Quim. Nova, Vol. 26, No. 3, 401-406, 2003

Divulgação

ASPECTOS TÉCNICOS ENVOLVIDOS NA CONSTRUÇÃO DE UM "CLUSTER BEOWULF"

Ataualpa Albert Carmo Braga*#
Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-970 Campinas - SP

Recebido em 19/2/02; aceito em 19/11/02

TECHNICAL ASPECTS OF BEOWULF CLUSTER CONSTRUCTION. Beowulf cluster is perhaps the cheapest way to construct a high performance computers. The strategy of reaching a high power computing isn't so difficult, but buying and configurating should be done carefully. Technical aspects of hardware components and message-passing libraries are considered, with some results given as examples.

Keywords: Beowulf cluster; computational chemistry; high-performance computing.

INTRODUÇÃO

"Beowulf": Poema épico com 3182 linhas, considerado o maior expoente da literatura anglo-saxônica; escrito por um anônimo cristão, provavelmente no século X. Narra as três batalhas do herói, que dá nome a obra, contra o gigante Grendel, a mãe de Grendel e o dragão que guarda um tesouro. Uma alegoria da luta entre o Bem e o Mal, retratando aspectos da cultura germânica na região onde hoje se encontram a Dinamarca e sul da Suécia.

"Beowulf": Conjunto de computadores pessoais (PC's), agrupados com o objetivo de obter a menor razão custo/benefício. Para isto são usados componentes de "hardware" disponíveis no mercado, independentes de fornecedores específicos, e baseados em sistemas operacionais e "softwares" gratuitos.

Estas são duas possíveis definições do termo "Beowulf". Existem várias outras, tanto para a versão literária[1-3] como para a computacional[4-6] . Em ambas "Beowulf" enfrenta um poderoso adversário, e vence. Enquanto a versão mítica triunfa sobre monstros e dragões, a real apresenta-se como uma alternativa barata à computação de alto desempenho, até então dominada pelo monopólio das grandes fornecedoras.

Thomas Sterling e Don Becker, pesquisadores do Goddard Space Flight Center (NASA), construíram o primeiro "cluster" de PC's em 1994. Uma alternativa barata e eficiente à limitação computacional da época. Mesmo a rica NASA não poderia fornecer individualmente a cada grupo de pesquisa recursos suficientes para a obtenção de um supercomputador. Assim, sem auxílio financeiro, Sterling teve a idéia de usar processadores de baixo custo (16 PC's com processadores 486DX4), com um sistema operacional gratuito (Linux) e placas de rede Ethernet. Estas máquinas foram montadas e programadas de forma a possibilitarem a paralelização/divisão das tarefas, buscando atingir-se um poder de processamento equivalente a um supercomputador da época, por uma fração do preço. Este primeiro "cluster" atingia 70 megaflops, com um custo estimado em 1/10 do valor cobrado pelo mercado para um sistema de desempenho similar. Tal projeto fez tanto sucesso que o termo "Beowulf" foi estendido a todos os "clusters" de PC's que viriam a surgir (Curiosamente este primeiro "cluster" levou o nome do amigo de "Beowulf", "Wiglaf"[7]). Com a evolução natural dos componentes de "hardware" e aperfeiçoamento do sistema operacional e "softwares" em geral, a NASA, em 1996, finalmente alcançou um poder de processamento por volta de 1 gigaflop (um bilhão de operações por segundo). Esta marca foi obtida com a construção de dois novos "clusters" formados por 16 Pentiuns Pro. O primeiro no Instituto Tecnológico da Califórnia, chamado de Hyglac e o segundo conhecido como Loki, construído no Laboratório Nacional de Los Alamos. Ambos apresentaram um custo de cerca de US$50.000, muito abaixo do valor de um milhão cobrado pelas grandes empresas especializadas[4,6].

Atualmente, entre as 500 máquinas mais rápidas registradas pelo Top500[8], 28 são "clusters Beowulf", estando o LosLobos, da Universidade do Novo México em octagésimo lugar, atingindo 237 gigaflops. E novos "clusters" vêm sendo projetados, cada vez mais sofisticados e especializados.

O trabalho em química teórica depende cada vez mais de grandes quantidades de recursos computacionais, exigindo dos pesquisadores conhecimento tanto de programas como de componentes de "hardware". Este trabalho visa trazer informações técnicas para facilitar a construção e configuração de um "cluster" de PC's, obtendo-se um equipamento de alto desempenho e de fácil administração, a um custo financeiro relativamente baixo. Morgon[9] discutiu de maneira ampla as aplicações e vantagens da construção de um "cluster", inclusive citando o histórico da experiência do nosso grupo desde 1997. Neste trabalho serão abordados com maior ênfase os aspectos técnicos envolvidos, especificamente relacionados à escolha e montagem dos equipamentos.

