Instituto de Física da Universidade de São Paulo

DEPARTAMENTO DE FÍSICA EXPERIMENTAL

Relatório Anual de Atividades
Programa PROCONTES
Processo Nº 03.1.27082.1.0

Rogério Luiz Iope

Nº USP 1712076

Dezembro 2007

Índice

1. Introdução

1.1. *e-Science*: Um novo modo de fazer ciência

As grandes colaborações em diversas áreas da ciência e da engenharia estão cada vez mais ganhando escala global, e por consequência tornando-se cada vez mais dependentes da tecnologia da informação. A ciência experimental não está mais limitada a laboratórios isolados, mas tem sido conduzida por dezenas ou mesmo centenas de laboratórios de pesquisa conectados entre si, tornando-se eminentemente colaborativa e globalizada.

Novos procedimentos experimentais estão em pleno desenvolvimento, viabilizando a instalação de sofisticados instrumentos de medida em localidades remotas, como imensos telescópios e aceleradores de partículas, de tal forma que os dados capturados ou coletados por eles podem ser enviados imediatamente para redução e análise a diversos centros de pesquisa simultaneamente, em diversos países. A análise dos dados pode ser conduzida em quaisquer desses centros, e os resultados podem ser guardados em outros centros, também remotos, em grandes silos de armazenamento. Devido à sofisticação crescente dos instrumentos de medida, os dados experimentais também estão sendo gerados e / ou coletados a taxas cada vez mais altas.

Esse avanço vem causando um forte impacto no modo como os cientistas analisam os dados e coordenam suas atividades de pesquisa. Duas razões tentam explicar o motivo dessa mudança de paradigma na forma como a ciência do século XXI está sendo conduzida:

- o avanço científico: novas descobertas dependem de instrumentos maiores, mais rápidos e de melhor definição. Da mesma forma, os experimentos estão se tornando extremamente complexos e caros demais para serem conduzidos por pesquisadores de um mesmo grupo, ou até mesmo de uma mesma nação.
- o avanço tecnológico: o aperfeiçoamento dos instrumentos de medida e de captura de dados, a rápida evolução dos sistemas de processamento, de armazenamento, e dos equipamentos de visualização de alta definição, e a incrível evolução das redes de comunicação de alta velocidade.

Os experimentos em plena operação na área de física de partículas, ou física de altas energias, como o CDF e o D0 no Fermilab ou o BaBar no SLAC (*Stanford Linear Accelerator Center*), já acumularam dados experimentais nos últimos anos que excedem o Petabyte (1 milhão de gigabytes). Numa escala ainda maior, os experimentos que serão realizados no LHC (*Large Hadron Collider*) do CERN a partir de 2008 irão acumular centenas de Petabytes de dados em poucos anos, que precisarão ser analisados pela comunidade mundial de pesquisadores da área. Outros projetos científicos de larga escala, como o SDSS (*Sloan Digital Sky Survey*), uma ambiciosa cooperação entre observatórios astronômicos para o mapeamento estelar, a colaboração LIGO (*Laser Interferometer Gravitational wave Observatory*) para detecção de ondas gravitacionais cósmicas, o NVO (*National Virtual Observatory*), que une os bancos de dados astronômicos de diversos observatórios e os torna acessíveis a astrônomos profissinais e amadores e estudantes, os experimentos de física nuclear nos laboratórios BNL (*Brookhaven National Laboratory*) e INFN (*Istituto Nazionale di Fisica Nucleare*), a colaboração BIRN (*Biomedical Informatics Research Network*) que viabiliza o compartilhamento em larga escala de dados na área da ciência biomédica, o GEO (*Global Earth Observation*) Grid, uma infra-estrutura computacional distribuída para acelerar o desenvolvimento das ciências geológicas, dentre inúmeros outros, todos eles enfrentam necessidades semelhantes de compartilhamento de complexas massas de dados entre colaborações em escala global.

Novas e sofisticadas infra-estruturas computacionais que permitem a integração de sistemas distribuídos em larga escala estão sendo implementadas para atender a essa crescente demanda por poder de processamento e armazenamento de dados da comunidade científica mundial. Essas infra-estruturas têm recebido a denominação genérica de “Grids computacionais”.

A computação em Grid teve início com a aplicação simultânea de recursos computacionais interligados em rede para a solução de um mesmo problema científico. De fato, a comunidade científica, da qual se destacam os físicos de altas energias, os astrônomos e os biomédicos, está liderando o desenvolvimento na área de Grids computacionais, devido à complexidade dos problemas que esses grupos têm que resolver. Os instrumentos científicos estão ficando cada vez maiores, os custos de construção altíssimos, e por isso estão sendo construídos de forma colaborativa por centenas – às vezes milhares – de pesquisadores espalhados ao redor do planeta. O acelerador de partículas LHC do CERN e os imensos telescópios do Projeto GEMINI são exemplos dessa tendência à construção de estruturas gigantescas e complexas por uma comunidade grande de pesquisadores que precisam colaborar entre si, trocar dados e informações, e ao mesmo tempo fazer uso compartilhado de um mesmo instrumento, construído em conjunto. Tais instrumentos, por sua vez, são capazes de capturar informações e gerar imensas massas de dados, que precisam ser analisados por sistemas computacionais de enorme capacidade de processamento, o que só se consegue se a tarefa de processamento for realizada de forma distribuída e colaborativa.

Assim, no contexto da computação de alto desempenho, O termo Grid refere-se a uma infra-estrutura de recursos computacionais distribuídos, operados e controlados por diferentes organizações, e integrados através de diversas camadas de *software* que habilitam o uso coordenado e colaborativo de seus diversos elementos constituintes: sistemas de processamento de alto desempenho, silos de armazenamento de alta capacidade, ambientes de visualização de alta resolução, redes ópticas de alta velocidade, redes de sensores distribuídos e instrumentos científicos de última geração. Aplicações que fazem uso de uma infra-estrutura desse porte em geral precisam movimentar grandes massas de dados, executar processamento intensivo e compartilhar recursos entre organizações distintas de forma descentralizada, segura e eficiente, usando protocolos e interfaces abertos e padronizados.

O conceito de Grids computacionais remonta a idéias estabelecidas desde a década de 60. O termo *computing Grid* foi introduzido em meados da década de 90 pelos pioneiros Ian Foster e Carl Kesselman, como substituto ao termo “meta-computação” usado na época, para descrever uma nova proposta de infra-estrutura de computação distribuída dedicada à ciência e à engenharia. O termo foi inspirado em outros tipos de redes, como as redes de distribuição elétrica (*power Grids*), as malhas ferroviárias (*railroad Grids*), e as redes telefônicas (*telephone Grids*), infra-estruturas que atendem o usuário final após serem distribuídas por meio de malhas de grande capilaridade.

Um aspecto fundamental dessas sofisticadas infra-estruturas está relacionado à infra-estrutura de rede, que dá suporte à comunicação entre as diversas entidades que compõem estes ambientes, sejam elas recursos de processamento, de armazenamento, de visualização, sensoriamento ou controle. A evolução e a adoção rápida das tecnologias de Grids computacionais estão ocorrendo justamente devido ao vertiginoso avanço das tecnologias de redes de comunicação de dados nos últimos anos, principalmente pela crescente adoção de sistemas baseados em tecnologia óptica. Entretanto, os mais recentes sistemas de Grids computacionais, que estão sendo construídos para viabilizar pesquisas científicas e tecnológicas de última geração, ainda dependem da infra-estrutura da rede mundial, a Internet, para a integração de seus recursos computacionais. Esta dependência está restringindo severamente a implantação, em escala global, desses sistemas computacionais extremamente avançados.

A Internet comercial, ou Internet *commodity*, embora alicerçada em uma tecnologia madura e extremamente bem sucedida, tem como premissa básica um modelo de transferência de informação baseado no melhor esforço. Nesse modelo, os fluxos de dados, divididos em pequenos pacotes, compartilham entre si a largura de banda total da rede. Quando ocorre congestionamento, pacotes de dados podem ser descartados, sem distinção. Não há garantia de que o serviço será realizado com sucesso, nem há garantia de desempenho. Este modelo tem se mostrado bastante adequado para transações baseadas em transferências intermitentes de pequenas massas de dados, porém é ineficiente para viabilizar a troca contínua de grandes volumes de dados, como os que estão sendo gerados em diversas aplicações científicas de última geração, ou no uso e controle eficiente de recursos remotos em tempo real a longas distâncias. Os experimentos em física de altas energias, por exemplo, estão produzindo mais dados do que é realisticamente possível armazenar e processar em uma única localidade geográfica. Não por acaso, experimentos científicos dessa natureza têm impulsionado o desenvolvimento das tecnologias de computação em *Grid* e de redes ópticas avançadas.

Nesse contexto, a formação de uma nova geração de cientistas e engenheiros capazes de integrar essas infra-estruturas emergentes nos processos profissionais, educacionais e criativos das diversas carreiras científicas e tecnológicas que estão fazendo uso dessas novas tecnologias torna-se essencial. Tais profissionais devem ser treinados para garantir que a infra-estrutura esteja disponível para uso pela comunidade científica, e para atuar no sentido de diminuir as barreiras relativas à aceitação e uso dessas novas tecnologias.

O termo "*cyberinfrastructure*", usado pela primeira vez em um comitê da NSF americana em 2003, parece expressar melhor essa tendência. A palavra *cyberinfrastructure* refere-se à integração, coordenação e disponibilização de tecnologia da informação e de recursos humanos para dar suporte aos complexos problemas envolvidos nas grandes colaborações científicas. Embutido nesse novo conceito está o reconhecimento de que as aplicações que fazem uso de uma infra-estrutura dessa natureza é que devem definir e dirigir a forma e a função das tecnologias de informação, e não o contrário, isto é, aplicações tendo que ser adaptadas às tecnologias existentes. Ou seja, o termo *cyberinfrastructure* engloba não só os sistemas de captura, processamento, armazenamento, transporte eficiente de dados e mecanismos de divulgação de informação e conhecimento, mas inclui os recursos humanos aptos a gerenciar e fazer uso efetivo de tais sistemas, ou seja, refere-se a toda infra-estrutura necessária para dar suporte à chamada *e-Science*. O objetivo primordial é prover ferramentas computacionais colaborativas que acelerem a taxa de novas descobertas mediante o uso de ferramental tecnológico altamente sofisticado. A premissa subjacente é que as pesquisas em diversas áreas do conhecimento humano estão se tornando excessivamente complexas e custosas para serem realizadas por pequenos grupos de cientistas em laboratórios individuais.

As iniciativas correntes em *cyberinfrastructure* estão primariamente focadas na construção de sofisticados Grids computacionais conectados entre si através de redes ópticas multi-gigabit, além de ferramentas de videoconferência e de tele-presença. Para que ocorra a adoção bem sucedida dessas novas tecnologias, esses esforços precisam ser complementados através do desenvolvimento de mecanismos e metodologias que permitam disseminar o seu uso eficiente à comunidade científica, através de programas educacionais e treinamento prático.

