ANÉIS E HIERARQUIAS DE ANÉIS COM INTERCONEXÕES ANSI/IEEE SCI

Sergio T. Kofuji, Carlos A. Paiva da Silva, Luís G. G. Kiatake, Marcelo H.S. Cintra , João A. Zuffo

Laboratório de Sistemas Integráveis, Escola Politécnica da Universidade de São Paulo
Av. Prof. Luciano Gualberto, trav. 3, 158. CEP 05508-900 - São Paulo, SP, Brasil
tel: (011) 818-5667 fax: (011) 211-4574. kofuji@lsi.usp.br

RESUMO

Tecnologias emergentes de comunicação ponto-a-ponto de alta velocidade, como o ATM, o SCI e o FibreChannel, abrem novos horizontes na implementação de sistemas de computação de alto desempenho. O SCI é um padrão ANSI/IEEE que provê recursos para a implementação de coerência de cache entre nós de processamento, permitindo a construção de multiprocessadores CC-NUMA com base em estações de trabalho. O SCI vem sendo estudado como opção para a interconexão de nós de processamento dentro do projeto SPADE, ora em desenvolvimento no LSI-EPUSP. Neste trabalho são feitas simulações do SPADE-I com topologias anel e hierarquias de anéis baseadas em ligações de 125 MBytes/s, que representam a tecnologia hoje disponível de produtos SCI a nível comercial. Em particular, é investigado o impacto do controle de fluxo e tamanho de filas em sistemas com interconexão SCI.

ABSTRACT

Emergent high speed point-to-point communication technologies, for instance, ATM, SCI and FibreChannel, open new perspectives in high performance computation. The SCI is a new IEEE/ANSI standard that specifies protocols for scalable cache coherence over point-to-point high speed, low latency links. The SCI is been studied at LSI-EPUSP as a alternative to implement the SPADE, a large scale multiprocessor system. In this work, simulations of the SPADE-I in ring and hierarchy of rings topologies based on links of 125MB/s are done in order to evaluate the impact of flow control and buffer size.

1 - INTRODUÇÃO

Tecnologias emergentes de interconexão ponto-a-ponto vem tornando possível a construção de aglomerados de estações com características comparáveis aos dos supercomputadores paralelos de larga escala, apresentando a vantagem de menor custo e tecnologia aberta. Diferente dos outros padrões, o padrão de interface **ANSI/IEEE-SCI** provê recursos para implementação de sistemas com memória compartilhada com caches coerentes.

O **SPADE-I** é um sistema paralelo em estudo no Laboratório de Sistemas Integráveis da **EPUSP** baseado em estações de trabalho e padrão **SCI**. O **SPADE-I** não se baseia em uma topologia específica. Diversas topologias podem ser implementadas com base nos anéis e chaves **SCI**: hierarquia de anéis, malha multidimensional, rede multiestágio, etc. O anel é a topologia mais simples e a forma padrão de interconexão com o padrão **SCI** e, pela sua simplicidade, é o candidato natural para a a implementação de multiprocessadores de pequena escala, com até algumas dezenas de processadores. Para a formação de multiprocessadores com um número maior de processadores, outras topologias com maior banda e menor latência devem ser empregadas. Pretende-se no **SPADE-I** dar ênfase a sistemas de interconexão que possibilitem a exploração de localidade de referências, como hierarquias de anéis, árvores gordas ("fat-trees") implementadas com redes multiestágio bidirecionais, e redes ***n-cúbicas k-árias***, como malhas multidimensionais.

Este trabalho faz uma avaliação de anéis e hierarquia de anéis através de simulações empregando o simulador **SCILAb** desenvolvido pelo **CERN**. Pretende-se investigar o impacto do controle de fluxo ("go-bit") e do tamanho das filas no desempenho destas topologias, que dada a sua simplicidade, são as configurações naturais de pequeno porte do **SPADE-I**. Diferente de outros estudos, o desempenho das redes é investigado com transações com resposta, que são as transações típicas de sistemas **SCI** com coerência de cache. Inicialmente, na seção 2, faz-se uma breve apresentação do padrão e dos trabalhos relacionados. Na seção 3 descreve-se o simulador e a metodologia empregada. Os resultados para anéis simples e hierarquia de anéis são apresentados na seção 4. Finalmente, na seção 5 temos as conclusões principais do trabalho.

2 - O PADRÃO ANSI/IEEE-SCI

O **SCI** é um padrão emergente **ANSI/IEEE** que especifica serviços típicos de barramento entre nós de processamento através de ligações unidirecionais ponto-a-ponto de até 1GBytes/s. A interface básica **SCI** consiste de uma única ligação de entrada e uma única ligação de saída. Pacotes passando através de um nó são submetidos a um pequeno atraso, da ordem de nanossegundos e, na ausência de contenção, o progresso de um pacote ao longo de um anel é largamente limitado pelo tempo de propagação no cabo e número de nós intermediários.