A construção de um "cluster" possui várias vantagens que, às vezes, passam desapercebidas. Por exemplo:

- uma configuração inicial pode ser facilmente atualizada, sem desperdício do investimento;
- a necessidade e criatividade do administrador limitam a forma com que as tarefas são divididas. É comum que o compromisso dos sistemas montados seja com o melhor aproveitamento dos recursos, evitando ociosidade, procurando satisfazer uma larga e diversificada clientela; enquanto em outra situação o desempenho passa a ser a motivação, o interesse neste caso está em obter-se o maior número de "flops" possível, implementando soluções por programação paralela;
- a manutenção geralmente fica por conta do grupo de usuários, que ganha com a experiência e economia de recursos;
- qualquer máquina considerada ultrapassada pode ser aproveitada, contanto que haja um estudo da melhor distribuição de tarefas, evitando que os nós mais sofisticados fiquem ociosos aguardando o final das tarefas;

*e-mail: atabraga@iqm.unicamp.br
#Dedicado à memória do Prof. Antonio Luiz Pires Valente

- nada se perde; mesmo que não exista interesse em manter determinado componente ou nó no "cluster", estes podem ser aproveitados como peças de reposição ou terminais;
- o desenvolvimento acelerado relacionado à evolução do sistema operacional e de programas específicos para a instalação e monitoramento dos "clusters" está intimamente ligado à influência da filosofia "open source" desde o início do primeiro projeto.

Para o melhor aproveitamento dos recursos disponíveis, deve-se sempre ter em mente critérios para a escolha do equipamento a ser adquirido. Tais critérios devem ser baseados no interesse específico de cada grupo de pesquisa. Por exemplo, cálculos em paralelo podem ser bastante úteis em extensas simulações, onde o tempo total dos cálculos pode inviabilizar a pesquisa. Neste caso, pode ser interessante investir em meios de comunicação mais sofisticados entre os nós, por exemplo, usando placas Myrinet[10].

Tendo-se claros os objetivos a serem alcançados deve-se escolher entre os modelos de cada componente que formarão os nós do sistema. O ideal seria poder testar cada item, usando-se programas específicos de avaliação, assim como cálculos científicos que reflitam a realidade das aplicações. Vazquez[11] descreve vários meios para tal avaliação, especificamente quanto ao desempenho de ponto flutuante do processador, ao sistema de acesso à memória e ao comportamento do sistema de discos.

Infelizmente nem sempre é possível ter-se disponíveis os equipamentos de interesse. Neste caso, o conhecimento técnico das virtudes e defeitos de cada componente de "hardware" pode diminuir muito as chances de cometer enganos.

A configuração e montagem inicial de um "cluster" são tarefas simples. Não exigem grandes conhecimentos técnicos, nem equipamentos que não possam ser encontrados em qualquer parte do país. Muitas vezes os equipamentos já estão disponíveis e não são aproveitados em um almoxarifado qualquer. Na sequência deste trabalho os conceitos básicos são apresentados, exemplificando-se com a montagem do "cluster" de nosso grupo. São colocados também detalhes técnicos sobre razões que podem ser consideradas na escolha do melhor processador, HD ou implementação de bibliotecas para cálculos paralelos. No entanto, este conhecimento não é indispensável para a construção e configuração de um "cluster". Mas, na nossa opinião, são aspectos importantes da contínua sofisticação do sistema.

METODOLOGIA

Nesta seção será apresentada uma breve descrição dos equipamentos que compõem os nós, bem como alguns comentários sobre a montagem e configuração do "cluster".

Componentes de "hardware" usados na construção de uma rede local:

- computador a ser usado como servidor, ou seja, o contato entre a rede local e a "internet", e como gerenciador das áreas dos usuários e dos programas científicos;
- duas placas de rede instaladas no servidor;
- contato físico entre o servidor e as máquinas da rede interna. Pode ser um "hub" ou uma "switch";
- cabos de rede entrelaçados (conector RJ45);
- instalação de placas de rede nas máquinas da rede interna. Sugere-se placas Intel ou 3Com.

A Figura 1 exemplifica a arquitetura básica de rede local implementada.

Montagem do "cluster"

O sistema operacional a ser usado deve ser escolhido de acordo com sua capacidade técnica, mas também levando-se em consideração o conhecimento do grupo de pesquisa. Existem vários "clusters" montados com distribuições Linux e com programas da Microsoft. No meio especializado os sistemas da Microsoft não são os mais indicados pela sua instabilidade e por ser de código fechado, o que não permite adaptações às necessidades específicas dos usuários. As distribuições Linux são as mais populares. Por este motivo existem várias ferramentas de gerenciamento, monitoramento e construção de "clusters".

