# Integração do Cluster Enterprise ao Grid Sinergia

Renato Luiz de Freitas Cunha[1] Alberto Ferreira De Souza

*Universidade Federal do Espírito Santo*
*{renatoc,albertodesouza}@gmail.com*

## Resumo

*Neste trabalho apresentaremos a experiência e os métodos utilizados para a integração do cluster de 64 processadores do Departamento de Informática da UFES ao grid Sinergia, que hoje disponibiliza os recursos computacionais do cluster Enterprise e agrega os recursos disponíveis nas instituições que o compõem.*

## 1. Introdução

A maioria dos sucessos alcançados ultimamente em computação em grade (*grid computing* [1]) está relacionada com problemas que requerem muito pouca comunicação entre processos, por exemplo, aplicações que precisam analisar uma grande quantidade de dados como Genoma [2], SETI@Home [3] etc. Essa classe de aplicações, contudo, é pequena (tanto em termos de número como de impacto econômico e industrial) em comparação com classes mais gerais em que aplicações requerem um grau bem maior de comunicação [4].

Este trabalho apresenta resultados parciais de um projeto de pesquisa cujo objetivo é facilitar a solução de uma classe maior de problemas que podem tirar proveito da computação em grids, mas também que podem fazer isso numa infra-estrutura de recursos computacionais já implantados, conectados por uma rede de alta velocidade.

Apresentaremos, também, soluções para os problemas encontrados durante o processo, como a utilização de um escalonador não suportado nativamente pelo *middleware* do Grid, integração entre os serviços de monitoração, adequação à política de segurança do Departamento de Informática da UFES, e utilização dos serviços do grid através de um firewall.

## 1.1 Cluster Enterprise

O cluster Enterprise é o cluster computacional do LCAD da UFES e é composto por 64 nós de processamento e um nó de administração (master). Com 256 MB de memória e 20 GB de armazenamento por nó, o cluster totaliza 16 GB de memória RAM e 1.2 TB de armazenamento.

## 1.2 Grid Sinergia

O grid Sinergia é uma iniciativa multi-institucional para projetar, desenvolver e disponibilizar uma infraestrutura de grid (recursos e *middleware*), explorando a rede óptica de alta performance da RNP (Rede GIGA) para a execução eficiente de aplicações científicas paralelas ou distribuídas. Mais especificamente, esse grid é composto pela seguintes instituições: IC/UFF, CAT/CBPF, LNCC, DI/PUC-Rio, IC/UNICAMP e DI/UFES.

O grid Sinergia também possui um portal de informações disponível em http://easygrid.ic.uff.br/GSin/GridSinergia.html.

## 2. Ferramentas

Esta seção descreve as principais ferramentas e técnicas utilizadas para a integração entre o cluster Enterprise e o grid Sinergia.

### 2.1. Globus Toolkit

Na integração do cluster Enterprise com o grid Sinergia, a principal ferramenta utilizada foi o Globus Toolkit [5]. O Globus Toolkit disponibiliza um conjunto de serviços que facilitam a construção de infra-estruturas para computação em grid. O Globus Toolkit possui três componentes principais, que são responsáveis pela Gerência de Recursos, Gerência de Serviços de Informação e Gerência de Dados [6].

**Gerência de Recursos**: O *Grid Resource Allocator Manager* (GRAM) e o *Global Access to Secondary Storage* (GASS) são os componentes primários de gerência de recursos em grids suportados pelo Globus Toolkit. O módulo GRAM, que é

1 Bolsista PIBIC-CNPq/UFES

implementado através de um *daemon* de nome *gatekeeper*, provê a execução remota e a gerência do estado de uma aplicação. Quando um *job* (requisição de um usuário por recursos computacionais) é submetido por um cliente, a requisição é enviada ao servidor remoto e tratada pelo *gatekeeper*. Este daemon é similar ao inetd [12] e serve para autenticar os pedidos de entrada utilizando a camada de segurança do GSI (*Grid Security Infrastructure*). A autenticação é feita mapeando o usuário remoto a uma conta de usuário local. Após a autenticação, o *gatekeeper* cria um gerenciador de *job* que inicia e monitora o *job* até o seu término. Quando o *job* é terminado, o gerenciador de *job* envia a informação de estado de volta ao cliente e termina.

- **Gerência de Serviços de Informação:** Baseados no *Lightweight Directory Access Protocol* (LDAP), os componentes *Grid Resource Information Service* (GRIS) e *Grid Index Information Service* (GIIS) são configurados de maneira hierárquica de modo a adquirir informação a respeito do grid e distribuí-la. A esses dois serviços dá-se o nome *Monitoring and Discovery Service* (MDS). A informação coletada pode ser tanto informação estática sobre as máquinas quanto informação dinâmica, como uso do processador ou atividade de disco. Esta informação é passada do GRIS para servidores GIIS no grid.
- **Gerência de Dados:** O *Grid File Transfer Protocol* (GridFTP) é um protocolo de transferência de dados seguro de alta performance baseado no FTP, otimizado para redes com alta taxa de transferência.