Um nó **SCI** geralmente contém uma interface **SCI**, zero ou mais processadores, e possivelmente memória e interface de **E/S** para dispositivos periféricos. O nó pode enviar e

receber pacotes. Como receptor, ele é capaz de reconhecer pacotes endereçados a ele, retirar o pacote da rede e enviar um pacote de eco avisando ao remetente o recebimento do pacote.

Uma chave é um nó contendo mais que uma interface **SCI**, com a capacidade para rotear pacotes de uma interface **SCI** para outra. Diversos anéis **SCI** podem ser interconectados através de chaves, permitindo configurações com grande número de processadores, sem a restrição de desempenho de um único anel. Várias topologias têm sido investigadas para interconectar uma grande quantidade de nós, como *cross-bar*, hierarquia de anéis, malhas **2D** e **3D**, redes multiestágio, etc.

2.1 ANÉIS E HIERARQUIA DE ANÉIS

Diversos sistemas multiprocessadores de larga escala têm sido propostos ou implementados baseados em anel, entre os quais o **Hector** [1] da Universidade de Toronto, o **Paradigm** [2] da Universidade de Stanford e o **KSR-1** da Kendall Square [3].

Existem pelo menos três tipos de soluções para controle de acesso em anéis: anéis com passagem de ficha ("token ring"), anéis com inserção de registros ("register insertion ring") e anéis com intervalos ("slotted ring"). Nos anéis com passagem de ficha um padrão de bits especial chamado ficha é passado de um nó a outro dando a permissão para transmitir; nos anéis com inserção de registros, como os anéis SCI, uma fila de transpasse armazena os pacotes que estiverem chegando ao nó enquanto este estiver transmitindo; e por fim, nos anéis com intervalos a banda passante do anel é dividida em intervalos de diferentes tamanhos, de modo que para um processador transmitir ele tem que esperar o próximo intervalo livre com o tamanho desejado.

Os anéis com passagem de ficha convencionais apresentam a desvantagem de permitir que apenas uma mensagem possa fluir no anel ao mesmo tempo, e não se presta à implementação de sistemas de alto desempenho, a não ser em casos onde o tamanho das mensagens é maior do que o atraso inerente do anel [4]. Devido à melhor adequabilidade ao tráfego com pacotes pequenos e à facilidade de implementação de protocolos de coerência de cache baseados em monitoração [5,6], os anéis com intervalos (em inglês "slotted ring"), que têm sido empregados há vários anos em redes locais [7,8], têm também sido utilizados como forma de interconexão em multiprocessadores de larga escala [9]. Uma das propostas de coerência de caches em anéis com intervalos é devida a Barroso e Dubois, que aproveitam a possibilidade de difusão global em um anel para impor a coerência por meio de simples monitoração. Simulações e modelos analíticos mostram que o esquema proposto é superior a outros baseados em anéis, como o protocolo diretório utilizado no **SCI** ou o protocolo de monitoração do **KSR-1**. É preciso ressaltar, todavia, que o protocolo proposto não provê o mesmo nível de escalabilidade proporcionado pelo **SCI**, e portanto não se presta à construção de sistemas muito grandes, com milhares de processadores.

Dentre os sistemas baseados em anel com intervalos pode-se citar o **KSR-1**, e como exemplos de sistemas com anel com passagem de ficha o **Hector** e o **MemNet**. O **KSR-1** é o primeiro multiprocessador de larga escala comercial a empregar anéis com intervalos, e diversas avaliações experimentais têm mostrado uma boa escalabilidade do sistema de interconexão e do esquema de coerência utilizado [10, 11]. O **Hector** é um multiprocessador construído pela Universidade de Toronto baseado em hierarquias de anéis com passagem de ficha, mas não inclui a coerência de caches. Já o **Memnet**, da Universidade de Delaware, inclui um esquema de coerência de memória entre estações de trabalho interconectadas através de anéis com passagem de ficha.

Os anéis com inserção de registros adotados pelo **SCI** foram estudados por Scott et al. [12] através de modelamento analítico e simulações para tráfegos uniforme e não uniforme. O modelo é baseado em uma solução iterativa aproximada da fila **M/G/1**, mas não leva em consideração o controle de fluxo. A eficácia do controle de fluxo proposto dentro do padrão **SCI** foi estudado através de simulações, e se mostrou importante como forma de prover um uso justo da banda passante do anel em condição de tráfego intenso, embora para tráfego baixo tenda a reduzir ligeiramente o desempenho da rede.

VAZÃO DA REDE ("THROUGHPUT")

