Measures of Latency on High Performance Processing Environments

André Barros Sales✍ Solange Teresinha Sari✍✍ Carlos Backer Westphall✍
andreb@inf.ufsc.br solange@npd.ufsc.br westphall@inf.ufsc.br

✍Network and Management Laboratory
Computer Science and Statistics Department
Phone +55 48 331-7559

✍✍Interoperability Laboratory
Data Processing Center
Phone +55 48 331-7535

Federal University of the Santa Catarina
Florianópolis – SC, Brazil.

Abstract

This article presents some performance evaluating results obtained by measuring computer clusters used for high computational capacity. For each cluster, named AIX, Linux, and SP2, measurements was taken comprising the data throughput in the transport layer (TCP and UDP). Using statistical results of the collected data, a linear relation between latency and segment size was determined. From the analysis of the clusters we observed that there is a strong relation between segment size and the latency for collision free environments. Analyzing the clusters, the smaller latency was shown by the Linux cluster with Fast-Ethernet. Although the AIX cluster throughput is greater than the Linux cluster throughput, other factors such as negotiation, encapsulating, etc, seems to consume much effort, increasing the overall latency.

Keywords:

Fault and Performance Management, High Performance Computing and Measures of Latency.

1. Introduction

As computational experiments grow in complexity more and more data is produced. Some of these experiments last a long time and sometimes they can not be completed with success. In most of the cases the computational simulation demands the execution of complex algorithms as well as the treatment of a very large volume of data, which needs high computational capacity. This situation came to a recent fast growth of the high performance computation – HPC. The Human Genome Project is an example, using about 700 processors for its calculations, which are broken into fragments of simple operations and sent to several processors to be executed.

At Federal University of the Santa Catarina, the HPC environment has been used for Meteorology, Numeric Calculations and Energy System applications, comprising a group of processors of several computers interconnected by a fast network,. This kind of arrangement is called *cluster*, allowing cooperation and data sharing among those processors and the execution performance depends on the number of processors and also on the network speed. More specifically it depends on the data transmission time among processors. Analytic models to evaluate HPC applications can be found in [HAA00], [WEL97] and [JOH94].

The aim of this article is to show how to evaluate HPC environments using measures of latency. Analyzing the variability of the end-to-end latency with segment size in the transport

layer, the goal is to optimize the configuration parameters of each cluster, as well as to make performance comparisons among them.

In the following sections some basic concepts of parallel processing and performance evaluation that support the work are presented. In section 4 the used measurement method is described, and in section 5, the collected values and resulting analysis are presented. Section 6 presents the conclusions.

2. Measurement Method

The objective of a measurement method is to find the characteristic latency values on each cluster being used for high performance computing. It starts with a description of the environment under study and the type of simulated process. Parameters for data collection are specified. The resulting data are then presented as a box and whiskers plot allowing the visualization of statistical values and the clusters performance evaluation. Starting from the samples of data a forecast model is defined through a linear regression where latency values are function of segment size. This measurement model is based on RFC 2544 [BRA99], where recommendations for latency measurement in interconnection equipment are given and also on some performance evaluation issues in Tanenbaum [TAN96].

2.1. Network Description

A research network at Federal University of Santa Catarina, called *redeCluster*, was used in our experiments. The *redeCluster* network comprises an ATM backbone connecting several Ethernet and Fast Ethernet subnets (Figure 2-1). The IBM and 3COM connecting equipments have interfaces whose bandwidths are 155 and 622 Mbps. The host's characteristics are depicted in Table 2-1 for each cluster. The LAN Emulation service is used for communication among several network types.