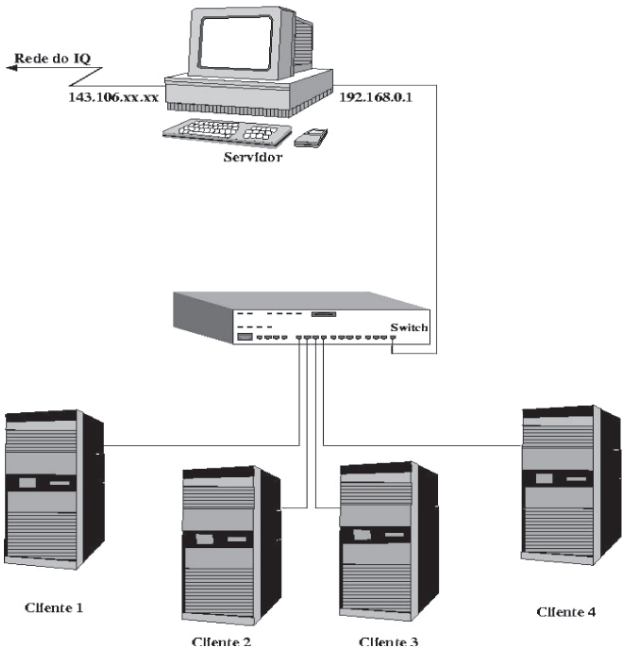

Figura 1. *Esquema básico de uma rede local usada na construção de um "cluster Beowulf"*

A escolha do FreeBSD[12] como sistema operacional deve-se não só a sua confiabilidade e estabilidade, mas também à facilidade com que pode construir e administrar redes e à familiariedade do grupo com este sistema. Especificamente neste trabalho, procurou-se confinar o "cluster" a uma rede local, oculto à "internet", procurando-se assim diminuir o tráfego, facilitando o controle quanto ao acesso remoto (segurança). A montagem do "cluster" teve como objetivo principal a resposta à demanda de nosso grupo, não procurando especificamente atingir o maior número de megaflops possível. A paralelização é usada como alternativa, não como único fim. A pesquisa de interesse não é quanto ao melhor balanceamento e eficiência do sistema, mas sim quanto à obtenção de resultados de forma prática, aproveitando ao máximo o sistema e evitando períodos de ociosidade. Para contribuir foi instituído um sistema de filas que gerencia a submissão e monitoramento de cálculos de forma automática, possibilitando a alteração do ordenamento das filas segundo os interesses do grupo. O gerenciador de filas DQS[13], inicialmente instalado, possui sérias deficiências. O programa foi originalmente escrito com limitação quanto ao tamanho dos arquivos a serem tratados. Esta deficiência vem da época dos "kernels" anteriores ao 2.4 do Linux atual, que impossibilitavam arquivos maiores que 2GB. Testando-se programas alternativos também disponíveis na rede, substituiu-se o gerenciador pelo OpenPBS[14]. Este não apresenta limitações quanto ao tamanho dos arquivos, possuindo também grande flexibilidade na configuração das filas, além de uma interface gráfica leve que facilita o uso pelo usuário e o monitoramente do administrador. Por seu código ser aberto permite que toda comunidade contribua para seu contínuo aperfeiçoamento.

Após a XI SBQT[15], onde este trabalho foi parcialmente apresentado, vários contatos foram feitos com perguntas práticas da montagem

da rede local. Aqui exemplifica-se a configuração do servidor. Excetuando-se os serviços que o servidor mantém para a sub-rede, as demais máquinas apresentam configuração semelhante. Serviços como:

- contato com a "internet", onde são tomados os cuidados com a segurança da rede;
- servidora dos programas a serem executados localmente por cada máquina e gerenciadora das áreas físicas onde os dados obtidos serão gravados (NFS);
- onde o OpenPBS monitora e gerencia as filas de submissão de cálculos às máquinas que estiverem sendo menos usadas;
- através do servidor as máquinas se mantêm sincronizadas, fator importante para o melhor desempenho ao compartilhar os arquivos dos usuários.

A configuração do servidor usando o FreeBSD é relativamente simples. Dentro do diretório /etc deve-se editar os arquivos hosts, hosts.equiv e rc.conf com informações sobre o endereçamento dos nós no "cluster". Exemplos destas configurações podem ser encontrados nos manuais disponíveis no próprio sistema. Depois disso deve-se, obviamente, criar contas com os mesmos ID para os usuários em todos os nós. As duas placas de rede devem ser configuradas como mostrado na Figura 1. Uma das placas estará em contato direto com a "internet", desta forma um endereço válido deve ser usado; neste exemplo usou-se 143.106.xx.xx do IQ/UNICAMP. A segunda, destinada à rede interna, não deve ser reconhecida fora do "cluster", o IP usado não é roteável pela "internet", segundo a sugestão da RFC 1918[16] . No "cluster" aqui apresentado foi usado como endereçamento uma rede C (192.168.xx.xx), com máscara 255.255.255.224.

Todas as máquinas envolvidas na implementação do "cluster" usado como exemplo usam FreeBSD–4.3[12] como sistema operacional. Existem, além do servidor, quatro tipos de configurações de "hardware" diferentes:

Servidor Processador Athlon 700 MHz, placa mãe Asus A7V com chipset Via KT133, 256MB de memória RAM genérica, HD IDE UDMA66 Quantum Fireball Plus AS-A com 7200rpm e 20.4GB, placa de rede 3Com 10/100 para a rede local e Intel Pro 10/100 para "internet".