1.2. A contribuição do SPRACE

O projeto SPRACE - São Paulo Regional Analysis Center - é um exemplo de como investimentos em ciência básica podem trazer benefícios para diversas outras áreas do conhecimento, bem como para o desenvolvimento de tecnologia. Devido à implementação bem sucedida do projeto e ao conhecimento acumulado nessa empreitada, o grupo de pesquisa, liderado pelo prof. Sérgio Novaes do IFT/UNESP, encaminhou um novo projeto à FINEP, com o objetivo de prover capacidade computacional a diversos grupos de pesquisa da UNESP, integrando assim diversos laboratórios espalhados pelo Estado de São Paulo em torno de uma estrutura computacional distribuída, porém acessível a todos os participantes. O projeto GridUNESP foi concebido, proposto e aprovado no decorrer de 2006. Em 2007, a especificação técnica foi aprimorada, as propostas dos fornecedores foram apresentadas, a melhor delas foi escolhida por um comitê independente, e a aquisição de todo o hardware deve ser completada até o final de dezembro.

O projeto GridUNESP busca justamente implementar o conceito de *cyberinfrastructure*, ou seja, uma infra-estrutura computacional espalhada por sete campi da UNESP, interligada através de um *middleware* de Grid, para uso compartilhado de pesquisadores que atuam nas mais diversas áreas: genômica, física do estado sólido, física de altas energias, bioinformática, bioquímica, geologia, química fundamental, nanotecnologia, e ciência da computação. O projeto GridUNESP permitirá, pela primeira vez no Brasil, o uso das tecnologias de Grids computacionais no âmbito acadêmico. Devido à integração com o Open Science Grid americano, os pesquisadores da UNESP terão acesso ao vasto poder computacional disponibilizado por essa nova infra-estrutura. Tão interessante quanto o hardware em si, está a intensa movimentação da UNESP no sentido de prover a capacitação de recursos humanos para operar essa nova infra-estrutura, e disseminar e popularizar o seu uso. A capacitação adequada de recursos humanos é um aspecto que não deve ser negligenciado.

Embora a computação seja a atividade mais visível do SPRACE, seu objetivo principal é contribuir na elucidação das grandes questões relativas à estrutura íntima da matéria e nas leis fundamentais da natureza. Os pesquisadores do SPRACE têm participado do grupo de análise de novos fenômenos da Colaboração DZero do Fermilab nos últimos anos, e à medida em que as atenções da comunidade de física de altas energias se deslocam do Fermilab para o CERN, o grupo prepara-se para deslocar seu foco na mesma direção. O grupo de pesquisa também tem feito um grande esforço para disseminar e popularizar o conhecimento adquirido nas últimas décadas na área de física de partículas entre os estudantes de segundo grau. Um projeto de divulgação científica, denominado “Estrutura elementar da matéria: Um cartaz em cada escola” [1], foi submetido à FINEP com essa finalidade, e está em pleno andamento.

O SPRACE iniciou suas operações em 2004. No mês de março daquele ano, começaram a ser processados, pela primeira vez em São Paulo, eventos produzidos pela Colaboração DØ do Fermilab, em Chicago, EUA. A Colaboração DØ opera um dos dois detectores do Tevatron, o acelerador de partículas com maior energia em operação atualmente. Ela conta atualmente com a participação de 73 instituições de 18 diferentes países. O Centro Regional de Análise de São Paulo (São Paulo Regional Analysis Center - SPRACE) foi implantando com apoio da FAPESP através do Projeto Temático “Física Experimental de Anéis de Colisão: SPRACE e HEPGrid/Brazil”. O projeto foi submetido em meados de 2003, e a aprovação ocorreu ao final daquele ano. Em fevereiro de 2004 a infra-estrutura básica necessária para a instalação dos primeiros servidores do centro de processamento já estava pronta, montada no Centro de Ensino e Pesquisa Aplicada do Instituto de Física da USP. Em agosto de 2005 o SPRACE passou a fazer parte do OSG (Open Science Grid), Em julho de 2006, o SPRACE foi incluído oficialmente na lista das Tiers-2 da Colaboração CMS.

2. Atividades realizadas no período

Exercendo a função de Especialista de Laboratório, as atividades profissionais que executo estão relacionadas ao suporte aos sistemas computacionais e à infra-estrutura de rede do centro de processamento de dados do SPRACE, e a de auxílio às atividades de pesquisa do grupo. As principais atividades que executei durante o período a que se refere este relatório são detalhadas a seguir.

2.1. Manutenção do SPRACE

O SPRACE mantém seu centro de processamento nas dependências do Centro de Ensino e Pesquisa Aplicada (CEPA) do Departamento de Física Experimental, no Instituto de Física da USP. A área ocupada, de cerca de 25 m^2, é mantida isolada através de divisórias, tem um quadro elétrico adequado com entrada de energia independente dos demais laboratórios, e é mantida sob condições de temperatura controlada, na faixa de 20 a 24 graus Celsius. O resfriamento é realizado por três aparelhos de condicionamento de ar, totalizando uma capacidade de refrigeração de cerca de 160 kBTU/h.

No final de 2006 a infra-estrutura computacional do SPRACE foi completada, com o término da implantação da fase 3. Assim, iniciamos o ano de 2007 com o cluster operando a plena capacidade – um poder de processamento de cerca de 300 kSPECint2000 – e 12,5 Terabytes de disco para armazenamento. Agora em sua versão final, o cluster conta com 84 servidores dual-processados dedicados exclusivamente ao processamento de jobs, mais 4 servidores principais. Os servidores principais realizam as funções de '*front-end*', 'OSG *gatekeeper*', controle de armazenamento centralizado, e controle dos '*pools*' de armazenamento distribuído. Os servidores de processamento são os que efetivamente executam o processamento de *jobs*, atuando como *number crunchers*. Dentre os nós de processamento, os 32 adquiridos no final de 2006 incorporam a tecnologia de núcleo duplo (*dual-core*). Logo, a somatória de núcleos de processamento chega a 240, cada um dos quais com 1 GB de memória disponível. O grupo vem também desenvolvendo análise física buscando alguma evidência de dimensões extras nos dados do DØ do Fermilab. O SPRACE também faz parte, desde agosto de 2005, do Open Science Grid (OSG), um Grid computacional que congrega dezenas de universidades americanas, sendo atualmente a segunda maior infra-estrutura dessa natureza em nível global, superada apenas pelo EGEE (*Enabling Grids for E-science in Europe*), uma infra-estrutura semelhante que congrega dezenas de universidades européias. A partir de julho de 2006 o SPRACE passou a integrar a estrutura de Grid do *Large Hadron Collider* (LHC) do CERN como uma unidade de processamento de classe Tier-2, processando dados do experimento CMS (*Compact Muon Detector*) do qual o grupo também faz parte, juntando-se assim a um seleto grupo de sites americanos de classe Tier-2 ligados ao Fermilab (um centro de processamento de classe Tier-1): Caltech, MIT, Purdue University, UCSD, University of Florida, University of Nebraska - Lincoln, University of Wisconsin, além da UERJ, no Rio de Janeiro. Obviamente, o CERN ocupa a posição hierárquica mais alta, sendo considerado a unidade de classe Tier-0.

As atividades de manutenção podem ser resumidas como as tarefas diárias para manter todo o conjunto de servidores em plena operação, tentando minimizar ao máximo o *downtime*, seja de um servidor único, no caso de falha no hardware, ou de todo o site, no caso de falha em algum serviço essencial do *middleware*. Podemos classificar essas tarefas em dois grandes grupos: a manutenção do hardware e a manutenção dos serviços (software). A manutenção do hardware engloba todas as ações

necessárias para manter os sistemas físicos em funcionamento. Isso inclui atuar nos servidores em caso de falha, efetuando o conserto ou chamando a assistência técnica se for necessário, atuar na infra-estrutura de rede para manter o centro conectado ao resto do mundo, manter a temperatura ambiente interna na faixa considerada adequada, inspecionar constantemente os sistemas de monitoração dos servidores para verificar se estão todos ativos e se nenhum excede a temperatura normal de operação, além de manter a limpeza e a organização do data center e da mesa de operação. Além da manutenção diária, os sistemas em geral precisam ser atualizados, partes e peças precisam ser adquiridas, o que requer um contato constante com diversos fornecedores da área de tecnologia da informação.

A manutenção do software é uma tarefa mais árdua, pois requer não apenas sólidos conhecimentos de administração de sistemas Linux, mas principalmente uma capacidade de aprendizado rápido, para permitir ao administrador acompanhar a evolução do complexo *middleware* que dá suporte à colaboração científica internacional. Grande parte da dificuldade ocorre devido ao fato de os subsistemas que compõem esse *middleware* estarem em pleno desenvolvimento, de modo que é muito comum surgirem problemas imprevistos e não documentados, além da dificuldade de manter os sistemas atualizados, dado que as versões dos pacotes de software são constantemente atualizadas.

Tais atividades requerem um cuidado diário, consumindo, dependendo da ocorrência, boa parte do dia de trabalho do administrador. O contato com os fornecedores é também tarefa constante, pois é muito comum ocorrerem falhas de hardware dos sistemas ou defeitos nos aparelhos de ar condicionado. A infra-estrutura de rede tem se mostrado bastante sólida, mas também não é imune a erros. Logo, o contato com o NARA – Núcleo de Apoio à Rede Acadêmica, responsável pelos elementos de rede que permitem a conexão do SPRACE ao NAP do Brasil em Barueri, onde está situado o *backbone* da rede ANSP (*Academic Network of São Paulo*), é também frequente.

Em outubro de 2007, iniciamos uma série de reuniões com o objetivo de redefinir o hardware de um dos pontos fracos do SPRACE no momento: o subsistema de armazenamento. Como resultado dessas reuniões, elaboramos uma proposta de atualização de baixo custo cujo objetivo é aumentar o desempenho desse subsistema. A figura 1 a seguir ilustra a configuração atual do subsistema de armazenamento do SPRACE.

