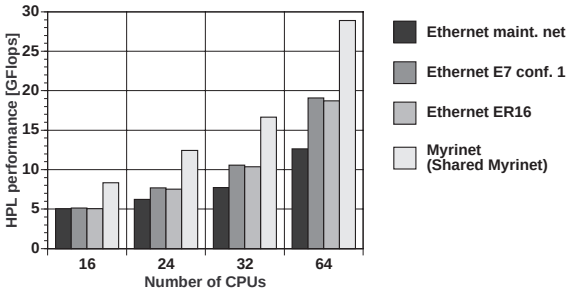

Figure 6: Performance comparison of cheap, medium and expensive networking architectures for the HPL benchmark in GFlops.

Since the MFlop counts of the HPL Benchmarks are well documented we can calculate a price/performance ratio for the different networking architectures based of the cost of $818 per port for the full bisection Ethernet ER16, $480 per port for the reduced bisection E7 and $145 per port for the maintenance Ethernet. For the different machine sizes the price per MFlop is roughly constant at about $1.60 for the E7 and $2.40 for the ER16 configuration and varies from about $0.50 on 16 nodes to about $0.75 on 64 nodes for the maintenance network. Thus, adding a better network increases the performance, but reduces the price performance ratio of a machine for HPL.

Mostly due to its much better latency Myrinet performs best in all the tests and scales nearly perfectly up to 64 nodes. The optimal scaling also holds for the 64 processor configuration where two CPUs share a motherboard and single network adapter. Due to the cost effective dual CPU

configuration the cost performance ratio is about $1.80 per MFlop with 2×$900 per node allocated to the interconnect.

6.3 QTPlan

QTPlan is a parallel program to model queuing in traffic micro-simulations [4]. In our benchmark the application simulated 6 hours of real-time traffic in Switzerland. The input comprised 50'000 and 990'000 automobiles respectively, on their way through a two lane tunnel of the single highway passage to the southern part of Switzerland. The road map is space partitioned in order to minimize the number of connections between the processor nodes. The 50 K cars case is a testing scenario for traffic jams, since all vehicles drive to the southern part of the country. The crossing of the partitions around the actual traffic bottleneck translate into a network bottleneck between the two machines that hold these partitions. The 990 K cars scenario is simulating a more balanced everyday scenario with most of the cars on the way to and from work all over Switzerland.

The QTPlan simulation has computing as well as communication intensive parts. The communication involves mostly small data packets at a fine granularity and requires a low latency interconnect. The performance results of the QTPlan application are shown in Figure 7.

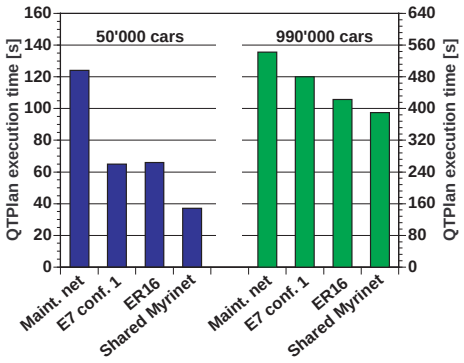

Figure 7: Runtime of QTPlan traffic simulations for 50'000 and 990'000 cars on 64 nodes/processors in seconds.

The execution times of the application tests are taken from a single run with 64 processors. The tests indicate a better performance on the Ethernet networks with central switches for small numbers of cars compared to the weaker maintenance network. The higher bandwidth and lower latency of the high performance shared Myrinet network results in nearly halved execution times. With a small working set there are less cars in each space partition (and therefore in each node) and the ratio of computation vs. communication drops. With a greater proportion of communication the factor network becomes more relevant resulting in a runtime that doubled on the minimal cost maintenance network. With many cars in the simulation, the performance difference between the different networking architectures becomes smaller. With large working-sets, the amount of local computation increases in every space partition resulting in a higher computation vs. communication ratio. The contribution of the factor network to the total runtime decreases and results in a smaller difference in runtime. Another factor that contributes to the higher runtime on the cheap network is the increased average latency because of the stacked switches.

The cost/performance analysis is more difficult with QTPlan since we do not have GFlops numbers. An evaluation of the execution times shows that the full bisection network offers about the same performance as the reduced bisection network, but both switched networks are a significant improvement over the minimal maintenance network. Myrinet can improve the results in the same amount as the ER16 improves the performance over the maintenance network.

