PADEX - Plataforma de Alto Desempenho EXpresso

Gustavo N. Dias[1], Alex S. Moura[1], Leandro N. Ciuffo[1], Iara Machado[1], Michael A. Stanton[1], Fabio Okamura[1], Eduardo Grizendi[1], Helmann S. Penze[1], Fernando Redigolo[2], Dino Magri[2]
1 Rede Nacional de Ensino e Pesquisa (RNP)
Rua Lauro Muller, 116 - Sala 1103 - 22.290-906 - Rio de Janeiro - RJ - Brazil
{gustavo.dias, alex.moura, leandro.ciuffo, iara.machado, michael.stanton, fabio.okamura, eduardo.grizendi, helmann.penze}@rnp.br
2 LARC - Laboratório de Arquitetura e Redes de Computadores - (PCS-EPUSP)
Universidade de São Paulo - Av. Prof. Luciano Gualberto, travessa 3, n.158, CEP 05508-010
{fernando, dino}@larc.usp.br

Resumo: O compartilhamento de recursos entre diferentes instituições de pesquisa foi uma das motivações para criação da internet. Interconexões de rede de alta capacidade entre organizações de grandes instrumentos científicos a centros de supercomputação são uma tendência que vem sendo chamadas de superinstalações. Este artigo apresenta um serviço especializado da RNP que combina infraestruturas de rede dedicadas com hardware e software otimizados para executar grandes transferências de dados em redes com alta latência, oferecendo o máximo desempenho das tecnologias para apoiar e acelerar os fluxos de trabalho das pesquisas científicas nacionais. Os resultados demonstram que alta sinergia de tecnologias selecionadas oferecem grandes benefícios, ainda desconhecidos por muitos pesquisadores.

Palavras-chave: Computação de Alto Desempenho, Supercomputação, DMZ Científica, Ciber-infraestrutura, Superinstalações.

Abstract: The sharing of resources between research organizations was one of the motivations for the advent of the internet. High speed interconnections between facilities with large scientific instruments with supercomputing centers is a trend known as super-facility. This work presents a specialized super-connectivity service from RNP that combines dedicated network infrastructure with optimized hardware and software for large data transfers in high latency networks, offering maximum performance of the technologies in order to support and accelerate workflows of national scientific research. The results confirm that the high synergy of selected technologies offer major benefits, yet unknown for many researchers.

Keywords: High Performance Computing, Supercomputing, Science DMZ, Connectivity, Cyberinfrastructure, Super-facility.

1 Introdução

A expressão "Ciberinfraestrutura" inicialmente usada em 2002 pela "Comissão Atkins" da National Science Foundation (NSF) dos Estados Unidos, criada para responder à pergunta: como a NSF, principal agência de fomento de pesquisa básica norte-americana, poderia eliminar barreiras à evolução de processamento, uso de dispositivos e instrumentos especiais, armazenamento, comunicação avançada e uso de dados, tornando esse ambiente acessível a todos os cientistas, engenheiros, estudiosos e cidadãos desse país?1 Este termo passou a ser utilizado para tratar da evolução destas tecnologias de informação e comunicação (TIC) na geração de conhecimento científico - a Comissão Europeia também o utiliza, sob a denominação de "e-infraestrutura" [Atkins, 2003].

O uso planejado de recursos de TIC é o que permite aos grupos de pesquisa, laboratórios e instituições realizarem trabalhos em ciência e tecnologia fortemente suportados pelas TICs. Chama-se de "e-Ciência" as atividades científicas em vários campos do conhecimento, que dependem criticamente do uso de Ciberinfraestrutura (CI) [NSF, 2004].

A Figura 1 resume como os componentes da CI se relacionam, para o desenvolvimento da pesquisa e educação.

Figura 1 - Componentes de Ciberinfraestrutura2.