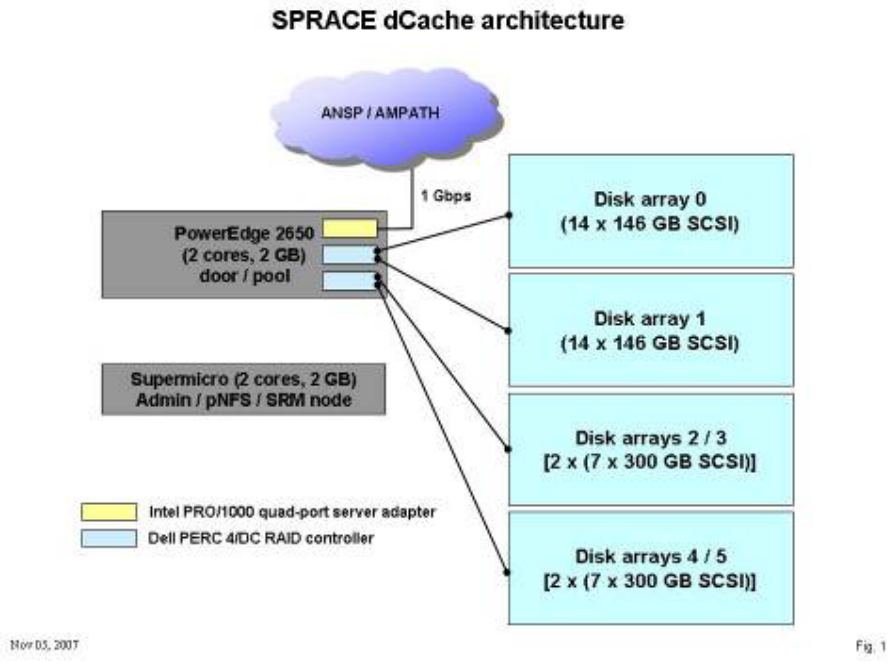

Figura 1: Arquitetura do subsistema de armazenamento do SPRACE, controlado pelo software dCache.

Através de uma série de testes identificamos que o servidor que controla os *pools* de armazenamento (o PowerEdge 2650, conforme mostra a figura acima) está operando além de sua capacidade, devido ao fato de estar controlando 4 unidades de armazenamento. A figura 2 apresenta a atualização proposta. Esta proposta foi aprovada em reunião e já está em plena execução. Estamos aguardando a chegada das partes e peças.

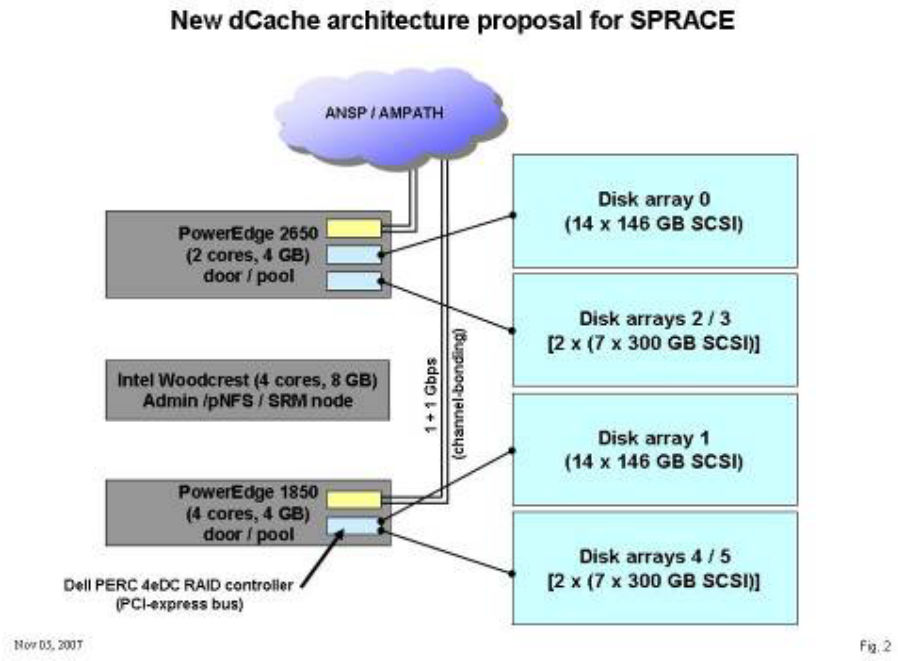

Figura 2: Proposta de reformulação da arquitetura do subsistema de armazenamento do SPRACE

2.2. Participação na colaboração CMS

Embora registrado como membro da colaboração CMS desde meados de 2006, somente a partir de 2007 comecei a participar mais ativamente de atividades relacionadas à colaboração. Em janeiro de 2007, registrei-me e passei a acompanhar mais de uma dezena de listas de discussão, sobre os mais variados temas, todos ligados à infra-estrutura computacional da colaboração. Ao longo do ano, também passei a assumir, aos poucos, a responsabilidade de representar o SPRACE nas reuniões quinzenais que ocorrem, via teleconferência através do sistema WebEx [2], entre todos os sites de classe Tier-2 ligados diretamente ao Fermilab. Essas reuniões são denominadas USCMS T2 *meetings*, e duram cerca de 1h30min. Nelas, cerca de 10 sites reportam os problemas que tiveram nos quinze dias anteriores, de forma resumida (em geral, cada site apresenta 2 ou 3 slides). As discussões incluem as dificuldades encontradas pelos diversos sites para manter sua respectiva infra-estrutura operacional. As experiências de sucesso também são reportadas e compartilhadas.

Uma tarefa importante para consolidar uma participação ainda mais efetiva na colaboração, e que deverá ser executada nos próximos dias, é o meu registro na VO (Virtual Organization) do CMS. Esse registro habilita o membro da colaboração a submeter jobs para execução na infra-estrutura

computacional do WLCG (Worldwide LHC Computing Grid). Para que o administrador de sistemas possa testar a capacidade operacional do site em toda a sua extensão, é preciso que ele disponha também da permissão de submissão de jobs para o Grid. O procedimento para minha inclusão ainda não havia sido disparado porque até há poucos meses não víamos a necessidade de se realizar testes que chegassem até o nível de submissão de jobs. Entretanto, com o aumento crescente da responsabilidade de manter o site ativo e em plenas condições de produção, estamos detectando a necessidade de efetuar testes de validação mais aprofundados. Além disso, o aprendizado dos comandos de submissão de jobs é necessário para que o administrador seja capaz de dar suporte adequado aos pesquisadores, que são efetivamente os usuários finais da infra-estrutura.

2.3. Participação em eventos

No decorrer de 2007 participei de diversos workshops e eventos relacionados à minha área de atuação, com o objetivo de aprofundar meus conhecimentos, aumentar a integração com os membros do próprio grupo, e interagir com outros grupos de pesquisa e com representantes da indústria de informática. A seguir apresento uma lista dos eventos que participei, incluindo uma breve descrição.

- I Workshop do SPRACE

Este workshop teve como objetivo proporcionar a todos os membros do grupo uma visão detalhada das atividades desenvolvidas no SPRACE em cada uma de suas frentes de atuação. Ocorreu de 27 a 30 de março de 2007em Barra do Una, São Sebastião – SP.

- I Workshop de Computação de Alto Desempenho

Realizado no auditório Prof. Francisco Romeu Landi na Escola Politécnica da USP, este evento discutiu o panorama da computação de alto desempenho na pesquisa avançada. Ocorreu em 10 de abril de 2007, em São Paulo – SP.

- NIDays 2007 – Conferência Mundial de Instrumentação Virtual

Ciclo de palestras técnicas e sessões práticas sobre a tecnologia de instrumentação virtual da National Instruments. Ocorreu em 12 de abril de 2007, em São Paulo – SP.

- Sun TechDays 2007

Congresso internacional destinado a desenvolvedores, arquitetos e executivos de informática, promovido pela Sun Microsystems, com o objetivo de discutir e disseminar as tecnologias desenvolvidas pela empresa. Ocorreu de 18 a 20 de abril de 2007 em São Paulo – SP.

- 7th IEEE International Symposium on Cluster Computing and the Grid – CCGrid'07

Um evento internacional bastante tradicional na área de computação baseada em clusters, que foi trazido para a América do Sul pela primeira vez, graças à iniciativa de Bruno Schulze, pesquisador do LNCC. Ocorreu de 14 a 17 de maio de 2007 na Barra da Tijuca, Rio de Janeiro – RJ.

- Workshop de P & D do Projeto GIGA / RNP

Workshop de apresentação dos resultados finais dos trabalhos de pesquisa da primeira fase do projeto GIGA, iniciado em 2004 por iniciativa da RNP em parceria com o CPqD, e patrocinado pelo FUNTEL

com apoio do Ministério da Ciência e Tecnologia. Ocorreu de 3 a 6 de setembro de 2007, na sede do LNCC em Petrópolis – RJ.

- 19th Symposium on Computer Architecture and High Performance Computing – SBAC-PAD'07

Simpósio anual da área de Arquitetura de Computadores e Processamento de Alto Desempenho, que reúne a comunidade de pesquisadores brasileiros que atuam nessas áreas. Ocorreu de 24 a 27 de outubro de 2007, em Gramado – RS.

2.4. Seminários e palestras técnicas

Durante o I Workshop do SPRACE, apresentei quatro palestras, três delas relatando algumas das atividades que participei no ano anterior (2006), e uma quarta descrevendo um projeto de *education & outreach* no qual estamos trabalhando. Foram apresentadas nos dias 28 e 29 de março de 2007.

- A brief overview fo the CERN School of Computing 2007
- The EELA Project and briefings of the 3rd EELA Tutorial
- SPRACE Network Overview and future plans
- A Grid-enabled infrastructure testbed for education and outreach in São Paulo

Além dessas, recebi o convite para apresentar uma palestra no III Workshop do GridUNESP, a ocorrer em 14 de dezembro próximo, cujo tema é:

- Programa de Treinamento e Formação de Recursos Humanos

2.5. Elaboração de projeto

No decorrer de 2007, trabalhei na concepção de um projeto relacionado a atividades de treinamento correlatas ao SPRACE. A proposta, que pode ser denominado de Grid educacional, corresponde à construção de uma infra-estrutura computacional, composta por 10 servidores de última geração, para dar apoio ao ensino e divulgação das tecnologias Grid. Estamos negociando a doação desses servidores com 4 empresas: a Intel, que deverá fornecer os processadores e ferramentas de software, a Sun Microsystems e a SGI, que deverão fornecer o *barebone* (hardware básico, excetuando processador, memória e disco), a Kingston, que deverá fornecer as memórias, e a Seagate, que entrará com os discos rígidos. As negociações estão ocorrendo através de representantes dessas empresas no Brasil, mas as doações serão feitas pelas respectivas matrizes nos EUA. Pretende-se instalar esses servidores em diversas Universidades na região metropolitana de São Paulo, de modo a formar uma infra-estrutura para treinamento de estudantes, administradores de sistemas e pesquisadores nas tecnologias de computação em Grid. Pretende-se também usar modernas técnicas de virtualização de servidores de modo a permitir que cada um dos 10 servidores reais acomodem diversos servidores virtuais, multiplicando a quantidade total de servidores disponíveis. Como se trata de uma plataforma educacional, que prescinde de alto desempenho, a técnica de virtualização torna-se bastante adequada. A figura 3 a seguir apresenta a proposta de distribuição dos 10 servidores por diversas universidades paulistas. Como mostra a figura, pretende-se, na medida do possível, instalar 2 deles no *NAP of Americas* em Miami, ponto onde chega nossa conexão de rede com os EUA. Essa possibilidade ainda

está em discussão e negociação, devido à dificuldade de manter os servidores operacionais em local tão distante. Caso esta opção se mostre inviável, os dois servidores extras serão encaminhados para uma outra Universidade (possivelmente a Universidade Federal de São Carlos).

