Universidade de São Paulo
Escola de Artes, Ciências e Humanidades
ACH 2055 - Arquitetura de Computadores

Melina Brilhadori Alves
Murilo Galvão Honorio

Supercomputador Pleiades

São Paulo, SP
2010

Índice

1. Introdução

1.1. Considerações Gerais

O presente trabalho visa descrever as características do supercomputador Pleiades, da NASA (Agência Espacial Norte-Americana). Seu nome de batismo origina-se de um aglomerado estelar aberto localizado na constelação de Touro.

O Pleiades é considerado o sexto computador mais rápido do planeta, conforme o último ranking TOP500. O projeto TOP500 começou em 1993 com o objetivo de fornecer uma base confiável para monitoramento e detecção de tendências em computação de alto desempenho (TOP500, 2010). O anúncio da colocação no ranking foi divulgado em 28 de maio de 2010, na 25ª Conferência Internacional Supercomputação (ISC'10), realizada na cidade de Hamburgo, Alemanha (NASA, 2010). O TOP500 lista os principais supercomputadores de propósito geral do mundo.

Cabe ressaltar que o Pleiades é um projeto em andamento, assim como a maior parte dos supercomputadores.

Debutou em terceiro lugar no ranking do TOP500, em meados de 2008. No momento da pesquisa, a máquina já possuía mais que o dobro da capacidade de processamento original (CNET, 2010). Há perspectivas de crescimento, devido à disponibilidade de espaço para a adição de novos gabinetes, segundo Rupak Biswas, chefe da Divisão de Supercomputação Avançada da NASA (NAS) (CNET, 2010).

O trabalho está dividido em quatro partes: introdução, desenvolvimento, conclusão e bibliografia. No desenvolvimento, primeiramente será feita uma descrição da metodologia utilizada no TOP500, seus critérios e uma explicação do benchmark LINPACK. Em seguida, uma breve discussão da organização e arquitetura do sistema, incluindo uma descrição dos processadores e barramentos de interconexão utilizados. A terceira seção trata do histórico do projeto. A última seção é dedicada ao sistema operacional, linguagens, aplicações e ferramentas utilizadas no sistema. Teceremos os comentários finais apresentando resultados na conclusão e, por fim, indicaremos as referências utilizadas na bibliografia.

Apresenta-se uma síntese do sistema Pleiades na seção seguinte.

1.2. Características Básicas

As características básicas do sistema podem ser vistas abaixo, na tabela 1, extraídas diretamente do sítio da organização TOP500 e acrescidas de detalhes onde se fez necessário:

Características básicas do sistema de computação Pleiades em junho de 2010.	
Nome	Pleiades
Localização	NASA / Centro de Pesquisa Ames / NAS
Família	SGI Altix
Modelo	SGI Altix ICE 8200EX/8400EX
Computadores	SGI Altix ICE 8200EX/8400EX, Xeon HT QuadCore 3.0 GHz / Xeon Westmere 2.93 GHz
Vendedor	SGI
Área de aplicação	Pesquisa
Memória principal	126970 GB
Ano de instalação	2010
Sistema Operacional	Novell SuSe Linux Enterprise Edition Pro Pack + SGI5
Interconexão	Infiniband
Processador	Intel Xeon EM64T E54xx (Harpertown) 3000 MHz (12 Glops)

Tabela 1

2. Desenvolvimento

2.1. Metodologia do TOP500

2.1.1. Descrição

O projeto TOP500 foi iniciado em 1993 para fornecer uma base confiável para monitoramento e detecção de tendências em computação de alto desempenho (DONGARRA, 2009). Duas vezes por ano, uma lista dos locais de funcionamento dos 500 sistemas de computadores comerciais mais poderosos é montada e liberada. O melhor desempenho no benchmark LINPACK é utilizado como medida de desempenho para a classificação dos sistemas informáticos. A lista contém uma variedade de informações, incluindo as especificações do sistema e suas principais áreas de aplicações.

A primeira lista TOP500 é liberada em junho, na Conferência Internacional de Supercomputadores, em Dresden, Alemanha. O prazo para submissão dos trabalhos é 15 de abril. A segunda lista TOP500 é lançada em Novembro, a tempo para a Conferência Internacional de Supercomputadores. O prazo para submissão dos trabalhos é 01 de outubro.

O objetivo principal da lista TOP500 é fornecer uma lista ordenada dos sistemas de propósito geral que são comumente utilizadas para aplicações de alto nível. Os autores do TOP500 se reservam o direito de verificar de forma independente os resultados LINPACK apresentados e excluir da lista os sistemas que não são válidos ou que não são de propósito geral por natureza. Por sistema de propósito geral, entende-se que o sistema de computador deve ser capaz de ser usado para resolver uma série de problemas científicos. Qualquer sistema projetado especificamente para resolver o problema do benchmark LINPACK ou que tenha como finalidade principal o objetivo de obter uma alta posição no ranking TOP500 é desclassificado.

Os sistemas na lista TOP500 deverão ser persistentes e disponíveis para utilização por um período prolongado de tempo. A lista TOP500 é compilada por Hans Meuer da Universidade de Mannheim, Alemanha; Erich Strohmaier e Horst Simon do NERSC/Laboratório Nacional Lawrence Berkeley, Estados Unidos; e Jack Dongarra da Universidade do Tennessee, Estados Unidos (SGI, 2009).

2.1.2. Inscrições

As inscrições devem conter as seguintes informações:

Informações requeridas e suas descrições para participar da avaliação para o TOP500.	
Fabricante	Fabricante ou fornecedor
Sistema informático	Tipo indicado pelo fabricante ou fornecedor
Local das instalações	Nome completo da empresa ou instituição
Local	Cidade e país
Ano	Ano de instalação/última grande atualização
Campo de Aplicação	Tipo de cliente (universidade, governo, indústria, etc.), bem como aplicação típica.
Processadores	Número de processadores (e os tipos)
Memória	Memória principal da configuração
Tabela 2	

2.1.3. Critérios de Avaliação de Desempenho

Usa-se a melhor performance, conforme medida pelo LINPACK Benchmark, como critério de desempenho no TOP500. As métricas utilizadas são as seguintes:

- Rmax - Máxima performance alcançada no LINPACK;
- Rpeak - Máximo desempenho teórico;
- Nmax - Tamanho do problema para alcançar Rmax;
- N1/2 - Tamanho do problema para alcançar metade do Rmax;

As métricas são obtidas através da execução do benchmark LINPACK, explicado em mais detalhes na seção seguinte. Conforme TOP500 (2010) é utilizada uma versão do benchmark que permite ao usuário dimensionar o tamanho do problema e otimizar o software, a fim de alcançar o melhor desempenho para uma dada máquina.