Os componentes da CI adequados para atender as demandas de uma grande comunidade de pesquisadores em ciência e tecnologia, tipicamente incluem:

Instrumentos geradores de dados (e.g.: observatórios, aceleradores, microscópios e laboratórios)

Recursos computacionais de grande capacidade, munidos com softwares de controle e aplicativos. Organizados nas seguintes categorias computacionais:

---

[1] Atkins et al., "Revolutionizing Science and Engineering Through Cyberinfrastructure", NSF, 2003, Obtido em http://www.nsf.gov/cise/sci/reports/atkins.pdf

[2] Adaptado de Cyberinfrastructure for Environmental Research and Education: Obtido em http://www.cyrdas.org/related.documents/reports/cyber_report_new.pdf http://www.ncar.ucar.edu/cyber/cyberreport.pdf (NSF)

Processamento de Alto Desempenho (PAD) - Supercomputadores, como o Santos Dumont do LNCC.
Computação de Alta Vazão (CAV) - tipicamente utilizando clusters e grades computacionais integrados em nuvem.
Armazenamento distribuído de grande capacidade, para guarda segura dos dados científicos em todos os passos da sua vida útil.
Visualização, que é a tradução de dados em imagens ou animações que facilitam a compreensão dos fenômenos sendo medidos ou modelados.
Gestão dos dados científicos, compreendendo a preservação e curadoria a longo prazo de coleções de dados científicos e seus metadados, bem como de outras informações, documentos e objetos digitais relacionados.
Conectividade global entre todos os componentes descritos anteriormente, provida por meio de redes de comunicação robustas, seguras, de capacidade adequada. Deve-se observar que estas são redes dedicadas às atividades de pesquisa e educação e usualmente interconectadas às de outros países, onde se encontram grupos e instituições parceiros de pesquisadores brasileiros.

Este artigo está relacionado a uma solução desenvolvida na linha de Processamento de Alto Desempenho (PAD).

2 Motivação

O Projeto PADEX surgiu de um acordo de colaboração firmado entre as lideranças da Rede Nacional de Ensino e Pesquisa (RNP), Laboratório Nacional de Computação Científica (LNCC), e Laboratório Nacional de Luz Sincrotron (LNLS), junto ao Ministério da Ciência, Tecnologia e Comunicações (MCTIC), para prestação de um serviço de supercomputação sob demanda para demandas de pesquisa de outras organizações.
A partir da identificação de uma demanda do Laboratório LNLS para aquisição de um supercomputador para suprir necessidades computacionais atuais e futuras (por exemplo, a nova fonte de luz sincrotron, Sirius, que atualmente está em construção e tem previsão de entrar em operação em 2018) os diretores da RNP, LNCC e CNPEM/LNLS acordaram com o MCTIC assegurar recursos para implantar uma infraestrutura que forneça ao LNLS um acesso de alto desempenho e qualidade à Plataforma de Alto Desempenho (PAD) do LNCC, o supercomputador conhecido como Santos Dumont (SD).
Nas áreas de pesquisa e educação, particularmente as apoiadas por recursos públicos, o compartilhamento e acesso a recursos e capacidades computacionais vem sendo uma das principais características da Internet, desde sua concepção e implantação, apesar de não ser uma característica frequentemente articulada e exposta de forma explícita. Descrever as bases desse compartilhamento é fundamentalmente importante para compreensão do Projeto PADEX (Plataforma de Alto Desempenho EXpresso).
O PADEX representa um modelo econômico efetivo de compartilhamento de recursos que busca o melhor aproveitamento dos investimentos públicos apoiados pelo MCTIC para suas Unidades de Pesquisa (UPs).

3 Visão do Serviço PADEX

O objetivo do Serviço PADEX é oferecer acesso otimizado através da rede acadêmica nacional à plataforma de processamento de alto desempenho (PAD) do LNCC, através do compartilhamento seguro de recursos do supercomputador Santos Dumont, para atender demandas de organizações cujas atividades de pesquisa requeiram processamento de alto desempenho (High Performance Computing, HPC) para execução de suas atividades científicas. O acesso otimizado é fornecido pela RNP através de seu backbone e por redes parceiras, entre a instituição de origem e o LNCC, sendo o LNLS a primeira organização atendida.