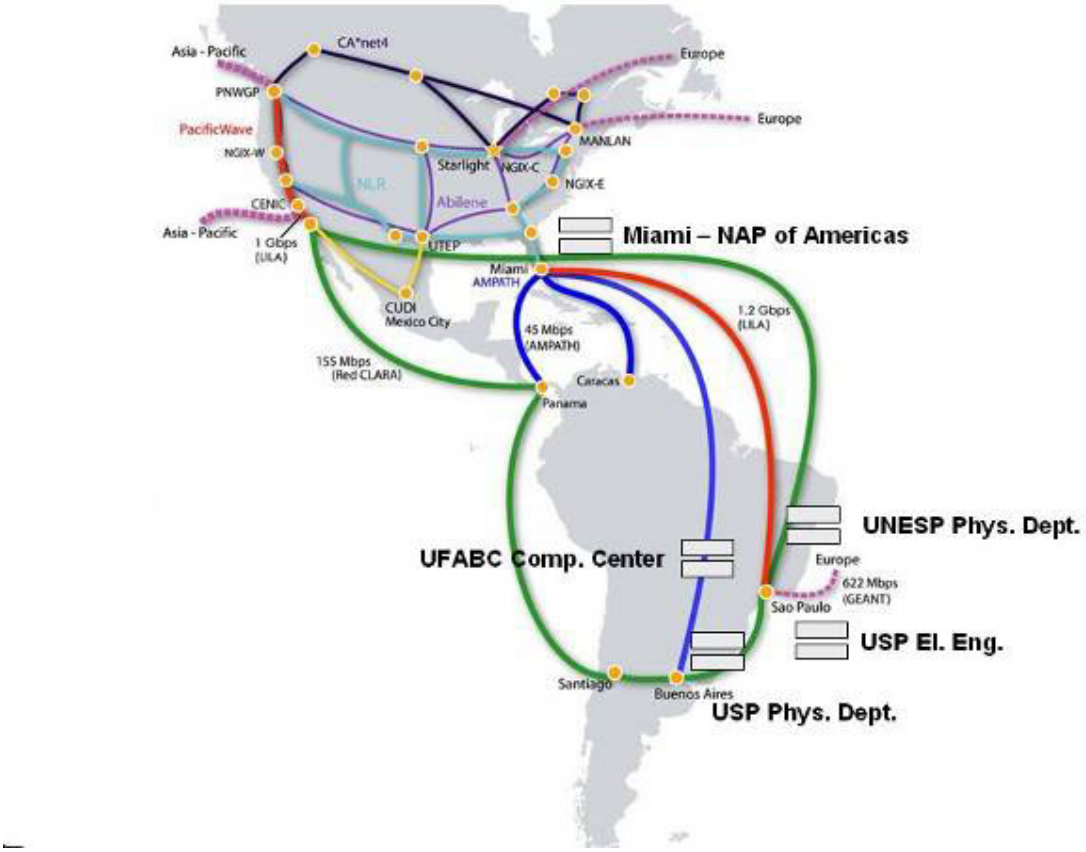

Figura 3: Proposta de projeto de uma infra-estrutura de Grid voltada para educação e disseminação

2.6. Atividades de pesquisa

No dia 11 de agosto de 2007 prestei o Exame de Qualificação no programa de pós-graduação em Engenharia Elétrica da Escola Politécnica da USP. Nesta ocasião discorri sobre o meu tema de pesquisa e detalhei algumas propostas de implementação. O título da apresentação foi:

- "Uma estratégia de escalonamento dinâmico em Grid Computacionais sobre redes ópticas WDM utilizando inteligência de enxame"

O tema trata de estratégias de controle de caminhos ópticos em redes baseadas na tecnologia WDM, considerando-se a viabilidade de usar tais redes como arcabouço de comunicação para uma infra-estrutura de Grid computacional construída sobre ela. Um dos objetivos é permitir que aplicações executando sobre uma infra-estrutura construída dessa forma sejam capazes de requerer conexões fim-a-fim, cujos pontos extremos sejam recursos de processamento ou de armazenamento, através de mecanismos de intermediação entre o plano de controle óptico da rede e o *middleware* de Grid. Em

última instância, deseja-se tornar a rede de comunicação um recurso escalonável, capaz de ser controlado por um *middleware* de Grid.

Embora estando à época matriculado como aluno de mestrado, a relevância do tema, bem como as estratégias de implementação propostas, levaram-me a conquistar uma qualificação para o programa de doutorado direto.

2.7. Outras atividades

- Elaboração de relatório detalhado sobre a infra-estrutura de rede do SPRACE, em parceria com Jorge Marcos de Almeida, analista do NARA. Parte deste documento foi incorporado no UltraLight Project Annual Report 2006-2007. O documento está reproduzido na íntegra no Apêndice I.

- Participação nas reuniões técnicas de discussão da implantação do projeto GridUNESP, como membro convidado.

- Consultoria na atualização e detalhamento da especificação técnica do projeto GridUNESP, atuando na articulação junto a seis fornecedores de soluções HPC, que permitiu a eles enviar propostas comerciais em perfeita adequação às especificações técnicas definidas pela equipe de implantação do projeto.

- Participação na Comissão Técnica que deliberou sobre a escolha da proposta vencedora para o fornecimento do hardware do GridUNESP, a convite do coordenador técnico do projeto, prof. Sérgio Novaes.

- Articulação junto à Sun Microsystems do Brasil para promover a inclusão do Instituto de Física da USP no programa SAI – *Sun Academic Initiative*, que propicia a estudantes e funcionários o acesso a farto material de treinamento nas tecnologias Sun (Java, Open Solaris, Open SPARC e NetBeans).

- Instalação e testes de dois novos servidores Intel de última geração (dual quad-core com 8 GB de memória) no SPRACE, doados pelo Caltech, com o objetivo de servir de plataforma de testes do link internacional.

- Aquisição de mobiliário para uma nova sala de operações do SPRACE, em espaço disponibilizado num edifício recém construído no Instituto de Física da USP.

- Reuniões com diversos fornecedores para conhecer novas tecnologias e auxiliar a definir com mais detalhes a atualização do parque computacional do SPRACE, que deverá ocorrer no próximo ano.

- Contatos técnicos permanentes com analistas e técnicos de informática do Instituto de Física e de outras unidades da USP, e participação em feiras e eventos relacionados à minha área de atuação, para atualização tecnológica.

- Submissão de artigo científico ao evento "12th International Conference on Optical Networking Design and Modeling - ONDM 2008", com o título *Classes of Service for Scheduling in Lambda Grids using Ant Colony Optimization*, a ser realizado de 12 a 14 de março de 2008, na Espanha (aguardando notificação de aceitação).

- Organizador e "*chairman*" de um Workshop sobre HPC (*High Performance Computing*) que ocorreu no Instituto de Física da USP em 18 de outubro de 2007, cujo tema foi: *Strategic Issues in Cyberinfrastructure for the 21st Century Science & Engineering Research*. A programação do Workshop, incluindo os temas dos seminários, é apresentada no Apêndice II.

3. Planejamento para o próximo período

Exercendo a função de Especialista de Laboratório do SPRACE, as atividades profissionais que executo estão relacionadas ao suporte aos sistemas computacionais e à infra-estrutura de rede do centro de processamento de dados o SPRACE, e a de auxílio às atividades de pesquisa do grupo. As principais atividades que executei durante o período a que se refere este relatório são detalhadas a seguir.

3.1. Ampliação do SPRACE

Um novo projeto temático será submetido à FAPESP em janeiro do próximo ano, o qual deverá incluir, dentre diversas outras particularidades, um pedido de ampliação do hardware do SPRACE. Pretende-se aumentar o poder de processamento atual em pelo menos 50%, e aumentar a capacidade de armazenamento dos atuais 12,5 TB para pelo menos 100 TB. Além disso, será incluído o pedido de aquisição de um roteador de rede mais robusto, superior ao roteador que o SPRACE usa atualmente.

Essa ampliação vem sendo estudada ao longo de 2007. Durante as reuniões quinzenais com as demais unidades Tier-2 ligadas ao Fermilab, tem-se notado que todas estão passando por sucessivos processos de ampliação, de forma que o parque computacional do SPRACE está ficando aquém do mínimo exigido para se operar como um centro do porte de uma Tier-2.

Quanto ao aumento do poder de processamento, estamos estudando a possibilidade de usar servidores do tipo '*blade*' (unidades de processamento montadas em lâminas verticais e acondicionadas em uma caixa metálica com apenas 2 fontes de alimentação redundantes), mais compactos e mais econômicos (consomem menos energia elétrica e dissipam menos calor) que os servidores atuais. O aumento da capacidade de armazenamento também tem sido discutido com diversas empresas que fornecem soluções na área, incluindo SGI, Dell, IBM, e Sun Microsystems. Contatamos também pequenos fornecedores, na busca de soluções customizadas e de baixo custo.

A aquisição de novo roteador de rede está vinculada à proposta de ampliação do link internacional dos atuais 2,5 Gbps para 10 Gbps, previsto para ocorrer até 2009. O equipamento de rede que usamos atualmente é incapaz de operar a taxas superiores a 1 Gbps, e já está operando no limite máximo de sua capacidade. Em função disso, tivemos várias reuniões ao longo desse ano com representantes da Foundry e da Cisco, com o objetivo de definir o melhor equipamento a ser adquirido no próximo ano. O roteador que usamos hoje continuará operacional, e será usado como contingência em caso de falha no roteador principal.

3.2. Participação na colaboração CMS

Pretende-se dar continuidade aos trabalhos de participação na colaboração CMS, através do acompanhamento da reuniões quinzenais das Tiers-2 ligadas ao Fermilab. Uma participação mais efetiva está se tornando imprescindível, devido à proximidade do início de operação do LHC, previsto

para meados de 2008. Os principais subsistemas do *middleware*, como o dCache (sistema de catálogo de arquivos distribuído) e o PhEDex (sistema utilizado pela colaboração CMS para a movimentação de dados entre os diversos elementos de processamento e armazenamento), precisam ser compreendidos com mais profundidade, e estratégias mais eficientes de monitoração, tanto de sistemas quanto de serviços, precisam ser urgentemente implementadas. Para auxiliar na monitoração dos serviços, estamos estudando o uso do sistema Nagios [3], e esperamos implementá-lo no início de 2008. Para monitoração do hardware dos servidores, estamos considerando a possibilidade de uso da ferramenta IPMItool [4]. Um sistema de apoio à operação, que inclua um wiki para documentação e um sistema de *log* eletrônico para catalogação de ocorrências e serviços executados, como por exemplo o Trac [5],está também sendo intensamente discutido.