O número obtido não reflete o desempenho geral da máquina, mas sim o desempenho de um sistema dedicado para a resolução de um sistema denso de equações lineares.

Considerando que o problema é muito regular, a performance alcançada é bastante alta, e os números de desempenho dão com boa aproximação o pico de desempenho.

Ao medir o desempenho real para n tamanhos diferentes do problema, um usuário pode obter não só o nível máximo de desempenho atingido (Rmax) para o problema do tamanho (Nmax), mas também o tamanho do problema N1/2 onde a metade da performance Rmax é alcançado. Estes números, juntamente com o pico de desempenho teórico Rpeak são os números apresentados no TOP500.

Para a composição da lista os sistemas são ordenados pelo maior valor Rmax. Em caso de empate, avalia-se o Rpeak. Para locais com computadores idênticos, a ordenação é dada pelo tamanho da memória e por fim, ordem alfabética.

2.1.4. Benchmark LINPACK

Um Benchmark é um programa de teste de desempenho que analisa as características de processamento e de movimentação de dados de um sistema de computação com o objetivo de medir ou prever seu desempenho e relevar os pontos fortes e fracos de sua arquitetura. Benchmarks podem ser classificados de acordo com a classe de aplicação para a qual são voltados como, por exemplo, computação científica, serviços de rede, aplicações multimídia, processamento de sinais, etc.

O benchmark Linpack é uma medida da taxa de execução de operações de ponto flutuante por um computador. Publicado em 1979, o projeto foi concebido originalmente com o objetivo de dar aos usuários do pacote chamado LINPACK um indicativo de quanto tempo levariam para resolver determinados problemas matriciais.

Há três benchmarks no LINPACK, que são os seguintes: Linpack Fortran n = 100, Linpack n = 1000 e Highly Parallel LINPACK. O último deles, também chamado HPL, é utilizado para a elaboração do ranking TOP500. O método de avaliação utilizado no consiste em resolver um sistema denso de equações lineares. As matrizes utilizadas nos cálculos são geradas usando um gerador de números pseudo-aleatórios.

O pacote LINPACK é uma coleção de subrotinas Fortran para solução de vários sistemas de equações lineares. O software é baseado em uma abordagem decomposicional numérica da álgebra linear. A idéia geral é a seguinte: dado um problema envolvendo uma matriz A, ela é fatorada ou decomposta em um produto de

matrizes simples e bem estruturadas, que podem ser manipuladas facilmente para resolver o problema original (DONGARRA, 2001). O próprio pacote é baseado em outro pacote, chamado de Level 1 Basic Linear Algebra Subroutines (BLAS), que resolve simples operações vetoriais.

A maioria das operações em ponto-flutuante requeridas pelos algoritmos LINPACK é efetuada pelo BLAS, o que torna possível tirar vantagem de hardware computacional específico sem ter que modificar o algoritmo subjacente, permitindo obter portabilidade e clareza ao software sem sacrificar a confiabilidade, de acordo com (DONGARRA, 2001).

O HPL afere a escalabilidade, tentando explorar a vantagem de haver grandes recursos de computação unificados para a solução de grandes problemas com a mesma eficiência. Para o desempenho do benchmark HPL NxN, contribuem a eficiência do código serial executado sobre uma única CPU bem como o algoritmo paralelo que faz todas as CPUs cooperarem. No artigo "The LINPACK Benchmark: Past, Present, and Future", Dongarra, Luszczeky e Petitetz discutem técnicas de otimização (pp. 11) que podem ser utilizadas. O HPL depende do algoritmo escolhido pelo fabricante e da quantidade total de memória disponível para no sistema.

2.2. Organização e Arquitetura

O supercomputador Pleiades é um sistema SGI Altix ICE, fabricado pela Silicon Graphics, composto de processadores Intel rodando cerca de 544 trilhões de operações de ponto flutuante por segundo, como indica o benchmark LINPACK (NASA, 2009). O Pleiades também possui a maior interconexão de rede InfiniBand do mundo, com cerca de 9 Km de cabos (NASA, 2009). É também um dos supercomputadores mais eficientes em termos de consumo de energia, segundo o NAS Datasheet (2010), usando um total de 2.35 megawatts, ou 232 megaflops por watt.

O sistema está organizado na forma de cluster, implementando a arquitetura de servidores blade, com configuração definida por Stallings (2002) como servidor secundário ativo, pois todos os nós são usados para processamento. O método de organização é a abordagem de discos compartilhados, de forma que cada computador tem acesso a todos os volumes de todos os discos.

Abaixo, na tabela 2, temos os dados do sistema em 2010, apresentados de forma sintética, conforme divulgados pela Divisão de Supercomputação Avançada da NASA (NAS, 2010):

Arquitetura do sistema

• 144 gabinetes de computadores (9.216 nós)
• 973.4 Tflop/s peak cluster
• 772.7 Tflop/s LINPACK rating
• Total de cores: 81.920
• Memória total: 127TB
• Nós
  ○ 2.048 nós
  ▪ 2 processadores 6-core por nó
  ▪ Processadores Xeon X5670 (Westmere)
  ▪ Velocidade do processador: 2.93GHz
  ▪ Cache - 12MB para 6 cores
  ▪ Tipo de Memória - DDR3 FB-DIMMs
  ▪ 2GB por core, 24GB por nó
  ○ 1.280 nós
  ▪ 2 processadores quad-core por nó
  ▪ Processadores Xeon X5570 (Nehalem)
  ▪ Velocidade do processador: 2.93GHz
  ▪ Cache: 8MB para 4 cores
  ▪ Tipo de Memória - DDR3 FB-DIMMs
  ▪ 3GB por core, 24GB por nó
  ○ 5.888 nós
  ▪ 2 processadores quad-core por nó
  ▪ Processadores Xeon E5472 (Harpertown)
  ▪ Velocidade do processador - 3GHz
  ▪ Cache - 6MB por par de cores
  ▪ Tipo de Memória - DDR2 FB-DIMMs
  ▪ 1GB per core, 8GB por nó
• Subsistemas
  ○ 8 nós "front-end"
  ○ 1 servidor PBS
• Interconexões
  ○ Entre nós: InfiniBand, 9.216 nós computacionais formando hipercubo parcial 11D
  ○ Duas topologias InfiniBand distintas
  ○ Infiniband DDR, QDR (taxa de transferência dupla e quádrupla)
  ○ Gerenciamento de rede Gigabit Ethernet
• Armazenamento
  ○ Sistema de arquivos local Nexis 9000
  ○ 8 DDN 9900 RAIDs - total de 3.1 PB
  ○ 5 sistemas de arquivo Lustre, contendo:
  ▪ 8 Servidores de Armazenamento de Objetos (OSS)
  ▪ 1 servidor de Metadados (MDS)