À primeira vista, a ampliação da capacidade da rede fim a fim entre LNCC e LNLS, inclusive com largura de banda garantida, parece ser suficiente para promover o acesso remoto em alta velocidade ao supercomputador do LNCC. Entretanto, a questão é mais complexa do que pode parecer, uma vez que há outros aspectos que não somente a velocidade do enlace de comunicação contribuem com o aumento do tempo de transferência de dados, como configurações de software ou mesmo características dos equipamentos envolvidos.

Frequentemente, instituições que dispõem de enlaces de alta capacidade só conseguem utilizar uma fração da capacidade total disponível. O problema de desempenho decorre do fato que os dispositivos e as arquiteturas da rede, bem como as configurações de equipamentos e servidores são previstas para fluxos tradicionais de redes corporativas, que suportam uma grande diversidade de aplicações, tais como aplicações Internet (e.g., web, e-mail, voz sobre IP, transferências peer-to-peer, transmissão de vídeo) e sistemas corporativos (e.g., softwares de gestão ERP, gerenciamento de matrículas, controle de ponto, circuitos de vigilância). Além disso, estas aplicações podem atualmente ser acessadas e utilizadas por diferentes tipos de dispositivos (e.g., desktops, notebooks, smartphones, tablets) com diferentes sistemas operacionais e diferentes formas de conectividade, escalonamento e interrupção (e.g., redes Ethernet de 100Mbps a 10 Gbps, redes Wi-Fi e/ou de telefonia celular), sem mencionar, na falta de determinismo dos comportamentos, até do mesmo controlador, memória, versão, layout e fornecedor de componentes de hardware dependendo do cenário. A variedade de opções, seja de aplicações ou de tipos de dispositivos e tipos de acesso, traduz-se em mecanismos de segurança complexos (firewalls, proxies, sistemas de detecção de intrusão, entre outros), capazes de lidar com tal diversidade de cenários e suportados por configurações genéricas, para atender grande diversidade de cenários. Por sua vez, as aplicações científicas constituem fluxos relativamente simples, essencialmente de transferência de grandes volumes de dados entre servidores com conexões de rede de alta velocidade [Magri 2014].

Para se alcançar a máxima performance, é fundamental que as aplicações científicas sejam tratadas de forma diferenciada na rede, com recursos, dispositivos e configurações otimizados para o seu uso, minimizando-se os fatores causadores de perdas de pacotes e redução de desempenho.

3.1 Demanda e fluxo de trabalho do LNLS

Conforme descrito pelo CNPEM, o LNLS possui demanda reprimida por não haver disponibilidade de computação adequada à sua necessidade de processamento. Em

especial, há dois tipos de demandas mapeadas: processamento em CPUs para projetos relacionados ao Sirius e Processamento em GPUs para uma estação de pesquisa (também chamado de linha de luz) de tomografia que trabalha com imagens em 3D.

A Figura 3 ilustra o fluxo de processamento dos dados gerados pelo LNLS.

Na rede local (LAN) do LNLS foram registrados picos de aproximadamente 1,7Gbps para a linha de luz de tomografia durante os experimentos e transferência de arquivos.

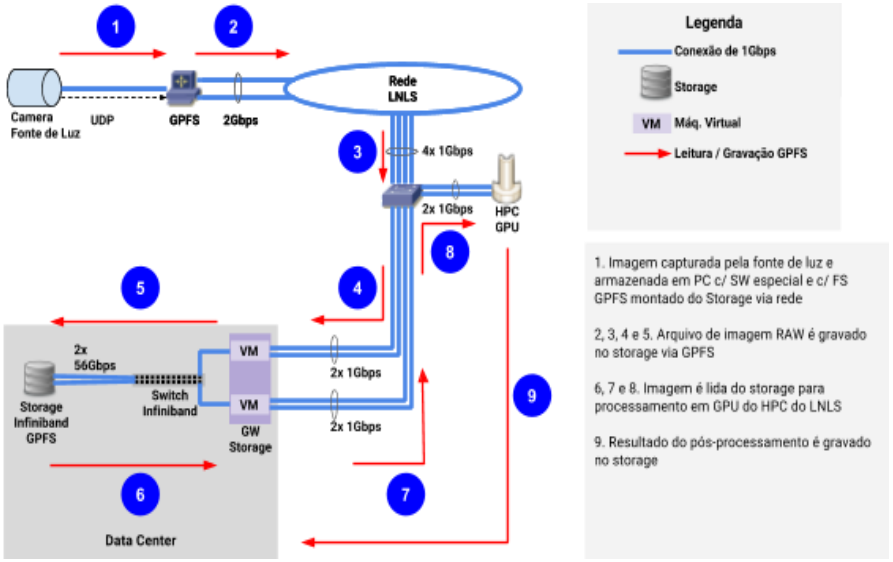

Figura 3 - Fluxo de processamento dos dados gerados pelo LNLS.