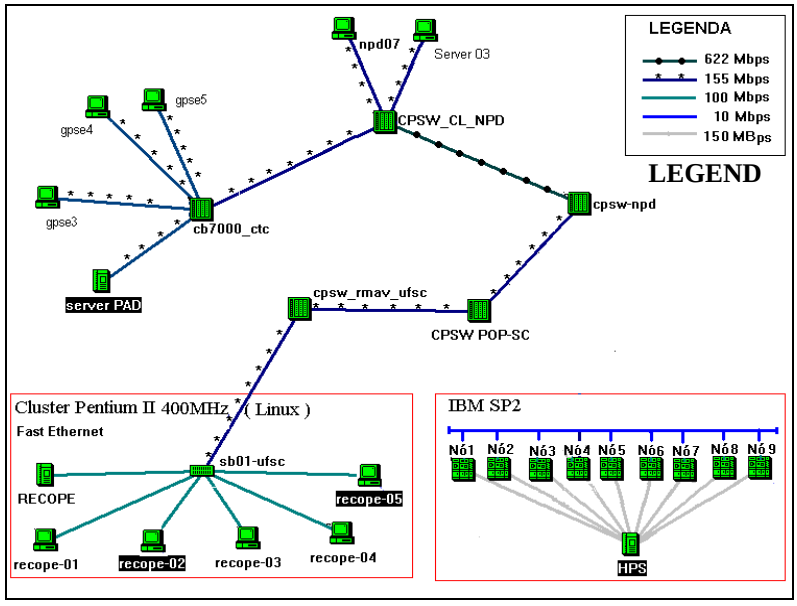


Figure 2-1 – The *redeCluster* Network Topology.

Table 2-1- Host in each cluster used for evaluation.

Clusters	Hosts	Network type	Hardware	OS
AIX	Gpse3, Gpse4, Gpse5 e Server Pad	ATM 155Mbit/s	Risc 6000 - 43P 200 MHz 64 MB	AIX versão 4.3
Linux	Recope, Recope1, Recope2, Recope3, Recope4, Recope5	Fast-Ethernet	Pentium II-400MHz 233 MHz 32MB	Linux–RedHat v.6.1 Kernel 2.12
SP2	SP2 (com 9 Nós)	Ethernet	IBM 9076, Risc 6000 66-135 MHz 128-512 MB	AIX versão 4.1

Functional parallelism applications are simulated, so that traffic is generated simultaneously among the stations.

2.2. Data Collection

In our experiments Netperf generates TCP /UDP traffic and measures the latency values among hosts. Four pairs of hosts are selected in each cluster using as parameters:

- segment size – 64, 128, 256, 512, 1024, 1280 e 1518 bytes, according to RFC 2544;
- execution time - 70 seconds, according RFC 2544;
- repetitions numbers - 20 time, representative amount to get average, according to RFC 2544;
- sequential measure - during the week and weekend so that the sample is significant, according to [TAN96];
- maximum transmission unit (MTU) - 1500 *octets*; and
- TCP window size - default operating system value, according to RFC1323 *"Large Windows*".

2.3. Performance Evaluation

The resulting data is analyzed through box and whisker for each protocol. These graphics represent the central tendency and latency variability.

The box plot describes the central tendency of the latency in terms of the median of the values, represented by the smallest box in the plot ( ). The spread (variability) in the latency value are represented by quartiles (the 25th and 75th percentiles, larger box in the plot, ( )). The minimum ( ) and maximum values ( ) of the latency are represented by "whiskers" in the plot.

2.4. Prediction

The prediction model is based on regression analysis. In the regression equation, the dependent value is the latency and the independent value is the segment size. The amount of common variation between the two values is given by the coefficient of determination, R^2 [BAR98] expressed in [eq 2-1]. The dispersion of the sample data related to the linear line is shown graphically.

$$R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\text{explained variability}}{\text{total variability}}$$

Where: $\hat{Y}$ = predict value; $\bar{Y}$ = aritmetic medium; Y = expected value [eq 2-1]

3. Results Analysis

To analyze the collected data some initial considerations are needed:

?? The segment size is in the 64 and 1.518 bytes range, and the latency is in milliseconds;

?? The host latency is not taken into account.