Tabela 3 (dados em 20/06/2010)

2.2.1. Processadores

O sistema composto por SGI Altix ICE é um cluster que inclui (NAS, 2010):

Componentes do cluster Pleiades				
Nro de Gabinetes	Processador	Total de Cores	Memória	Data da instalação
32	Intel 6-core Xeon 2.93 GHz X5670 Westmere	24.576	2GB por core	Maio/2010
20	Intel quad-core Xeon 2.93 GHz X5570 Nehalem-EP	10.240	3GB por core	Entre Novembro/2009 e Março/2010
92	Intel quad-core Xeon 3.0 GHz E5472 Harpertown	47.104	1GB por core	Setembro/2008

Tabela 4

A tabela ilustra um dos requisitos de projeto da organização em clusters, a escalabilidade incremental (STALLINGS, 2002), pois ele foi sendo expandido de forma incremental e conforme discutido anteriormente ainda há planos de adicionar mais nós. Isso foi facilitado por uma das características dos gabinetes ICE, que é permitir sua conexão em poucas horas, pois vem pré-testado de fábrica (SGI, 2010).

Todos os processadores são chips multicore, o que beneficia a velocidade de execução quando propicia o processamento paralelo de instruções. Muitas das aplicações do cluster são independentes umas das outras, então os servidores podem fazer as numerosas transações e execuções em paralelo. Aplicações muiltithreading também podem ser exploradas com ótimos resultados em multiprocessadores.

Pleiades que, como já visto, é composto por três séries do processador Intel Xeon. Cada série possui as seguintes especificações divulgadas pela Intel:

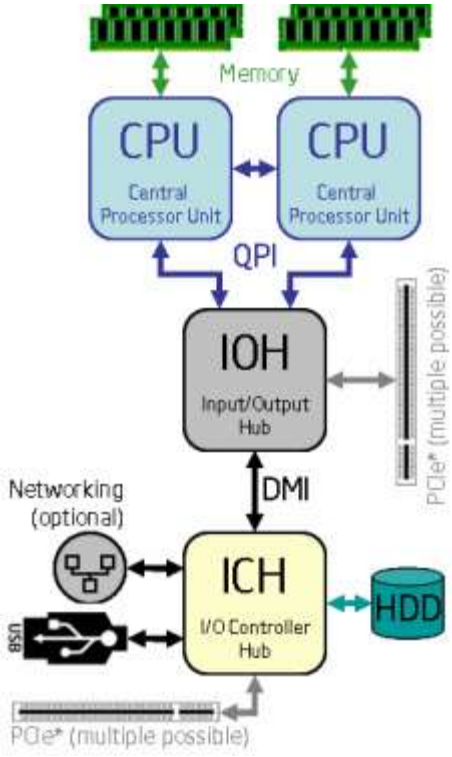
Características do processador Westmere	
Westmere	X5670
# of Cores	6
# of Threads	12
Clock Speed	2.93 GHz
Max Turbo Frequency	3.33 GHz
Intel® Smart Cache	12 MB
Bus/Core Ratio	22
Intel® QPI Speed	6.4 GT/s
# of QPI Links	2
Instruction Set	64-bit
Instruction Set Extensions	SSE4.2

Tabela 5

Fonte: http://ark.intel.com/Product.aspx?id=47920

Características do processador Nehalem	
Nehalem-EP	X5570
# of Cores	4
# of Threads	8
Clock Speed	2.93 GHz
Max Turbo Frequency	3.333 GHz
Intel® Smart Cache	8 MB
Bus/Core Ratio	22
Intel® QPI Speed	6.4 GT/s
# of QPI Links	2
Instruction Set	64-bit

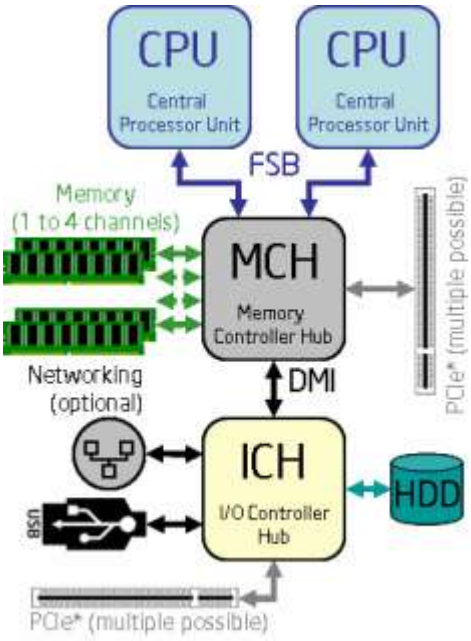
Tabela 6



Fonte: http://ark.intel.com/Product.aspx?id=37111

Características do processador Harpertown	
Harpertown	E5472
# of Cores	4
# of Threads	4
Clock Speed	3 GHz
L2 Cache	12 MB
Bus/Core Ratio	7.5
FSB Speed	1600 MHz
FSB Parity	
Instruction Set	64-bit

