Proposta para Grupo de Trabalho

GT-PID: Plataforma IaaS Distribuída para e-ciência

Luís Henrique M. K. Costa

09-08-2013

1. Título

GT-PID: Plataforma IaaS Distribuída para e-ciência

2. Coordenador

Coordenador: Luís Henrique Maciel Kosmalski Costa
Instituição: Universidade Federal do Rio de Janeiro
Escola Politécnica / DEL - COPPE / Programa de Engenharia Elétrica
Grupo de Teleinformática e Automação (GTA)
Lattes: http://lattes.cnpq.br/4193860852876287
Contato: E-mail: luish@gta.ufrj.br Tel.: +21 2562-8619 Fax: +21 2562-8627

Coordenador Adjunto: Miguel Elias Mitre Campista
Instituição: Universidade Federal do Rio de Janeiro
Escola Politécnica / DEL - COPPE / Programa de Engenharia Elétrica

Grupo de Teleinformática e Automação (GTA)
Lattes: http://lattes.cnpq.br/4256483085616956
Contato: E-mail: miguel@gta.ufrj.br Tel.: +21 2562-8635 Fax: +21 2562-8627

Coordenador Adjunto: Marcelo Gonçalves Rubinstein
Instituição: Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências, Faculdade de Engenharia
Departamento de Eletrônica e Telecomunicações
Lattes: http://lattes.cnpq.br/2787725227134746
Contato: E-mail: rubi@uerj.br Tel.: +21 2334-2165

3. Resumo

O objetivo do GT-PID é a implementação de uma plataforma colaborativa que permita a laboratórios de pesquisa brasileiros a ampliação de suas capacidades de processamento. Combinando o backbone de comunicação da RNP com uma camada de virtualização em nuvem usando o modelo de infraestrutura como serviço (IaaS), a ideia é que as instituições de pesquisa servidas pela RNP possam utilizar os recursos computacionais de todas as instituições em conjunto. O modelo possibilita a uma instituição específica ter acesso a muito mais recursos do que seria viável isoladamente, ao mesmo tempo em que reduz a ociosidade dos recursos computacionais globalmente. Diferente de um modelo de nuvem computacional comercial, os pesquisadores brasileiros terão acesso aos recursos de forma gratuita.

4. Abstract

The goal of GT-PID is to implement a collaborative platform to allow Brazilian research labs to enlarge their processing power. Combining RNP's communication backbone and a virtualization layer for cloud computing using the Infrastructure as a Service (IaaS), the idea is to foster the research institutes served by RNP to share their computational resources as a whole. The model permits a specific institution to have access to much more resources than it would be possible alone and, at the same time, reducing the idleness of the computational resources at a global scale. Unlike the commercial model of cloud computation, Brazilian researchers will have free access to all resources.

5. Parcerias

Parceria Nacional

Marcelo Gonçalves Rubinstein
E-mail: rubi@uerj.br
Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências, Faculdade de Engenharia
Departamento de Eletrônica e Telecomunicações
Rua São Francisco Xavier, 524 – Maracanã
CEP 20550013 - Rio de Janeiro, RJ - Brasil
Telefone: (21) 23342165
http://www.lee.eng.uerj.br/~rubi

Parceria Internacional

Stefano Secci
Email: stefano.secci@upmc.fr
Université Pierre et Marie Curie - LIP6/CNRS
Boite Courriel 169, Bureau 318:25-26
4, place Jussieu
75252 PARIS Cedex 05 – França
Tel.: +33 (0)1 4427 3678 Fax:
+33 (0)1 4427 8783
http://www-phare.lip6.fr/~secci/

6. Duração do projeto

A duração deste projeto é de 12 meses. O cronograma físico seguirá o Edital do Programa de Grupos de Trabalho da RNP 2013-2014.

