Cloud Computing

Andrei S. Braga, Geraldo M. Silva, e Marcos C. Barros
RAs: 079713, 079740, 820650
Instituto de Computação - Universidade Estadual de Campinas
Av. Albert Einstein, 1251
Campinas, São Paulo, Brazil - 13083-852
{andreisampaio, geraldoms007}@gmail.com, marcos.barros@freescale.com

ABSTRACT

Cloud computing surgiu recentemente como um novo paradigma para a indústria da informática com relação à disponibilidade e acesso a recursos através da Internet. Ela tem transformado a maneira como consumidores, donos de negócios e empresas da área de tecnologia da informação se relacionam, na mesma medida em que tem alterado a forma como plataformas de hardware e software são agrupadas, interagem entre si e como são comercializadas. Embora algumas das tecnologias consideradas como fundamentos de cloud computing já estejam disponíveis há algum tempo, como por exemplo virtualização e utility computing, cloud computing ainda está em sua "infância" e possui uma série de desafios em aberto, como padronização, provisionamento de recursos, entre outros. Neste trabalho, apresentamos uma visão geral dos conceitos e características referentes à cloud computing, como ela se diferencia de outras tecnologias semelhantes, como ela está influenciando a área de arquitetura de computadores e quais desafios a serem superados para que ela se estabeleça disponibilizando todo o seu potencial.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Management, Performance, Security, Standardization

Keywords

Cloud computing, Computação em Nuvem, Data centers, Virtualização, Escalabilidade, Arquitetura de Computadores, Multiprocessadores, Consumo de Energia, Network-on-chip

1. INTRODUÇÃO

Em 1969, Kleinrock disse (em uma tradução livre): "Agora, as redes de computadores ainda estão na sua infância, mas, ao modo que amadureçam e tornem-se mais sofisticadas, nós provavelmente veremos a difusão de utilidades computacionais que, como a eletricidade e a telefonia de hoje, irão servir lares e escritórios por todo o país". De fato, algumas décadas depois, foi presenciado, não somente nos Estados Unidos, mas em todo o mundo, o surgimento de um novo modelo computacional, chamado cloud computing, onde recursos computacionais são disponibilizados como utilidades que podem ser alocadas a usuários ou desalocadas através da Internet e sob demanda.

O surgimento deste novo modelo computacional provocou um tremendo impacto sobre a indústria de Tecnologia da Informação (IT, Information Technology). Nos anos recentes, empresas como Google, Amazon e Microsoft têm se esforçado ao máximo para prover plataformas de cloud de maior poder computacional, mais confiáveis e de menor custo. Em geral, as grandes organizações têm buscado reformular os seus modelos de negócios para se beneficiar desta nova tecnologia.

Cloud computing é uma tecnologia bastante interessante para provedores de serviços por apresentar várias características muito atrativas. Entre essas características estão: não requerer investimento inicial, possuir baixo custo operacional, ser altamente escalável, ser de fácil acesso e reduzir riscos de negócios e despesas de manutenção.

O objetivo deste trabalho é prover um bom entendimento de cloud computing, mostrando como a arquitetura de computadores está sendo influenciada por esta tecnologia e apontando os principais desafios desta já consolidada área de pesquisa.

O restante deste trabalho está organizado da seguinte forma. Na Seção 2, são apresentados os principais conceitos de cloud computing: a arquitetura e o modelo de negócios adotados e as características mais relevantes. Na Seção 3, é traçado um paralelo com duas das tecnologias mais intimamente relacionadas: grid e cluster computing. Um exemplo de um chip criado com o intuito atender aos requisitos de um ambiente de cloud computing é dado na Seção 4. Na Seção 5, são comentados os desafios existentes em um ambiente de cloud computing e, por fim, na Seção 6, são expostas as conclusões obtidas.

2. CLOUD COMPUTING

O Instituto Nacional de Padrões e Tecnologia dos Estados Unidos (NIST, National Institute of Standards and Technology) adota a definição abaixo para cloud computing [12]. Esta definição é fruto de um esforço feito para a padronização da definição de cloud computing, que somente tem sido concretizada recentemente.

Cloud computing é um modelo para prover acesso ubíquo, conveniente e sob demanda, via rede, a um conjunto compartilhado de recursos computacionais configuráveis (e.g., redes, servidores, armazenamento, aplicações e serviços) que podem ser alocados e desalocados de maneira rápida e com

mínimo esforço de gerenciamento ou interação do provedor do serviço.

A principal ideia que fundamenta cloud computing não é nova. Como dito antes, nos anos 60, já se anteviu que facilidades computacionais seriam disponibilizadas para o público em geral como uma utilidade. No entanto, foi depois de o CEO da Google, Eric Schmidt, em 2006, usar a palavra cloud para descrever o modelo de negócios de prover serviços pela Internet que o termo cloud computing passou a realmente ganhar popularidade.