Cliente 1 Processador Athlon 1.2 GHz, placa mãe Asus A7V com chipset Via KT133A, 768MB de memória RAM Itaucom, HD IDE UDMA100 Quantum AS-A com 7200rpm e 30GB, placa de rede 3Com 10/100.

Cliente 2 Processador Athlon 1.0 GHz, placa mãe Asus A7V com chipset Via KT133A, 768MB de memória RAM Itaucom, HD Ultra160 SCSI Seagate Cheetah 18XL com 10.033rpm e 18.4 GB, placa de rede 3Com 10/100.

Cliente 3 Processador Athlon 900 MHz, placa mãe Asus A7V com chipset Via KT133, 512MB de memória RAM Genérica, HD Ultra160 SCSI Quantum Atlas V com 7200 rpm e 18.0GB, placa de rede 3Com 10/100.

Cliente 4 Processador Pentium Pro 200 MHz, placa mãe Asus P/I–P6NP5, com chipset Intel 440FX, 128MB de memória RAM Genérica, HD Seagate SCSI-2 Medalist Pro com 5400 rpm e 2.17GB, placa de rede Digital 10/100.

As configurações mais recentes escolhidas na montagem do "cluster" demonstram preferência pelo uso de processadores Athlon Thunderbird, fabricados pela AMD. Para tomar a decisão o grupo fez, na época da aquisição do equipamento, uma análise técnica, pesquisou "benchmarks" e solicitou vários orçamentos. Alguns dos aspectos discutidos são apresentados a seguir.

Escolha do processador

"Benchmarks" e análises[17,18] direcionados a aplicativos de escritório, jogos 3D e ferramentes gráficas de modo geral indicam uma pequena superioridade do Pentium 3 Coopermine em comparação ao Athlon original de mesmo "clock". Pode-se atribuir este comportamento ao cache L2 (512Kb, soldados ao processador) "half-speed" usado pelo Athlon original, enquanto o Pentium 3 Coopermine possui 256Kb de L2 "full-speed" (junto ao núcleo do processador).

Nos cálculos teóricos em programas como o Gaussian/98, a vantagem se inverte a favor do Athlon. A Tabela 1 (construída a partir de resultados fornecidos pelo Prof. P. Vazquez, IQ/Unicamp) apresenta o desempenho de um processador Athlon 700MHz frente a um Pentium 3/700MHz Coopermine. Considere ainda que a partir do Athlon Thunderbird o cache L2 também é "full-speed", o que proporcionou uma diferença ainda maior de desempenho.

Tabela 1. Análise do tempo de processamento para cálculos usando-se o programa Gaussian/98, para três processadores diferentes: Pentium 3 Coopermine 700MHz, Athlon 700 MHz e Alpha/DS20.

Cálculo	Funções de Base	CPUs	Tempo	Processador
Otimização (HF)	144	2	31min 32s	Pentium3 700
	144	1	51min 27s	Pentium3 700
	144	1	35min 40s	K7/Athlon700
	144	2	12min 28s	DS20
	144	1	23min 14s	DS20
Frequência (HF)	108	2	10min 13s	Pentium3 700
	108	1	19min 20s	Pentium3 700
	108	1	14min 38s	K7/Athlon700
	108	2	04min 40s	DS20 600
	108	1	08min 34s	DS20 600
CISD	108	2	15min 39s	Pentium3 700
	108	1	21min 13s	Pentium3 700
	108	1	10min 29s	K7/Athlon700
	108	2	05min 43s	DS20 600
	108	1	06min 10s	DS20 600
CCSD(T)	90	2	22min 33s	Pentium3 700
	90	1	30min 02s	Pentium3 700
	90	1	14min 23s	K7/Athlon700
	90	2	11min 36s	DS20 600
	90	1	11min 13s	DS20 600

Enquanto a AMD investiu em um novo projeto, a Intel manteve-se com o mesmo coprocessador, outrora líder imbatível no mercado. A arquitetura como um todo não evoluiu significativamente desde os primeiros Pentium Pro até o Pentium 3. O Pentium 4 veio como uma proposta de inovação da Intel, o que não convenceu de imediato.

O Pentium 4 Willamette possui um "pipeline" muito extenso, com 20 estágios. Como muitas vezes têm-se instruções dependentes sendo processadas ao mesmo tempo, o processador precisa "escolher" quais destas serão consideradas. Por exemplo, tem-se em certo programa uma tomada de decisão: se A > B, C = 10; senão C = 0. O processador pode comparar se A é maior que B ao mesmo tempo que usa um dos resultados para C (10 ou 0) no decorrer da tarefa. A cada decisão errada, no caso do Pentium 4, há uma perda de 20 ciclos.