7. Sumário executivo

Há um consenso entre pesquisadores trabalhando em instituições de pesquisa de que muitas vezes os recursos computacionais de um determinado laboratório permanecem ociosos por longos períodos, enquanto em outros momentos esses recursos não são suficientes. Por exemplo, na proximidade do prazo final de submissão de uma conferência, muitos usuários do mesmo laboratório tendem a compartilhar os recursos computacionais disponíveis, muitas vezes de forma pouco coordenada. Tendo isso em vista, a utilização do poder de processamento das máquinas ocorre, de alguma forma, em rajadas, o que dificulta o planejamento prévio dos recursos que pode ser feito tanto pelo pico quanto pela média em um determinado período. Partindo dessa premissa, seria de extrema utilidade uma plataforma que desse suporte à utilização de recursos computacionais de forma compartilhada. Dessa forma, o planejamento levaria a uma utilização mais eficiente, já que evitaria tanto períodos longos de ociosidade quanto períodos de saturação. *A ideia do projeto GT-PID é promover o compartilhamento de recursos computacionais entre centros de pesquisa a partir do compartilhamento em nuvem de todos os recursos existentes.* Assim, durante períodos de ociosidade, os recursos estariam disponibilizados para outros laboratórios, enquanto que, em períodos de necessidade crítica, eles poderiam ser usados de maneira distribuída pelos diferentes usuários dos laboratórios participantes [1]. Por um lado, provê-se uma capacidade global do sistema superior à oferecida localmente. Por outro lado, diminui-se a ociosidade dos recursos computacionais, aumentando a eficiência e o retorno do investimento financeiro aportado na pesquisa.

A infraestrutura computacional será organizada em torno de um centro de dados (*data center* - DC) distribuído, que interligará todas as universidades brasileiras e estrangeiras participantes. No entanto, é importante ressaltar que, mais que um DC distribuído, o GT-PID visa construir uma plataforma computacional distribuída, agregando capacidade de processamento e de armazenamento. Este aspecto é importante, pois frequentemente em atividades de pesquisa os processos executados requerem alto poder de processamento e de armazenamento [2], tomese por exemplo a simulação de protocolos de redes sem-fio ou a análise de dados experimentais em física. Assim, diferentemente de muitas das arquiteturas de DC atuais, os nós da rede não podem ser equipamentos de baixo custo, mas máquinas com maior poder de processamento. No Brasil, as universidades e demais instituições de pesquisa poderão utilizar a infraestrutura existente da rede Ipê (http://www.rnp.br/ipe/) para se interligarem, aproveitando um serviço já prestado pela RNP. Já a infraestrutura física da plataforma computacional distribuída será formada pelos recursos computacionais cedidos voluntariamente por cada laboratório participante. Sendo assim, um determinado laboratório deverá disponibilizar uma quantidade mínima de recursos computacionais e, em troca, poderá executar suas simulações na plataforma distribuída.

O principal papel da plataforma computacional distribuída é permitir a execução de simulações e/ou experimentos de pesquisadores dos laboratórios envolvidos. Para tanto, a infraestrutura provida poderá ser utilizada para experimentos e simulações das mais diversas áreas de pesquisa e não somente da área de computação. Para isso, será utilizado o conceito de máquinas virtuais (MVs) que serão oferecidas a cada solicitação de utilização da infraestrutura [3, 4]. As MVs serão disponibilizadas em um conjunto responsável por executar a simulação desejada. Assim, o pesquisador possuirá acesso em nível de administrador a essas MVs e poderá instalar as ferramentas necessárias para sua simulação ou experimento. O pesquisador, então, possuirá total flexibilidade na escolha de suas ferramentas de simulação. Ao término da simulação, os resultados poderão ser coletados e as máquinas virtuais criadas poderão ser simplesmente removidas da plataforma [5].

Outro fator importante a ser ressaltado é a transparência da localização das MVs para os usuários, caracterizando um serviço em nuvem [6, 7]. Pretende-se criar uma interface que liste os recursos disponíveis e as máquinas para que os usuários sejam capazes de escolher o poder de processamento desejado e as máquinas onde querem executar os experimentos. Essa interface será capaz ainda de bloquear acessos indevidos, permitir o início dos experimentos, criar máquinas virtuais personalizadas e retornar os resultados. Caso haja alguma alteração ao longo de um experimento, o usuário poderá ainda migrar as máquinas virtuais para outras máquinas físicas que tenham recursos disponíveis. Assim, a arquitetura da plataforma computacional é basicamente formada por quatro elementos: recursos físicos distribuídos e virtualizados, servidor de máquinas virtuais, interface gráfica e usuário, conforme mostrado na Figura 1. Note que os usuários acessam a plataforma através de uma interface gráfica, que acessa um servidor de máquinas virtuais responsável por todas as ações administrativas da plataforma proposta. Naturalmente, o acesso ao sistema deve ser autenticado para garantir a segurança e controle do acesso.