Tabela 7

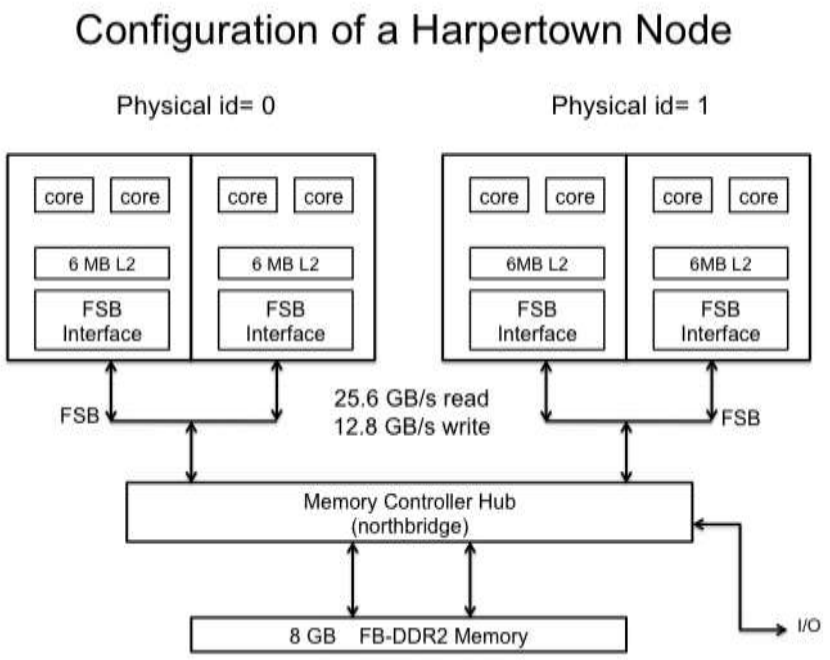


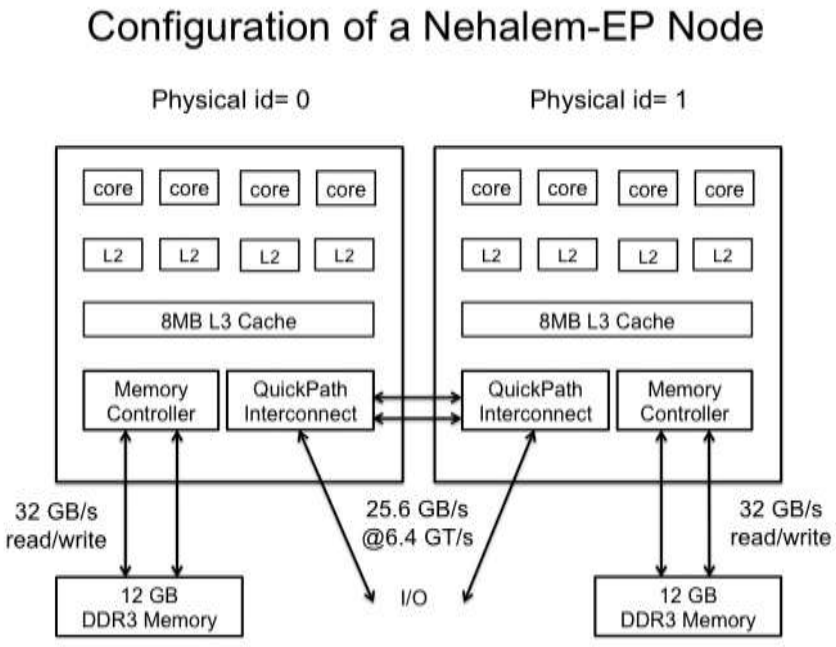
Fonte: http://ark.intel.com/Product.aspx?id=33085

Westmere e Nehalem-Ep implementam Hyper-Threading, já o Harpertown não. A tecnologia Intel Hyper-Threading (HT) oferece paralelismo em nível de thread sobre cada processador, resultando em um uso mais eficiente dos recursos do processador - maior rendimento de processamento - e melhoria no desempenho do software multithreaded. O quick-patch interconnect (QPI) é um caminho de interconexão rápida (STALLINGS, pp. 571) entre os compoentes.

2.2.1.1. Configuração dos nós

Conforme os diagramas abaixo:





Figuras 1 e 2. Fonte:

http://www.nas.nasa.gov/Users/Documentation/Ice/nehalem_quickstart.html

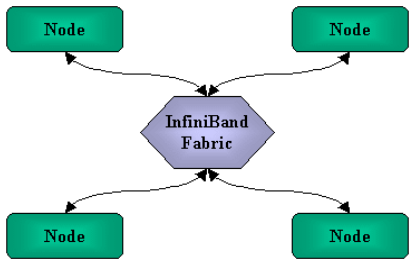
2.2.1.2. Desempenho

Para aplicações que são sensíveis à frequencia de clock e amigáveis às técnicas de cache (como o Linpack), não há ganho de performance no NehalemEP sobre o Harpertown. Por outro lado, aplicações sensíveis à largura de banda (bandwidth) da memória se beneficiam mais da arquitetura Nehalem-EP, segundo dados da SGI (2010).

2.2.2. Interconexões

O InfiniBand oferece conexão serial bidirecional ponto-a-ponto, destinado à conexão de processadores com periféricos de alta velocidade, tais como discos. Suporta taxa de transferência de dados dupla (DDR) e quádrupla (QDR), que oferecem 5 Gbit/s ou 10 Gbits/s respectivamente, para a mesma taxa de ciclo de dados. A latência dos swithes é de 140 nanosegundos.

Infiniband, em oposição aos barramentos compartilhados tradicionais, exibe uma arquitetura de nós interligados através de switches, conforme o diagrama abaixo (PENTAKALOS, 2002), que mostra uma instalação básica:



Dois ou mais nós conectam-se uns aos outros através da estrutura. Um nó pode representar um dispositivo host como um servidor ou um dispositivo de E/S, por exemplo um subsistema RAID. A estrutura em si pode consistir de um único switch até uma coleção de swtiches e roteadores interconectados. Não foi possível obter detalhes da topologia do Pleiades, além dos já apresentados.

2.2.3. Armazenamento

O sistema de arquivos local é o Nexis 9000, fornecido pela própria fabricante dos servidores, a SGI, sendo uma alternativa barata às implantações em fibra ótica.

Trabalha sobre a Infiniband, escalando até 1 GB por segundo na escrita e 3 GB por segundo na leitura. Para a infraestrutura de armazenamento, no Pleiades são utilizadas 8 plataformas de armazenamento da DirectData Network (DDN 9900). São cerca de 3 PB ao todo. O DDN 9900 é capaz de escrever na mesma velocidade da leitura, oferece 6 GB/s de rendimento e paridade RAID 6. Possui uma combinação de SSD, SAS e SATA no mesmo compartimento de disco. Cada um dos 8 sistemas gerencia cerca de 1200 drives, em 2 gabinetes por sistema. Segundo a fabricante, o acesso a dados tem latência zero, ou seja, é em tempo real.