O termo cloud computing, ao contrário de outros termos técnicos, representa, então, não uma nova tecnologia, mas sim um novo modelo operacional que reúne um conjunto de tecnologias já existentes para conduzir negócios de uma maneira diferente. É por este fato que existem diferentes percepções do que seja cloud computing, o que torna a padronização da sua definição uma tarefa difícil.

Nas subseções seguintes, são descritas as principais características e a arquitetura e o modelo de negócios adotados em cloud computing.

2.1 Arquitetura

Zhang *et al.* [18] explicam que, de uma forma genérica, a arquitetura de um ambiente de cloud computing pode ser dividida em quatro camadas: a camada de hardware, a camada de infraestrutura, a camada de plataforma e a camada de aplicação. Essas camadas são descritas a seguir. A Figura 1, adaptada de [18], em inglês, ilustra a arquitetura citada.

Camada de hardware: Nesta camada, é feito o gerenciamento dos recursos físicos da cloud, como servidores, roteadores, switches e sistemas de energia e resfriamento. A camada de hardware é comumente implementada em um data center e questões tipicamente tratadas são: configuração de hardware, tolerância a falhas, gerenciamento de tráfego e gerenciamento de energia e resfriamento de equipamentos.

Camada de infraestrutura: Também conhecida como camada de virtualização, nesta camada é criado um conjunto de recursos computacionais por meio do particionamento dos recursos físicos usando tecnologias de virtualização como Xen, KVM, e VMWare. A camada de infraestrutura é fundamental para o ambiente de cloud computing, pois é a virtualização empregada nessa camada que permite a alocação dinâmica dos recursos da cloud.

Camada de plataforma: Esta camada está logicamente sobre a camada de infraestrutura e consiste em sistemas operacionais e frameworks de aplicações. O propósito da camada de plataformas é facilitar a implamantação de aplicações em máquinas virtuais (VMs, Virtual Machine).

Camada de aplicação: Esta camada está no nível mais alto da hierarquia e consiste nas aplicações de cloud computing propriamente ditas.

Compara a arquiteturas de ambientes de hospedagem de serviços como server farms, a arquitetura de uma cloud computing é mais modular. Cada camada está fracamente ligada a outra, sendo o esquema parecido com o modelo OSI para protocolos de rede. Esta arquitetura modular propicia o suporte a uma grande variedade de requisitos de aplicações ao mesmo tempo que facilita o gerenciamento e a manutenção.

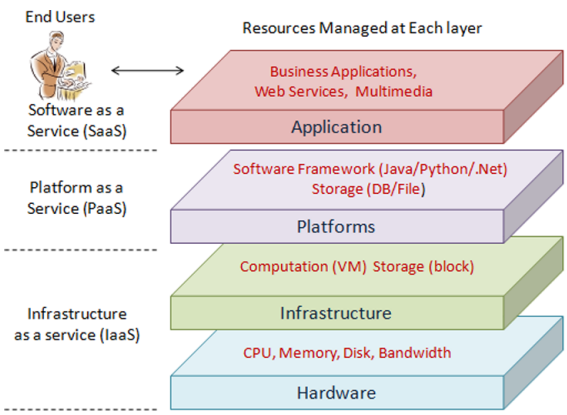

Figura 1: Arquitetura de um ambiente de cloud computing (adaptação de [18])

2.2 Modelo de negócios

Cloud computing emprega um modelo de negócios orientado a serviços. Assim, os recursos de uma cloud são disponibilizados como serviços e sob demanda. Conceitualmente, cada camada da arquitetura de uma cloud pode ser implementada como um serviço para a camada acima. Por outro lado, cada camada pode ser vista como um cliente da camada abaixo. No entanto, na prática, os serviços oferecidos em uma cloud podem ser agrupados em três categorias: Software como um serviço (SaaS), Plataforma como um serviço (PaaS) e Infraestrutura como um serviço (IaaS). Essas categorias estão representadas na Figura 1 e são detalhadas abaixo.

Infraestrutura como um serviço: IaaS consiste na disponibilização de recursos infraestruturais sob demanda, comumente em termos de VMs.

Plataforma como um serviço: PaaS consiste na disponibilização de recursos da camada de plataforma, como frameworks de desenvolvimento de software.

Software como um serviço: SaaS consiste na disponibilização sob demanda de aplicações através da Internet.

Como provedores de IaaS e PaaS são geralmente partes de uma mesma empresa, estes são denominados um só provedor: provedor de infraestrutura. Na Figura 2, retirada de [18], em inglês, é esboçado o modelo de negócios adotado em uma cloud.