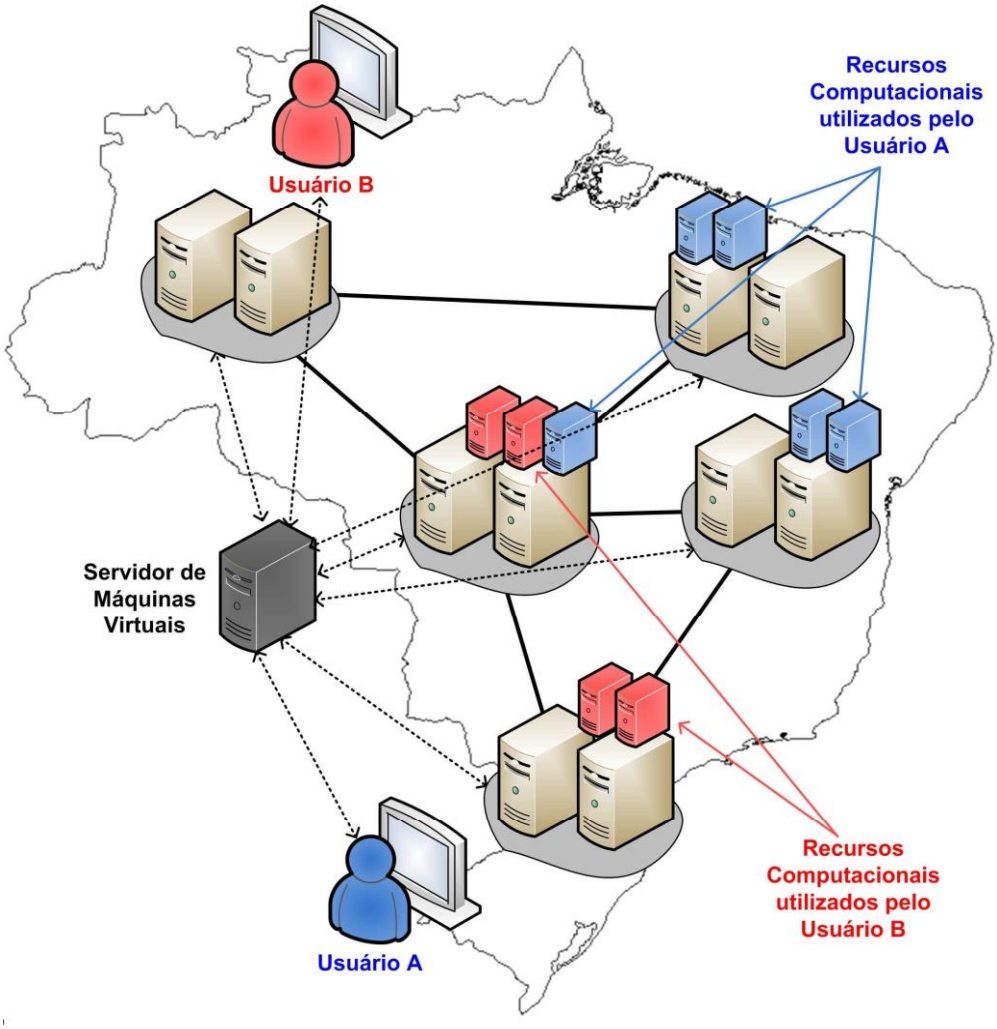

Figura 1: Arquitetura do DC distribuído e do acesso aos recursos computacionais compartilhados.

De maneira geral, a plataforma computacional distribuída requer o desenvolvimento técnico de três componentes importantes: a autenticação dos usuários, o gerenciamento de recursos computacionais e o endereçamento IP. Esses componentes estão detalhados a seguir:

- **Autenticação**: O acesso à infraestrutura será permitido apenas aos laboratórios que possuem recursos disponibilizados no DC distribuído, de forma a incentivar a colaboração das instituições de pesquisa. Assim, a alocação de máquinas virtuais (MVs) para um determinado usuário será precedida por uma etapa de autenticação. O sistema de autenticação deverá então verificar se a credencial fornecida está associada a algum laboratório doador.

