

Proposta para Grupo de Trabalho

GT-Computação em Nuvem para Ciência:
Armazenamento de Dados

Roberto Samarone dos Santos Araujo

Agosto/2011

1 Título

GT-Computação em Nuvem para Ciência: Armazenamento de Dados.

2 Coordenador

Coordenador	Roberto Samarone dos Santos Araujo, Dr.
Curriculum Lattes	http://lattes.cnpq.br/6894507054383644
Email	rsa@ufpa.br
Instituição	<i>Universidade Federal do Pará</i> (UFPA) Instituto de Ciências Exatas e Naturais Faculdade de Computação
Endereço	Campus Universitário do Guamá Rua Augusto Corrêa, n.1 - Campus Básico Belém/Pará - CEP: 66075-110

3 Resumo

A computação em nuvem vem se tornando cada vez mais popular e atrativa. Um dos motivos é a facilidade de acesso a serviços como o armazenamento de dados. No entanto, serviços em nuvem são em geral disponibilizados por empresas e muitas vezes possuem um custo financeiro associado. Esta proposta visa realizar um estudo sobre os conceitos, metodologias e ferramentas de armazenamento em nuvem de baixo custo disponíveis e que se adequem aos serviços e a infraestrutura da RNP. Além disto, o projeto propõem-se a implantar um protótipo deste serviço no ambiente da RNP. Este serviço seria disponibilizado a professores e pesquisadores em caráter experimental e temporário.

4 Abstract

Cloud computing is becoming very popular and attractive. One of the reasons is that it is very easy to access services as data storage. However, cloud services are in general made available by companies and usually have a cost. This proposal aims at performing a study about the concepts, the methodologies, and the low cost cloud storage tools available that could be used in the RNP infrastructure. In addition, the project aims at installing a cloud storage service in the RNP environment. This service would be available to professors and researchers in an temporary and experimental environment.

5 Parcerias

O GT terá como parceiro o Laboratório de Segurança (LabSEC) da Universidade Federal de Santa Catarina. Este laboratório tem como coordenador o professor e pesquisador

Dr. Ricardo Felipe Custódio.

6 Duração do Projeto

O projeto terá duração de 12 meses.

7 Sumário Executivo

O avanço da tecnologia tem contribuído para a geração e disseminação de um volume de dados cada vez maior. Para comportar este volume, dispositivos com grande capacidade de armazenamento como discos rígidos são cada vez mais comuns e acessíveis. Embora os recursos atuais possibilitem o armazenamento de grandes quantidades de dados, a disponibilidade destes dados é um requisito cada vez mais presente em nosso cotidiano. Se antes era suficiente armazenar dados e acessá-los em um único computador, hoje é cada vez mais comum o acesso a dados de qualquer lugar e utilizando variados tipos de computadores (e.g. notebooks, celulares, tablets).

Um dos recursos muito empregados atualmente para realizar o armazenamento de dados com alta disponibilidade é a computação em nuvem. A ideia desta tecnologia é o compartilhamento de recursos computacionais através de redes de alta velocidade para prover serviços como armazenamento de dados. A nuvem se dá pelo desconhecimento dos usuários dos detalhes da infraestrutura que fornece o serviço; usuários simplesmente solicitam o serviço e a nuvem se encarrega de atender suas solicitações. Desta forma, por exemplo, ao invés de armazenar um arquivo em seu computador, um indivíduo poderia enviar o arquivo para a nuvem e esta se encarregaria de armazená-lo em um de seus servidores. Esta tecnologia, no entanto, possibilita outros tipos de serviços como o processamento de dados e o uso programas. O armazenamento em nuvem é o serviço de computação em nuvem que trata do armazenamento de dados.