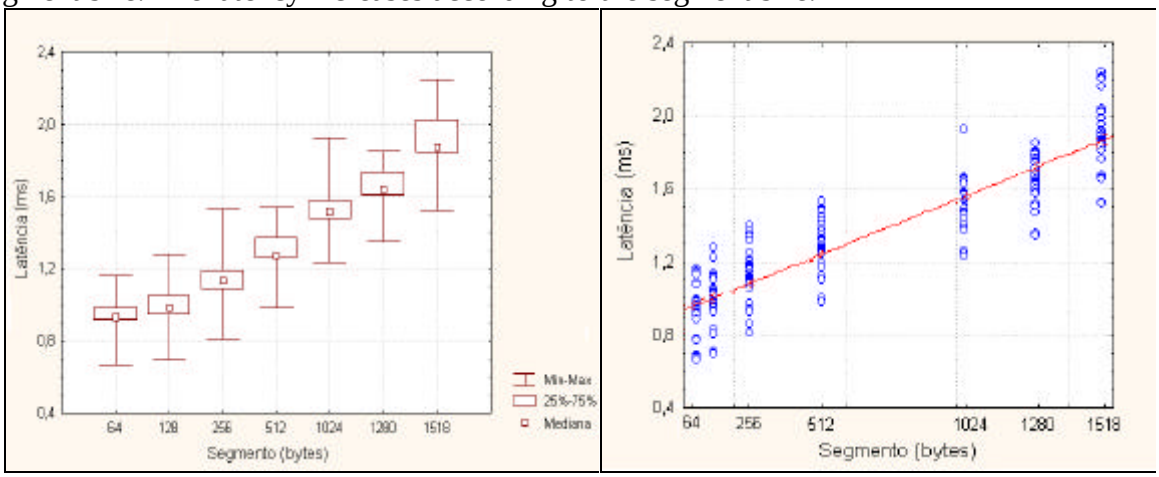
?? Preliminaries tests results with the loop-back interface showed that the internal latency is smaller than the end-to-end latency;

?? Although the measurement tool allows a variable MTU, in this case a fixed value of 1500 bytes is used.

3.1. Analysis for the AIX Cluster

3.1.1. TCP Traffic

In the box plot of Figure 3-1(a) the minimum and maximum values are closed to each other (i.e. 128 bytes: ? 0,7 ms and T 1,3 ms) and a dependence between the latency and TCP segment size. The latency increases according to the segment size.



(a) (b)

Figure 3-1 – Relation between latency (ms) and TCP segment size (bytes) for the Cluster AIX (a) box plot e (b) scatterplot.

Starting from the sample data, as shown in Figure 3-1(b), is defined the regression equation [eq. 3-1] with determination coefficient equal $R^2 = 0{,}834197$.

$$Lat\_TCP \;?\; 0{,}925607 \;?\; 0{,}000615 * TCP\_segment\_size \qquad \text{[eq. 3-1]}$$

Considering that the determination coefficient is approximately 1, it follows that this equation represents the variability behavior of the latency for the net. It can be explained by 83% due to variation of segment size and 17% caused by other factors.

The UDP traffic behavior follows the same tendency observed for the TCP traffic. For example, the UDP segment of 64 bytes has ? = 0,6 ms and T = 1,2 ms. The inferior and superior quartiles are between 0,9 and 1,0 ms, presenting a regular behavior.

The regression equation [eq. 3-2] has determination coefficient $R^2 = 0{,}83279$, representing the high dependence of the latency in relation to the UDP segment size, according to the sample data of **Erro! A origem da referência não foi encontrada.**(b).

$$Lat\_UDP \;?\; 0{,}894871 \;?\; 0{,}000636 * UDP\_segment\_size \qquad \text{[eq. 3-2]}$$

The determination coefficient indicates that the equation [eq. 3-2] represents the latency's variability and can be explained in 83%.

3.2. Analysis of the Linux cluster

Due to incompatibilities between the operating system with the Netperf tool version used it was not possible to gather data for the UDP segment in the Linux cluster. Analyzing the collected data in the Linux cluster one can observe an irregular behavior of the latency for a segment size equal to 1.580 bytes (latency graphic, Figure 3-2).

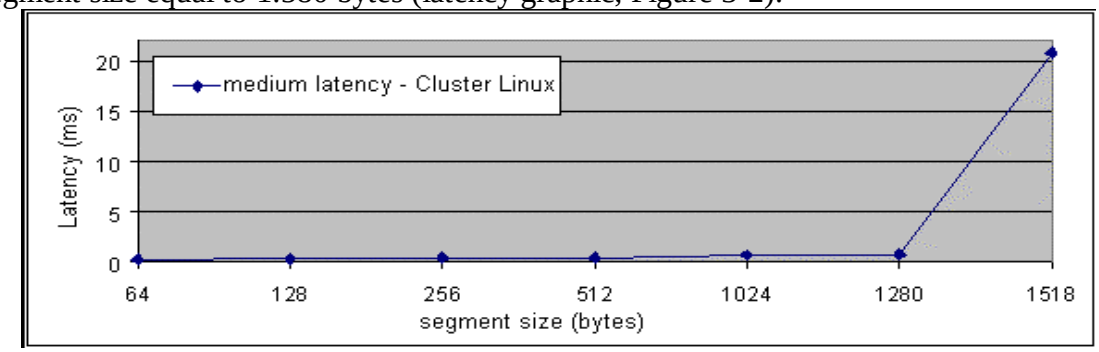


Figure 3-2 – Relation between latency (ms) and TCP segment size (bytes) for the Linux cluster. (a) box plot e (b) scatterplot.

