Análise do custo de comunicação em aglomerados de estações.[1]

Hermes Senger[¥]

[0]PCS- Departamento de Engenharia de Computação e Sistemas Digitais
e-mail : liria@pcs.usp.br

Liria Matsumoto Sato[0]

[¥]LSI-Laboratório de Sistemas Integráveis
e-mail: hermes@lsi.usp.br

Escola Politécnica da Universidade de São Paulo
Av. Prof. Luciano Gualberto, trav. 3, nº 158
05508-900 - São Paulo, SP
Tel. (011) 818-55589/818-5662

Abstract

This paper presents some features related to communication cost on clusters of workstations, used in distributed processing. The impact of changes in the network throughput and message size over communication costs are analyzed.

Resumo

Este trabalho apresenta algumas características relacionadas com o custo de comunicação em aglomerados de estações de trabalho, utilizadas para o processamento distribuído. A influência do aumento da largura de banda e do uso de diferentes tamanhos de mensagens sobre o custo de comunicação é avaliada.

1. INTRODUÇÃO

A tecnologia na área da microeletrônica tem conseguido quadruplicar a velocidade (*clock*) dos processadores baseados em tecnologia CMOS a cada três anos [2,3], aproximadamente. Da mesma forma, a capacidade dos chips de memória também é quadruplicada em um período semelhante. Esse avanço proporcionou um rápido aumento na capacidade de processamento das estações de trabalho e micro-computadores disponíveis comercialmente. No âmbito da tecnologia de redes de computadores, a largura de banda também aumenta cerca de dez vezes [3] a cada período de 10 anos. Mais recentemente, devido aos grandes investimentos nas áreas de telecomunicações tem melhorado ainda mais o panorama das redes de alta velocidade [4,5].

Por outro lado, nas últimas décadas o avanço da ciência em diversas áreas do conhecimento humano tem promovido uma demanda crescente de processamento, motivando a pesquisa em sistemas de computação de alto desempenho. Neste sentido, inúmeras propostas de arquiteturas paralelas têm sido lançadas com o objetivo de

[1] Este trabalho foi financiado pelo projeto FINEP/PAD processo nº 5.6.94.0260.00, e parcialmente financiado pelo projeto FINEP/RECOPE., processo nº 3607/96.

oferecer grande capacidade de processamento. De um modo geral, tais arquiteturas baseiam-se na conexão de diversos processadores através de algum tipo de sistema de interconexão, utilizando diversas formas de agrupamento e topologias, que vão desde os multiprocessadores e multicomputadores limitados a um único gabinete, até redes de longa distância, passando por inúmeras soluções intermediárias.

Neste trabalho, pretende-se abordar alguns aspectos relativos à velocidade de comunicação oferecida pelos sistemas que de alguma forma interconectam processadores. Mais precisamente, referímo-nos às redes de computadores, deixando de lado dispositivos como os barramentos e vias internas dos computadores.

As aplicações destinadas às arquiteturas de processamento paralelo e distribuído, igualmente apresentam-se em grande variedade e quantidade, incluindo desde aplicações tradicionais como NFS (*Network File System*), RPC (*Remote Procedure Call*), até aquelas baseadas em algoritmos projetados para resolver problemas específicos. Assim, este trabalho não irá caracterizar os requisitos desta ou daquela aplicação em particular, mas sim discutir alguns aspectos de comunicação em função de alguns parâmetros simples, como por exemplo o tamanho das mensagens trocadas por aplicações.

2. VELOCIDADE DE PROCESSAMENTO E VELOCIDADE DE COMUNICAÇÃO

O aumento na velocidade dos processadores traz consigo a necessidade de melhoria nos periféricos de um modo geral, inclusive periféricos de comunicação. A lei de Amdhal estabelece que, para cada bit processado na unidade de tempo, um bit deve ser tratado pelos dispositivos de entrada e saída em intervalo equivalente. Admita, hipoteticamente, a existência de um processador com as seguintes características:

- trata palavras de 64 bits e executa uma instrução a cada 1 nanossegundo; e
- devido ao tipo de aplicação, cada instrução também gera uma operação de E/S.

Portanto seria possível estimar, em cálculos aproximados, que tal processador geraria uma demanda da ordem de 64 Gigabits/s sobre seus periféricos. É bem verdade que a perspectiva de aumento nas taxas de comunicação das redes é bastante promissora [4,5], entretanto, mesmo uma rede capaz de oferecer uma taxa de 1 Gigabit/s não teria um desempenho plenamente satisfatório para tal computador.