- **Gerenciamento de recursos computacionais**: No contexto da infraestrutura proposta, cada servidor fornecido por um laboratório será considerado como um nó físico do DC distribuído. A cada requisição de utilização da plataforma, um mecanismo de alocação de recursos será responsável por distribuir as MVs solicitadas entre os diferentes nós físicos. Para realizar a alocação, o mecanismo deverá levar em conta a classe do serviço exigida pelo usuário das MVs. Para a primeira fase do projeto, serão definidas duas classes de serviços. A primeira, denominada “Alta Confiabilidade”, prioriza a distribuição geográfica das MVs por diversos nós físicos. Assim, caso haja uma pane em um determinado nó físico, nem todos os nós estariam comprometidos, o que evita maiores prejuízos a um grande processo de simulação, por exemplo. A segunda classe, denominada “Baixa Latência”, atua em sentido contrário à primeira classe, priorizando a latência de rede. Assim, o mecanismo de alocação priorizará alocar todas as MVs em um mesmo nó físico. Se não for possível, o mecanismo realiza a alocação entre nós físicos que possuam baixa latência entre si. Além da alocação de recursos, um mecanismo de migração das MVs será disponibilizado. Esse

mecanismo terá atuação local e global. A local permite que um administrador de um nó físico migre as MVs desse nó para outros nós do DC distribuído. Dentre as razões para a migração, pode ser citada uma manutenção programada no laboratório hospedeiro. A atuação global será realizada apenas pelo administrador do DC distribuído e será executada periodicamente a fim de melhorar a latência ou confiabilidade das simulações em execução, dependendo da classe de serviço solicitada.

- **Endereçamento IP:** De forma a possibilitar a migração de máquinas virtuais entre nós físicos, a infraestrutura possuirá roteadores executando o protocolo LISP *(Location ID Separation Protocol*) [8]. Esse protocolo permite a separação entre a identificação e localização de uma MV, permitindo que ela mantenha seu identificador mesmo após sofrer migração. O roteador LISP pode ser um PC comum, executando por exemplo a implementação de código aberto OpenLISP (http://www.openlisp.org/). O LISP não exige modificações de endereçamento nas MVs e permite que máquinas não-LISP da Internet acessem as MVs.

Diferentemente de iniciativas anteriores, a plataforma computacional em nuvem do GT-PID se concentra na execução de processos científicos personalizados, tais como simulações ou cálculos matemáticos complexos. Não se trata de uma plataforma de experimentação de e avaliação de redes avançadas, como o PlanetLab (http://www.rnp.br/pd/planetlab/), FITS (http://www.gta.ufrj.br/fits/) ou Fibre (http://www.fibre-ict.eu/). Além disso, essa estrutura será nacional, gratuita e de software livre, voltada para a comunidade científica brasileira, diferentemente de um serviço de computação em nuvem comercial, tal como a Amazon Cloud ou UOL Cloud.

7.1 Serviço Provido pela RNP

O serviço provido pela RNP através da plataforma IaaS distribuída para e-ciência consiste no acesso a um parque computacional científico de grande porte. A RNP irá orquestrar o acesso ao DC distribuído para processamento, oferecendo às instituições de ensino e pesquisa nacionais a possibilidade de executarem processos complexos. O aumento da capacidade de processamento é de suma importância para o processamento de grandes massas de dados. Mesmo adotando como critério para entrada das instituições a colaboração com recursos, a RNP proverá um serviço que trará como benefício a popularização e utilização eficiente de recursos computacionais que por muitas vezes são, ao mesmo tempo, ociosos em grandes centros e escassos em muitas instituições brasileiras.