7 Conclusion

In this study we considered four networks as alternatives to connect the compute nodes of a PC cluster: (1) Myrinet 2000, a dedicated, high-performance interconnect, (2) a high-end Fast Ethernet based on a switch delivering full bisection bandwidth, (3) a lower cost Fast Ethernet based on a central switch but with reduced bisection bandwidth and (4) a minimal cost Fast Ethernet designed as a secondary network for maintenance purposes.

Looking at the costs of these networks we see that depending on the performance requirement the total fraction of costs allocated to the network in a cluster can range from 4% for a lowest cost Fast Ethernet up to 35% for a multi Gigabit high-performance interconnect. The intermediate cost of 20% is reached for a full bisection, Fast Ethernet using an expensive high-end switch. It is also noted that there is an entirely different allocation of equipment costs between switches, cables and interface hardware in commodity networks (Fast Ethernet) and in the non-commodity high-performance networks (Myrinet). With Ethernet the cost is typically in the switches while with dedicated high-performance interconnects the high cost is in the sophisticated network adapters.

With the help of Switchbench, our congestion controlled all-to-all personalized communication (AAPC) micro-benchmark and some isolated pairwise communication patterns we managed to analyze the performance and the non-performance of different Fast Ethernet configurations built from off-the-shelf LAN communication equipment. For the performance analysis we developed two different views to see AAPC and state its performance. The first view gradually increases the logical communication distance in the different phases of the AAPC patterns revealing strength and weaknesses of the switches in full speed inter-module communication. The second view uses histograms over all possible source/destination pairs and routes to track down congested routes that slow down particular communication patterns and lead to poor performance in AAPC.

The performance and the non-performance of our different Ethernet switches gives a very interesting architectural insight, as we are trying to answer the question of whether commodity Ethernet components used in LAN networking can provide a fully scalable, full bisection interconnect cheaply and efficiently for a mid-sized PC cluster. But only a fairly expensive and powerful switch can reach near full bisection bandwidth on Fast Ethernet. Considering the cost/performance ratio it appears that the best cost/performance tradeoffs are at the low end of the Fast Ethernet interconnects built from low cost switches as used in our Xibalba secondary maintenance network or

then alternatively at the high end of the non-commodity high-speed interconnects. Spending large sums on a high-performance Ethernet switch to achieve full bisection bandwidth seems to increase costs unnecessarily. Our Xibalba cluster received such a top-of-the line switch only due to a vendor claim in the bidding process that finally lead to a free equipment upgrade.

Unlike with the microprocessor used for computing power the technical requirements for interconnects in mainstream networking with LANs remains sufficiently different from the setting of a high-performance PC cluster to warrant non-commodity hardware. Therefore the idea of using solely commodity networking components for a cluster interconnect remains questionable.

In addition to the raw performance figures measured by our micro-benchmarks we also managed to gain an idea of the overall performance impact of the interconnect on (a) a storage system utility for high-speed disk cloning using partition cast, on (b) the HPL benchmark used in the competition for the Top 500 list and (c) on a vehicular traffic simulator application that is used regularly on the cluster.

Our micro-benchmarks exercise the communication system and therefore the performance of the different interconnects appears highly significant at a first sight. However looking at the performance impact of networking on application codes a bit more carefully the differences are less than expected. Good network performance can be very helpful to get some high-performance codes up and running more quickly, but ultimately, most application codes can be rewritten to accommodate reduced communication bandwidth. While an Ethernet with full bisection bandwidth is a nice feature to have in a cluster, it is probably not the most critical issue for the success of a Beowulf-type system.

Acknowledgements

We would like to thank Nurhan Cetin and Kai Nagel for their help with the QTPlan application and the opportunity to use their application code for our experiments.

References

[1] D. J. Becker, T. Sterling, D. Savarese, J. E. Dorband, U. A. Ranawake, and C. V. Packer. Beowulf: A Parallel Workstation for Scientific Computation. In *Proceedings of International Conference on Parallel Processing*. CRC Press, Boca Raton, FL, USA, August 1995.

[2] Guy E. Blelloch, Charles E. Leiserson, Bruce M. Maggs, C. Greg Plaxton, Stephen J. Smith, and Marco Zagha. A comparison of sorting algorithms for the connection machine CM-2. In *ACM Symposium on Parallel Algorithms and Architectures*, pages 3–16, 1991.

[3] Nanette J. Boden, Robert E. Felderman, Alan E. Kulawik, Charles L. Seitz, Jakov N. Seizovic, and Wen-King Su. Myrinet — A Gigabit per Second Local Area Network. *IEEE-Micro*, 15(1):29–36, February 1995.