Os benefícios da computação em nuvem, em particular o armazenamento em nuvem, são inúmeros quando comparados a métodos tradicionais. Ela é altamente escalável possibilitando o aumento da capacidade de armazenamento se necessário. Usuários não estão mais restritos aos arquivos armazenados em seus computadores. Com os arquivos armazenados na "nuvem", eles podem acessá-los de qualquer computador conectado a rede (e.g. Internet). Qualquer falha nos dispositivos de armazenamento local dos usuários não resultarão em comprometimento dos arquivos armazenados na "nuvem". Além disso, arquivos podem ser compartilhados com outros usuários. A infraestrutura que provê o armazenamento em nuvem também possui recursos de tolerância a falhas. Isto possibilita melhores garantias do armazenamento e disponibilidade aos arquivos.

Devido a suas vantagens inerentes, o armazenamento em nuvem tem atraído um número cada vez maior de empresas a oferecerem este serviço. Dentre estas empresas, destacam-se: a Amazon que oferece o Amazon Simple Storage Service (Amazon S3) (1), o Dropbox (2), o Cloudsafe (3) e o Google Docs(4) que oferece armazenamento e aplicativos. Estes serviços, no entanto, diferem-se entre si pela capacidade de armazenamento, limite e transferência de dados, segurança, tamanho máximo de arquivos e protocolo de

acesso aos dados. Além de empresas provedoras de serviços em nuvem, existem ferramentas para instalação de nuvens privadas como a solução da Oracle (5) e projetos de código aberto como o Nimbus (6) e o Open Eucalyptus(7).

Apesar das vantagens do armazenamento em nuvem e da disponibilidade deste serviço por empresas, a segurança deste serviço é um ponto que merece atenção. Ao enviar arquivos para a nuvem, estes são armazenados em servidores de terceiros. Estes servidores são responsáveis não somente pelo acesso aos arquivos posteriormente, mas também pela segurança destes. No entanto, grande parte das soluções disponíveis no mercado não fornece recursos de segurança adequados para proteção destes arquivos. A segurança de soluções como o Dropbox tem sido muito contestada (8), por exemplo. Em muitas soluções, os arquivos são armazenados em texto plano, sem a proteção de mecanismos criptográficos. Em outras, os arquivos são criptografados pelos próprios servidores que os armazenam. Ambos os casos possibilitam aos servidores o livre acesso aos dados dos usuários. Além disso, existem outros problemas inerentes a estas soluções como autenticação para acesso a arquivos. Em um ambiente ideal, os arquivos devem ser criptografados no computador do usuário antes do envio para os computadores que formam a nuvem.

A tecnologia de armazenamento em nuvem pode trazer grandes benefícios ao meio acadêmico. Através dela, professores e pesquisadores poderiam armazenar seus projetos, artigos, etc. em uma plataforma segura, eficiente e altamente disponível. Dessa forma, estes usuários poderiam ter acesso a seus arquivos de qualquer computador conectado a Internet, poderiam utilizar a nuvem para realizar *backup* de seus dados ou até mesmo compartilhar dados com outros usuários. A fim de possibilitar o acesso a nuvem no meio acadêmico com segurança, eficiência e disponibilidade, este projeto propõem-se a realizar um estudo comparativo entre as ferramentas e tecnologias existentes que permitam criar uma plataforma de computação de armazenamento em nuvem de baixo custo. Adicionalmente, o projeto propõem-se a implantação de um protótipo de serviço de armazenamento em nuvem como prova do conceito e da ferramenta que possa ser utilizada por professores e pesquisadores de instituições de pesquisa e ensino.

O projeto objetiva o armazenamento de baixo custo de dados em nuvens dentro do ambiente de rede da RNP. Dessa forma, ele é voltado para soluções de nuvens que ofereçam infraestrutura como serviço (IaaS) para armazenamento privado. A fim de prover o ambiente de nuvens dentro da RNP, será realizado primeiramente uma comparação entre as ferramentas comerciais e opensource. Dentre as ferramentas de código aberto utilizadas nesta comparação estão o Syncany (9), o OpenNebula (10, 11), Eucalyptus (7, 12) e o Nimbus (6). O Nimbus, em particular, possui um módulo específico para armazenamento em nuvens, o Cumulus (13). A comparação levará em conta critérios como segurança, capacidade de armazenamento, limite de transferência de dados e critérios como os apresentados em (14). Esta comparação visa definir as ferramentas que mais se adequam ao ambiente da RNP.