3.2 Fluxo de trabalho para processamento dos dados do LNLS via PADEX

Para a aplicação de processamento remoto foi proposto o seguinte fluxo de trabalho:

Importação dos dados brutos: os dados gerados pelo instrumento, que se encontram armazenados no Storage do LNLS, são copiados para o Storage do LNCC, através dos DTNs das instituições.
Processamento de dados: a) o pesquisador se conecta ao servidor de VPN do LNCC; b) O pesquisador efetua o login via SSH no nó apropriado do LNCC e dispara os comandos para o processamento local dos dados copiados em 1; c) O pesquisador verifica o resultado do processamento, verificando anecessidade de novos processamentos (e.g.:, uso de novos parâmetros de processamento, como diferentes filtros) e interagindo com o ambiente de processamento até a definição final do resultado.
Exportação dos dados processados: os dados processados são copiados de volta para o Storage do LNLS, através dos DTNs das instituições.

3.3 Componentes do Serviço PADEX

O diagrama da Figura 4 ilustra os principais componentes do Serviço PADEX:

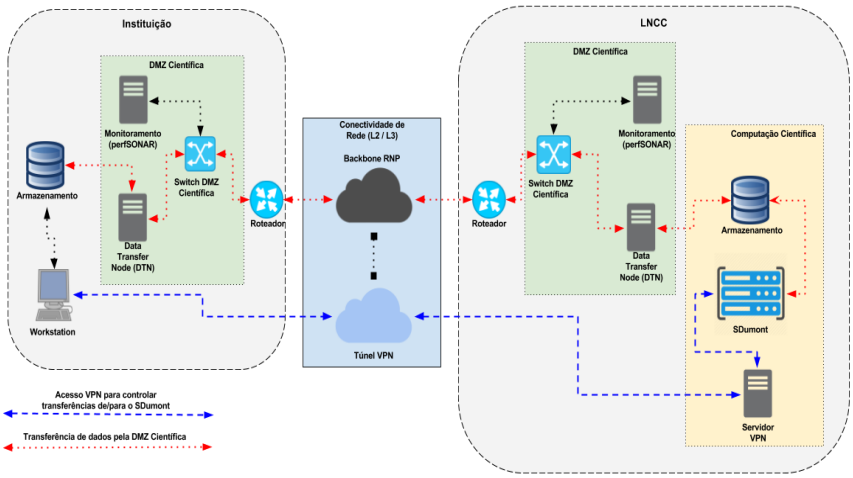

Figura 4 - Diagrama geral da solução PADEX.

3.3.1 DMZ Científica

O modelo de Zona Desmilitarizada Científica (DMZ Científica) sugere que a conectividade das aplicações científicas seja "desmilitarizada" através da implementação de uma segregação da rede de produção do campus. Em 2013 a RNP iniciou o projeto "Science DMZ", que teve como objetivo implantar o modelo de DMZ Científica em um conjunto de instituições selecionadas. O projeto especificou o seguinte conjunto mínimo de equipamentos necessários para a implantação do modelo:

Servidor de Transferência Otimizado - DTN (Data Transfer Node)
Switch especializado (com deep buffer)
Servidores de Monitoramento (originalmente duas máquinas: uma para monitoramento da largura de banda máxima alcançável e outra para o monitoramento da latência).