2.2.3.1. Sistema de arquivos

São cinco sistemas de arquivos Lustre, com um tamanho total aproximado de 2400 TB, montados conforme a tabela abaixo:

Montagem do sistema Lustre	
Sistema de Arquivos	Tamanho
/nobackupp10	880 TB
/nobackupp20	440 TB
/nobackupp30	440 TB
/nobackupp40	219 TB
/nobackupp50	440 TB

Tabela 8

Fonte: http://www.nas.nasa.gov/Users/Documentation/Ice/nehalem_quickstart.html#filesystems

O Lustre é um sistema de arquivos para clusters baseado em objetos. Foi desenvolvido como projeto de pesquisa por Peter Braam em 1999. A kernel do sistema operacional do Pleiades (SUSE Linux) suporta este sistema de arquivos sem alterações. O nome lustre é uma composição de Linux e cluster, segundo o sítio do projeto.

Um sistema de arquivos Lustre tem três áreas funcionais:

- Um alvo de metadados (metadata target - MDT) único por sistema de arquivos que armazena metadados, como nomes de arquivos, diretórios, permissões, e layout de arquivos, em um servidor de metadados (metadata server - MDS);

- Um ou mais servidores de armazenamento de objetos (object storage server - OSS), que armazenam dados dos arquivos em um ou mais alvos de armazenamento de objetos (object storage target - *OST)*. Um OSS serve de dois a oito alvos, cada alvo tem um sistema de arquivos local de tamanho de até oito terabytes (TB). A capacidade de cada sistema de arquivos Lustre é a soma das capacidades dos alvos;

- Clientes que acessam e utilizam os dados. Lustre fornece a todos os clientes uma semântica padrão POSIX e acesso à leitura e escrita concorrente aos arquivos.

O módulo driver do sistema de arquivos Lustre é carregado na kernel e o sistema de arquivos é montado como qualquer outro sistema de arquivos local ou em rede. Aplicações clientes enxergam um único sistema de arquivos unificado, embora ele seja composto por milhares de servidores individuais e sistemas de arquivos MDT/OST.

No Pleiades, a rede conectando servidores e clientes provê comunicação. O armazenamento de disco é conectado através aos sistemas de arquivos MDSs e OSSs usando tecnologias storage area network (SAN) convencionais. Há um robusto mecanismo de failover e restauração, que torna as falhas de servidores e reinicializações transparentes aos clientes. São possíveis atualizações e reinicializações enquanto jobs ativos continuam rodando, meramente havendo atraso.

2.2.3.2. Transferência de Arquivos de/para o sistema

Além dos comandos básicos, limitados por uma conexão de largura de banda limitada, existem dois sistemas adicionais, chamados bridge1 e bridge2, cada um com uma conexão de alta velocidade de 10-Gbit.

2.2.4. Balanceamento de carga

Há oito nós frontais no supercomputador Pleiades, chamados pfe1 a pfe8. Visando prover um ambiente balanceado para todos os usuários do Pleiades, foi criado um mecanismo que automaticamente seleciona o sistema menos carregado durante o login. A carga do sistema é uma função ponderada das cargas dos discos, processadores e rede.

2.3. Histórico do Projeto, Aspectos de Implementação e Utilização

O Pleiades foi concebido com o objetivo de atender ao requerimentos de computação da Agência Espacial Norte-Americana (NASA), provendo capacidade computacional para todas as suas missões (investigação aeronáutica, sistemas de exploração, ciência e operações espaciais) (NAS, 2008).

A seguir temos os principais aspectos relacionados ao histórico do projeto e sua implementação. Mostraremos os órgãos e empresas envolvidos no desenvolvimento e financiamento do Pleiades, bem como sua localização e ciclo de vida. Também apresentaremos uma linha de tempo do projeto e por fim as pesquisas nas quais ele é utilizado.

2.3.1. Órgãos e Empresas Envolvidos

Temos como órgão financiador e principal interessado, a Agência Espacial Norte-Americana (NASA).

O principal fornecedor de hardware é a Silicon Graphics, Inc (SGI), que desenvolve e vende uma grande linha de servidores e soluções de data-storage, desde os de baixo custo até os de alta performance. Os servidores que compõe o Pleiades possuem processadores fornecidos pela empresa Intel.

O cluster encontra-se alocado na Divisão Supercomputação Avançada da NASA (NAS), que dedica-se a fornecer aos cientistas e engenheiros os recursos de supercomputação e ferramentas de simulação necessários para a realização das missões críticas da NASA e a fazer novas descobertas científicas em benefício da humanidade.

NASA, SGI e Intel trabalham de maneira muito próxima, com o objetivo de incrementar as capacidades computacionais do Centro de Pesquisa Ames, principalmente acrescentando nós ao Pleiades. Trabalham através de um "Space Act Agreement", que são um tipo de acordo legal diferenciado, dentro das leis federais de licitação americanas.

2.3.2. Localização

Localização do supercomputador Pleiades	
Nome do Local	NASA / Centro de Pesquisa Ames (ARC) / NAS
Cidade	Mountain View
Estado	Califórnia
País	Estados Unidos

Tabela 9

O Centro de Pesquisa Ames (Ames Research Center) foi fundado em 1939 e concentra-se em pesquisar novas tecnologias para viabilizar as missões da NASA. Essas tecnologias incluem biologia espacial, nanotecnologia, biotecnologia, sistemas de proteção a calor para espaçonaves e aeronaves, tecnologia da informação e astrobiologia.

2.3.3. Custo

Por tratar-se de projeto estratégico, torna-se difícil encontrar dados precisos com relação aos custos de implantação e operação. Não foi possível estimar os custos de aquisição envolvidos com base nos dados publicados pela SGI e Intel. Além disso, é política das empresas divulgar esses dados apenas mediante requisição de clientes interessados em adquirir soluções de computação em alta performance.

2.3.4. Ciclo de vida

Supercomputadores são como a maioria das outras classes de máquinas. Sendo assim, o projeto Pleiades foi concebido dentro de um ciclo de vida limitado a cerca de três anos. Segundo Biswas, chefe da divisão de computação da NASA (CNET, 2010), após esse período, devido ao rápido avanço do setor de tecnologia e das inovações, não há expectativa de se obter uma relação de custo-benefício em operá-lo.