Após a comparação das soluções, será definido a ferramenta mais adequada e iniciada implantação do protótipo de serviço de armazenamento em nuvem. O protótipo será instalado em servidores dispostos na UFSC e na UFPA. Esses servidores formarão uma nuvem de armazenamento dentro da estrutura da RNP. A nuvem possibilitará aos demais participantes da rede o armazenamento de dados com alta disponibilidade e segurança.

A Figura 1 ilustra o armazenamento em nuvem dentro da estrutura de redes da RNP.

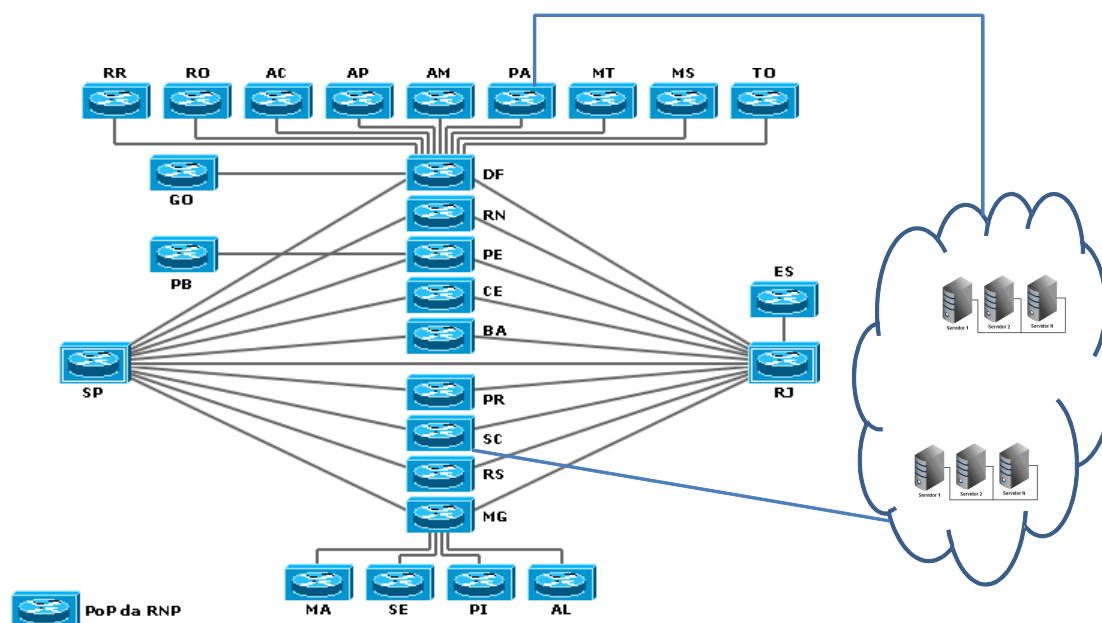


Figura 1: Armazenamento em nuvem dentro da estrutura de redes da RNP. Os servidores do protótipo encontram-se na UFSC e na UFPA. Eles formam a nuvem que fornece o serviço de armazenamento aos professores e pesquisadores dentro da estrutura da RNP.

Nossa proposta está relacionada a GTs anteriores como o GT-Digital Preservation – Preservação Digital com Armazenamento Distribuído e o GT Armazenamento em rede. O primeiro objetiva a preservação de conteúdo digital a longo prazo por meio de um sistema de armazenamento distribuído baseado no conceito de redes Peer-to-Peer. O segundo, visa implantar uma infraestrutura de armazenamento temporário de dados com interfaces tanto para usuários (via browser) quanto para aplicações (via API). Nossa proposta, no entanto, objetiva o armazenamento de dados utilizando a tecnologia de computação em nuvens.