Pode-se calcular a vazão máxima de em um anel **SCI** com **N** nós assumindo-se que cada nó envia pacotes somente ao seu vizinho seguinte. Se onde cada nó pode injetar até **G** Gbytes/s, a vazão máxima é então **N.G** GBytes/s. Para se obter uma estimativa mais realística da vazão, pode-se assumir um tráfego onde os pacotes são endereçados aleatóriamente de forma uniforme. Nestas condições, os pacotes percorrem em média **N/2** nós até atingir o nó destino e a vazão média máxima pode ser calculada simplesmente como [15]:

$$Vazão = G.\frac{N}{\frac{N}{2}} = 2.G \qquad (1)$$

A equação 6.1 mostra que a vazão para tráfego uniforme não depende do tamanho do anel. Como parte da banda passante é consumida em pacotes de eco e nos cabeçalhos, a vazão efetiva de dados é menor. Por exemplo, em um pacote **dmove64**, apenas 70% do total de símbolos associados à transação contém dados. Assim, a vazão efetiva para pacotes **dmove64** para um anel com tráfego baixo é:

$$VazãoEfetiva \cong 70\%.2.G \cong 1,4G \qquad (2)$$

LATÊNCIA

Pode-se também obter uma expressão simples para a latência em um anel **SCI** com tráfego baixo. Sendo **N** o número de nós, $\mathbf{t_{pass}}$ o tempo de passagem em um nó, $\mathbf{t_{cabo}}$ o tempo de propagação nos cabos, e **S** o tamanho do pacote em símbolos, têm-se [15]:

$$LatênciaMédiaMínima = \frac{N}{2}.(t_{cabo} + t_{pass}) + S / G \qquad (3)$$

2.2 REDES K-ÁRIAS N-CÚBICAS

Redes **k-árias n-cúbicas** empregando técnicas de bombeamento semelhante ao do **SCI** foram estudadas por Scott & Goodman [13] visando avaliar o impacto do bombeamento[1] na dimensionalidade ótima de redes. O bombeamento permite que o tempo de ciclo da rede possa ser desvinculado do comprimento das ligações, e, apesar de ser um técnica relativamente comum em redes de longas distâncias, não tem sido, ou tem sido empregado de uma forma limitada em interconexão de sistemas paralelos. Scott & Goodman mostraram que a dimensionalidade ótima de redes bombeadas é maior de que a de redes não bombeadas para diversas restrições de construção, e que redes com bombeamento apresentam menor latência e maior banda passante, especialmente para redes com dimensionalidade elevada.

Johnson & Goodman [14] estudaram diversas topologias baseadas em anéis **SCI**, sendo uma delas as redes **k-árias n-cúbicas** implementadas com **chaves-N**. As outras topologias estudadas foram: ***i***) rede de um único estágio "embaralha-troca" (em inglês "*shuffle-exchange*"); ***ii***) rede multiestágio **omega**; e ***iii***) rede *crossbar*. Para a maioria delas foram obtidas expressões aproximadas dos parâmetros mais importantes, como vazão e latência, tanto para condição de tráfego leve como para tráfego intenso.

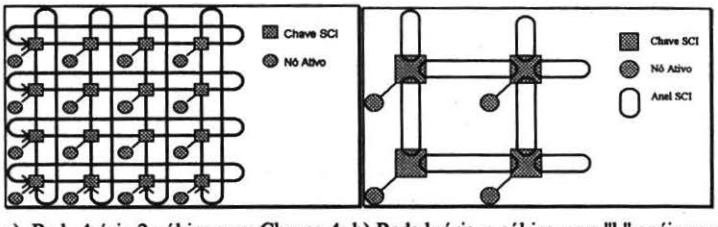

Fig. 1 - a) Rede 4-ária 2-cúbica com Chaves-4; b) Rede k-ária n-cúbica com "k" anéis nas arestas

A figura (1.a) mostra uma rede **4-ária 2-cúbica** onde há uma chave em cada interseção da rede. A desvantagem dessa rede é o fato do pacote de eco ter de percorrer muitos nós (agentes) se o

[1] "pipelining"

"k" for elevado. Uma forma alternativa de implementação de redes k-árias n-cúbicas é mostrada na figura (1.b). Em vez de se empregar um único anel para interconectar os "k" processadores de cada aresta, pode-se empregar "k" anéis. Assim que o pacote chegar no primeiro agente, um pacote de eco é logo devolvido liberando a fila de saída do nó que deu origem ao pacote. Infelizmente esta configuração exige chaves de maior tamanho (maior número de ligações de entrada/saída), mais caras e possivelmente mais lentas. Se as chaves tiverem uma direção preferencial (ou seja, é mais rápido continuar no anel do que comutar para outro) a necessidade de trocar muitas vezes de anel pode implicar em um tempo elevado de envio de um pacote de um nó para outro. Além disso, a configuração da figura (1.a) permite uma dimensionalidade maior do que a figura (1.b), o que significa que um nó pode ser ligado a um número maior de nós adjacentes, provendo um diâmetro de rede menor.

Bothner & Hulaas [15] estudaram redes **n-cúbicas k-árias** construídas com pontes **(chaves-2) SCI,** ao invés das chaves-N utilizadas por Johnson&Goodman. A figura 2 mostra uma rede 4-ária 2-cúbica implementada com pontes e onde se tem em cada vértice um anel (anel de canto) com 1 ou mais nós ativos. Através de simulações, foram determinadas a vazão e latência variando-se o número de nós ativos por vértice/aresta, estratégia de chaveamento nas chaves, tamanho dos armazenadores, e localidade de comunicação.