3.3. Participação em eventos

Em 2008, está prevista minha participação nos seguintes eventos:

- Florida International Grid School 2008, de 23 a 25 de janeiro de 2008, em Miami, FL.
- Network Operations & Management Symposium – NOMS 2008, de 7 a 11 de abril de 2008, em Salvador, BA.
- 26º Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, de 26 a 30 de maio de 2008, no Rio de Janeiro, RJ.
- 20th International Symposium on Computer Architecture and High Performance Computing, em outubro de 2008, em Mato Grosso do Sul.

3.4. Projeto Grid educacional

Pretende-se dar continuidade ao projeto do Grid educacional, com a chegada dos servidores prevista para os primeiros meses de 2008. Já dispomos de um pequeno servidor, em parte doado pela Intel Brasil, para realizar os primeiros testes de instalação e configuração de servidores virtualizados, usando ferramentas de virtualização *open source* (Xen e OpenVZ).

Em paralelo, devemos dar início à geração de material didático, incluindo manuais técnicos que descreverão o arcabouço teórico, além de atividades práticas e apresentações visuais para serem usados nos futuros seminários de treinamento.

3.5. Estágio de treinamento no CERN

Em outubro passado, um pedido de bolsa foi submetido ao HELEN – *High Energy Physics Latinamerican-European Network* [6], um bem-sucedido programa de cooperação acadêmica entre a União Européia e a América Latina. O pedido foi acompanhado de um plano de trabalho conciso, e o objetivo foi concorrer a uma bolsa de treinamento especializado de curta duraçao (STT – *Short-term*

training) no experimento LHC do CERN. O pedido foi aprovado, de forma que estamos reservando o período compreendido entre setembro e novembro de 2008 para uma permanência no CERN, sob supervisão do Dr. Matthias Kasemann, para executar um treinamento intensivo nas tecnologias Grid relacionadas ao experimento CMS. O palno de trabalho submetido ao programa HELEN está reproduzido no Apêndice III.

3.6. Outras atividades

As atividades do dia-a-dia, ou seja, a manutenção da operação dos sistemas de forma ininterrupta, o processo de monitoração dos servidores e serviços, as manutenções eventuais e programadas, sendo atividades rotineiras e permanentes, deverão ocorrer no próximo período de forma muito semelhante ao que foi realizado no decorrer de 2007.

Minha participação no projeto de implantação do GridUNESP deverá se tornar ainda mais intensa no próximo ano, bem como minha participação e contribuições nas discussões referentes à futura implantação de uma infra-estrutura de Grid computacional na Universidade de São Paulo.

4. Apêndices

Nas páginas seguintes são apresentados os apêndices mencionados ao longo do texto.

Apêndice I

SPRACE-ANSP Network Update

http://hep.ift.unesp.br/SPRACE/

Submitted by R. L. Iope (rliope@usp.br) and J. M. de Almeida (jrgmrcs@ansp.br)
on behalf of the SPRACE Collaboration

January 2007

Introduction

Since it was first commissioned in 2004, the SPRACE Tier2 cluster that is part of the worldwide CMS Grid, has made remarkable progress in its networking connectivity over the last three years. This has been possible through the efforts of the Academic Network at São Paulo (ANSP), and the Brazil's National Research and Education Network (RNP), and the support of UltraLight project members, Caltech, Florida International University, and Fermilab. Starting in 2004 when there was a single main server connected at 100 Mbps to a small Layer 2 access switch at University of São Paulo (USP), the SPRACE cluster has grown to full Tier2-scale with 240 processors and one data server housing 14 Terabytes of RAID and 6 Terabytes of distributed storage. All of the production and R&D data servers are integrated through a full Layer 3 switch and connected to two dedicated 1 Gbps wavelengths right at the heart of the São Paulo state R&E network provider. During these years SPRACE has achieved some important goals and has benefited from the recent investments in networking made by ANSP.

The Academic Network at São Paulo was created in 1989 as a special program of FAPESP, the São Paulo State Research Foundation. ANSP was the first NAP in Brazil, a neutral and secure environment where São Paulo state universities and research centers, Internet access providers and local Telecom companies started exchanging traffic, storing data and keeping secure information, among other functional features. The ANSP Network's NAP worked on the third floor of FAPESP's building until 2002, sharing the same space with the Foundation's data center. In 2002, it was transferred to the Terremark data center, the NAP of Brazil in Barueri, a small city in the São Paulo metropolitan area.

ANSP now leads the provisioning of advanced Internet to São Paulo state public and private universities, government and research institutes. ANSP is in a constant process of updating end-to-end connections across its network backbone, being co-responsible for the WHREN-LILA initiative [1], the single most important international link for research and education in Brazil.

The NAP of Brazil has emerged from an agreement between FAPESP and Terremark Latin America Ltda., a subsidiary of Terremark Worldwide, Inc, responsible for the NAP of the Americas [2] operation. NAP of the Americas houses AMPATH, a virtual network maintained by Florida International University in collaboration with Global Crossing, and connects researchers from several Universities and research Institutes of South and Central America, Mexico and the Caribbean. AMPATH is the Latin America access gateway to Internet2 and other international advanced research and education networks.

The agreement, established in 2002, transferred to Terremark the task of operating, maintaining and marketing the Network Access Point operated so far by ANSP. Prior to the partnership with Terremark, FAPESP was solely responsible for the operation and management of the entire peering point. Nowadays, Terremark operates the peering and houses ANSP core routers, whereas ANSP is responsible for network configuration and project planning for expansion and improvements on the São Paulo academic network.

SPRACE network connectivity – Historical overview

SPRACE started operating in March 2004. In those early days, we had only one main server connected to USP Physics Institute through a 100 Mbps port in a shared Layer 2 switch. A few months later the connection evolved to 1 Gbps, but remained connected to the local Physics Institute network, which owns a single 1 Gbps connection to USP Computer Center that is shared by hundreds of desktops and servers. By June 2005, we managed to connect SPRACE servers to one of the routers of ANSP network, through a Cisco Catalyst 3750 switch-router donated by Caltech. We used a Finisar FTR-1519 long-haul GBIC transceiver module loaned by ANSP at one end, and a Cisco 1000Base-ZX SFP module donated by Caltech at the other end. Signal strengths generated by both transceivers were enough to cover a distance of more than 40 km between USP Physics Institute and NAP of Brazil in Barueri.

Until January 2006, we were ourselves lighting one of the fiber pairs that connects USP and NAP of Brazil, while waiting for the new WDM boxes acquired by ANSP but retained in the Customs. By the end of 2005, however, a new telecom provider, CTBC Telecom, took over the contract with ANSP, and they had to start using a longer fiber pair connecting USP and NAP of Brazil. Unfortunately, the transceivers' signal strengths were not enough to overcome fiber attenuation and we could not reach Barueri anymore. This problem was temporarily solved by connecting SPRACE gateway, the Cisco 3750, directly into USP network backbone. SPRACE servers were kept attached to one of USP core routers until October 1st, 2006, when we finally managed to connect it again directly to NAP of Brazil, now using ANSP newly installed WDM MAN network.

Figure 1 below shows the SPRACE network connection as it is in these days, as well as the interconnection with our partners in Rio de Janeiro. Green lines in Figure 1 indicate the path SPRACE network traffic follows to reach the WHREN-LILA link, whereas dashed lines indicate the temporary network path we used to reach AMPATH from January to October 2006, through USP network backbone.

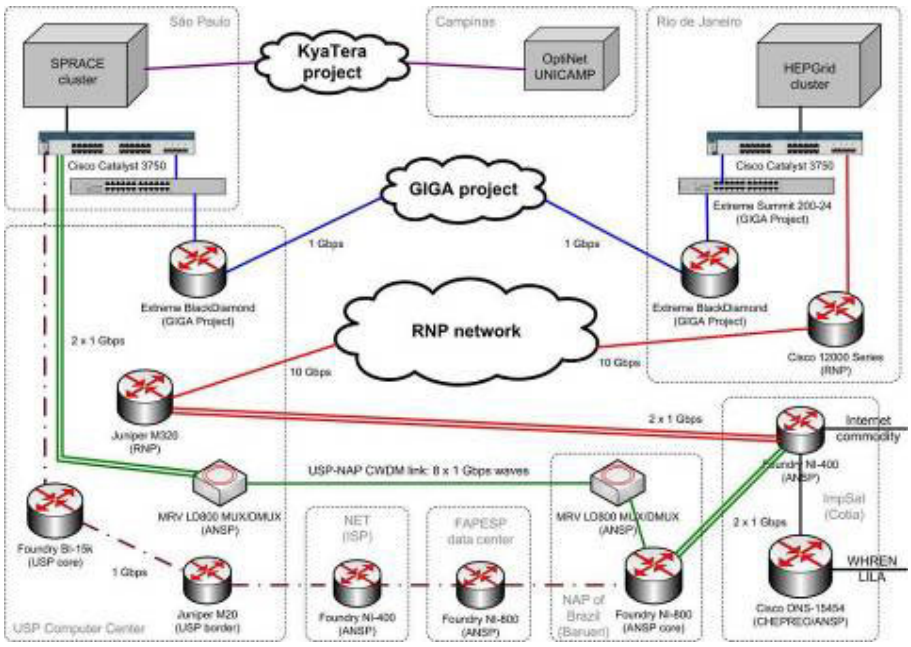

Figure 1: São Paulo and Rio de Janeiro HEP facilities' network connectivity

By the end of October 2006, during preparation for SuperComputing 2006 participation, we managed to connect a second 1 Gbps link between SPRACE and NAP of Brazil for experimental purposes. Due to the success of these experiments, ANSP engineers decided to include a 2 X 1 Gbps link in the first upgrade of USP-NAP-Impsat connection.

USP – NAP of Brazil – Impsat connectivity improvement

ANSP engineers are continuously working to provide enhanced connectivity to São Paulo public and private Universities and Research Institutions. One of those recent enhancements is directly related to SPRACE connectivity, as its main goal is to leverage existing bandwidth between USP and ANSP core routers in Barueri. USP Computer Center is a very important peering point, housing RNP and RedCLARA Points of Presence in São Paulo state and being connected with several telecom carriers and backbone providers. ANSP project goal is to aggregate traffic coming from different sources using WDM technology over the dark fiber pairs that connect USP and NAP of Brazil in Barueri, and those that connect NAP of Brazil and Impsat data center in Cotia. Although the core of ANSP network is located in Barueri, Impsat in Cotia has the access point to AMPATH through LANautilus [3] (São Paulo–Miami optical cable provider) and thus houses the gateway for the WHREN-LILA link, the Cisco ONS-15454 SDH MSPP. Barueri and Cotia are two small cities bordering on São Paulo, considered part of the São Paulo metropolitan area.