2.3 Características

Cloud computing apresenta várias características notáveis que a diferenciam de outras tecnologias. Entre essas, Zhang *et al.* [18] citam as características abaixo como as principais.

Multi-tenancy: Em um ambiente de cloud computing, serviços de provedores distintos estão localizados em um mesmo

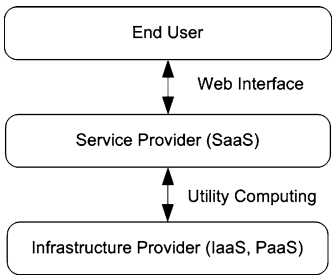

Figura 2: Modelo de negócios adotado em uma cloud (figura retirada de [18])

data center. Assim, questões sobre desempenho e manutenção desses serviços são compartilhadas entre os provedores dos serviços juntamente com o provedor de infraestrutura.

Conjunto de recursos compartilhados: O provedor de infraestrutura disponibiliza um conjunto de recursos computacionais que podem ser alocados dinamicamente a consumidores distintos. A capacidade de realizar essa alocação dinâmica propicia bastante flexibilidade ao provedor de infraestrutura para gerenciar o uso dos recursos e os custos operacionais. Por exemplo, a técnica de migração de VM pode ser empregada para maximizar a utilização de recursos e minimizar custos com energia e resfriamento de equipamentos (esse é um dos desafios citados na Seção 5).

Geodistribuição e acesso ubíquo: Clouds são, em geral, acessíveis por meio da Internet. Assim, é comumente possível acessar, em qualquer lugar, a partir de qual qualquer dispositivo com conexão à Internet, os serviços de uma cloud. Além disso, para atinigir um alto desempenho, muitas clouds consistem, atualmente, em vários data centers localizados em locais distintos do planeta. Com esta geodistribuição, um provedor de serviços pode alcançar a máxima utilização dos serviços.

Orientação a serviços: Cloud computing adota um modelo operacional orientado a serviços. Então, há um ênfase no gerenciamento de serviços. Em uma cloud, os provedores disponibilizam seus serviços respeitando um Acordo a Nível de Serviços, negociado com os seus clientes.

Alocação dinâmica de recursos: Uma das principais características de cloud computing é que os recursos computacionais podem ser alocados e desalocados *on-the-fly*. Essa característica permite que provedores de serviços tenham um menor custo operacional. Isto, porque, ao contrário do que ocorre com outras tecnologias onde serviços são disponibilizados segundo uma demanda de pico, provedores podem adquirir, no ambiente de cloud computing, recursos de acordo com a demanda corrente.

Auto-organizável: Como, em uma cloud, o gerenciamento dos recursos é automatizado e os recursos podem ser alocados e desalocados sob demanda, os provedores de serviços conseguem responder rapidamente a mudanças bruscas na demanda de serviços.

Preço baseada em utilização: A implementação pode variar de um serviço para outro, mas, em uma cloud, é adotado o modelo de cobrança *pay-per-use*. Este modelo de cobrança propicia um menor custo operacional de um serviço, já que a cobrança de um cliente é baseada no uso do serviço.

3. TECNOLOGIAS RELACIONADAS

Cloud computing é um fenômeno recente e por esse fato grande confusão é gerada sobre sua relação com outras tecnologias, como grid computing e cluster computing. Nesta seção, é dada uma visão geral sobre grid e cluster e uma breve comparação com cloud computing.

Um cluster é uma coleção de máquinas distribuídas ou paralelas conectadas entre si trabalhando juntos na execução de tarefas que precisam de computação intensiva. Clusters fornecem alta disponibilidade e balanceamento de carga. Se comparado com cloud, clusters são voltados a tarefas científicas enquanto cloud são setadas para aplicações de negócios. Cloud possui características como fraco acoplamento, elasticidade (devido a virtualização), computação sob demanda, facilidade de utilização, não presentes em cluster. Por outro lado, tecnologias em cluster são padronizadas e possuem grande maturidade [13].

Todavia, o maior ponto de confusão está relacionado à semelhança com a tecnologia de grid computing, que possui várias características em comum com cloud. Grid envolve a integração, gestão de recursos computacionais fracamente acoplados, heterogêneos, e geograficamente distribuídos. Ian Foster fornece uma descrição amplamente adotada presente em [4], na qual, tecnologias em grid são compostas por:

- Recursos coordenados que não estão sujeitos a controles centralizados;
- Padrões abertos de interfaces e protocolos de propósito geral; e
- Qualidade de serviço não triviais.