Considering that the latency values are smaller than 1.280 bytes and that it has exponential increasing after 1.518 bytes, we can list some possible causes:

- configuration error of the MTU size in the hosts;
- configuration error of the MTU size in the *switch*;
- implementations errors in the transport protocol in the hosts; or
- problems with the *switch*.

New tests were conducted for end-to-end and loopback latency trying to reveal the erroneous results. It was observed that the latency has the same behavior presented in the Figure 3-3. In [DIE98] an implementation error was recognized in the Nagle algorithmic of the TCP protocol for the Kernel 2.12.

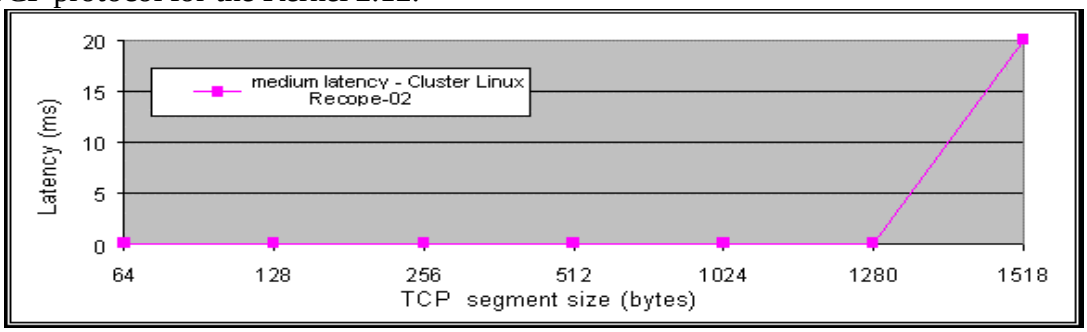


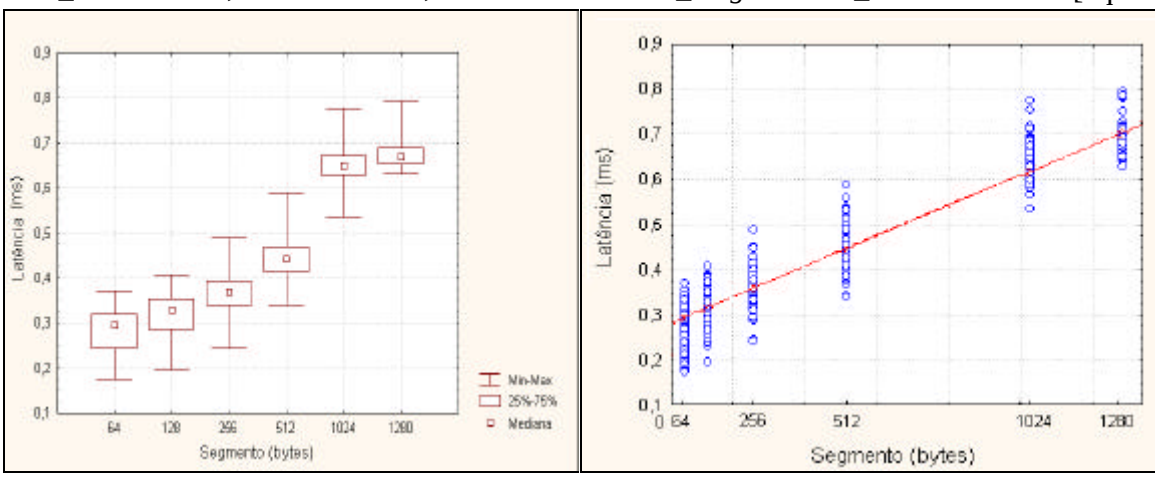
Figure 3-3 – Relation between latency (ms) and TCP segment size (bytes) for the loopback interface in the host Recope02.