Para aumentar o número de processadores no sistema, mantidos "n" e "k" constantes, Bothner & Hulaas estudaram as seguintes alternativas: ***i***) aumentar o número de nós ativos em cada anel de canto; ***ii***) colocar nós ativos nos anéis de aresta; e ***iii***) dispor os nós ativos em um anel adicional ligado ao nó de canto através de uma ponte (esta alternativa corresponde a utilizar uma pseudo-chave em cada vértice). As simulações mostraram que exceto para dimensionalidades muito elevada, a inserção de nós nos anéis de canto é preferível a colocá-los em anéis adicionais. Isto pode ser explicado pelo fato de trocar de anel ser mais caro do que continuar nele. No entanto, aumentando-se a dimensionalidade, o atraso introduzido pelas pontes dentro dos anéis de canto pode tornar vantajoso trocar de anel.

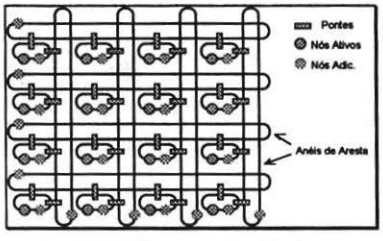

Fig. 2- Uma rede **4-ária 2-cúbica** com 1 ou 2 nós ativos em cada vértice e 1 nó ativo em cada aresta

3 - SIMULAÇÕES

3.1 - O SIMULADOR SCILAB3.0

O SCILab3.0 é um pacote de simulação implementado no **CERN** com vistas ao desenvolvimento de um sistema de aquisição de dados baseado no **SCI**. É composto basicamente de dois módulos: ***i)*** **SCIMP** - Programa de Modelamento **SCI**; e ***ii)*** **TopoEngine** - Gerador de Topologias **SCI.**

O SCIMP é um simulador de sistemas **SCI** escrito em grande parte na linguagem de simulação **MODSIM-II**. A versão presentemente distribuída inclui apenas os protocolos de comunicação do **SCI**.

O TopoEngine é um programa (escrito em "C") que faz a geração dos arquivos de descrição da topologia, tabelas de roteamento, e parâmetros de simulação do **SCIMP** a partir de opções na linha de comandos. As topologias presentemente suportadas são o anel, malhas **2D** e **3D** de anéis, e rede multiestágio. Uma vez que está escrito em "C", o **TopoEngine** pode ser facilmente modificado para incorporar a geração de outras topologias.

A figura (3.a) mostra os modelos básicos implementados no **SCILab**, a saber, nós e chaves. Cada nó pode ser uma memória, um gerador, ou ambos. Quatro modelos de chaves são providos no **SCILab**: o **cross-switch**, o **switch-link**, e o barramento (**b-link**).

O SCILab permite que se possa especificar uma série de parâmetros de implementação de um sistema **SCI**. Os principais parâmetros estão mostrados na figura (3.b).

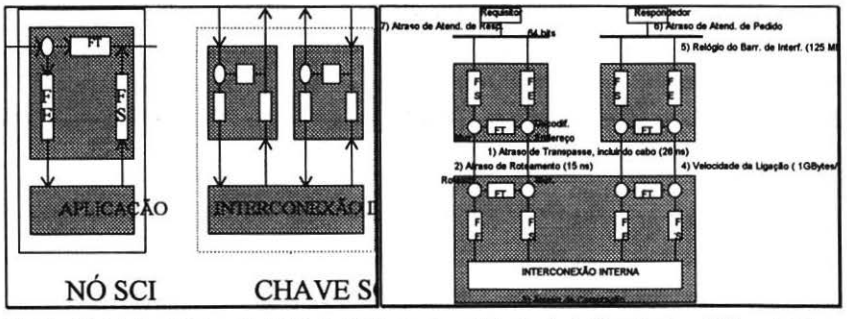

Fig. 3- a) Modelos Básicos do SCILab; b) Parâmetros Principais de Simulação e Valores Típicos

3.2 - METODOLOGIA

3.2.1 - TOPOLOGIAS

Os anéis e anéis interconectados por chave são soluções simples e econômicas para a formação de sistemas paralelos de pequeno e médio porte à base de estações de trabalho. Utilizando o simulador

SCILAB3.0 desenvolvido pelo **CERN**, serão feitas comparações das duas topologias para vários números de nós ativo, utilizando pacotes de comandos com resposta. Uma vez que o objetivo é a construção de sistemas multiprocessadores com coerência de memória, a latência dos comandos com resposta e a sua dependência com os demais parâmetros do sistema é um importante parâmetro na avaliação de topologias.

Embora as topologias em questão sejam restridas a um número limitado de processadores, elas podem ser utilizadas como aglomerados de um sistema de maior porte como mostra a figura 4. A análise de desempenho destes sistemas é um tópico ativo de pesquisa, e será objeto de pesquisa futura dentro do Projeto **SPADE**.