Se comparado a sistemas em grid, cloud apresenta consideráveis vantagens. Em resumo, tecnologias em cloud são fáceis de utilizar, permitindo rápido desenvolvimento e necessitando pouca customização. Por outro lado, sistemas em grid precisam adaptar aplicações com camadas adicionais em seu ambiente ("gridified"). Cloud geralmente incorpora virtualização, o que gera grandes vantagens como migração e escalabilidade em nível de hardware. Grid comumente utiliza o princípio no qual uma única instância serve a múltiplos clientes (*multi-tenancy*). Em cloud, uma relativa economia de gastos na infraestrutura e suporte é observada, devido principalmente à centralização dos recursos, economia esta nem sempre alcançada em sistemas em grid. Qualidade de serviço (QoS) é outra inerente vantagem de cloud, pois sistemas em grid apresentam apenas um melhor esforço [17, 3].

A Tabela 1 resume as principais diferenças entre cloud, grid e cluster [13].

Os gráficos da Figura 3 exibem a popularidade para os termos "Cloud Computing", "Grid Computing" e "Cluster Com-

Tabela 1: Cloud, Grid e Cluster

Características	Cloud	Grid	Cluster
Usabilidade	Sim	Parcial	Não
Virtualização	Sim	Parcial	Parcial
Padronização	Não	Sim	Sim
Multi-tenancy	Sim	Sim	Não
Self-service	Sim	Sim	Não
Escalabilidade	Sim	Parcial	Não
Interoperabilidade	Parcial	Sim	Sim
Segurança	Não	Parcial	Sim
Computação	Sob Demanda	Alta	Alta

puting" mensurada a partir do Google Search Trends entre 2004 e 2012. Os gráficos são construídos baseado na média de tráfego no volume de indexação em pesquisas pelo Google Search e o número de vezes em que estes termos aparecem em informações na ferramenta de notícias *Google News*. Pode-se observar, como exposto por [2], que cluster foi um termo emergente durante a década de 90, em seguida, a partir do ano 2000, grid se tornou popular e, a partir de 2007, cloud se transformou na "palavra da moda" (*buzz word*).

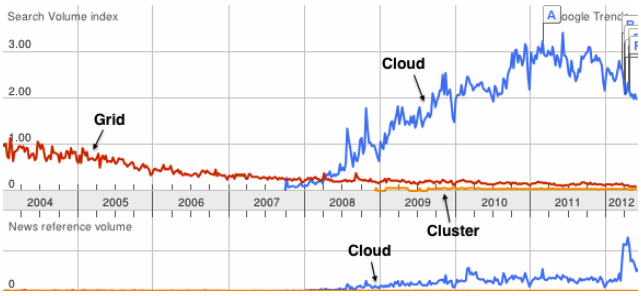

Figura 3: Cloud, Grid e Cluster(Fonte:*Google Trends*)

Na imagem acima, coloca-se em evidência também que tecnologias em cloud não são um substituto imediato aos sistemas de grid ou cluster, o que pode ser observado pelo declínio normal da tecnologia de grid, e pela constante no termo cluster. Geralmente tecnologias como grid, cluster e virtualização, podem fazer parte da estrutura de uma cloud.

4. CLOUD COMPUTING E ARQUITETURA DE COMPUTADORES

Os requisitos dos sistemas de cloud computing de alta escalabilidade, flexibilidade, alto desempenho, execução de aplicações em paralelo, mantendo sob controle o consumo de potência e energia, têm criado desafios para a área de arquitetura de computadores. Tendo em vista facilitar a pesquisa e o desenvolvimento de novas soluções de arquitetura que atendam a estes requisitos, a Intel Labs criou um chip protótipo chamado Single-Chip Cloud Computer (48-core SCC) para ser utilizado pela comunidade científica para explorar futuras arquiteturas de mais de 100 cores num mesmo chip, e as maneiras de como conectá-los e como criar programas para estas arquiteturas conforme descrito em [6, 8, 11].

4.1 Arquitetura geral do Intel SCC

Este chip, cuja arquitetura de topo é mostrada na Figura 4 e é construído em tecnologia 45nm high K CMOS com aproximadamente 1,3 bilhões de transistores e área de silício de 567mm^2, possui 48 Pentium (P54C execução em ordem) divididos em 24 bancos, de 2 cores cada, conectados por uma rede interna (NOC, Network On Chip) de topologia em malha 2D retangular 6×4. Em cada nó da rede há um roteador de 5 portas compartilhado entre os 2 cores. Possui ainda 4 controladores de memória *double data rate* tipo 3 (DDR3) conectados na periferia da rede permitindo acesso a memória de sistema de até 64GB e um bloco de interface de sistema que se conecta a um dispositivo FPGA externo que realiza a conversão do protocolo de rede para os protocolos PCI-e e Ethernet MACs para realizar acessos de I/O a outros dispositivos do sistema.