Além disso existem alguns fatores que suavizam essa diferença entre a demanda em potencial e a oferta real de comunicação. Por exemplo, mesmo em aplicações que utilizam exaustivamente os serviços de comunicação oferecidos pela rede, apenas um número reduzido de instruções executadas irá gerar dados a serem transmitidos. Ainda assim, a grosso modo pode-se dizer que uma taxa de 1Gbps atenderia apenas de forma modesta a demanda criada por aplicações distribuídas. Certamente continuariam sendo necessários alguns cuidados ao se projetar aplicações para esse sistema, tais como minimizar a quantidade de mensagens trocadas pelos processadores através da escolha de uma distribuição adequada dos dados, escalonamento das tarefas, escolha dos algoritmos apropriados, etc.

Por outro lado, as redes de alta velocidade já são realidade, e mais que isso, estão ganhando popularidade, à medida que seus custos caem. Redes FDDI (*Fiber Distributed Data Interface*) e 100VG-AnyLAN [5], por exemplo, oferecem velocidade

típicas de 100Mbps. Redes ATM [4,5] (*Assynchronous Transfer Mode*) operam a velocidades típicas de 155 e 622Mbps atualmente, mas já estão sendo discutidos e implementados os padrões de 1.2 e 2.5 Gbps, e outros acima deverão vir a seguir. As populares redes Ethernet de 10 Mbps, por exemplo, estão ficando ultrapassadas, pois é comum encontrarmos equipamentos Fast Ethernet de 100Mbps em redes de universidades, empresas, etc. Em breve teremos equipamentos Giga Ethernet disponíveis no mercado, com velocidades de 1 Gbps.

As tecnologias que utilizam fibra ótica como meio de transmissão poderão atingir taxas ainda maiores. Caso se consiga elaborar dispositivos que aproveitem algo bem próximo da largura de banda total oferecida pela fibra, existe um limite físico que fica entre 50 e 75 terabits por segundo [4]. Apesar desse número ter proporções astronômicas para os dias de hoje, gostaríamos de analisar de maneira bastante simplista, até que ponto se pode tirar proveito de tal panorama em favor do aumento de desempenho de aplicações e ferramentas de processamento distribuídos. Para isso, consideremos alguns aspectos da comunicação entre processadores distintos, interconectados por algum sistema de comunicação. Caberiam aqui, desde os sistemas com nós de processamento (compostos de um ou mais processadores associados a uma memória local) contidos dentro de um único gabinete, até computadores independentes, ligados através de uma rede local.

De maneira bastante simplificada, cada vez que uma **mensagem** é enviada, seus bytes são agrupados para compor um **pacote** (camada 3 do modelo OSI), que por sua vez irá gerar um ou mais **quadros** (camada 2 do modelo OSI), para então serem transmitidos. Dependendo do tamanho, mensagens grandes podem gerar mais de um quadro, que por sua vez, podem gerar um ou mais quadros. Tudo depende dos protocolos utilizados nas camadas de enlace e de rede. Cada vez que um quadro (na camada de enlace) é injetado no meio físico, é necessário aguardar algum tempo, até que seus bits estejam disponíveis no nó destino. Por **tempo de transmissão** entenda-se aqui, como sendo a somatória de todos os tempos necessários para se obter acesso ao meio físico, injetar o sinal (bits codificados) no meio, tempo de propagação do sinal através do meio físico, e finalmente, o tempo que o destinatário levará para remontar o quadro. Existe ainda uma série de outros fatores que não serão aqui considerados - tais como a ocorrência de erros (tempo de retransmissão), tempo para se obter acesso ao meio físico que pode ser compartilhado (e neste caso depende da carga da rede), atrasos inseridos por equipamentos (comutadores, pontes, roteadores), bits de *overhead* gerados pelos protocolos, etc.; uma vez que o que se pretende discutir aqui é algo bem mais simples, que dispensa tal complexidade.