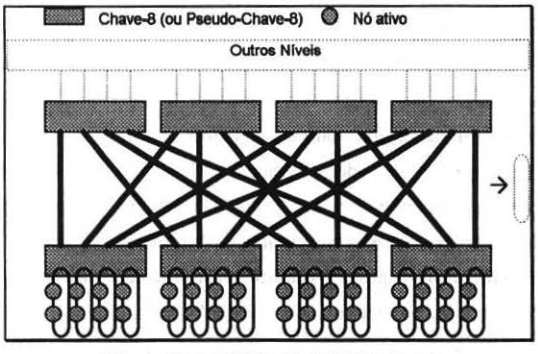

Fig. 4 - Rede Multiestágio Bidirecional

3.2.2 - VELOCIDADE DAS LIGAÇÕES

As simulações são feitas com ligações de 125 MBytes/s, que corresponde às implementação serial prevista no padrão **SCI**.

3.2.3 - TIPOS DE PACOTES

As simulação são feitas com transações com resposta, visando avaliar a rede em condições semelhantes à encontradas em sistemas multiprocessadores com coerência de cache, onde as transações são tipicamente com resposta.

3.2.4 - GERAÇÃO DE PACOTES

Visando verificar o comportamento da rede em condições de tráfego alto a tráfego baixo, o tempo médio de injeção de pacotes será variado entre 100, 1000 e 10.000 ciclos de relógio entre transações[2]. Estes valores foram escolhidos de forma a estimular o sistema de interconexão em

[2] uma transação somente pode ser enviada se o nó tiver recebido a subação de resposta da anterior

condições semelhantes aos de sistemas multiprocessadores com e sem coerência de cache. Isto é, os intervalos de injeção de pacotes correspondem aos intervalos entre acessos compartilhados em arquiteturas multiprocessadoras. Admite-se que o intervalo típico entre acessos à variáveis compartilhadas é de 10 ciclos de relógio [16], e cada instrução é executada em um único ciclo de relógio. O tempo entre transações de 100 ciclos visa avaliar uma arquitetura **NUMA** onde para cada 10 acessos locais, temos 1 acesso remoto. Os demais tempos de injeção correspondem a arquiteturas **CC-NUMA** onde uma percentagem maior dos acessos é satisfeita localmente, sem exigir transações pela rede.

3.2.5 - CONTROLE DE FLUXO

As simulações serão realizadas com controle de fluxo habilitado e desabilitado, visando avaliar o impacto do controle de fluxo.

3.2.6 - TAMANHO DAS FILAS

As simulações serão feitas com dois tamanhos de fila: *i*) 32 pacotes; e *ii*) 1 pacote. O objetivo é verificar o impacto da hipótese de filas infinitas, adotada na implementação do simulador do **SPADE-I**.

3.2.7 - PARÂMETROS DE SIMULAÇÃO

Na tabela 1 tem-se os parâmetros de simulação empregados. O valores da tabela 3 foram baseados em dados fornecidos pela Doplhin, fabricante de componentes e subsistemas **SCIs** [17].

Tab. 1 Parâmetros de Simulação de Ligações Seriais

Velocidade da Ligação	125 MBytes/s (16ns/símbolo)
Atraso na Ligação (cabo + CI)	32 ns
Tempo Atendimento de Pedidos	128 ns
Tempo de Atendimento de Respostas	16 ns
Tamanho das Filas de Pedido e Resposta	16 pacotes
Tempo de Injeção de Pacotes após o recebimento da resposta (distr. expon.)	100, 1000 e 10.000 ciclos (16ns)
Tempo de Chaveamento	128 ns
Atraso de Roteamento	16 ns
Tipo de Pacote	NREAD64
Controle de Fluxo	Bit "go" desabilitado

3.2.8 - INTERVALO DE SIMULAÇÃO

Para as simulações com ligações de 125 Mbytes/s, o intervalo total de simulação adotado é de 10 milissegundos, sendo que após os primeiros 100 microssegundos é feita uma inicilalização das estatísticas visando minimizar efeitos transitórios.

4 - RESULTADOS

4.1 - DEFINIÇÕES

- Tempo Médio entre Transações (τ): é o tempo médio (exponencialmente distribuído) transcorrido entre o término de uma transação e o início de uma nova;

- Vazão Bruta de Recepção: número total de bytes recebidos por um nó (ou, para ser mais exato, pela aplicação), incluindo cabeçalho e eco, dividido pelo tempo total de simulação corrigido do tempo de inicialização;

$$VBR = \frac{NúmeroTotaldeBytes\ \mathrm{Re}cebidos}{TempoTotaldeSimulação - TempodeInicialização} \quad (4)$$

- Vazão Efetiva de Recepção: é o número total de bytes recebidos por um nó, descontado dos bytes de cabeçalho e de eco, dividido pelo total de tempo de simulação corrigido do tempo de inicialização;

$$VER = \frac{NúmeroTotaldeBytes\ \mathrm{Re}cebidos - BytesCabeçalhos - BytesEco}{TempoTotaldeSimulação - TempodeInicialização} \quad (5)$$