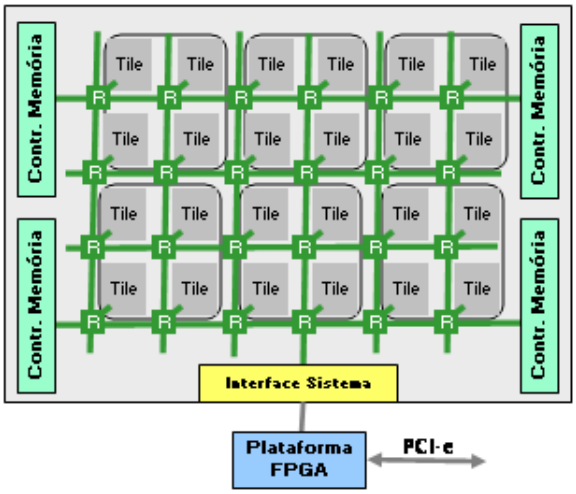

Figura 4: Arquitetura Intel SCC -24 bancos de processadores conectados em rede malha 2D

4.2 Arquitetura do banco de processadores

Cada um dos 24 bancos possui 2 cores, uma memória de passagem de mensagem (MPB, Message Passing Buffer) SRAM de 16KB e uma unidade interface de rede unificada (MIU, Mesh Interface Unity), que se conecta a uma das portas do roteador da rede. Cada core possui 16KB de cache L1 para instrução e 16KB para dados e 256KB de cache L2 unificada como mostrado na Figura 5.

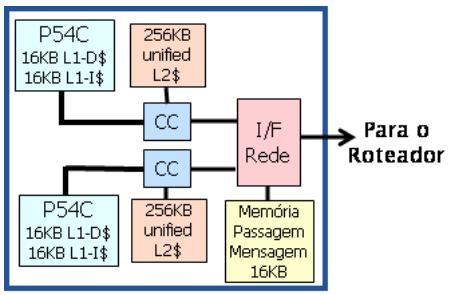

Figura 5: Banco de processador Intel SCC - 2 processadores P54C

Foram adicionadas funções específicas para estudar duas

questões críticas quando se aumenta significativamente o número de cores em um mesmo chip.

A primeira questão de acordo com [8], diz respeito a manter caches coerentes por mecanismos de hardware. O ganho computacional atingido pelo aumento do número de cores pode ser perdido pelo overhead do protocolo necessário para manter as caches coerentes. A arquitetura do SCC prevê que a consistência dos dados compartilhados pelos processadores deve ser mantida por software, e foram adicionados recursos em hardware para facilitar este controle como um novo tipo de memória, MPB, para permitir a comunicação em rede dos processadores por troca de mensagens e registradores tipo *test-&-set* cuja operações são atômicas suportando a semântica de lock necessária para garantir acessos mutuamente exclusivos à áreas de memória compartilhada.

A segunda questão, de acordo com [11], diz respeito a eficiência de consumo de energia. Para explorar técnicas de ajuste dinâmico de tensão e frequência, foram adicionados controladores de reguladores de tensão para controle de 8 domínios de tensão independentes: um domínio para rede, um para os circuitos de interface e outros 6, um para cada 4 cores. Para cada domínio, pode-se ajustar a tensão de 0 a 1.3V em passos de 6.25mV em questão de milisegundos. Com isto, pode-se desligar completamente um conjunto de 4 cores, ou colocá-los em modo de espera de baixo consumo a 0.7V preservando o conteúdo dos registradores e memória cache. Células de isolação e conversão de nível de tensão são utilizadas para integrar os diferentes domínios de tensão. Além disto, foram adicionados reguladores de frequência para controle de 28 domínios de frequência independentes: um para cada core, um para a rede, um para os controladores de memória DDR3, um para a interface do sistema e um para os controladores de reguladores de tensão. Filas determinísticas do tipo FIFO são utilizadas para sincronizar os diferentes domínios de frequência. A frequência máxima para os cores é de 1GHz a 1.1.V e para a rede é de 2GHz e pode ser ajustada para valores menores por divisores inteiros que podem ir de 1 a 16 em questão de nanosegundos. Através dos ajustes de tensão e frequência, o processador pode consumir de 25W a 125W.

4.3 Hierarquia de memória

Cada core possui 16KB de cache L1 para instrução e 16KB para dados e 256KB de cache L2 unificada.

Cada um dos 24 bancos possui ainda 16KB de memória interna para passagem de mensagem, totalizando 384KB para este tipo de memória interna ao chip. Qualquer um dos 48 processadores pode acessar qualquer um dos 24 blocos de 16KB de memória distribuídos pelo chip.

Cada controlador de memória DDR3 pode ser conectado a até dois módulos de memória tipo dual-inline de 8GB sendo possível atingir 64GB de memória de sistema. Como cada processador possui apenas 32 bits de endereço, permitindo acesso somente a 4GB de memória, para permitir que os 48 processadores utilizem os 64GB de memória de sistema disponíveis, este espaço de 32 bits de endereçamento é dividido em 256 páginas de 16MB e uma tabela de configuração de página mapeia o endereço de cada página de 16MB em um endereço de 34 bits e indica qual o controlador DDR3 deve ser utilizado pela página indicada.