3.2.1. TCP Traffic

The Figure 3-4(a) shows a small variation of the latency for each TCP segment size in the Linux cluster, as well as the median position and the inferior and superior quartiles. For the segment size of 128 bytes is ? = 0,2 ms and T = 0,4. The data is not dispersed as shown

in the Figure 3-4(b). The equation [eq 3-1] expresses the linear relationship through the determination coefficient $R^2 = 0{,}913747$.

$$Lat\_TCP \;?\; 0{,}273835 \;?\; 0{,}000336 * TCP\_segment\_size \quad \text{[eq 3-1]}$$



(a) (b)

Figure 3-4– Relation between latency (ms) and TCP segment size (bytes) for the Linux cluster (a) box plot e (b) scatterplot.

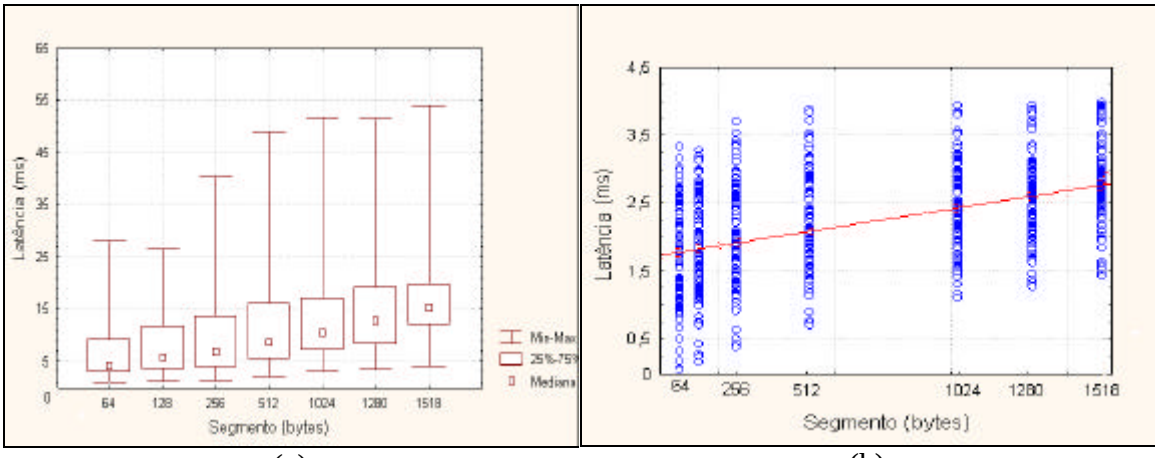
3.3. Analysis for the SP 2

3.3.1. TCP Traffic

Figure 3-5(a) shows the increase of median latency related to the segment size, but there is a big variation between minimal and maximum latency. For example, the TCP segment size of 64 bytes has ? = 1 ms and T = 27 ms. This behavior is maintained for other segment sizes, indicating the possibility of another cause affecting the latency value in the SP2 besides the segment size.

The low determination coefficient, $R^2 = 0.231976$, shows that the regression equation [eq 3-2] can not be used for latency prediction. The great dispersion of the sample data is shown in the Figure 3-5(b).

$$Lat\_TCP \;?\; 1{,}735339 \;?\; 0{,}000677 * TCP\_segment\_size \quad \text{[eq 3-2]}$$



(a) (b)

Figure 3-5- Relation between latency (ms) and TCP segment size (bytes) for the SP2. (a) box plot e (b) scatterplot.

New tests with fewer loads were conducted in the network, with a reduced number of processors and no simultaneous applications being run. This resulted in a different behavior for a reduced number of processors, indicating that the collision rate is another factor affecting latency on the Ethernet bus.

The UDP traffic behavior is similar to TCP, also presenting an irregular area. The great distance between the inferior and superior quartiles indicates collision affecting latency. The regression equation [eq 3-3 has low determination coefficient, $R^2 = 0.01587371$.

*Lat _ UDP ? 2,977304 ? 0,000409 * UDP _ segment _ size* [eq 3-3

The sample data dispersion confirms for an Ethernet bus that the latency variability does not depend only on the size of the segment but also on collisions that happen in the medium access.

3.4. Latency Results Comparison

Table 5.1 presents the medium's latency values for each protocol type in each *cluster* (and determination coefficient respectively) related to several segment sizes. It is observed thar the Linux cluster has the smallest latency, followed by AIX and SP2 clusters, for all segment sizes analyzed.

