# Projeto de uma Biblioteca de Comunicação Infiniband *

Rodrigo da Rosa Righi†, Philippe Olivier Alexandre Navaux†, Marcelo Pasin‡

† PPGC - UFRGS, Av. Bento Gonçalves 9500, Bloco IV, Porto Alegre/RS
‡ UFSM, Informática CT, Campus UFSM, Santa Maria/RS
{rrrighi, navaux}@inf.ufrgs.br, pasin@inf.ufsm.br

## Resumo

Os aglomerados de computadores (*clusters*) são freqüentemente utilizados como suporte ao processamento de alto desempenho. Com o intuito de maximizar o desempenho dessas máquinas, são propostas novas arquiteturas e protocolos de conexão entre os nós que as compõem. Nesse contexto, surgiram a Arquitetura de Interface Virtual (VIA) e a Arquitetura Infiniband (IBA). Esta última é uma evolução de VIA e apresenta a característica de operar sobre uma malha de chaveadores, provendo uma vazão superior a 10 Gbps. O presente trabalho tem por objetivo apresentar o projeto de construção de uma biblioteca de comunicação para dispositivos Infiniband.
**Palavras-chave: biblioteca de comunicação, aglomerados, interconexão**

## Introdução

A máquina paralela baseada em aglomerado de computadores é composta de computadores comuns, também chamados de nós, e uma rede de interconexão dedicada e rápida [Pasin and Kreutz, 2003]. Em uma aplicação desenvolvida para executar em um aglomerado, os nós necessitam trocar mensagens e realizar sincronizações e, tais tarefas são realizadas através da rede de interconexão dos nós. Com o intuito de maximizar o desempenho de aglomerados, existem pesquisa na área que compreende a troca de mensagens, pois, para certas aplicações, estas operações podem ser mais onerosas que o tempo de processamento.

Nesse contexto, foi desenvolvida, em 1997, a especificação da Arquitetura de Interface Virtual, ou VIA, com o intuito de padronizar uma série de protocolos de comunicação de alto desempenho. A principal característica de VIA é a sua capacidade de realizar a troca de mensagens sem a intervenção do sistema operacional. Além disso, VIA minimiza a quantidade de cópias intermediárias de memória nestas operações [Speight et al., 1999]. Ela fornece redução na carga de processamento local, baixa latência de comunicação e alta largura de banda. O projeto da arquitetura VIA foi redirecionado para a construção de uma nova tecnologia de interconexão, chamada de arquitetura Infiniband, ou IBA (*Infiniband Architecture*) [InfiniBand Trade Association, 2001].

*Financiamento: CNPq (processo 552256/02-1) e projeto Labtec-DELL

## Desenvolvimento de uma Biblioteca de Comunicação

A arquitetura Infiniband é uma nova tecnologia para servidores de E/S e especifica métodos para conectar servidores, dispositivos de E/S e servidores com dispositivos de E/S. Além destas capacidades, Infiniband especifica uma infra-estrutura para ligação de alto desempenho dentro de estações de trabalho, substituindo, por exemplo, a tecnologia PCI (*Peripheral Component Interconnect*) [Barcellos and Gaspary, 2003]. O projeto da tecnologia Infiniband foi totalmente baseado nas caracteríticas de alto desempenho de VIA, como o agendamento de DMA (*Direct Memory Access*), operações DMA remotas (RDMA), cópia zero e utilização de filas de trabalho para processar a troca de mensagens. IBA substitui a estrutura de barramento compartilhado por uma outra com comunicação chaveada de alta velocidade e ligações ponto-a-ponto. Ela possui muitas capacidades que são previamente encontradas na arquitetura de máquinas *mainframes* e fornece alto desempenho nas operações de troca de mensagens entre os nós de um aglomerado.

Atualmente, não existe uma padronização quanto a interface de programação para Infiniband. As empresas Mellanox e IBM possuem as suas próprias interfaces de programação para esta tecnologia. Para operações RDMA, existe uma tentativa de padronização através do projeto DAPL (*Direct Access Provider Library*) [DAPL Project, 2003].

Para realizar pesquisa na área inerente a arquitetura Infiniband, o Instituto de Informática da UFRGS adquiriu um conjunto de *hardware* Infiniband da empresa Mellanox, composto de 4 adaptadores de rede MTEK23108B-C02, cada qual com duas portas de entrada, 8 cabos de interconexão e um chaveador de 8 portas MTEK43132. Para a execução de testes, foi instalado e configurado o conjunto de *hardware* Infiniband no aglomerado de computadores do Laboratório de Tecnologia em Clusters (LabTeC).

Como um dos principais pontos de pesquisa para redes Infiniband é a interface de programação, está sendo construída uma biblioteca de comunicação. Num primeiro momento, foi realizado um estudo sobre as capacidades da tecnologia e sobre as interfaces de programação Infiniband disponibilizadas para o usuário. A biblioteca de comunicação que está em desenvolvimento objetiva disponibilizar uma interface de programação de alto nível para a escrita de aplicações Infiniband para aglomerados. Os programas escritos com essa biblioteca podem usufruir de alta largura de banda nas operações de troca de mensagens, pois a malha Infiniband proporciona uma vazão superior a 10 Gbps e características como múltiplos caminhos (*multipath*) e tolerância a falhas.

## Referências

Barcellos, M. P. and Gaspary, L. P. (2003). Tecnologias de rede para processamento de alto desempenho. In *Escola Regional de Alto Desempenho*, volume 3, pages 67–98, Santa Maria, RS.

DAPL Project (2003). Atualização: Set. 2003 http://sourceforge.net/projects/dapl.

InfiniBand Trade Association, I. B. T. A. (2001). InifiniBand Architecture Specification Volume 1, Release 1.0.a. Especificação do padrão. Disponível em http://www.infinibandta.org/specs. 2001.

Pasin, M. and Kreutz, D. L. (2003). Arquitetura e administração de aglomerados. In *Escola Regional de Alto Desempenho*, volume 3, pages 3–31, Santa Maria, RS.

Speight, E., Abdel-Shafi, H., and Bennett, J. K. (1999). Realizing the Performance Potential of the Virtual Interface Architecture. In *International Conference on Supercomputing*, pages 184–192.