- Vazão Total de Recepção do Sistema: é a soma das vazões de recepção de cada nó, em MBytes/s;

$$VRS = \sum_{NúmeroTotaldeNós} VBR \quad (6)$$

- Vazão Efetiva de Recepção do Sistema: é a soma das vazões efetivas de recepção de cada nó, em MBytes/s;

$$VES = \sum_{NúmeroTotaldeNós} VER \quad (7)$$

- Latência de Transação: é o tempo médio transcorrido entre a injeção do primeiro byte do pacote de pedido na rede e a retirada do último byte do pacote de resposta da rede;

$$LT = \frac{\sum_{NúmeroTotaldeTransações} LatênciadeTransação}{NúmeroTotaldeTransações} \quad (8)$$

- Latência Média de Transação do Sistema: é a média das Latências de Transação de cada nó, em nanossegundos;

$$LTS = \frac{\sum LT_{NúmeroTotaldeNós}}{NúmeroTotaldeNós} \quad (9)$$

- Latência Mínima de Transação: é a menor latência de transação de um nó;

$$LTm = \min_{deTodasasTransações}(LatênciasdeTransação) \quad (10)$$

- Latência Mínima de Transação do Sistema: é a menor das Latências Mínima de Transação dos nós, em nanossegundos;

$$LTmS = \min_{deTodososNós}(LTm) \quad (11)$$

- Latência Máxima de Transação: é a maior latência de Transação de um nó, em nanossegundos;

$$LTM = \max_{deTodosasTransações}(LatênciadeTransação) \quad (12)$$

- Latência Máxima de Transação do Sistema: é a maior das Latências Máxima de Transação dos nós, em nanossegundos.

$$LTMS = \max_{deTodososNós}(LTM) \quad (13)$$

4.2 ANEL

4.2.1 EFEITO DO CONTROLE DE FLUXO

A figura (5.a) mostra a Vazão Total de Recepção do Sistema para um anel baseado em ligações de 125 Mbytes/s sem controle de fluxo, para diferentes valores de tempo de injeção de pacotes -τ- e número de processadores.

Observa-se que para um intervalo de tempo longo entre pacotes (τ=10.000 ciclos de relógio de rede), a vazão total cresce com o aumento do número de processadores e tende assintóticamente a um valor de saturação, em torno de 180 MBytes/s. Para τ=1000 ciclos, esta saturação ocorre antes, para anéis com mais de 10 processadores. Para τ=100 ciclos, o anel já está saturado com 2 processadores.

A diferença entre o valor calculado pela expressão 2 e o obtido na figura (5.a) através de simulações pode ser explicada pelo fato da expressão 2 não levar em consideração os pacotes de eco e os demais custos associados à comunicação ao longo do anel, como tempo de atraso nos cabos e CIs de interface, e pelo fato da transação em questão envolver dois tipos de pacotes, um de pedido, e outro de resposta, de tamanhos diferentes. A figura (5.b) mostra um gráfico da vazão para transações sem-resposta DMOVE64, para um anel com ligações de 125 MBytes/s, filas de 1 único pacote, e τ= 100 ciclos de relógio. O valor de vazão total de saturação obtido, em torno de 212 Mbytes/s, coincide aproximadamente com o obtido através da expressão 2 corrigido do custo do pacote de eco (6 símbolos) e outros custos adicionais de comunicação(2 símbolos):

$$VazãoTotal[DMOVE\ 64] \cong \frac{42}{42+6+2}.2.125MBytes/s \cong 210MByes/s$$

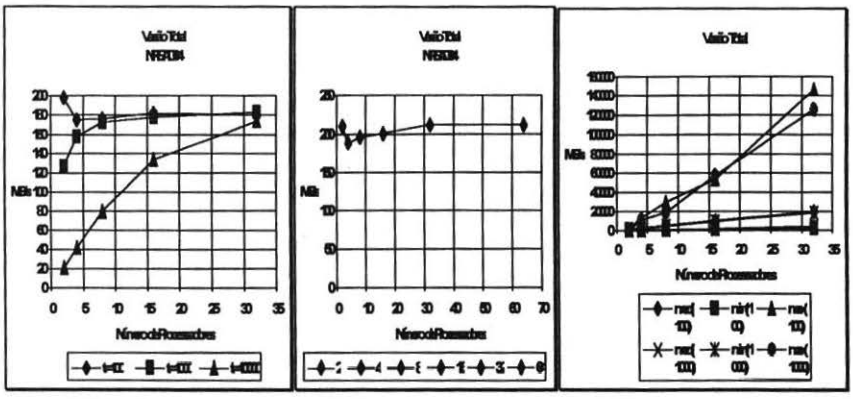

Fig. 5 - a) Vazão Total de Recepção do Sistema com Ligações de 125MBytes/s, com Controle de Fluxo Desabilitado; b) Vazão Total com Ligações de 125MBytes/s p/ DMOVE64, com Controle de Fluxo Desabilitado; c) Latências de Transação do Sistema com Ligações de 125MBytes/s, com Controle de Fluxo Desabilitado

A figura (5.c) mostra as latências de sistema para o anel de 125 MBytes/s. O tempo mínimo segue aproximadamente o valor calculado com base na expressão 3 acrescido de componentes adicionais de atraso relativos aos modelos do SCILab. As curvas de latência média permitem avaliar o impacto da contenção na rede na latência de transação. Para 32 processadores e condição de tráfego elevado (τ= 100 ciclos de relógio), o atraso médio é cerca de uma ordem de grandeza maior do que o atraso médio. A curva da latência máxima evidencia o não determinismo da latência, que é uma característica inerente aos anéis com inserção de registros. A latência máxima é cerca de 63 vezes maior do que a latência mínima para 32 processadores com τ=100 ciclos.