ANSP project intends to establish six bidirectional 1 Gbps channels between USP and NAP of Brazil and four 1 Gbps channels between NAP of Brazil and Impsat. Figure 2 below shows the available fiber pairs linking each site and their approximate lengths.

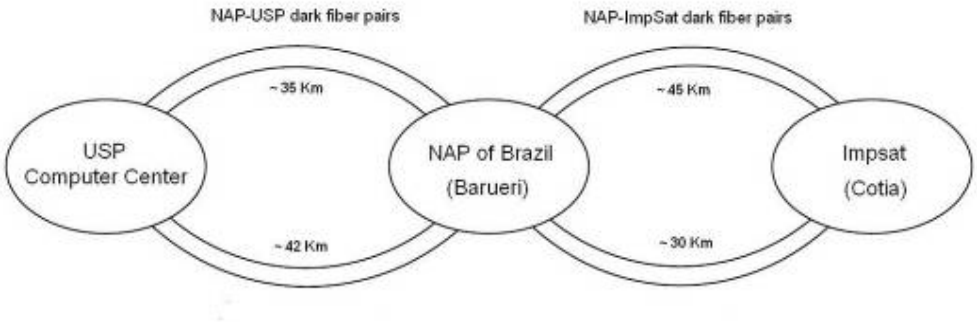

Figure 2: Dark fiber pairs managed by ANSP

This project is already partially implemented, using compact, modular WDM systems provided by MRV Communications Inc [4]. Since October 2006 there are four 1 Gbps links between USP and NAP of Brazil. One of these lambdas is being used by SPRACE. Figure 3 shows a detailed drawing of USP-NAP connection at the present time.

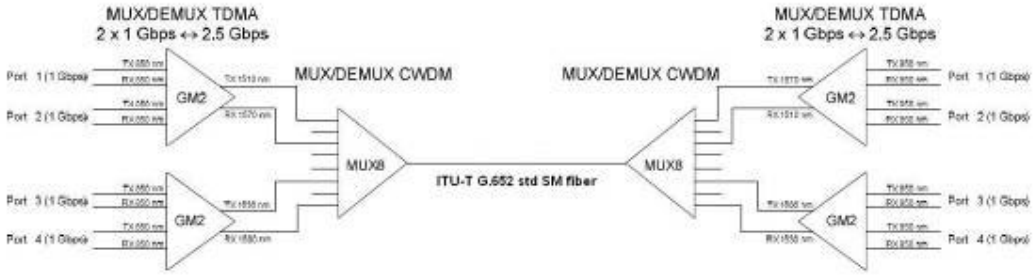

Figure 3: WDM link between NAP of Brazil and USP Computer Center

An MRV Lambda Driver LD800 chassis was installed at each site. According to Figure 3, the EM2009-GM2 module multiplex two 1 Gbps channels into one single 2.5 Gbps channel, which is then plugged into the EM800-MUX8/CW CWDM multiplex-demultiplex module. A management module and a pair of redundant power supplies complete the configuration of each LD800 chassis. As one of the lambdas is still available, during SC'06 we used a second lambda to connect SPRACE to NAP of Brazil, for performing bandwidth/latency tests. More recently we have used both lambdas for SPRACE-NAP link troubleshooting.

A third MRV chassis is planned to be installed at Impsat, as well as some extra modules at each site. The complete USP–NAP–Impsat WDM network is planned to be as shown in Figure 4.

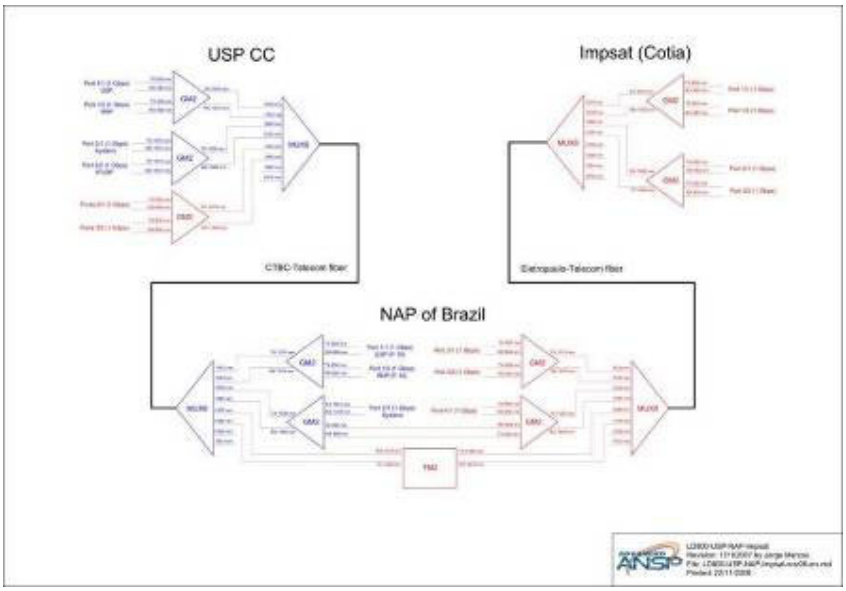

Figure 4: USP-NAP-Impsat WDM link upgrade planning

As can be seen in Figure 4, elements of the project already implemented are shown in blue, whereas the remaining pieces, to be soon implemented, are highlighted in red. MRV module TM2-SFP will act as a repeater to boost the signal between USP and Impsat and is planned to be used by SPRACE. The use of this module prevents SPRACE channels from being de-multiplexed and re-multiplexed at NAP of Brazil, since the main target is the Cisco ONS-15454 installed in Cotia.

In search for the best timely and cost-effective solution for a MAN topology, ANSP engineers are also studying two other alternatives to the project detailed above. The first one considers using WDM equipments provided by PadTec [5], a local company which develops similar WDM systems that can be used instead of MRV boxes, possibly at a fraction of the price. The second alternative is still in its early development stages and is highly dependent on an agreement among ANSP, Terremark, AMPATH and LANautilus. It consists of transferring the AMPATH access point from Impsat data center in Cotia to NAP of Brazil in Barueri. It is worth to say that this is not a technical issue, so we avoid diving into this discussion. However, this alternative, if succeeded, would be undoubtedly beneficial to SPRACE, as traffic exchange between our main servers and the ONS-15454 would travel only 40 km. This would possibly allow us to reach Abilene in only three hops, a quite remarkable achievement.

Another very interesting feature related to the MAN connectivity devised by ANSP is its flexibility to support higher bandwidth demands. A prominent example of this flexibility is demonstrated by the fact that we can easily upgrade the entire link between SPRACE and the ONS-15454, by just plugging MRV model TM-DXFP 10 Gbps transponders directly into the LD800 boxes, thus enabling SPRACE to fully utilize the WHREN-LILA link upgrade to 10 Gbps, being planned to occur in 2008.

SPRACE and the UltraLight Collaboration

Due to the data-intensive computing challenges imposed by its complex experiments and the nature of its extremely large international collaboration spread around the globe, the High Energy Physics community has long been a demanding user of leading edge, long-range networks. As this tradition is set

to continue in the future, the development of advanced network services is thus an essential task to be included in the community research & development schedule. High-performance network links are the foundation of the Grid systems that will drive the scientific discoveries in the next decades.

Although networks are essential to Grid-enabled systems, network resources have not yet been considered full participants, implemented as directly addressable and reconfigurable resources in those environments, but instead used as generic, external resources. The challenge of implementing core network nodes that can be fully monitored and administered by Grid processes is addressed by the UltraLight project.

SPRACE has been participating as a South American active member site of the UltraLight project since 2005. As described earlier in the historical overview, SPRACE international network connection has been suffering successive improvements due to the combined efforts of ANSP, RNP, Caltech, Florida International University and Fermilab. The only reasonable way for SPRACE to address the challenges imposed by the extremely fast network advancements is to work in a close partnership with leading researchers in the area. In this sense, partnership with UltraLight has been of primary importance to us, and UltraLight staff members and researchers have been providing an exceptional support to local researchers. This can be easily seen by the results we have achieved during recent bandwidth demonstrations: thanks to the UltraLight researchers' strong support, Brazilian research groups are being able to participate in such global demonstrations for the first time. As a remarkable example, during the Bandwidth Challenge at SuperComputing 2004 event, a new record was set for a Latin American R&E network, where data transfer throughputs reached 2.94 Gbps (1.96 incoming + 0.98 outgoing) and were sustained for almost 1 hour between the conference center floor in Pittsburg and servers located in São Paulo and Rio de Janeiro. By that time current network infrastructure between ANSP and USP was not yet available, so we had to collocate three of our compute nodes at NAP of Brazil. Nowadays, data transfer rates with such magnitudes can be easily achieved directly from SPRACE data center, thanks to the flexible metropolitan area network built by ANSP.

Figure 5 below shows a detailed view of the network connections between SPRACE facilities and AMPATH, our entry point to the UltraLight network. The ANSP metropolitan area network infrastructure, as described earlier and shown in more detail in this figure, directs SPRACE traffic through its routers up to the wide area network provided by the WHREN-LILA link, which directly connects São Paulo to Miami. For the sake of clarity, in this figure we have kept the same color patterns used in Figure 1: green lines indicate network path followed by SPRACE facilities to reach the WHREN-LILA link, whereas red lines indicate the network path followed by HEP facilities located in Rio Janeiro. Network traffic between São Paulo and Rio de Janeiro flows through the RNP backbone.