3.3.2 Serviço de Transporte e Vazão assegurada

Para a interconexão feita entre as DMZ Científicas instaladas no LNCC e LNLS foi configurado um canal dedicado dentro da infraestrutura de conectividade do backbone RNP. O canal passa pelas redes metropolitanas no Rio de Janeiro e em Campinas, que fornecem conexões de última milha sobre infraestrutura óptica de 10Gbps em anel. O canal é independente de outro circuito de 1Gb/s utilizado para tráfego de produção do LNCC, e foi implementado usando o serviço MPLS (Multi-Protocol Label Switching) do tipo Ethernet Virtual Private Line (EVPL), restrito a 2Gb/s no domínio do backbone Ipê, interligando os PoPs de São Paulo e Rio de Janeiro e encapsulando o tráfego

científico com vazão assegurada com Quality of Service (QoS), na classe de serviço Assured Forwarding (AF) e limitação de banda do provedor do LNCC de 1Gbps para 3Gbps com posterior secessão de 2Gb/s exclusivo ao tráfego científico do Supercomputador Santos Dumont.

4 Resultados

Para comprovação da eficácia da solução PADEX foram realizados alguns testes. Nos testes um conjunto de 41 arquivos foi utilizado para o download e upload o que juntos totalizaram 58Gbytes de volume de dados. O tamanho do conjunto de arquivos representa o caso médio esperado para transferências entre o LNLS e LNCC em cada execução do workflow.

A seguir mais detalhes acerca dos testes e seus respectivos resultados.

4.1 Teste de transferência 1 (baseline):

Upload e *download* de arquivos utilizando infraestrutura do LNLS e do LNCC antes das modificações realizadas pelo projeto PADEX, sem considerar melhorias de conectividade ou otimizações (Figura 5) para estabelecer a base para comparação.

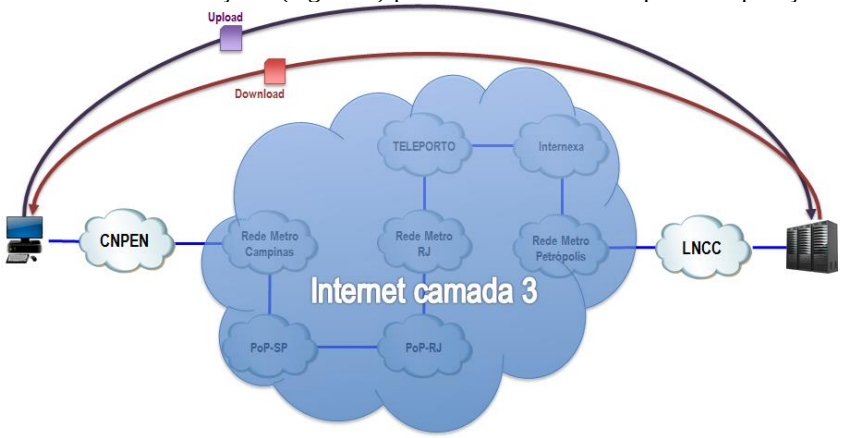

Figura 5 - *Uso do Supercomputador Santos Dumont por pesquisadores do CNPEM/LNLS*

O teste foi realizado no dia 7/11/2016. O melhor resultado alcançado foi uma taxa de vazão aproximada de 12 MBytes/s (96 Mbps); e a transferência do total de 58Gbytes de dados consumiu 83 minutos.

4.2 Teste de transferência 2 (com modelo SDMZ e VLA*N):*

Envio (*upload*) e recebimento (*download*) de arquivos em ambos sentidos, utilizando o servidor DTN conectado à conectado à rede "desmilitarizada" (DMZ) e a

rede dedicada camada 2 (Figura 6). O objetivo do teste é observar o desempenho com uso da VLAN fim a fim com garantia de banda (QoS) entre os servidores DTN.