As figuras (6.a) e (6.b) mostram o comportamento do anel com o controle de fluxo habilitado. Nota-se na figura (6.a) que, de fato, o controle de fluxo contribui para reduzir a latência máxima de comunicação, provendo uma utilização mais justa da banda passante. Confirmando outros estudos [12], verifica-se, todavia, que o preço pago é uma redução na vazão máxima, que no caso atingiu cerca de 10%. Os gráficos mostram que para tráfego uniforme a latência média não é significativamente afetada.

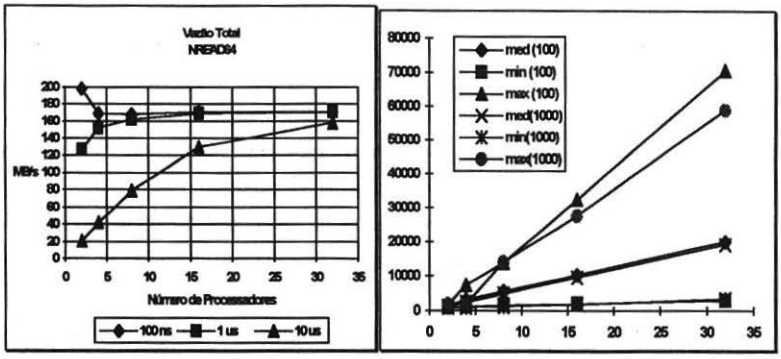

Fig. 6 - a) Vazão Total de Recepção do Sistema com Ligações de 125MBytes/s, com Controle de Fluxo Habilitado; b) Latências de Transação do Sistema com ligações de 125 MBytes/s, com Controle de Fluxo Habilitado.

4.2.2 TAMANHO DAS FILAS

As simulações anteriores foram feitas com filas de 16 pacotes. As figura (7.a) e (7.b) mostram a vazão total e a latência média para filas de 1 pacote, com τ=1000 ciclos. Os gráficos mostram que a latência média e vazão total não mudam significativamente para sistemas com filas pequenas ou grandes para transações com resposta, o que pode ser explicado pela restrição imposta de haver apenas uma transação por processador por vez na rede.

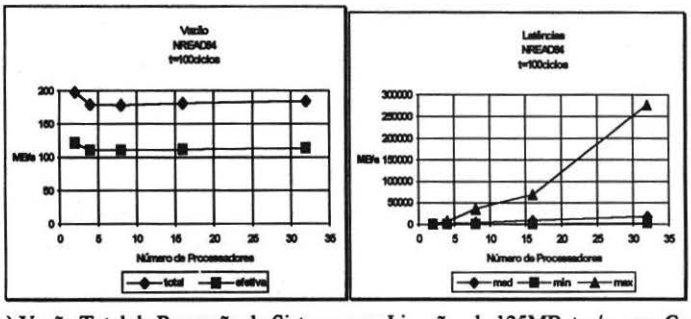

Fig. 7 - a) Vazão Total de Recepção do Sistema com Ligações de 125MBytes/s, com Controle de Fluxo Desabilitado e Filas de 1 pacote; b) Latências do Sistema com Ligações de 125MBytes/s, com Controle de Fluxo Desabilitado e Filas de 1 pacote

4.3 HIERARQUIA DE ANÉIS

Os gráficos (8.a) e (8.b) mostram a vazão e a latência para hierarquias de anéis (vários anéis primários com nós de processamento interconectados através de um único anel secundário) com 32 processadores, variando o número de anéis e número de processadores por anel.

Os gráficos mostram que a hierarquia de anéis provê um melhor desempenho, tanto em termos de vazão como de latência, do que anéis simples. Para 32 nós de processamento, a configuração com 4 anéis de 8 processadores mostrou melhor resultado.

Comparando os gráficos com e sem controle de fluxo, verifica-se que, confirmando os resultados referentes a anéis simples, que, em geral, para tráfego uniforme, o impacto não é significativo, com uma ligeira redução na vazão total e uma melhora na latência máxima. Na configuração com melhor desempenho, a influência do controle de fluxo não é significativa para nenhum parâmetro.