Além disto, existe um bit na tabela de configuração de cada página para indicar se a página pode ser espelhada na cache e um bit para indicar se a página é do tipo de memória de passagem de mensagem.

A MIU realiza o controle de acesso à memória externa e a MPB. No caso de acesso à memória externa ela captura misses de acesso na cache L1 e na cache L2 e realiza a conversão de endereço de 32 bits para 34 bits e redireciona o acesso ao controlador de memória externa. No caso de acesso à MPB quando ocorrer miss na cache L1, o controlador de cache cancela o acesso a cache L2 e envia o acesso à MIU que redireciona o acesso ao bloco de MPB do próprio banco, ou redireciona o acesso ao roteador para realizar o acesso a uma MPB existente em outro banco.

Com isto é possível configurar três tipo de memória para uso no SCC:

Memória DRAM externa exclusiva associada a cada core: A página é mapeada na tabela de configuração para uma região exclusiva no espaço de memória externa podendo ser acessada por apenas um core e com permissão de espelho em cache, correspondendo a uma memória convencional de um computador.

Memória DRAM externa compartilhada: A página é mapeada na tabela de configuração de todos os cores que acessam esta página para o mesmo espaço de memória externa, e a permissão de espelho em cache é desabilitada, sendo que a coerência dos dados deve ser mantida por software com o auxílio dos registradores de *test-&-set* para garantir acessos mutuamente exclusivos.

Memória SRAM interna compartilhada (MPB): A página é marcada como MPB através do bit apropriado em todos os cores que devem trocar mensagens entre si através desta memória. Neste caso os acessos são redirecionados à memória SRAM interna em vez de acessar a memória DRAM externa.

4.4 Comunicação por passagem de mensagem

A MPB funciona como uma memória compartilhada com coerência mantida por software, porém com a vantagem de acesso muito mais rápido pois é uma memória SRAM interna ao chip. Quando uma mensagem deve ser enviada de um programa que executa num core à um programa que executa em um outro core, o programa que executa no core que envia a mensagem deve primeiro invalidar o conteúdo de todas as linhas da cache L1 correspondentes a MPB através da instrução INVDMB, uma nova instrução adicionada ao conjunto de instruções para este propósito, e então escrever no endereço correspondente na MPB. Isto evita que o dado seja escrito somente na cache ficando incoerente com o conteúdo da memória. O programa que receber a mensagem deve também utilizar esta instrução antes de realizar a leitura no endereço de memória compartilhada garantindo que o dado mais atualizado escrito na MPB será lido.

4.5 Arquiteturas alternativas

A arquitetura do Intel SCC é classificada como "light-weight" many-cores, pois possui muitos processadores não comple-

xos, de execução em ordem e sem técnicas sofisticadas de pre-fetch ou especulação. Esta arquitetura foi comparada em [15] com outras arquiteturas que podem ser utilizadas em servidores de cloud computing como:

- Arquitetura "heavy-weight" multi-cores representada pelo processador Intel Core i7 Nehalem com 4 cores, hyperthreading de até 8 threads, frequência de 2.8 GHz e cada processador com 32KB de cache L1 de instrução, 32 KB de cachê L1 de dados, 256KB de cache L2 unificada e 8M de cache L3 compartilhada entre os quatro processadores.

- Arquitetura de baixo consumo representada pelo processador Intel Atom D525 com 2 cores, hyperthreading de até 4 threads, frequência de 1.8GHz, cache 512KB, sem reordenamento de instruções, execução especulativa ou renomeação de registros.

- Arquitetura altamente paralela semelhante ao SIMD (Single Instruction Multiple Data) do processadores GPGPUs (General-Purpose Graphics Processing Units) representada pelo processador Nvidia GT218 com 16 CUDA cores e freqüência de 475MHz.

Os resultados do estudo comparativo mostraram que apesar de nenhuma das arquiteturas citadas ser superior em todos os aspectos analisados, a arquitetura do Intel SCC parece ser uma alternativa que fornece um bom resultado de compromisso entre desempenho e consumo, sendo mais veloz que os processadores de baixo consumo, consumindo menos que os processadores heavy-weight, tendo maior compatibilidade com programas existentes e maior facilidade de programação que os processadores GPGPUs.