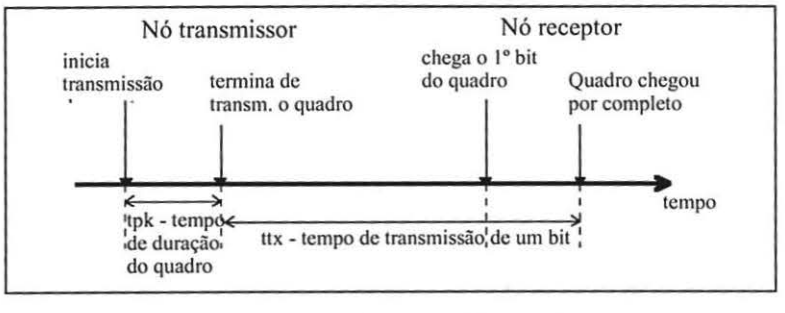

Figura 1 - Característica temporal de uma transmissão

A Figura 1 ilustra as métricas que se pretende utilizar nesta análise:

- *tpk* : é o tempo gasto para que os bits de um quadro possam ser injetados no meio físico. Em outras palavras, é o tempo de duração de um quadro. Esse parâmetro pode ser calculado da seguinte forma:
  tpk = *tbit* * *nbits*, onde *tbit* é o tempo de duração de um bit, e *nbits* é o número de bits contidos no quadro.
- *ttx* : É o tempo que um único bit leva até chegar ao nó destino.

Para uma rede do tipo Ethernet por exemplo, o parâmetro *tbit* pode ser facilmente calculado como sendo o inverso da taxa nominal de transmissão da rede, isto é, 10 Mbps. Assim, obtêm-se o valor $tbit = 10^{-7}$ segundos, ou seja, 100 ηs (nanossegundos). Analogamente, obtém-se *tbit* = 10 ηs para uma rede Fast Ethernet e *tbit* = 1 ηs para Giga Ethernet. Dessa forma, pode-se prosseguir estimando o valor desse mesmo parâmetro para cada tecnologia que se conheça a taxa de transmissão.

Já o parâmetro *ttx* independe, por exemplo, da natureza do meio físico. A velocidade de propagação de um sinal em um fio de cobre não é significativamente diferente da velocidade de propagação da luz na fibra, ou seja, aproximadamente 2×10^8 m/s para ambos os casos [4,5]. Assim, pode-se obter $ttx = d / c$, onde c é a velocidade de propagação do sinal, e d é a distância entre o nó origem e o destino de uma transmissão. Neste caso, a distância entre as estações envolvidas é relevante.

3. ANÁLISE DO CUSTO DE COMUNICAÇÃO

Vamos tomar como exemplo inicial, o caso de uma rede cuja taxa de transmissão é de 100 Mbps (*tbit* = 10 ηs). Uma vez que o tempo de transmissão depende da distância, pode-se considerar três situações : *ttx* = 5 ηs (1 m de distância entre as duas estações); *ttx* = 50 ηs (10 m de distância) e *ttx* = 500 ηs (100 m de distância). Assim, obtêm-se o tempo total necessário para a transmissão de quadros de variados tamanhos $tt = tbit \times nbits + ttx$. A partir do tempo total gasto para transmissão do quadro, pode-se obter o tempo médio de espera por bit (e assim sucessivamente, para byte ou quadro) transmitido. Esse parâmetro ilustra uma relação custo/benefício, na qual o custo está associado ao tempo de espera, e o benefício significa a quantidade de dados(bytes) transmitidos. Assim, o termo **custo de comunicação** será aqui empregado para indicar o tempo médio de espera para cada byte transmitido.

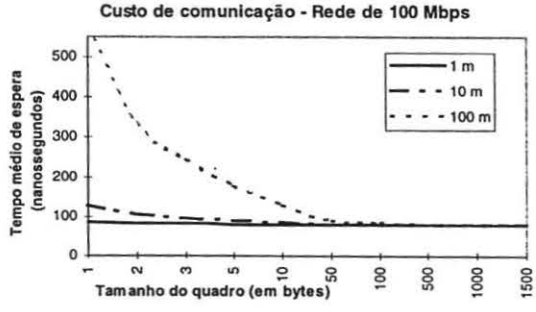

Figura 2 - Tempo médio de espera por byte transmitido.

Na Figura 2 pode-se observar que o custo de comunicação é fortemente influenciado pela distância entre os nós. No caso de uma distância de 100 m entre as estações, quadros muito pequenos como o de 1 byte por exemplo, pagam um custo bastante alto, em torno de 580 ηs. Contudo, para quadros de 10 bytes o custo de comunicação torna-se razoável em torno de 130 ηs, atingindo níveis realmente interessantes para os quadros com tamanho a partir de 50 bytes, em torno de 90 ηs. A partir desse ponto, a distância entre as estações pode ter seu impacto minimizado, uma vez que o tempo de transmissão será diluído na quantidade de informações transmitidas.