Um "cache" L1 é muito pequeno (8K) se comparado ao L1 de um Athlon moderno (128K), assim o processador a todo momento precisa de dados do cache L2 e da memória para continuar. "Benchmarks" mostram que acessos à memória, com dados de tamanho médio de 8K, são bastante rápidos no Pentium 4, acima disso (16 ou 32K) um Pentium III 650MHz praticamente empata; os decodificadores extras foram cortados, o que leva a um consumo muito maior de ciclos para decodificar e processar as instruções x86 (3 bytes). Por exemplo, uma instrução de 64 bytes precisaria de 21 ciclos para ser decodificada e mais 21 para ser processada, em um Athlon ter-se-ia de 7 a 11 ciclos.

Mesmo com estas e outras alterações discutíveis na arquitetura, o número de ciclos que o processador consegue atingir compensa as deficiências. Esta capacidade de atingir altos "clocks" deve-se em grande parte à técnica empregada de produzir contatos de cobre com 0,13 mícrons. A AMD ainda usa 0,18 mícrons, com contatos por alumínio, o que dificulta o aumento do "clock" sem provocar aquecimento exagerado. Atualmente (janeiro de 2002), enquanto o Pentium 4 baseado no "core" Northwood (com 512MB de cache L2), atinge 2,2 GHz, o AthlonXP 2000+, com core Palomino (256MB de cache L2), não ultrapassa 1,67 GHz. Mesmo com esta diferença por volta de 32% no número de ciclos por minuto, o AthlonXP 2000+ consegue ganhar do novo Pentium em alguns ítens[18-20].

O próximo projeto da AMD consiste em aliar a qualidade técnica da arquitetura Athlon com a fabricação de componentes do processador com 0,13 mícron, o que permitiria maior número de ciclos por segundo. Este novo processador receberá o nome de Thoroughbred, e está programado para ser lançado ainda no primeiro semestre de 2002; uma pequena demonstração deste processador foi apresentada, em uma sala fechada durante o último Comdex (novembro/2001) em Las Vegas[21,22]. A Intel em poucos meses disponibilizará novas placas mães com FSP ("Frontside Bus", o barramento do processador com a memória RAM) de 133 MHz, o que permitirá às memórias Rambus, que podem transmitir quatro instruções por ciclo, um barramento de 533 MHz[18].

Até para inovar, as filosofias da Intel e AMD são diferentes. A Intel lança novos componentes, um de cada vez, exigindo uma reformulação quase que completa em cada "upgrade". Por exemplo, ao comprar uma processador Northwood, hoje, o cliente fica impossibilitado de usá-lo em novas placas mãe que serão lançadas em dois ou três meses. Já a AMD garante que até meados de 2003 nenhuma novidade quanto aos processadores impedirá um "upgrade" específico, possibilitando o aproveitamento dos componentes de "hardware" atuais.

Em resumo, desde o lançamento da arquitetura Athlon, a AMD vem contribuindo para o desenvolvimento tecnológico de processadores, competindo ponto-a-ponto com a Intel. No momento da montagem do "cluster" apresentado neste trabalho, um sistema completo com Pentium 4 custaria mais que o dobro se usássemos Athlons Thunderbird, enquanto que o desempenho, usando compilador da Gnu, praticamente empatava. Após a Intel disponilizar um compilador específico, otimizado à arquitetura do Pentium 4, o desempenho em cálculos científicos melhorou bastante. Pela atual proximidade nos preços, a aquisição de processadores da Intel passou a ser viável. A área de atuação de cada grupo de pesquisa, a possibilidade de consultar "benchmarks" confiáveis e a possibilidade de "upgrade" são ítens que devem ser abordados. Hoje, possuindo o compilador da Intel, a escolha do grupo é diferente e a preferência seria dada aos processadores da Intel. Entretanto, mesmo com o melhor desempenho do Pentium 4 em cálculos científicos, seria prudente aguardar o lançamento do Athlon Thoroughbred, que poderia ser usado em placas mãe já disponíveis e estaria sendo usado durante um longo período de tempo, sempre com "clocks" crescentes com possíveis "upgrades". Se esta última expectativa não se concretizar, a Intel volta à sua posição de líder no mercado.

Testes de desempenho

Em recente artigo, o Prof. Morgon[9] descreve as arquiteturas de "hardware" presentes no parque computacional do CENAPAD/SP e apresenta alguns cálculos comparando os tempos de processamento com os processos desenvolvidos no "cluster" de nosso grupo de pesquisa. Os resultados são bastante interessantes, tanto quanto ao custo/benefício como à disponibilidade das máquinas do "cluster" em detrimento à longa espera das filas do CENAPAD/SP. Por exemplo, durante o mês de maio de 2002 o tempo médio de espera nas filas, com tempo de CPU acima de 24 h, foi em torno de 75 h nas máquinas seriais e 55 h nas filas de processamento paralelo. Como não existem filas intermediárias para cálculos mais longos, a espera pode inviabilizar alguns projetos. Já no "cluster" temos um número reduzido de usuários, com grande flexibilidade operacional. Enquanto cada nova máquina que forma o "cluster", dependendo dos componentes escolhidos, tem um custo médio de R$ 3.000,00, uma máquina da série RISC/6000 SP (POWER3-II/POWER4) da IBM pode variar de R$ 3.200,00 a R$ 280.000,00[7].