2.3.5. Linha do tempo

O sistema Pleiades foi construído e testado durante o verão de 2008, e tornou-se disponível para os usuários para computação produtiva no mês de Dezembro de 2008 (NAS, 2009). Os seguintes fatos marcaram o desenvolvimento do projeto:

- 23/05/08 - Primeira entrega de gabinetes.
- 06/06/08 - Primeiros 11 gabinetes instalados. Cada gabinete contém 512 cores e 512 GB de memória.
- 17/06/08 - Segunda entrega de gabinetes. Ferramentas básicas de monitoramento disponibilizadas.
- 27/06/08 - Todos os cores aprontados. Baterias de testes de aplicação não apresentaram erros. Infiniband integrado. Conexões 1-GigE instaladas no switch de rede.
- 03/07/08 - Atualização do sistema de resfriamento (instalação de Chiller de 450 toneladas). Continuação dos testes de aplicação.
- 17/07/08 - Mais trabalhos nas instalações, com novo sistema de energia. Mais testes de aplicação.
- 25/07/08 - Testes nos roteadores InfiniBand
- 01/08/08 - Chegada de mais 32 gabinetes. Servidor de metadados com 300GB. Fase de testes próxima da conclusão. Segunda compra de 44 gabinetes SGI ICE.
- 18/08/08 - Fase de testes de hardware. Testes de estabilidade nos primeiros 8 gabinetes. Integração do servidor de metadados Lustre ao Pleiades.
- 02/09/08 - 40 gabinetes originais alocados na NAS em uma configuração 8 (P1) e 32 (P2). Aplicações de performance e escalabilidade executadas exaustivamente.
- 22/09/08 - P1 estáveis e em modo "produção". Adição de memória ao P2. Todos os 16 gabinetes do P3 instalados. Grupo de Performance e Produtividade de Aplicativos realiza testes de stress.
- 29/09/08 - Pleiades configurado com 64 gabinetes, conectando P2 e P3 através de 512 cabos InfiniBand.
- 06/10/08 - Necessário um sistema de treliças para interligar os cabos de fibra ótica dos gabinetes.
- 13/10/08 - Completada execução do benchmark LINPACK em 92 gabinetes.

- 10/11/08 - Ajustes finos. Criado um sistema de arquivos Lustre com 219 terabytes.
- 17/11/08 - Pleiades obtém terceiro lugar no TOP500 e o vigésimo segundo lugar no Green500, em eficiência energética.
- 11/12/08 - Cerimônia de entrega do sistema.

2.3.6. Ranking TOP500

Pleiades estreou no ranking TOP500 em novembro de 2008, tendo ocupado desde então as seguintes posições:

Posições do sistema Pleiades no TOP500					
Lista	Cassificação	Cores	Rmax	Rpeak	Energia
06/10	6	81920	772.70	973.29	3096
11/09	6	56320	544.30	673.23	2348
06/09	4	51200	487.01	608.83	2090
11/08	3	51200	487.01	608.83	2090

Tabela 10

Observação: Rmax e Rpeak em TFLOPS. Energia em KW.

O Pleiades entrou no ranking na terceira posição. Observa-se que em apenas um ano e meio, apesar do aumento expressivo no número de cores, surgiram novos supercomputadores competidores que o ultrapassaram. Esse incremento em cores e memória instalada apenas manteve o Pleiades na mesma posição no ranking.

2.3.7. Utilização

Dentre outros projetos de engenharia e científicos, o Pleiades apóia as seguintes trabalhos:

- Simular o comportamento de espaçonaves e sistemas durante a reentrada na atmosfera terrestre e projetar sistemas de segurança para futuras espaçonaves;
- Simulações extensivas dos grandes problemas computacionais para o projeto do futuro veículo espacial;

- Desenvolvimento de modelos cada vez mais detalhados de grandes halos de matéria escura e evolução de galáxias;
- Execução de modelos oceano-atmosféricos acoplados para avaliar habilidade de previsão climática decadal do Painel Intergovernamental sobre Mudança Climática;
- Cálculos para avaliar hipóteses de astrobiologistas de que muitas das reações químicas prebióticas foram catalisadas por superfícies minerais localizadas em aberturas hidrotermais durante o período inicial da história da Terra;
- Simulações de dinâmica de fluidos de aeronaves complexas, para melhora da performance dos veículos e redução dos impactos ambientais;
- Estudos do comportamento de ligas metálicas com características de memorização de formato para aplicações de altíssima-temperatura.

2.4. Ferramentas Computacionais

O ambiente operacional do sistema Pleiades é descrito nas próximas seções. Existe um sistema chamado "modules" que centraliza a localização dos produtos licenciados e domínio público instalados no Pleiades.

2.4.1. Sistema Operacional

A distribuição base de sistema operacional instalado é o SUSE Linux Enterprise Server 10 (SLES10), da empresa Novell. Segundo a empresa SGI, os sistemas ICE, como o Pleiades, utilizam-se desse sistema operacional para obterem uma ambiente Linux robusto, que suporta servidores de computação de alto desempenho (High Performance Computing - HPC) altamente escaláveis. Além disso, a própria empresa SGI fornece suporte para o sistema operacional.

A Novell fornece kit de desenvolvimento de software (SDK) que apóia o desenvolvimento e a migração de aplicações para essa versão do SUSE Linux.

O Pleiades está configurado para haver tolerância a falhas. Para a realização de failover, as máquinas possuem o software Linux FailSafe, que provê redundância de processadores e controladoras de I/O. Os serviços são monitorados pelo software, se uma falha é detectada, o processo de failover é iniciado de acordo com políticas definidas pelos administradores do sistema.

2.4.2. Middleware

O sistema Pleiades possui o SGI ProPack6 SP3 instalado junto ao sistema operacional, cujo foco são recursos para acelerar aplicações, para possibilitar o desenvolvimento de aplicações paralelas e em tempo-real e para administração dos servidores, com mecanismos de balanceamento de carga.

2.4.3. Escalonador de Jobs

O Portable Batch System (PBS) fornece o ambiente para submissão de jobs em lote. A versão 10 é a que está sendo usada no Pleiades. Os nós de computação são gerenciados por um scheduler que aloca blocos de nós para os jobs a fim de fornecer acesso exclusivo. Os usuários apresentam os batch jobs para serem executados em um ou mais nós usando o comando 'qsub' a partir de uma sessão interativa em um dos nós de front-end.