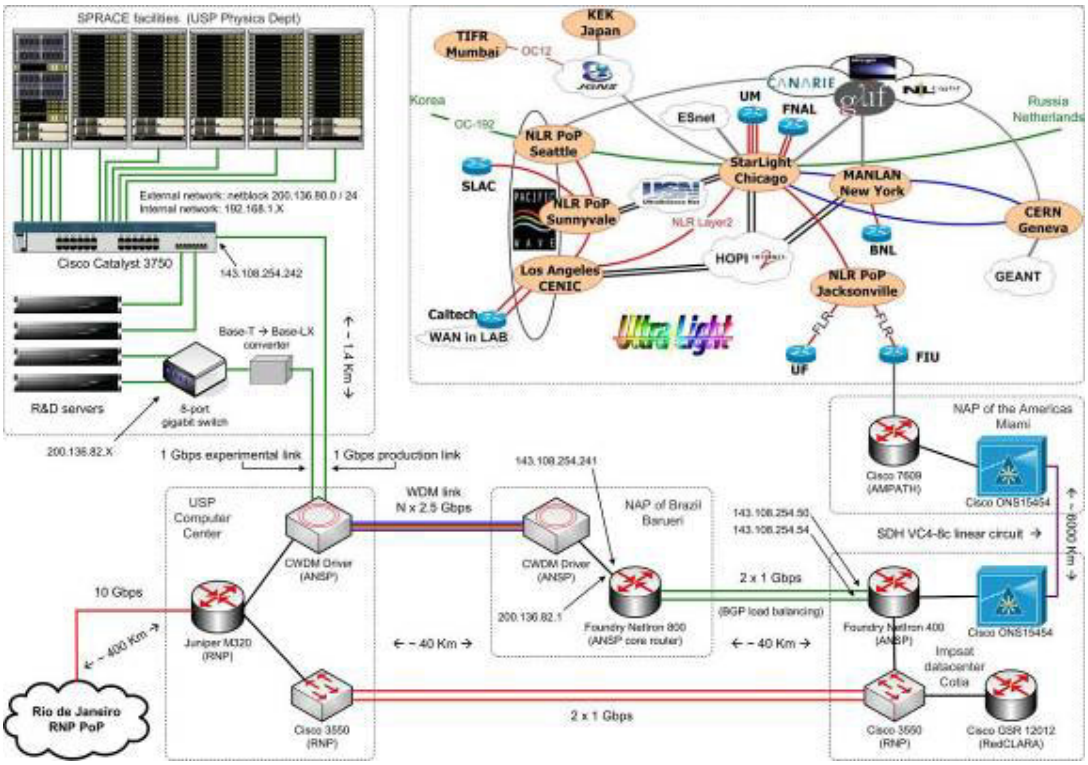

Figure 5: Detailed drawing of the network path from SPRACE to UltraLight

Figure 5 also shows a detailed view of SPRACE internal connections. The Cisco Catalyst switch provides connectivity to almost all servers in our data center. It is logically connected to one of ANSP core routers (Foundry NetIron 800) through a point-to-point network provided by ANSP (143.108.254.240/30). ANSP has also provided us with a full Class C netblock (200.136.80.0/24), which is resolved by the Cisco Catalyst and used by all our servers. The cluster production servers and storage systems are installed into six 23U racks, whereas the R&D servers are installed into a separate 37U rack. Thanks to a strong and fruitful partnership with Intel Brazil, we were able to get some extra servers for research and development, apart from the SPRACE production servers. Available R&D servers include (as of Dec/06):

- Intel dual-Xeon 2.4 GHz / 1 MB L2 cache server, 2 GB memory, 1 x 36 GB SCSI disk (Supermicro X5DPA-8GG platform)

- Intel dual-Xeon 2.8 GHz / 2 MB L2 cache server, 1 GB memory 2 x 160 GB SATA disks (Intel SE7520JR2 platform)

- Intel dual-Itanium2 1.4 GHz / 8 MB L3 cache, 2 GB memory, 2 x 36 GB SCSI disks (Intel SR870BH2 platform), PCI-X 10 Gbps Neterion XFrame II SR adapter

- Intel Pentium D dual-core / 2 MB L2 cache, 2 GB DDR2 667 MHz memory, 4 x 160 GB SATA disks, integrated PCI-e 1 Gbps adapter (Intel SE7230NH1-E platform)

As can be seen in Figure 5, two of these servers (the dual-Xeon ones) are directly attached to the production link, whereas the other two are attached to the experimental link through a small 8-port gigabit Ethernet switch and a 1000Base-T/1000Base-LX converter. This setup is quite flexible and has enabled us to establish network connectivity tests through a closed path that includes SPRACE Cisco Catalyst, two different lambdas in the WDM Layer 2 link, and ANSP core router at NAP of Brazil. Performance tests through the USP-NAP link is routinely performed, enabling us to check for eventual bottlenecks and helping us on setting suitable values for the Linux kernel servers' TCP tuning parameters, without disturbing the production servers. One of those R&D servers also runs the Fast Data Transfer - FDT service [6], a Java application developed by Dr. Iosif Legrand from Caltech that uses Java network I/O libraries to achieve efficient data transfers at disk speed coordinated with smooth data flow across wide-area networks.

Figure 6 below shows a schematic drawing of the SPRACE data center.

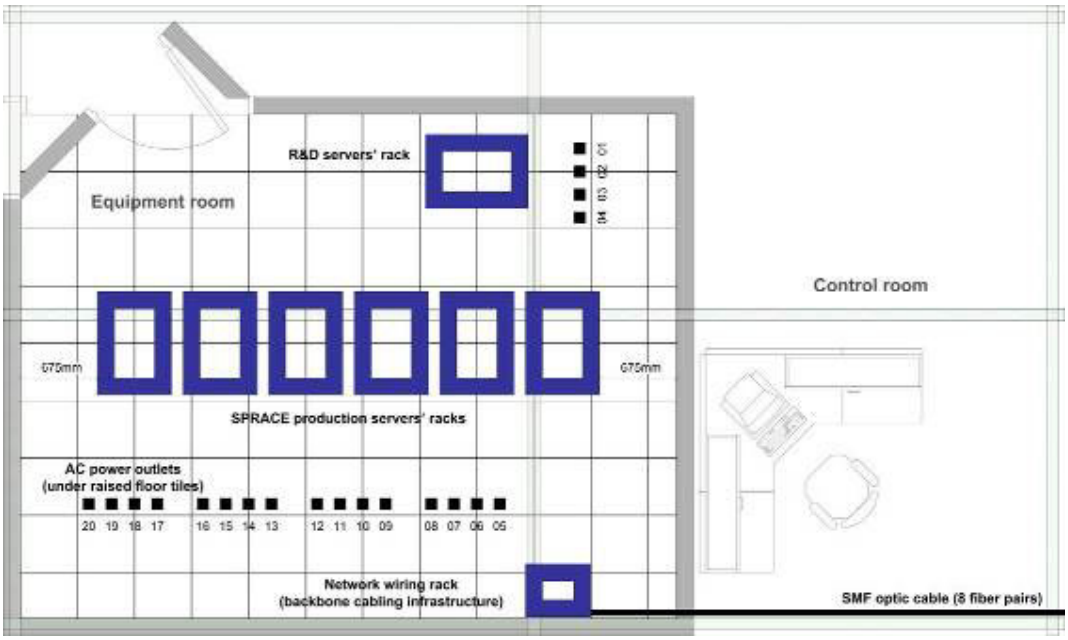

Figure 6: Schematic drawing of SPRACE data center

Another server not mentioned so far is the 'spruna' dual-Xeon 2.4 GHz server, the first donation we got from Intel, in 2004. Although this machine is installed together with the production servers, it is considered a hybrid production and R&D server: it is treated as a production system (stable Linux kernel, up and running all the time, and so on), but mainly used for research purposes. This server is totally dedicated to UltraLight project (this can be seen by the peculiar name we gave to it, an acronym which stands for SPrace Research server for hosting Ultralight Network related Applications).

Although 'spruna' is considered an R&D server, it has the status of a production server, as its health and the services it runs are being constantly monitored. This is because 'spruna' runs a series of services that should be up all the time, such as the MonALISA UltraLight service [7], the PingER service (developed at the Stanford Linear Accelerator Center - SLAC, to monitor end-to-end performance of

Internet links) [8], and a web traceroute service (also developed at SLAC) [9]. This server is also being configured to run the RANCID network management application tool to track and record the configuration changes we usually make in our Cisco Catalyst switch. Such a configuration is already in its final stages.

Besides its essential role as the SPRACE gateway, working as our border router for both ANSP network and the GIGA project network, the Cisco Catalyst switch is also a key player in the UltraLight project, as its configuration setup includes SNMP community 'ultralight' enabled on it. This was accomplished in July 2005. Since then, SPRACE link connectivity is being constantly monitored and our traffic load can be seen at the MonALISA UltraLight Repository web page. Although not yet done, our plans also include enabling syslog monitoring on the Catalyst, for making the task of network problem tracking and diagnosing easier than it is nowadays.

As soon as the WHREN-LILA upgrade to 10 Gbps occurs, by 2008, a significant achievement will be the establishment of a 10 Gbps connection between SPRACE and UltraLight sites in US, using a dedicated lightpath through ANSP optical network infrastructure, terminating at SPRACE in the form of a 10 gigabit Ethernet port in a module installed at the Cisco Catalyst switch. Although not an easy goal, the path to this accomplishment is already being paved.

References

[1] http://whren.ampath.net/

[2] http://www.napoftheamericas.net/

[3] http://www.lanautilus.com/

[4] http://www.mrv.com/

[5] http://www.padtec.com.br/eng/

[6] http://monalisa.cern.ch/FDT/

[7] http://monalisa-ul.caltech.edu:8080/

[8] http://www-iepm.slac.stanford.edu/pinger/

[9] http://www.slac.stanford.edu/comp/net/wan-mon/traceroute-srv.html

Apêndice II

"Strategic Issues in Cyberinfrastructure for the 21st Century Science & Engineering Research"

October 18, 2007

08h45 - 09h15:	***Opening Session*** Prof. Dr. Gil da Costa Marques *Instituto de Física - USP* *Coordenador da CTI/USP* Profa. Dra. Tereza Cristina M. B. Carvalho *Depto. de Engenharia de Computação e Sistema Digitais - EPUSP* *Diretora do CCE/USP*
09h15 - 10h00:	***"Evolução de Galáxias e Estruturas em Grandes Escalas: Desafios Computacionais"*** Prof. Dr. Laerte Sodré Júnior *Depto. de Astronomia* *Instituto de Astronomia, Geofísica e Ciências Atmosféricas - USP*
10h00 - 10h45:	***"São Paulo Regional Grid"*** Prof. Dr. Sérgio Ferraz Novaes *Instituto de Física Teórica - UNESP* *Coordenador dos projetos SPRACE e GridUNESP*
10h45 - 11h00:	***Coffee Break***
11h00 - 11h45:	***"Aplicações de HPC na Indústria Offshore e de Óleo e Gás"*** Prof. Dr. Kazuo Nishimoto *Depto. de Engenharia Naval e Oceânica - EPUSP* *Coordenador do Laboratório TPN (Tanque de Provas Numérico)*
11h45 - 12h30:	***"Modelos Climáticos de Simulação Global"*** Profa. Dra. Ilana Elazari Klein Coaracy Wainer *Depto. de Oceanografia Física* *Instituto Oceanográfico - USP*
12h30 - 14h00:	***Lunch***
14h00 - 15h00:	***"Sun Strategies for Education and Research in High Performance Computing"*** Sun Microsystems Presentation
15h00 - 15h30:	**Coffee Break**
15h30 - 17h00:	***"IBM High Performance Computing Solutions: BlueGene, Cell BE and Linux Clusters"*** IBM Presentation
17h00 - 18h00:	***"A Multi-application, Multi-workflow Approach to High Performance Computing"*** SGI Presentation
18h00 - 18h30:	***Closing remarks & Informal discussions***

Apêndice III

Program: HELEN
High Energy Physics LatinAmerican - European Network
THIRD YEAR: August 1st, 2007 - July 31st, 2008

FIRST CALL

APPLICATION FOR GRANTS
Deadline: October 20 , 2007

Introduction

The LHC experiments at CERN will need to exploit Petabytes of information in the incoming years. This huge requirement has been pushing physicists, engineers and computer scientists to develop new concepts, services, and tools for enabling the reliable sharing of distributed processing power and storage capacity. As a result, many developments have been achieved on the data acquisition, storage and analysis of the LHC data, as well as on data transmission and distributed processing in the recent years. The most prominent examples of this worldwide effort are the EGEE (Enabling Grids for E-sciencE) and the OSG (Open Science Grid), the two largest, worldwide distributed computing infrastructures, based on the most recent advances in Grid technology. Both initiatives are focused at developing shared computational resources distributed in a global scale that can be viewed as a single computing infrastructure for serving the most demanding applications. The Grid is thus the answer to the need of the world particle physics community for the coming years.