A seguir é apresentado um resumo dos objetivos deste projeto:

- Realizar um estudo comparativo entre as ferramentas de armazenamento em nuvem;
- Realizar testes nas ferramentas disponíveis;
- Instalar um protótipo de serviço de armazenamento em nuvens;
- Analisar o desempenho do protótipo utilizando como plataforma computacional o Windows, o Linux e o Mac OS, e dispositivos móveis como o Android e o iOS da Apple;
- Realizar testes de segurança no protótipo.

A primeira etapa deste projeto é voltada para comparação de ferramentas e a implantação do protótipo de serviço de armazenamento em nuvem. No entanto, o projeto pode ganhar futuramente novas etapas com a sua integração com a Comunidade Acadêmica Federada (CAFe) e com a Infraestrutura de Chaves Públicas para Ensino e Pesquisa (ICPEdu). Além disso, o projeto pode ser estendido para disponibilização de outros recursos de computação em nuvem como máquinas virtuais.

8 Ambiente para o Teste do Protótipo

Os testes no protótipo ocorrerão em duas fases. Na primeira fase serão realizados testes entre as duas instituições (UFPA/UFSC). Nesta fase será realizado testes preliminares do protótipo. Na segunda fase o protótipo será replicado para máquinas de serviços disponíveis nos POPs (Pontos de Presença) da RNP e/ou dentro do ambiente do PlanetLab (<http://www.rnp.br/pd/planetlab/>), onde será analisado critérios como escalabilidade, disponibilidade, segurança, integração, conectividade entre outros.

Referências

- 1 AMAZON Simple Storage Service. Agosto 2011. Disponível em: <http://aws.amazon.com/s3/>.
- 2 DROPBOX. Agosto 2011. Disponível em: <http://www.dropbox.com/>.
- 3 CLOUDSAFE. Agosto 2011. Disponível em: <https://secure.cloudsafe.com/login/>.
- 4 GOOGLE Docs. Agosto 2011. Disponível em: <https://docs.google.com>.
- 5 ORACLE Cloud CComputing. Agosto 2011. Disponível em: <http://www.oracle.com/us/technologies/cloud/index.html>.
- 6 NIMBUS Project. Agosto 2011. Disponível em: <http://www.nimbusproject.org/>.
- 7 EUCALYPTUS - The Open Source Cloud Platform. Agosto 2011. Disponível em: <http://open.eucalyptus.com/>.
- 8 CENTER PcWorld Business. *Dropbox: Insecure by Design*. Agosto 2011. Disponível em: http://www.pcworld.com/businesscenter/article/224857/dropbox_insecure_by_design.html.
- 9 SYNCANY. Agosto 2011. Disponível em: <http://www.syncany.org/>.
- 10 THE Open Source Toolkit for Cloud Computing. Agosto 2011. Disponível em: <http://sparkleshare.org/>.
- 11 SOTOMAYOR, B. et al. Virtual infrastructure management in private and hybrid clouds. *IEEE Internet Computing*, IEEE Computer Society, Los Alamitos, CA, USA, v. 13, p. 14–22, 2009. ISSN 1089-7801.

- 12 NURMI, D. et al. The eucalyptus open-source cloud-computing system. In: *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*. Washington, DC, USA: IEEE Computer Society, 2009. (CCGRID '09), p. 124–131. ISBN 978-0-7695-3622-4. Disponível em: <<http://dx.doi.org/10.1109/CCGRID.2009.93>>.
- 13 BRESNAHAN, J. et al. Cumulus: an open source storage cloud for science. In: *Proceedings of the 2nd international workshop on Scientific cloud computing*. New York, NY, USA: ACM, 2011. (ScienceCloud '11), p. 25–32. ISBN 978-1-4503-0699-7. Disponível em: <<http://doi.acm.org/10.1145/1996109.1996115>>.
- 14 SEMPOLINSKI, P.; THAIN, D. A comparison and critique of eucalyptus, opennebula and nimbus. In: *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science*. Washington, DC, USA: IEEE Computer Society, 2010. (CLOUDCOM '10), p. 417–426. ISBN 978-0-7695-4302-4. Disponível em: <<http://dx.doi.org/10.1109/CloudCom.2010.42>>.