Muito embora o custo de comunicação apresente uma queda mais acentuada com distâncias maiores entre as estações, preferiu-se ilustrar, na Figura 2, o caso em que a distância entre estações é de 10 metros, por acreditar-se que isto ocorra com maior freqüência do que distâncias de 100 metros em aglomerados de estações. Isto, de forma nenhuma prejudica a amplitude desta análise, pois, no caso de distâncias maiores, conforme já foi ilustrado, o custo pode cair ainda mais à medida que se aumenta o tamanho dos quadros. Tanto no caso de redes de 100 Mbps quanto de 1 Gbps, pode-se conseguir um custo de comunicação bastante razoável para quadros com tamanho a partir de 10 bytes. Apesar do custo no caso das redes de 100 Mbps e 1 Gbps ser bastante diferente em valores absolutos, observa-se que nos dois casos o custo cai significativamente à medida que reduz-se o tamanho dos quadros até por volta de 50 bytes, aproximadamente. A partir daí, o custo praticamente se estabiliza.

4. CONCLUSÕES

O trabalho apresentou uma discussão sobre algumas características importantes de desempenho de redes de comunicação, que afetam o desempenho de aplicações distribuídas. Constatou-se que as redes de alta velocidade não trazem benefícios significativos com relação à latência de comunicação. Por exemplo, o tempo necessário para que um único bit de informação chegue ao seu destino é o mesmo, sem levar em conta a taxa de ocupação do meio. Entretanto, o custo de comunicação pode ser reduzido drasticamente, pois, uma vez transcorrida a latência inicial de propagação do sinal, os dados chegarão com uma taxa diretamente proporcional à velocidade da rede. Desta forma, o custo de comunicação cai bastante com o uso de quadros maiores, permitindo assim, o aproveitamento efetivo da largura de banda oferecida pelas redes de alta velocidade em favor do aumento de desempenho das aplicações distribuídas com alta demanda de comunicação. No caso de distâncias muito curtas, como a de 1 metro por exemplo, o tamanho dos quadros não demonstrou ser um fator tão crítico, pois a latência de propagação do sinal é relativamente pequena. Por outro lado, este fato pode ser um forte motivo para trabalhar-se com distâncias tão curtas quanto possível, quando o objetivo é o alto desempenho.

Como conseqüência, é recomendável agrupar uma quantidade maior de dados, para só então solicitar sua transmissão. Nesse sentido, pode-se considerar a possibilidade de adiantar o envio de alguns dados, ou explorar melhor algumas características como *pipelining*, por exemplo, visando compensar a latência de comunicação. Evidentemente, esse é apenas um fator adicional a ser considerado na implementação de aplicações em ambientes de processamento distribuído, devendo-se encontrar um ponto de equilíbrio entre todos os fatores envolvidos.

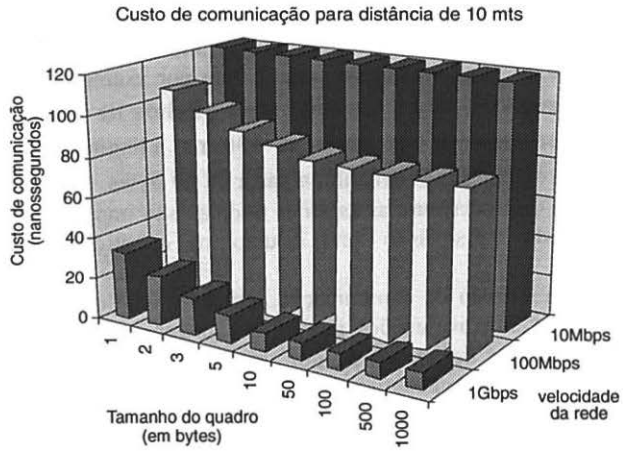

Figura 3 - Custo de comunicação, com 10 mts de distância entre as estações.

5. BIBLIOGRAFIA

[1] BELL, G. **"Scalable, Parallel Computers: Alternatives, Issues and Challenges"**. *Int. Journal of Parallel Programming*. Vol. 22. N°1, 1994.

[2] HENNESSY, J. & PATTERSON, D. **Computer Architectures: A Quantitative Approach**. Morgan Kaufman, San Mateo, California 1990.

[3] DUKE, D.W; ELIAS, D.; LIVNY M., TURCOTTE, L. **Clustered Workstation Environments.** Tutorials of the Supercomputing 93. ACM SIGARCH & IEEE Computer Society. Nov. 1993.

[4] PARTRIDGE, C. **Gigabit Networking.** Reading, Addison Wesley, 1993.

[5] STALLINGS, W. **Local & Metropolitan Area Networks.** Prentice-Hall, 1997.