2.4.4. Compiladores, bibliotecas e linguagens

No Pleiades, não há um compilador padrão definido. Estão disponíveis diversas versões de compiladores Intel. A NAS recomenda a versão contida em *comp/intel/10.1.021_32,* mas outras podem ser testadas pelos usuários para avaliar o impacto no desempenho.

A biblioteca chamada Intel Math Kernel Library (MKL) é disponibilizada. Ela é composta de funções matemáticas altamente otimizadas, para aplicações matemáticas e de engenharia que requeiram alto desempenho em plataformas Intel. As áreas funcionais do código da biblioteca incluem álgebra linear, consistindo de LAPACK e BLAS (usadas, por exemplo, pelo benchmark LINPACK e já descritas na seção dedicada a ele), além de algoritmos para aritmética de números complexos (FFT), e funções vetoriais transcedentais, entre outros.

Está disponível também a biblioteca MPT (Message Passing Toolkit) da SGI. Segundo a NAS, a versão mpi/mpt.1.23.nas, é particularmente importante para executar jobs muito grandes (que usem mais de 4000 CPUs).

MPT é um pacote de software que suporta o intercâmbio de dados entre processos para aplicações que usam processos concorrentes cooperativos em um único host ou

múltiplos hosts. A troca de dados é feita através de troca de mensagens, as quais usam chamadas de biblioteca para solicitar entrega de dados de um processo para outro ou entre grupos de processos.

O pacote MPT contem os seguintes subcomponentes: a Message Passing Interface (MPI) e a biblioteca de programação SHMEM. O componente MPI do MPT inclui uma biblioteca de comunicação e funções de sincronização que são necessárias para escrever aplicações distribuídas. Por exemplo, podem-se adicionar chamadas que ocasionam que uma tarefa envie uma mensagem para outra, ou receba uma mensagem, ou aguarde até que outra tarefa esteja concluída.

MPI é uma especificação padrão para bibliotecas de troca de mensagens, permitindo programas de intercâmbio de mensagens portáteis em linguagem Fortran e C. O modelo de programação SHMEM é outra forma de programação distribuída. Difere das trocas de mensagem MPI, que usam comunicação unilateral. Um elemento de processamento (PE) pode enviar ou receber dados de um PE remoto sem a participação direta do PE remoto.

Foi possível deduzir a partir da pesquisa que os compiladores GNU gcc e gfortran estão disponíveis, pois na documentação são citados comandos referentes a eles. Além disso, segundo o gerente do projeto William Thigpen, uma das características do sistema é permitir a migração de códigos desktop de maneira facilitada.

2.4.5. Aplicações

Na tabela 11, apresenta-se uma relação de algumas das aplicações executadas no Pleiades, obtidas do site da NAS, além de um breve resumo sobre elas:

Exemplos de aplicações do Pleiades	
Aplicação	Descrição
CART3D	Realiza design aerodinâmico preliminar e conceitual. Permite que os usuários realizem análise computacional automatizada de dinâmica de fluídos sobre geometrias complexas.
debris	Um pacote que inclui programas para estimar trajetórias dos ônibus espaciais, especialmente durante os lançamentos.

ECCO	Aplicação utilizada para conduzir modelagem e análise em larga-escala e alta resolução do larga-escala do oceano, no contexto do consórcio ECCO, que envolve o Laboratório de Propulsão a Jato da NASA (JPL), o Instituto de Tecnologia de Massachusetts (MIT), e o Instituto de Oceanografia Scripps. Cientistas estão utilizando as estimativas para conduzir estudos sobre o clima global.
fvGCM	É um modelo de previsão de clima e tempo globais usado tradicionalmente para simulações de longo prazo a uma resolução horizontal de aproximadamente 100 km.
INS3D	Estudos relacionados ao impacto da microgravidade na circulação sangüínea, através da solução de equações de Navier-Stokes.
Overflow	Um programa de dinâmica de fluídos para resolver problemas de fluxo complexos. Muito usado pela NASA e pela indústria para projetar lançamento e reentrada de veículos, navios, aviões comerciais, entre outros.
POP	POP é o componente oceânico do Community Climate System Model (CCSM), um modelo climático global altamente acoplado que permite simulações precisas dos estados climáticos no passado, presente e futuro.

Tabela 11

Além das aplicações descritas acima, a NAS oferece uma série de softwares para download em seu sítio, com explicações.

2.4.6. Uso do sistema

Na utilização dos sistemas ICE os recursos são alocados para os usuários em múltiplos nós (com oito CPUs em cada nó) e a alocação é baseada no número de nós (não importando quantos processadores ou nós são realmente usados) utilizados por um job. Quando são atribuídos recursos a um usuário, este terá acesso exclusivo aos recursos disponibilizados, até que o job chegue ao fim ou exceda o tempo solicitado (wall-clock).

A unidade de contabilidade de uso é um *System Billing Unit* (SBU), que é medido como:

SBU =(Wall_Clock_Hours_Used * Número de nós * 8) * CPU_factor

O fator de CPU para os sistemas ICE é descrito abaixo:

Fator de CPU para os sistemas ICE	
Host	**CPU_Factor**
Nós Pleiades Harpertown	0.90
Nós Pleiades Nehalem-EP	2.2 (por core físico)
Nós Pleiades Westmere-EP	1.8 (por core físico)
Tabela 12	

É permitido rodar jobs entre os três tipos de processadores. Isso pode ser útil em dois cenários:

1. Quando não houver nós suficientes livres de um modelo para o job;
2. Quando alguns dos processos submetidos precisam de mais memória enquanto outros precisam de menos.

Tipicamente, um usuário que deseje usar o Pleiades efetuara login através do Secure Front End (sfe1 ou sfe2). Quando obtiver acesso, precisará logar em um dos Pleiades Font Ends (pfe1 a pfe8). São permitidas operações remotas de jobs autônomos e scripts em hosts específicos, desde que se obtenha uma chave para o Secure Unattended Proxy (SUP).

Quanto aos programas que rodam sobre o Pleiades, conforme discutido anteriormente, é possível executar facilmente qualquer código que funcionam em desktops Linux. O sistema operacional, middleware e pacote MPT proveem paralelismo e permitem otimizar o uso do sistema.