7.2 Aplicação de Prova de Conceito

Os professores Luís Henrique Costa, Miguel Elias Campista e Marcelo Rubinstein contam com uma equipe com grande experiência na área de virtualização de máquinas e de redes, adquirida em projetos nacionais e internacionais que atuaram como participantes ou líderes, como por exemplo o projeto Horizon – Um Novo Horizonte para a Internet, financiado pela ANR na França e FINEP no Brasil (http://www.gta.ufrj.br/horizon/) e ReVir – Redes Virtuais na Internet do Futuro, financiando pelo CTIC/RNP (http://www.gta.ufrj.br/revir/). Esse conhecimento será utilizado na criação de um protótipo com poucos nós físicos, distribuídos inicialmente nas instituições parceiras do projeto, a UFRJ, a UERJ e posteriormente o LIP6 na França. Após os primeiros testes, testes de escala um pouco maior serão realizados em outras instituições de ensino e pesquisa servidas pela RNP, e com as quais já são mantidas relações de cooperação em projetos anteriores com o GTA, como por exemplo a UFF, a Unicamp, ou a UFRGS.

Além das instituições nacionais, haverá uma cooperação mais efetiva com um parceiro internacional, a Université Pierre et Marie Curie (UPMC), na França, que irá instalar nós do sistema distribuído como contrapartida ao projeto. É de interesse que haja a participação da UPMC já que, atualmente, já existe uma rede de testes experimental entre UFRJ e UPMC que executa o LISP, um dos protocolos a serem utilizados na plataforma computacional distribuída. Em resumo, o software

de virtualização será o Xen, com máquinas virtuais Linux (a distribuição do Linux pode ser qualquer). O servidor de máquinas virtuais será a evolução de um servidor já implementado em trabalhos anteriores da equipe [5]. Esse servidor poderá oferecer diversas distribuições de Linux para que o usuário possa escolher a que melhor lhe convier em termos de facilidade de administração e instalação de ferramentas necessárias para os experimentos.

O software necessário para o gerenciamento de recursos será desenvolvido em software livre e terá como objetivo alocar máquinas virtuais nos nós físicos após cada requisição, segundo os requisitos de cada simulação. A alocação será baseada na informação da localização geográfica dos nós físicos e estatísticas da rede, calculadas a partir de medições periódicas de ferramentas como o Ping, Traceroute e o Iperf. Além disso, periodicamente o gerenciamento de recursos reposicionará as máquinas virtuais para realização de manutenção programada ou para melhoria de desempenho. O reposicionamento será calculado utilizando as mesmas informações de localização e estáticas citadas anteriormente, e será executado a partir do mecanismo de migração padrão do Xen.

7.3 Viabilidade Técnica

Atualmente existem plataformas de virtualização gratuitas e APIs para programação nesse tipo de ambiente. Além disso, o endereçamento IP deixa de ser problema já que existe o protocolo LISP que separa identificação de um nó de sua posição geográfica, facilitando a migração de MVs. Por exemplo, o LISP permite a migração de uma MV em uma máquina física local para outra máquina física instalada remotamente. Já a autenticação dos usuários pode ser realizada aproveitando os conhecimentos adquiridos durante a participação no projeto Eduroam-BR pelos membros da equipe do GTA/COPPE/UFRJ. Por fim, a implementação da interface gráfica não apresenta maiores obstáculos. A ideia é construir uma interface web simples, que liste todas as possibilidades de recursos disponíveis.

7.4 Produto Vislumbrado para o 2° Ano do GT-PID

Na eventual recondução do GT-PID por um segundo ano, o objetivo principal será a expansão do protótipo desenvolvido na primeira fase, através da inclusão de mais nós físicos pertencentes a diversos laboratórios e universidades. Para isso, serão propostas parcerias com equipes de pesquisa de diferentes áreas do conhecimento. De forma a incentivar a inclusão de recursos na plataforma, um mecanismo de créditos será desenvolvido, no qual a utilização permitida para cada instituição usuária será proporcional à quantidade de recursos doados pela mesma. Dessa forma, as máquinas virtuais serão monitoradas e possuirão seus recursos limitados. A limitação e monitoramento serão realizadas através de primitivas já disponíveis no Xen, como o gerenciamento da capacidade de CPU utilizada por uma máquina virtual [9]. O número maior de usuários e a existência do sistema de créditos implicarão na necessidade de um mecanismo de autenticação mais sofisticado. Por exemplo, cada usuário do sistema será relacionado a uma instituição doadora e sua capacidade de utilização da plataforma deverá ser proporcional à máxima permitida à sua instituição. Dessa forma, cada instituição poderá definir limites de utilização para seus usuários. Além disso, planeja-se utilizar o serviço da RNP CAFe (http://portal.rnp.br/web/servicos/cafe) para tornar a autenticação mais eficiente, permitindo que o usuário utilize na plataforma as mesmas credenciais já utilizadas em outros serviços de sua instituição.