Os processadores heavy-weight apresentaram o melhor desempenho em programas irregulares, porém são os que têm o consumo mais elevado. Os processadores de baixo consumo possuem o menor desempenho e nem sempre o menor consumo e os processadores GPGPUs possuem o melhor desempenho para programas regulares que apresentam alto nível de paralelismo, sem um consumo de energia elevado, porém tem a desvantagem de apresentar maior dificuldade de programação e não compatibilidade com programas já existentes no mercado, requerendo que muitas das aplicações sejam reescritas.

Outros resultados de desempenho e consumo de energia para o SCC podem ser encontrados em [7, 14, 1, 9].

Estudos referentes a desenvolvimento de software para realizar a comunicação entre os processadores utilizando os mecanismos definidos pelo padrão MPI (Message Passing Interface) e para facilitar a programação paralela de processadores podem ser encontradas em [10, 5, 16].

5. DESAFIOS

Cloud computing é um paradigma recente, ainda em sua "infância" e com muitos desafios em aberto. Nesta seção listamos alguns destes desafios [13, 18].

Padronização: Como cada organização utiliza diferentes APIs e protocolos, a integração e a interoperabilidade de todos os serviços e aplicações é um grande desafio.

Alocação dinâmica de serviços: O objetivo de um provedor de serviço é alocar e desalocar recursos a partir da cloud, minimizando seu custo operacional. Todavia, não é trivial o mapeamento desse objetivo em termos de requisitos de QoS, CPU e memória.

Migração de máquinas virtuais: Virtualização permite a migração de maquinas virtuais e o balanceamento de carga entre data centers. Assim o seu principal benefício é evitar pontos carregados (*hotspots*), o que não é uma tarefa simples pois falta agilidade para responder a rápidas modificações da carga de trabalho.

Consolidação de servidores: Migração de máquinas virtuais é frequentemente utilizado para consolidar multiplas máquinas virtuais a partir de servidores subutilizados em um único, onde o restante dos servidores pode minimizar seu consumo de energia. A formulação do problema em questão é NP-Difícil. Adicionalmente, a consolidação não pode ferir o desempenho da aplicação e dependências entre máquinas virtuais como por exemplo, requisitos de comunicação.

Gerência de energia (green computing): O objetivo não é apenas cortar os custos de energia em data centers, mas também manter regulamentações governamentais e ambientais. O desafio chave é alcançar um bom equilíbrio entre desempenho da aplicação e redução no consumo de energia.

Análise e gerência de tráfego: Conhecimento sobre o tráfego através da rede é importante a fim de tomar decisões sobre gestão e planejamento. Nesse sentido, existem diversos desafios na medição de trafego e métodos de análise na Internet e para empresas aplicarem ao data center.

Mecanismos de segurança de informações: Uma vez que o provedor de serviços não possui acesso ao sistema de segurança físico, ele depende do provedor de infraestrutura para fornecer completa segurança dos dados. Objetivos como confiabilidade e auditabilidade são fundamentais em provedores de infrastrutura. Adicionalmente, o ambiente dinâmicos (a partir da migração de máquinas virtuais) disponíveis em cloud torna a tarefa ainda mais complexa.

Frameworks de software: Ao mitigar o gargalo de acesso aos recursos, o tempo de execução das aplicações pode ser reduzido significamente. Os desafios incluem modelagem de desempenho de *frameworks* tais como Hadoop para escalabilidade e tolerância a falhas.

Tecnologias para armazenamento e gerência de informações: Provedores tipicamente utilizam diferentes sistemas de arquivos (como GFS e HDFS). Estes sistemas de arquivos são diferentes dos sistema de arquivos distribuídos tradicionais em termos de estrutura de armazenamento, padrões de acesso e da API. Nesse contexto, questões de compatibilidade são importantes uma vez que elas não implementam o padrão POSIX.

Novas arquiteturas para cloud: Hoje, implementações de clouds comerciais enfrentam limitações como altos custos em energia elétrica e alto investimento inicial. Nesse contexto, data centers de portes menores podem trazer vantagens. Adicionalmente, a característica de geodiversidade é frequentemente desejável para por exemplo, reduzir o tempo de resposta de serviços críticos.

6. CONCLUSÃO

Cloud computing surgiu recentemente como um novo e atraente paradigma para gerenciamento e disponibilização de serviços através da Internet. Este novo paradigma segue mudando de maneira rápida o universo da tecnologia da informação, tornando a antiga promessa de utiliy computing uma realidade.

No entanto, junto com os inúmeros benefícios, cloud computing traz também muitos desafios. Um dos desafios importantes é o de gerenciamento de energia, conectado ao conceito de sustentabilidade, tão evidente nos dias de hoje. É evidente, também, que as implementações correntes de cloud ainda não concretizam todo o potencial da tecnologia.