As it is widely known, the LHC Computing Grid model consists of a hierarchical arrangement of tiered regional centers, starting from Tier-0 at CERN, to major Tier-1 centers in several Countries, to smaller Tier-2 and Tier-3 centers in Institutes and Departments of collaborating members. It is expected that the Tier-2 regional centers will play a very important role in this arrangement. However, the dramatic growth in complexity of the software and the underlying hardware needed to operate a Tier-2 class regional center has been pushing several research group's system management capabilities to its limits, bringing them the responsibility to urgently understand more deeply how the internal gearings of those complex computing infrastructures really work. Also, it has been prominent to develop efficient training tools to efficiently and painlessly share and disseminate the knowledge of using such infrastructures with first-time users and site administrators.

The role of a Tier 2 regional center is thus to enable the physicists at its Institute to perform their science. However, it does not make sense to own huge amounts of computing and storage resources if the software infrastructure required by the experiments is not well supported. The provision of core

services is crucial to the existence of a Tier-2 center. It needs to be robust and reliable. It needs an adequate support team who is able to define its operational goals and ensure that those goals are implemented and deployed in an effectively and timely manner. Our own experience has shown us that a team of two or three under-trained people is absolutely not enough to maintain the core services of a Tier-2 center.

Motivation and work plan

SPRACE is a typical Tier-2 site, with Grid-enabled computing and storage elements, and the minimum network bandwidth necessary to operate adequately. SPRACE facilities include a 240 processor computing farm, with around 14 TB of raw disk space, that was initially designed for Monte Carlo production and event reprocessing for the Fermilab's D0 Collaboration. However, its resources and technical expertise have been, slowly but consistently, upgraded to allow the full exploitation of the facility for the CMS Collaboration. One important goal was achieved in August 2005, when we joined the Open Science Grid community. Almost one year later, on July 2006, SPRACE was included in the list of Tier-2 sites for the CMS Collaboration, associated to the American Fermilab Tier1 site. Thus, mainly for historical reasons, SPRACE is strongly connected to Fermilab's Tier-1 center. Indeed, the Brazilian HEP community is the main responsible for the existence of a 2.5 Gbps academic link between Sao Paulo and Miami, which is planned to grow up to 10 Gbps in one or two years. All SPRACE members are well committed to direct efforts to engage the center into the Worldwide LHC Grid Computing initiative, in order to be ready to fully contribute for the CMS collaboration by 2008.

Up to now, most of SPRACE resources have been administered by a couple of physicists working as system managers and me, a graduate student in computer engineering with a background in physics. None of us have background in computer science. Thus, this tiny but brave group has been suffering increasing strain due to the increasing complexity of the middleware management and the constant infrastructure interventions required as the size of the farm increases. Examples of our computer-related day-to-day activities include network design and deployment, system administration and security, maintenance of the communication assets such as the multimedia collaborative resources, planning and management of the hardware and software lifecycle, as well as the planning, installation and administration of the Grid infrastructure middleware.

It is becoming clear that such under-provision of system support is unsustainable in the long run. If it was not enough, a significant amount of computing resources is about to be deployed at São Paulo State University – UNESP, due to the so called GridUNESP project. This will provide a 256-quad-processor-node central cluster plus seven 16 quad-processor-node peripheral clusters, and a 100 TB central storage system. The Grid infrastructure is planned to be based on OSG middleware, with internal batch scheduling based on a combination of Condor and Sun Grid Engine. Currently, UNESP has no technical staff with enough expertise to handle this huge enterprise, so new staff should be selected and trained. It is our group's responsibility to take care of this project. At the same time, SPRACE Tier-2 also needs to grow up, so we are planning to double its processing capability and surpass the 200 TB storage capacity within the next two years. It is absolutely clear that our research group claims for professionally-qualified and well-trained Grid computing experts.

IT should be viewed as an instrument, not a goal, for generating, processing and distributing information to aid the research group in achieving its goals. Thus, a valued IT professional should not only understand the technical aspects of the infrastructure, but also have a common sense of the organizational facets concerning the whole collaboration, and at least a broad view of the huge

complexity involved in the data acquisition and analysis of the physical experiments. This knowledge should help him to grasp the extent of information technology dependency of modern e-Science experiments. I personally think that it simply is not possible to quickly acquire such a broad view without an immersive stay at CERN.

In fact, not only the experiments, but the whole collaboration is completely dependent on information technology. We need to have appropriate systems in the correct places and in time for things to happen effectively. Those systems must work properly and securely, as well as maintained, upgraded and replaced as appropriate. The collaboration researchers' require support from IT staff who must deeply understand computer systems and the software they run. Above all, IT staff should be committed to solving whatever computer-related problems the group might have. They should be trained to address those needs, and I think the best and quickest way is to immerse them in the collaboration, until they acquire the right combination of knowledge and practical hands-on expertise to take care of the research group's computing infrastructure and be helpful to the researchers who effectively use it in their analyses. Because of the complexities involved, computing professionals need to be hard-trained to become experts in the broad field of emerging Grid technologies for contemporary e-Science.

The computer specialist should play a key role in determining the requirements for the research group's information systems and should be active in the hardware and software specification, design, deployment, support and maintenance. As a result, such professional requires a deep understanding of the organizational principles and practices of the whole collaboration, so he can serve as an effective bridge between scientists and the technological tools they need to do better what they were trained to do: Science.

The main goal of my work at SPRACE nowadays has been targeted at trying to answer the challenge I have imposed to myself: as a graduate student in computer engineering, what strategies can I use to become a more effective and efficient collaborator? To assume the responsibility to work as a Tier-2 site operations' manager, who will be primarily responsible for the research team's computer infrastructure, the main goal of my stay at CERN will then be focused at working on a Tier-2 Center operations' planning document. It is imperative, to the survival of SPRACE as a regional center in the long run, to rigidly define operational goals, and ensure that those predefined goals are implemented and deployed in a timely and effective manner.

Research topics

Several areas of research in science, medicine, and engineering in general, and HEP in particular, are nowadays requiring higher and higher data transfer rates between widely separated sites. There is an increasing need for moving substantial volumes of data between hosts over long periods of time, from several hours to even days. These requirements are rather difficult to achieve using conventional packet-switched networks such as the Internet. In such special cases, it is much better to establish dedicated circuits between each pair of interacting hosts during the data transfer period. To meet those requirements, several NRENs are collaborating to create intelligent, dynamically provisioned optical network infrastructures. A prominent example is the Global Lambda Infrastructure Facility (GLIF), an experimental optical network testbed to leverage the use of circuit-switched networks. GLIF project is intended for experimental research rather than production. As a Ph.D. student, this research theme is the main focus of my graduate studies, which is, to design, implement and test control plane tools and services to enable application-provisioned dynamic circuits of required bandwidth, that can be adjusted between various end-points of the network, for mission-oriented priority data transfers.

One possible drawback of dedicated circuits is their administrative overhead. Thus, by using dynamically provisioned circuits (i.e., circuits that are requested at the application level), the network user can concentrate on its primary task - data transfers - without having to be concerned with network provisioning details or transport protocol efficiency.

Thinking of Grid infrastructures not only as a bunch of hardware and software components distributed over wide areas, but instead as a single entity, or a cyberinfrastructure, we need to effectively integrate their constituent elements with sophisticated internetworking elements. Projects led by Caltech, such as the development of FastTCP and Fast Data Transport (FDT), allied to their vast experience with next generation standards-based SONET/SDH protocols, such as the Virtual Concatenation (VCAT), the Link Capacity Adjustment Scheme (LCAS), the General Framing Protocol (GFP), and others, are really effective to long distance international links and towards maximizing their utilization. As SPRACE is more than 6000 Km far from its hierarchically superior Tier – Fermilab – this kind of knowledge is of prime importance to us. As an example, an interesting planning could be to work collaboratively to extend data flows that go from and to Sao Paulo over the international link through Miami, and across the "Atlantic Wave" peering infrastructure, to reach the IRNC-awarded links across the Atlantic Ocean to and from CERN.

Both the UltraLight[1] and the US LHCNet[2] networks are planning to deploy dynamically provisioned circuits across their infrastructures. UltraLight has already demonstrated dynamic circuits using MonALISA/VINCI, using programmable optical switches. US LHCNet will take the task further by deploying adjustable bandwidth dynamic circuits over its newly-deployed VCAT/LCAS/GFP capable infrastructure. Through an active participation in these projects, the Brazilian HEP community can effectively contribute to the hard task of delivering high-speed data flows to South America in the coming years. This will ultimately contribute to guarantee the participation of Latin American institutions in the LHC experiments[3] .

[1] UltraLight – a hybrid network testbed for data-intensive research – http://www.ultralight.org

[2] US LHCNet – the transatlantic network linking Geneva to Chicago and New York for supporting the US researchers working on CERN LHC experiments

[3] Almost "ipsis literes" citation of Professor Harvey Newman's presentations, speeches, and talks

5. Referências

Bibliografia consultada

- Foster, I., “What is the Grid? A three point checklist”, Grid Today, Jul 2002.
- Foster, I., “Service-Oriented Science”, Science, Vol. 308, No. 5723, pp 814-17, 2005.
- Berman, F., Hey, A. J. G., Fox, G. C., “Grid Computing: Making the Global Infrastructure a Reality”, John Wiley & Sons, 2003.
- Foster, I., Kesselman, K.,“The Grid: Blueprint for a New Computing Infrastructure”, 2nd ed., Morgan Kaufmann, 2004.
- Novaes, S. F., Gregores, E. M., “Da Internet ao Grid: A globalização do processamento”, Editora UNESP, 2004.

Citações

[1] http://www.sprace.org.br/eem/

[2] http://www.webex.com/

[3] http://www.nagios.org/

[4] http://ipmitool.sourceforge.net/

[5] http://trac.edgewall.org/

[6] http://www.roma1.infn.it/exp/helen/

São Paulo, 4 de dezembro de 2007

_____

Rogério Luiz Iope

Funcional Nº 1712076

_____

Prof. Gil da Costa Marques
Coordenador
Centro de Ensino e Pesquisa Aplicada
Departamento de Física Experimental
Universidade de São Paulo