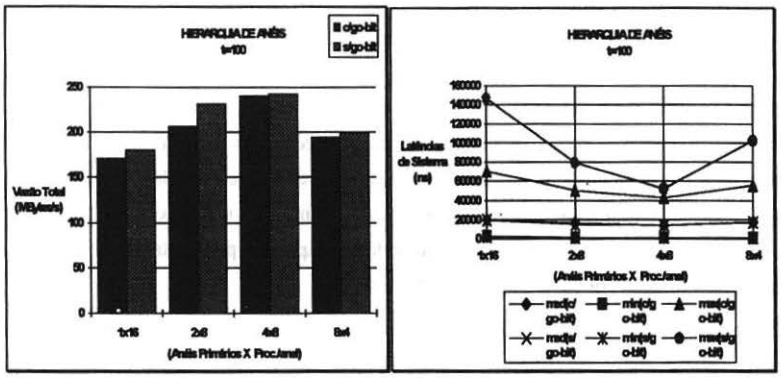

Fig. 8 - a) Vazão Total de Sistema para Hierarquias de Anéis; b) Latências de Sistema para Hierarquias de Anéis

5 - SUMÁRIO

Neste trabalho foram discutidos tópicos relacionados com a interconexão do **SPADE-I**, . a saber anéis e hierarquias de anéis baseados no padrão SCI. Foram feitas simulações com anéis e hierarquia de anéis com e sem controle de fluxo e avaliando o efeito do tamanho das filas. Os resultados mostram que para tráfego uniforme e injeção não "posted", o impacto nas topologias estudadas estes parâmentros não é significativo, principalmente em condições de tráfego baixo, típicas de sistemas multiprocessadores **CC-NUMA** executando aplicações com alta localidade de referências.

Comparando anéis e hieraquias de anéis, verificou-se que para um número pequeno de processadores (32), hierarquias de anéis em 2 níveis constituem uma boa alternativa aos anéis

simples, com a vantagem de permitir uma melhor exploração da localidade de referências a nível de anéis.

AGRADECIMENTOS

Este trabalho foi desenvolvido com apoio da **FINEP** e **CNPq-RHAE**. Agradecemos ao **CERN** pela permisão de uso do simulador **SCILab**.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] VRANESIC Z.G. et al. "*Hector: A hierarchicallly structured shared-memory multiprocessor*". IEEE Computer, p. 72-78, Jan. 1991.

[2] CHERITON, David R. et al. "*Paradigm: A highly Scalable Shared-Memory Multicomputer Architecture*". IEEE COMPUTER, v.24, n.2, p.33-48, Feb. 1991.

[3] Kendall Square Research: Technical Summary, 1992.

[4] KING, P.J.B. & MITRANI, I. "*Modeling a slotted ring local area networks*". IEEE Transactions on Computers, v. C-36, n.5, p.554-561, May 1987.

[5] BARROSO, L. & DUBOIS, M. "*The Performance of Cache-Coherent Ring-based Multiprocessors*". Proceedings of the 20th Annual International Symposium on Computer Architecture, p. 268-277, May 1993.

[6] FARKAS, K. et al. "*Cache Consistency in hierachical-ring-based multi-processors*". Proceedings of the Supercomputing 92, Nov.1992.

[7] KAMAL, A.E. & HAMACHER, V.C. "*Utilizing Bandwidth Sharing in the Slotted Ring*". IEEE Transactions on Computer, v. 39, n. 3, p. 289-299, March 1990.

[8] ZAFIROVIC-VUKOTIC, M. & NIEMEGEERS. "*A Performance Modeling and Evaluation of Cambridge Fast Ring*". IEEE Transactions on Computers, v. 41, n. 9, p. 1110-1125, Sept. 1992.

[9] HOLLIDAY, Mark & STUMM, Michael. "*Performance Evaluation of Hierarchical Ring-Based Shared Memory Multiprocessors*". IEEE Transactions on Computers, v.43, n.1, p.52-67, Jan. 1994.

[10] RAMACHANDRAN, Umakishore et al.: "*Scalability Study of the KSR-1*". Georgia Institute of Technology, GIT-CC 93/03, 1993.

[11] ZHANG,Xiadong & YAN, Yong. "*Latency Analyses of CC-NUMA and COMA Rings*". High Performance Computing and Software Laboratory, University of Texas at San Antonio, 1994.

[12] SCOTT, S.L. et al. "*Performance of the SCI Ring*". Proceedings of the 19th Annual International Symposium on Computer Architecture". p. 403-414, May 1992.

[13] SCOTT, S.L. & GOODMAN, J.R. "*The Impact of Pipelined Channels on k-ary n-Cube Networks*". IEEE Transactions on Parallel and Distributed Sytems, v.5, n.1, p. 2-16, Jan. 1994.

[14] JOHNSON, R. & GOODMAN, J. "*Synthesizing General Topologies from Rings*". Proceedings of the 1992 International Conference on Parallel Processing. 1992.

[15] BOTHNER, John E. & HULAAS, Trond I. "*Topologies for SCI-based systems with up to a few hundred nodes*". Department of Informatics, University of Oslo. Feb. 1992. Candidatus Scientiarum Thesis.

[16] LENOSKY, Daniel. "*THE DESIGN AND ANALYSIS OF DASH: A SCALABLE DIRECTORY-BASED MULTIPROCESSOR*". Computer Science Laboratory, Stanford University, Tech. Rep. No. CSL-TR-92-507, Feb. 1992. PhD Thesis.

[17] BRYHNI, Haakon & WU, Bin. "*Initial studies of SCI LAN topologies for local area clustering*". Proceedings of the First International Workshop on SCI-Based Low- Cost/High-Performance Computing, 1994.