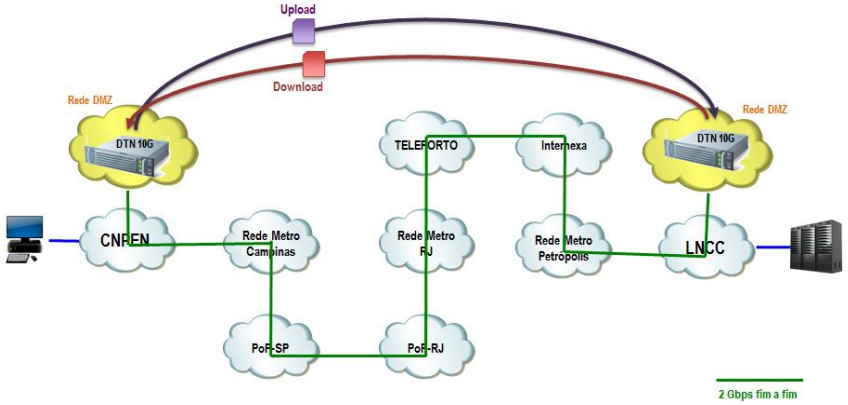

Figura 6 - *Uso do Supercomputador Santos Dumont após implantação do PADEX.*

O Gráfico 1 apresenta os tempos *download* e *upload* de 58Gbytes de dados durante 10 dias.

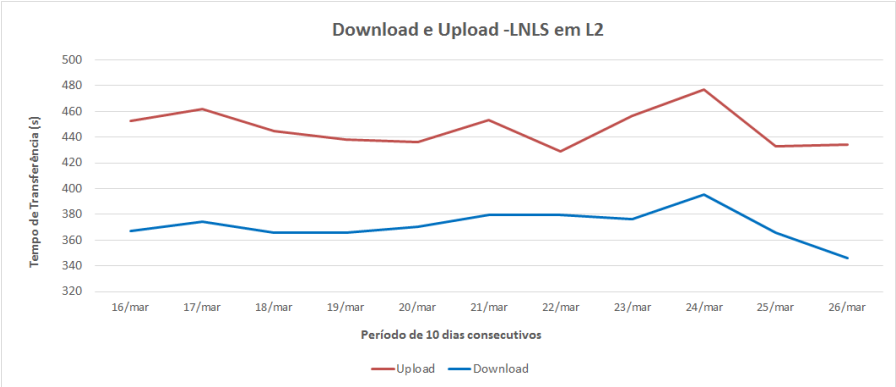

Gráfico 1 - Tempo para transferência de 58 Gbytes em período de 10 dias.

Com a solução do PADEX os pesquisadores do LNLS agora levam 7 minutos para transferência de 58Gbytes de dados, frente ao 83 minutos que levava no cenário anterior, um ganho excepcional de 1.186%.

4 Conclusão

Os resultados obtidos demonstraram que a combinação de soluções selecionadas e adequadamente otimizadas para transferências de grandes volumes de dados com conectividade de rede dedicada e o uso de arquitetura de hardware especializada para otimização de tarefas críticas, como leitura e escrita de dados em disco, traz benefícios valiosos ao suporte e apoio tecnológico a infraestruturas

disponíveis para supercomputação, como mostra esse estudo de caso do uso do Supercomputador Santos Dumont em pesquisas feitas pelo CNPEM/LNLS.

Mostra também que o compartilhamento destas soluções tecnológicas mesmo a quilômetros de distâncias traz excelentes resultados e viabilizam a pesquisa nacional que vive momentos de limitação orçamentária e cortes frequentes.

Referências

1. [Atkins, 2003] Atkins et al., "*Revolutionizing Science and Engineering Through Cyberinfrastructure*", NSF, 2003, Acesso em 17/04/2017 - http://www.nsf.gov/cise/sci/reports/atkins.pdf

2. [NSF, 2004] Adaptado de *Cyberinfrastructure for Environmental Research and Education*: Acesso em 17/09/2016 - http://www.cyrdas.org/related.documents/reports/cyber_report_new.pdf http://www.ncar.ucar.edu/cyber/cyberreport.pdf

3. [Magri, 2014] Magri et al. "*Science DMZ: Support for e-Science in Brazil*", 2014 IEEE 10th International Conference on e-Science, página 75-78, 2014.

4. [Parkinson, 2014] Parkinson DY, Beattie K, Chen X, Correa J, Dart E, Daurer BJ, Deslippe JR, Hexemer A, Krishnan H, MacDowell AA, Maia FR. "*Real-time data-intensive computing*". AIP Conference Proceedings 2016, (Vol. 1741, No. 1, p. 050001). AIP Publishing. Acesso em 29/04/2017. http://dx.doi.org/10.1063/1.4952921