Nesta seção são apresentados alguns testes de desempenho para as diferentes máquinas que compõem o "cluster". São usadas várias espécies químicas diferentes: 1. metileno (singleto), 2. água, 3. fluoretileno, 4. 1-hexamina, 5. n-propilamina, 6. biciclopropenil-2, 7. fulereno C60 (icosaédrico), 8. 1-dodecamina, 9. amônia. As tabelas são apresentadas em função do tipo de método (DFT, HF, MP2, QCISD(T), CCSD(T)) e pelo tipo e número das funções de base. Na Tabela 2 tem-se alguns tempos comparativos para as três versões de máquinas Athlon que compõem o "cluster" (exceto o servidor). Em todos os testes, com vários tipos de metodologias diferentes, a máquina chamada de Cliente 1 (1,2 GHz com HD IDE Quantum AS-A) apresenta os menores tempos, seguida pelo Cliente 2 (1,0 GHz com HD SCSI Cheetah 18XL) e Cliente 3 (900 MHz com HD SCSI Quantum Atlas V).

Algumas considerações quanto ao custo/benefício podem ser feitas. Mesmo em cálculos onde há um maior acesso ao disco, como os métodos de correlação eletrônica, a vantagem do processador mais rápido é demonstrada, ficando em segundo plano a eficiência e velocidade do disco rígido mais sofisticado. O HD Cheetah, da Seagate, possui 10.033 rpm, é SCSI e apresenta-se como o mais rápido entre os discos de sua categoria (considerando a época dos testes, segundo semestre de 2001). O HD Quantum AS-A, de 7.200 rpm é também de alto desempenho, mas pelo seu preço mais acessível pode ser encontrado até mesmo em máquinas domésticas. O sistema chamado de Cliente 2 fica em torno de R$ 1.000,00 mais caro que a máquina Cliente 1, enquanto seu desempenho não justifica tal investimento, considerando as aplicações avaliadas. Estes dados demonstram que a aquisição de discos de boa qualidade é necessária, entretanto,

Tabela 2. Tempos de CPU (em s) de cálculos teóricos usando-se máquinas Athlon que compõem o "cluster" e o programa Gaussian/98

Cálculo	Método[a]	Sistemas	Máquinas			Nº de F. de Bases
			Cliente 1[b]	Cliente 2[c]	Cliente 3[d]	
Energia	DFT[e]	1	1344,6	1573,8	1756,1	72
	MP2/STO-3G	2	19,2	19,5	22,6	7
	NMR/DFT[f]	3	847,3	965,9	1099,0	108
	CCSD(T)/6-31G	4	4010,1	4428,4	4624,9	95
	QCISD(T)/6-31G	4	587,0	4021,4	4411,3	95
	MP2/6-31G(d,p)	5	542,9	642,4	658,4	185
Otimização	RHF/6-31G	6	127,5	140,7	162,1	66
	B3LYP/STO-3G	7	831,0	936,3	1076,2	300
	B3LYP/6-31+G(d,p)	4	25001,0	29078,5	31675,2	213
	B3LYP/6-31G	8	383,9	3915,7	4152,6	173
	HF/6-31G	8	1616,7	1825,0	2000,6	173
Freqüência	RHF/3-21G	9	20,3	21,5	25,5	15
	B3LYP/STO-3G	7	9290,3	10537,7	11819,8	300
	B3LYP/6-31G	8	9254,1	10890,8	11353,4	173
	HF/6-31G	8	4987,5	5633,8	6069,7	173

[a]A maioria dos cálculos apresentados são testes que acompanham a distribuição dos fontes do programa Gaussian/98; [b]processador de 1.2 GHz, com HD Quantum AS-A IDE, de 7.200 rpm; [c]processador de 1.0 GHz com HD Seagate Cheetah 18XL SCSI, de 10.033 rpm; [d]processador de 900 MHz com HD Quantum Atlas V SCSI de 7.200 rpm; [e]seqüência de cálculos RBLYP+HF e UBLYP+HF com bases 6-31+G(df,p) usando metodologia de convergência quadrática para o SCF; [f]simulação de espectro RMN com a metodologia NMR embutida no programa Gaussian/98 usando B3LYP/6-311G(df,pd)

para máquinas de alto desempenho científico; especificamente quanto aos testes feitos aqui, o processador ainda é o componente a receber os maiores cuidados na montagem de um equipamento novo ou em um "upgrade". Entretanto, é importante ressaltar que em cálculos onde o volume de dados não pode ser todo alocado na memória e com o acesso ao disco passando a ser bastante intenso, o HD passa a ser o gargalo do processo[11].