Table 3.1 – Latency values in the clusters for TCP and UDP several segment sizes.

	Linux	AIX		SP2	
Protocol	R^2 0,91	R^2 0,83	R^2 0,85	R^2 0,01	R^2 0,23
Segment	TCP	UDP	TCP	UDP	TCP
64	0.295339	0.964967	0.935575	1.778718	3.00348
128	0.316843	1.004327	0.976279	1.822046	3.029656
256	0.359851	1.083047	1.057687	1.908702	3.082008
512	0.445867	1.240487	1.220503	2.082014	3.186712
1.024	0.617899	1.555367	1.546135	2.428638	3.39612
1.280	0.703915	1.712807	1.708951	2.60195	3.500824
1.580	0.783883*	1.859177	1.860319	2.763076	3.598166

The low latency value in the Linux cluster is due to the combination of operating system characteristics with a Fast-Ethernet network (without collision). It guarantees smaller times of traffic on the net, confirming the recommendations given by [DIE98]. Another issue is the segment size, since for larger values of 1.500 bytes it is necessary a segmnet fragmentation, which could affect the latency directly.

Although the Aix cluster uses a faster ATM technology with its 155 Mbit/s compared to Fast-Ethernet 100 Mbit/s, it shows a greater latency due to message encapsulation time and negotiation of LANE service in the ATM network. These factors affect latency for segment sizes between 64 and 1.580 bytes.

The low determination latency coefficient and segment size variability for the SP2 suggest another factor not considered in the measurements, possibly resulting from collisions on the Ethernet bus. The HPS has not been used due to technical problems.

* Predictable time in [eq 3-1].

4. Conclusions

This work presents results of measuring network latency – the time a segment takes to travel between end-to-end applications in function of its size. The objective is the performance measurement of some clusters used for high performance computing (AIX, Linux, and SP2).

After the resulted analysis one can possibly conclude:

✍ For the AIX cluster – The linear model described express 85% and 83% of the variability between latency and segment size (TCP and UDP respectively);

✍ For the Linux cluster– The linear model described express 91% of the variability between latency and TCP segment size;

✍ For the SP2 cluster – The linear model is not representative, explaining only 23% and 1.5% of the variability between latency and segment size (TCP and UDP respectively).

From the analysis of the clusters we can conclude that there is a strong relation between segment size and the latency for collision free environments. Analyzing the clusters, the smaller latency was shown by the Linux cluster with Fast-Ethernet. Although the AIX cluster throughput is greater than the Linux cluster throughput, other factors such as negotiation, encapsulating, etc, seems to consume much effort, increasing the overall latency.

Bibliographical References

[BAR98] BARBETTA, Pedro Alberto, Estatística aplicada às Ciências Sociais, 2 ed. – Florianópolis: Editora da UFSC, 1998.

[BRA99] BRADNER, S., *Benchmarking Methodology for Network Interconnect Devices*, Network Working Group, Request for Comments: 2544, March 1999.

[DIE98] DIETZ, Hank, *Linux Parallel Processing HOWTO*, janeiro 1998. Url http://www.how2linux.com/, página visitada em 15 de junho de 2000.

[HAA00] HAAS, Reinaldo, AMBRIZZI, Tércio, FILHO, Augusto José Pereira, Comparação de Desempenho entre um Cluster PC-Linux e um SP 2 em simulações com o Modelo ARPS, XI Congresso Brasileiro de Meteorologia, Centro Cultural da UERJ - RJ, Outubro de 2000.

[JOH94] JOHNSON, Kenneth W. BAUER, Jeff, RICCARDI, Gregory A., XUE, Ming, DROEGEMEEIER, Kelvin K., *Distributed Processing of a Regional Prediction Model*, Monthly Weather Review, 122: 2558-2572, 1994.

[TAN96] TANENBAUM, Andrew S. *Redes de Computadores*, tradução [ds 3. Ed. Original] Insight Serviços de Informática. Rio de Janeiro: Campus, 1997.

[WEL97] WELSH, Matt, BASU, Anindya, EICKEN Thorsten von, *ATM and Fast Ethernet Network Interfaces for User-Level Communication*, Proceedings of High-Performance Computer Architecture 3, San Antonio, February 1997.