[4] Nurhan Cetin and Kai Nagel. Parallel Queue Model Approach to Traffic Microsimulations. In *Proceedings of Swiss Transportation Research Conference*, March 2002.

[5] William J. Dally. *A VLSI Architecture for Concurrent Data Structures*. PhD thesis, California Institute of Technology, Pasadena, 1986. Technical report 5209:TR:86.

[6] J. Dongarra, J. Bunch, C. Moler, and G. W. Stewart. Linpack users guide. SIAM, Philadelphia, PA, 1979.

[7] G. Hipper and D. Tavangarian. Advanced workstation cluster architectures for parallel computing. *Journal of Systems Architecture*, 44(3–4):207–226, December 1997.

[8] Ch. Kurmann and T. Stricker. A Comparison of Three Gigabit Technologies: SCI, Myrinet and SGI/Cray T3D. In *SCI Based Cluster Computing, H. Hellwagner and A. Reinefeld, eds.* Springer, Berlin, Spring 1999. An earlier version appeared in Proc. of the SCI Europe'98 Conference, EMMSEC'98, 28-30 Sept 1998, Bordeaux, France.

[9] Christian Kurmann. Switchbench benchmark, June 2001. `http://www.cs.inf.ethz.ch/CoPs/sbench/`.

[10] Christian Kurmann, Felix Rauch, and Thomas M. Stricker. Cost/Performance Tradeoffs in Network Interconnects for Clusters of PCs. Technical Report 391, Department of Computer Science, ETH Zürich, 2003. `http://www.inf.ethz.ch/`.

[11] Charles E. Leiserson, Zahi S. Abuhamdeh, David C. Douglas, Carl R. Feynman, Mahesh N. Ganmukhi, Jeffrey V. Hill, W. Daniel Hillis, Bradley C. Kuszmaul, Margaret A. St. Pierre, David S. Wells, Monica C. Wong, Shaw-Wen Yang, and Robert Zak. The Network Architecture of the Connection Machine CM-5. *Journal of Parallel and Distributed Computing*, 33(2):145–58, March 1996.

[12] Richard P. Martin, Amin M. Vahdat, David E. Culler, and Thomas E. Anderson. Effects of Communication Latency, Overhead, and Bandwidth in a Cluster Architecture. *Computer architecture news*, 25(2):85–97, May 1997. 24th Annual International Symposium on Computer Architecture, ISCA '97.

[13] Hans Meuer, Erich Strohmaier, Jack Dongarra, and Horst D. Simon. TOP500 Supercomputer Sites. `http://www.top500.org/`.

[14] Felix Rauch, Christian Kurmann, and Thomas M. Stricker. Partition Cast — Modelling and Optimizing the Distribution of Large Data Sets on PC Clusters. In Arndt Bode, Thomas Ludwig, Wolfgang Karl, and Roland Wismüller, editors, *Lecture Notes in Computer Science 1900, Euro-Par 2000 Parallel Processing, 6th International Euro-Par Conference Munich*, Munich, Germany, August 2000. Springer. Also available as Technical Report 343, Department of Computer Science, ETH Zürich, `http://www.inf.ethz.ch/`.

[15] Alexander Reinefeld and Volker Lindenstruth. How to Build a High-Performance Compute Cluster for the Grid. In *Proceedings of Intl. Workshop on Metacomputing Systems and Applications (MSA'2001)*, pages 221–227, September 2001.

[16] Chance Reschke, Thomas Sterling, Daniel Ridge, Daniel Savarese, Donald Becker, and Phillip Merkey. A Design Study of Alternative Network Topologies for the Beowulf Parallel Workstation. In *Proceedings of the Fifth IEEE International Symposium on High Performance Distributed Computing*, pages 626–635. IEEE Comput. Soc. Press, Los Alamitos, CA, USA, 1996.

[17] Thomas L. Sterling, John Salmon, Donald J. Becker, and Daniel F. Savarese. *How to Build a Beowulf: A Guide to Implementation and Application of PC Clusters*. MIT Press, Cambridge, USA, 1999.

[18] T. Stricker and J. Hardwick. From AAPC Algorithms to High Performance Permutation Routing and Sorting. In *Proc. SPAA'96*, pages 200–203, Padua, Italy, June 1996. ACM.

[19] Michael Stumm, Zvonko Vranesic, Ron White, Ron Unrau, and Keith I. Farkas. Experiences with the Hector Multiprocessor. In *Proceedings of the Seventh International Parallel Processing Symposium*, April 1994.