Outra avaliação superficial na escolha de HD's foi feita com a disponibilidade de uma nova máquina com mesmo conjunto do Cliente 2 (placa mãe, chipset, processador) mas com um disco IDE ATA-100 Quantum AS-A e 256MB de memória. O número de acessos ao disco para leitura/escrita (i/o) e o tempo gasto pelo sistema operacional para paginação, i/o entre outras operações, foram obtidos simplesmente com o comando "time", embutido na "shell" (csh) do sistema operacional[25]. Os dados são apresentados na Tabela 3.

Poder-se-ía esperar, com o conjunto formado por HD SCSI e 768MB de memória Itaucom, ganho no tempo de processamento que justificasse todo o investimento feito, mesmo que para cálculos menos sofisticados. Entretanto o ganho médio em tempo de CPU da máquina com HD mais sofisticado (Seagate/Cheetah) foi, em média, de apenas 2,0%. Não se espera com estes exemplos apresentar uma avaliação completa e inequívoca, mas sim testar alguns "inputs" rotineiros do programa Gaussian/98, bastante popular em computação científica. Análises, por exemplo, de trajetórias em uma simulação por dinâmica molecular, poderiam fornecer resultados bastante diferentes[11,25]. Mesmo com programas específicos (Iozone[23] e Bonnie[24])

Tabela 3. Tempos de CPU (em s) de cálculos teóricos usando-se máquinas Athlon de 1.0 GHz e o programa Gaussian/98

Cálculo	Método	Sistemas	Cheetah 18XL[a]		Quantum AS-A[b]		Nº de F. de Base
Energia	DFT[c]	1	1573,8	(1361:11,2)[d]	1607,4	(2455: 23,4)	72
	MP2/STO-3G	2	19,5	(3184: 6,9)	21,0	(4902: 8,4)	7
	NMR/DFT[e]	3	965,9	(759: 7,5)	983,7	(1422: 10,7)	108
Otimização	RHF/6-31G	6	140,7	(1681:11,6)	144,3	(1975: 12,2)	66
	B3LYP/STO-3G	7	936,3	(8850: 9,8)	974,5	(11230: 10,8)	300
	B3LYP/6-31G	8	3915,7	(2032:9,2)	3960,9	(2261: 10,3)	173
	HF/6-31G	8	1825	(2103:9,7)	1869,3	(3282: 11,2)	173
Freqüência	RHF/3-21G	9	21,5	(1148:8,7)	22,5	(1439: 9,2)	15
	B3LYP/STO-3G	7	10537,7	(33293:18,3)	10746,9	(55419: 21,8)	300
	B3LYP/6-31G	8	10890,8	(23150:77,3)	11089,3	(38622: 93,4)	173
	HF/6-31G	8	5633,8	(18574:104,9)	5757,3	(55832: 121,5)	173

[a]SCSI, 10.033 rpm. Sistema com 768 MB de memória RAM; [b]IDE, 7.200 rpm. Sistema com 256 MB de memória RAM; [c]seqüência de cálculos RBLYP+HF e UBLYP+HF com bases 6-31+G(df,p) usando metodologia de convergência quadrática para o SCF; [d]entre parênteses (I/O: Tempo) número de operações de leitura e escrita: tempo gasto pelo sistema operacional, em segundos; [e]simulação de espectro RMN com a metodologia NMR embutida no programa Gaussian/98 usando B3LYP/6-311G(df,pd)

a análise sistemática do desempenho na velocidade de leitura/escrita em função do tamanho dos blocos, ou tamanho dos arquivos, discutido anteriormente nesta revista[11], apresenta-se como uma das mais complexas entre os componentes principais de um computador, não podendo ser reduzida a um simples parâmetro de comparação.

Estes resultados fornecem informações sobre a atual capacidade de cada nó para cálculos científicos, e a relação custo/benefício entre o sistema de discos e o processador. Outro fator importante na construção de um "cluster" é o desempenho da implementação dos programas de interesse em paralelo. Para isto são empregadas bibliotecas que permitam troca de mensagens entre os nós. Por exemplo, o programa NWChem[26] pode usar TCGMSG ou MPI. O TCGMSG é automaticamente gerado pelo NWChem, já as bibliotecas MPI podem ser escolhidas a partir de diferentes implementações (MPICH[27,28] ou LAM/MPI[29], por exemplo). Os primeiros testes com MPICH apresentaram problemas quanto à alocação de memória, assim optou-se por testar o LAM/MPI[29].

Estes são apenas os primeiros de uma série de testes que farão parte de um trabalho sobre as melhores configurações paralelas para um "cluster" computacional com ênfase em química. Pelos resultados apresentados na Tabela 4, a implementação LAM/MPI é cerca de 30% mais rápida no tempo de CPU e 24% no tempo total do que usando TCGMSG.

Tabela 4. Tempos de CPU e total da otimização de geometria de uma alquilamina ($C_6H_{16}N$), usando funções de Pople (6-31G) em um cálculo DFT/B3LYP. Com memória compartilhada máxima de 128 MB usando o programa NWChem[26] com LAM/MPI e TCGMSG

Nós	Tempo de CPU (s)		Tempo Total (s)	
	LAM/MPI	TCGMSG	LAM/MPI	TCGMSG
1	9454.5	11153.5	19225.3	22750.1
2	3907.5	5580.3	9824.3	12867.4
SPEED-UP[a]	2.419	1.998	1.957	1.768

[a]Razão entre os tempos obtidos para 1 e 2 nós.