3. Conclusão

O Pleiades representa um significativo avanço de engenharia de várias maneiras. Além de seu enorme poder e do recorde de escala da estrutura InfiniBand, ele é capaz, segundo o gerente do projeto William Thigpen, de executar os códigos da NASA com modificações mínimas, sendo também compatível com estações de trabalho de engenharia desktop, permitindo que os usuários migrem seus códigos facilmente dos

desktops. Todas as missões da NASA têm à disposição um enorme recurso computacional para atingir seus objetivos críticos.

Stallings (2002, pp. 671) relaciona benefícios da organização de cluster. Podemos observar, no trabalho, a escalabilidade absoluta, pois o cluster Pleiades é constituído de dezenas de máquinas multiprocessador. A escalabilidade incremental foi discutida na seção 2.2.2. Há alta disponibilidade, pois nos relatos das fases de implementação e baterias de testes, partes do sistema permaneceram disponíveis. Embora não tenha sido possível estimar os custos do projeto, um fato relevante é que foram usados elementos (gabinetes de nós) disponíveis comercialmente.

Avaliar o desempenho de um sistema é uma questão complexa, possuindo uma enorme quantidade funções interrelacionadas. Os resultados dos Benchmarks não devem ser utilizados como medidas de desempenho total do sistema, mas sim como pontos de referências (vide discussão em 2.2.1.2). O Linpack provê o referencial do projeto TOP500, pois há grande interesse em estatísticas por parte de fabricantes, organizações, usuários e potenciais clientes. Além disso, pelo fato do projeto reunir dados de onde os sistemas se localizam e como são utilizados, ele facilita o estabelecimento de colaborações e troca de dados e software, visando a evolução do mercado de alta performance.

Conforme visto em 2.3.6, desde a entrada no ranking TOP500 o Pleiades perdeu três posições, embora tenham sido adicionados milhares de nós com processadores 6-core e mais memória. Isso ilustra os avanços constantes e a competitividade acirrada do setor de supercomputadores. Uma breve análise do ranking mostra, por exemplo, que a China vem ser tornando uma potência emergente nesse setor, com diversas máquinas despontando na última divulgação do ranking, ou seja, os interesses das potências científicas, econômicas, políticas e militares estão voltados para essa área da computação. Considerando que as principais máquinas no topo do ranking são clusters, confirma-se que esses arranjos computacionais são uma das mais atrativas soluções para a computação de alto desempenho.

4. Bibliografia

DONGARRA, J; LUSZCZEKY, P; PETITETZ, A. The LINPACK Benchmark: Past, Present, and Future. 2001. Disponível em:

<http://www.netlib.org/utk/people/JackDongarra/PAPERS/hplpaper.pdf>. Acesso em 15 jun. 2010.

GREEN500. The Green500 List - November 2008. Disponível em: <http://www.green500.org/lists/2008/11/top/list.php>. Acesso em: 31 mai. 2010.

MEWHINNEY, M; DUNBAR, J. Nasa Supercomputer Ranks Among World's Fatest. NASA, 2008. Disponível em: <http://www.nas.nasa.gov/News/Releases/2008/11-18-08.html>. Acesso em: 12 jun. 2010.

MEWHINNEY, M; DUNBAR, J. Nasa Supercomputer Ranks Among World's Fatest. NASA, 2009. Disponível em: <http://www.nas.nasa.gov/News/Releases/2009/11-18-09.html>. Acesso em: 30 mai. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. Nehalem-EP Quickstart Guide. Disponível em: <http://www.nas.nasa.gov/Users/Documentation/Ice/nehalem_quickstart.html>. Acesso em: 3 jun.2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. Pleiades Construction. Disponível em: <http://www.nas.nasa.gov/About/Projects/Pleiades/pleiades_build.html>. Acesso em: 31 mai. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. Pleiades Datasheet. Disponível em: <http://www.nas.nasa.gov/Resources/Systems/PDF/pleiades_datasheet.pdf>. Acesso em 30 mai. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. Pleiades Hardware. Disponível em: <http://www.nas.nasa.gov/Users/Documentation/Ice/hardware_pleiades.html>. Acesso em: 1 jun. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. Pleiades Software. Disponível em: <http://www.nas.nasa.gov/Users/Documentation/Ice/software.html>. Acesso em: 1 jun. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. SGI Altix ICE Systems. Disponível em: <http://www.nas.nasa.gov/Users/Documentation/Ice/ice.html>. Acesso em: 31 mai. 2010.

NASA ADVANCED SUPERCOMPUTING DIVISION. SGI ICE - Pleiades Supercomputer. Disponível em: <http://www.nas.nasa.gov/Resources/Systems/pleiades.html>. Acesso em: 31 mai. 2010.

PENTAKALOS, O. An Introduction to the InfiniBand Architecture. 2002. Disponível em:<http://www.oreillynet.com/pub/a/network/2002/02/04/windows.html>. Acesso em 16 jun. 2010.

RIGUES, R. NASA inaugura novo supercomputador. IG, 2008. Disponível em: <http://tecnologia.ig.com.br/noticia/2008/11/19/nasa_inaugura_novo_supercomputador_2121643.html>. Acesso em: 30 mai. 2010.

SGI. Message Passing Toolkit (MPT) Release Notes MPT Release 1.20. Disponível em: <http://techpubs.sgi.com/library/tpl/cgi-bin/getdoc.cgi?coll=linux&db=relnotes&fname=/usr/relnotes/sgi-mpt-1.20>. Acesso em: 20 jun. 2010.

SGI. SGI Altix Ice: World Record Cluster Performance. Disponível em:<http://www.sgi.com/products/servers/altix/ice/>. Acesso em 20 jun. 2010.

STALLINGS, W. Arquitetura e Organização de Computadores: Projeto para o Desempenho. São Paulo: Prentice Hall, 2002.

STELIAS COMPUTING INC. Lustre Project Homepage. 1999. Disponível em: <http://web.archive.org/web/20000823132309/http://www.lustre.org/>. Acesso em: 15 jun. 2010.

TERDIMAN, D. Inside NASA's world-class supercomputer center. CNET, 2009. Disponível em: <http://news.cnet.com/8301-13772_3-20003333-52.html>. Acesso em: 30 mai. 2010.

TOP500. The TOP500 Project. Disponível em: <http://www.top500.org/project>. Acesso em: 30 mai. 2010.