8. Recursos financeiros

8.1 Equipamentos e softwares

Descrição	Quantidade	Valor unitário	Valor total
Estação Intel Core I7 3930K Torre ATX e fonte real 400W Coolermaster, placa-mãe Intel Intel DX79 (som e rede on-board), processador Intel Core I7 3930K, memória DDR3 1333 8Gb, vídeo GeForce GT 610, HD SATA 1Tb, gravador DVD-RW SATA, mouse e teclado Microsoft.	5	R$ 4,650.00	R$ 23,250.00
Monitor LCD 22 pol. (Ex. Monitor LG 22EA53T 21.5 Led IPS full HD 1920X1080 VGA/DVI)	3	R$ 585.00	R$ 1,755.00
TOTAL:		R$ 5,235.00	R$ 25,005.00

9. Ambiente para testes do protótipo

Para a implementação e primeiros testes do protótipo, serão utilizados os computadores adquiridos no GT, equipamentos de rede e o acesso à Internet disponível nos laboratórios. Para os testes em maior escala, poderão ser utilizadas máquinas de serviço máquinas de serviço disponíveis nos PoPs (Pontos de Presença) da RNP. Para a posterior implantação, as máquinas utilizadas serão as disponibilizadas pelas próprias instituições de ensino e pesquisa servidas pela RNP e participantes da plataforma PISE.

10. Referências

[1] Bari, M. F., Boutaba, R., Esteves, R., Granville, L. Z., Podlesny, M., Rabbani, M. G., Qi, Zhang, Zhani, M. F., "*Data Center Network Virtualization: A Survey,*", em IEEE Communications Surveys & Tutorials, vol.15, no.2, pp.909-928, segundo trimestre de 2013.

[2] Costa, L. H. M. K., Amorim, M. D., Campista, M. E. M., Rubinstein, M. G., Florissi, P., and Duarte, O. C. M. B., "*Grandes Massas de Dados na Nuvem: Desafios e Técnicas para Inovação*", em Minicursos do Simpósio Brasileiro de Redes de Computadores - SBRC'2012, Ouro Preto, MG, Brasil, maio de 2012.

[3] Khan, A., Zugenmaier, A., Jurca, D., Kellerer, W., "*Network virtualization: a hypervisor for the Internet?",* em IEEE Communications Magazine, vol.50, no.1, pp.136-143, janeiro de 2012.

[4] Duan, Q., Yan, Y., Vasilakos, A. V., "*A Survey on Service-Oriented Network Virtualization Toward Convergence of Networking and Cloud Computing*", em IEEE Transactions on Network and Service Management, vol.9, no.4, pp.373-392, dezembro de 2012.

[5] Alves, R. S., Campista, M. E. M., Costa, L. H. M. K., and Duarte, O.C. M. B., "*Towards a Pluralist Internet Using a Virtual Machine Server for Network Customization*", em Asian Internet Engineering Conference (AINTEC'2012), pp. 9-16, Bangkok, Thailand, novembro de 2012.

[6] Moreno-Vozmediano, R., Montero, R. S., Llorente, I. M., "*IaaS Cloud Architecture: From Virtualized Datacenters to Federated Cloud Infrastructures*", em Computer, vol.45, no.12, pp.65-72, dezembro de 2012.

[7] Nguyen, K.-K., Cheriet, M., Lemay, M., "*Enabling infrastructure as a service (IaaS) on IP networks: from distributed to virtualized control plane*", em IEEE Communications Magazine, vol.51, no.1, pp.136-144, janeiro de 2013.

[8] Saucez, D., Iannone, L., Bonaventure, O., Farinacci, D., "Designing a Deployable Internet: The Locator/Identifier Separation Protocol", IEEE Internet Computing, vol.16, no.6, pp.14-21, novembro-dezembro de 2012.

[9] Couto, R. S, Campista, M. E. M., Costa, L. H. M. K, "*XTC: A Throughput Control Mechanism for Xen-based Virtualized Software Routers*", em IEEE Global Communications Conference (GLOBECOM'2011), Houston, Texas, EUA, dezembro de 2011.