CONCLUSÕES

A escolha criteriosa dos componentes de "hardware" e "software" para a construção de "clusters" do tipo Beowulf vem se mostrando importante na viabilização de meios para que a computação de alto desempenho seja também possível em instituições onde grupos de pesquisa não podem dispor de milhares de dólares para a compra de equipamentos, programas e suporte técnico. Pretende-se com este trabalho agregar conhecimento técnico a artigos já publicados nesta revista[9,11], que têm como objetivo apresentar a um público menos especializado algumas sugestões de ferramentas computacionais que podem ser úteis a várias áreas de pesquisa. Contudo, deve-se salientar que a montagem de "clusters" não é solução geral para todos os procedimentos em química computacional. Existem programas que não são compiláveis em ambiente "unix", ou que dependem de um compilador proprietário específico, ou ainda "softwares" que não são paralelizáveis, ou que são apenas parcialmente. Ou seja, dependendo do grupo, as sugestões apresentadas aqui podem não ser completamente aplicáveis. Mesmo assim, espera-se ter despertado no pesquisador a consciência da viabilidade, tanto financeira como técnica, da construção de um "cluster" computacional como alternativa barata, de alto desempenho, à resolução de problemas em computação científica.

AGRADECIMENTOS

O autor agradece a toda estrutura computacional do Instituto de Química/Unicamp. Ao Conselho Nacional de Desenvolvimento Científico (CNPq), à Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) e ao CTPetro por todo o suporte financeiro; à FAPESP, em especial, pela bolsa de doutorado. O autor agradece também as discussões, correções e incentivos dos Profs. P. Vazquez e N. H. Morgon, fundamentais à conclusão deste trabalho.

REFERÊNCIAS

1. Loyn, H. R.; *Dicionário da Idade Média,* Jorge Zahar: Rio de Janeiro, 1990.
2. Bunson, E. M.; *Encyclopedia of the Middle Ager,* Facts on File, Inc.: New York, 1995.
3. Dralble, M.; Stringer, J.; *The Concise Oxford Companion to English Literature,* Oxford: New York, 1996.
4. Hargrove, W. W.; Hoffan, F. M.; Sterling, T.; *Scientifica American* **2001**, *August,* 62.
5. http://www-hpc.jpl.nasa.gov/PUBS/Beowulf/report.html, acessada em Janeiro 2002.
6. http://www.beowulf.org, acessada em Janeiro 2002.
7. http://www.ibm.com, acessada em Janeiro 2002.
8. http://www.top500.org, acessada em Janeiro 2002.
9. Morgon, N. H.; *Quim. Nova* **2001**, *24,* 676.
10. http://www.myrinet.com, acessada em Janeiro 2002.
11. Vazquez, P. A. M.; *Quim. Nova* **2002**, *25,* 117.
12. http://www.freebsd.org, acessada em Janeiro 2002.
13. http://www.scri.fsu.edu/~pasko/dqs.html, acessada em Janeiro 2002.
14. http://www.openpbs.org, acessada em Janeiro 2002.
15. Braga, A. A. C.; Morgon, N. H.; *Resumos do XI Simpósio Brasileiro de Química Teórica,* Caxambu, Brasil, 2001.
16. http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1918.html, acessada em Janeiro 2002.
17. http://www.anandtech.com, acessada em Janeiro 2002.
18. http://www.tomshardware.com/, acessada em Janeiro 2002.
19. http://www.tech-report.com/reviews/2002q1/northwood-vs-2000/index.x?pg=%1, acessada em Janeiro 2002.
20. http://www.guiadohardware.net/, acessada em Janeiro 2002.
21. http://www.a1-electronics.co.uk/AMD_Section/CPUs/AMD_Roadmap.shtml, acessada em Janeiro 2002.
22. http://www.infoworld.com/articles/hn/xml/01/11/14/011114hnamd.xml, acessada em Janeiro 2002.
23. http://www.iozone.org/, acessada em Maio 2002.
24. ftp://ftp.sunet.se/pub.benchmark/Bonnie/Bonnie.tar.Z, acessada em Maio 2002.
25. Gomes, A. S. P; Martins, L. R.; Vazquez, P. A. M.; *Quim. Nova* **2002**, *25,* 465.
26. http://www.emsl.pnl.gov:2080/docs/nwchem/, acessada em Janeiro 2002.
27. Gropp, W.; Lusk, E.; Doss, N.; Skjellum, A.; *Parallel Computing* **1996**, *22,* 789.
28. http://www-unix.mcs.anl.gov/mpi/mpich/, acessada em Janeiro 2002.
29. http://www.lam-mpi.org, acessada em Janeiro 2002.