Neste trabalho, foi feito um apanhado dos pricipais conceitos e desafios relacionados a cloud computing, mostrando como esta tecnologia está afetando a arquitetura de computadores. Foi apresentado um chip, idealizado pela Intel Labs, criado com o intuito de atender aos requisitos de um ambiente de cloud computing. Além dessa, outras alternativas foram comentadas e, com isso, conclui-se, que cada vez mais será possível presenciar o impacto de cloud computing nas mais variadas áreas de IT, entre outras.

7. REFERÊNCIAS

[1] A. Bartolini, M. Sadri, J. Furst, A. Coskun, and L. Benini. Quantifying the impact of frequency scaling on the energy efficiency of the single-chip cloud computer. In *Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 181 –186, march 2012.

[2] R. Buyya, C. S. Yeo, and S. Venugopal. "*Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities*". In *High Performance Computing and Communications. HPCC '08. 10th IEEE International Conference on*, pages 5 –13, sept. 2008.

[3] T. Dillon, C. Wu, and E. Chang. "*Cloud Computing: Issues and Challenges*". In *Advanced Information Networking and Applications (AINA), 24th IEEE International Conference on*, pages 27 –33, 2010.

[4] I. Foster. "*What is the Grid? A Three Point Checklist*". *Argonne National Laboratory & University of Chicago*, July 20 2002.

[5] D. Gö andhringer, M. Hü andbner, L. Hugot-Derville, and J. Becker. Message passing interface support for the runtime adaptive multi-processor system-on-chip rampsoc. In *Embedded Computer Systems (SAMOS), International Conference on*, pages 357 –364, july 2010.

[6] M. Gries, U. Hoffmann, M. Konow, and M. Riepen. Scc: A flexible architecture for many-core platform research. *Computing in Science and Engg.*, 13(6):79–83, Nov. 2011.

[7] P. Gschwandtner, T. Fahringer, and R. Prodan. Performance analysis and benchmarking of the intel scc. In *Cluster Computing (CLUSTER), IEEE International Conference on*, pages 139–149, sept. 2011.

[8] J. Howard, S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom, F. Pailet, S. Jain, T. Jacob, S. Yada, S. Marella, P. Salihundam, V. Erraguntla, M. Konow, M. Riepen, G. Droege, J. Lindemann, M. Gries, T. Apel, K. Henriss, T. Lund-Larsen, S. Steibl, S. Borkar, V. De, R. Van Der Wijngaart, and T. Mattson. A 48-core ia-32 message-passing processor with dvfs in 45nm cmos. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), IEEE International*, pages 108 –109, feb. 2010.

[9] M. Korch, T. Rauber, and C. Scholtes. Memory-intensive applications on a many-core processor. In *High Performance Computing and Communications (HPCC), IEEE 13th International Conference on*, pages 126–134, sept. 2011.

[10] J. Lee, J. Kim, J. Kim, S. Seo, and J. Lee. An opencl framework for homogeneous manycores with no hardware cache coherence. In *Parallel Architectures and Compilation Techniques (PACT), 2011 International Conference on*, pages 56 –67, oct. 2011.

[11] T. Mattson, R. Van der Wijngaart, M. Riepen, T. Lehnig, P. Brett, W. Haas, P. Kennedy, J. Howard, S. Vangal, N. Borkar, G. Ruhl, and S. Dighe. The 48-core scc processor: the programmer's view. In *High Performance Computing, Networking, Storage and Analysis (SC), International Conference for*, pages 1–11, nov. 2010.

[12] P. M. Mell and T. Grance. Sp 800-145. the nist definition of cloud computing. Technical report, Gaithersburg, MD, United States, 2011.

[13] N. Sadashiv and S. Kumar. "*Cluster, grid and cloud computing: A detailed comparison*". In *Computer Science Education (ICCSE), 6th International Conference on*, pages 477 –482, aug. 2011.

[14] S. Sha, J. Zhou, C. Liu, and G. Quan. Power and energy analysis on intel single-chip cloud computer system. In *Southeastcon, Proceedings of IEEE*, pages 1 –6, march 2012.

[15] E. Totoni, B. Behzad, S. Ghike, and J. Torrellas. Comparing the power and performance of intel's scc to state-of-the-art cpus and gpus. In *Performance Analysis of Systems and Software (ISPASS), IEEE International Symposium on*, pages 78–87, april 2012.

[16] R. F. van der Wijngaart, T. G. Mattson, and W. Haas. Light-weight communications on intel's single-chip cloud computer processor. *SIGOPS Oper. Syst. Rev.*, 45(1):73–83, Feb. 2011.

[17] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner. "*A break in the Clouds: Towards a Cloud Definition*". *SIGCOMM Comput. Commun. Rev.*, 39:50–55, December 2008.

[18] Q. Zhang, L. Cheng, and R. Boutaba. "*Cloud computing: state-of-the-art and research challenges*". *Journal of Internet Services and Applications*, 1(1):7–18, May 2010.