Os três componentes discutidos acima constituem os pilares do Globus e foram construídos em cima da camada de segurança *Grid Security Infrastructure* (GSI). Ela provê funções de segurança para autenticação, comunicação confidencial, autorização de acesso a recursos, dentre outras. Esta infra-estrutura foi construída utilizando a *Secure Sockets Layer* (SSL) [7], criptografia de chave pública [8] e certificados X.509 [9].

## 2.2. SGE – Sun Grid Engine

Em sistemas com muitos usuários, é necessário que haja alguma forma de gerenciamento de recursos para que esses não sejam dominados por um único usuário e, no caso desses recursos serem computacionais, para que todos possam executar os seus jobs. Caso todos os recursos estejam ocupados, é necessário manter uma fila para que à medida que recursos vão sendo liberados, outros usuários possam ter acesso aos mesmos.

O cluster Enterprise utiliza o *Sun Grid Engine* (SGE [10]) para o balanceamento de carga. O SGE orquestra a entrega de poder computacional baseado em políticas definidas pela equipe de gerência do cluster, no caso, a equipe do Laboratório de Computação de Alto Desempenho (LCAD) do Departamento de Informática (DI) da UFES. O SGE utiliza as políticas definidas para examinar os recursos computacionais disponíveis dentro do cluster, obtém esses recursos e depois os aloca e disponibiliza automaticamente de uma forma que otimize o uso desses recursos.

É tarefa do SGE aceitar jobs submetidos para o cluster, colocá-los em uma área de espera até que eles possam ser executados, enviá-los da área de espera pela liberação do dispositivo de execução, gerenciá-los durante a execução e manter um registro da execução quando ele for concluído [10].

## 2.3. NAT – Network Address Translation

Nas redes de computadores, às vezes é necessário utilizar a tradução de endereço de rede (NAT), que envolve a reescrita dos endereços de origem e/ou destino dos pacotes IP quando eles passam por um roteador ou por um firewall. NAT geralmente é usada para permitir que vários computadores em uma rede privada possam ter acesso à internet utilizando um único endereço IP externo.

Normalmente, numa rede utilizando NAT, a rede local utiliza endereços IP pertencentes à classe das redes ditas privadas com um roteador/firewall inserido naquela rede. O roteador também fica conectado à internet utilizando um endereço IP público. À medida que o tráfego passa da rede local para a Internet, o endereço de origem de cada pacote tem seu valor traduzido de endereços IP da rede privada para o endereço IP público. O roteador, então, mantém dados a respeito de cada conexão ativa, armazenando particularmente o endereço e a porta de destino. Quando uma resposta é recebida, o roteador usa os dados registrados durante a fase de saída dos pacotes para determinar para que local da rede privada o pacote deve ser enviado e, novamente, traduz dinamicamente o endereço de destino do pacote para entregá-lo na rede interna. Para um observador de fora, parece que todo este tráfego de informação tem sua origem e destino no roteador.

## 2.4 Ganglia

Ganglia possui uma arquitetura distribuída que permite monitorar sistemas como clusters de larga escala e grids computacionais compostas por federações de clusters [11].

A implementação de Ganglia consiste em dois tipos de processos *daemons* (*gmond* e *gmetad*), um programa de linha de comando (*gmetric*) e uma biblioteca para implementação de clientes. A monitoração de um único cluster é realizada por *gmond* (*Ganglia Monitoring Daemon*), sendo que este deve estar presente em todos os nós. O processo *gmetad* é responsável por monitorar uma federação de clusters. Uma árvore de conexões entre vários daemons *gmetad* permite agregar as informações de vários clusters. O programa *gmetric*, por sua vez, permite estender as métricas monitoradas por Ganglia [11].

## 3. Métodos

### 3.1. Instalação e configuração do Ganglia

No cluster Enterprise, quando um usuário ou administrador desejava conhecer o *status* do cluster, este deveria utilizar uma ferramenta que apresentava estatísticas de nós de processamento ocupados com jobs, livres e/ou com erros (nesse caso uma lista contendo os nós com erro ou desligados era exibida). Outra alternativa era utilizar a ferramenta de linha de comando *qstat* do SGE para obter informações sobre o cluster, mas era necessário que o usuário conhecesse sua sintaxe.

Visando simplificar essa tarefa, foi instalado o Ganglia (o módulo *gmond*) nos nós do cluster e as informações obtidas são então publicadas na página do LCAD.

A escolha do Ganglia foi feita por ele ser um sistema de código aberto, gratuito, com uma grande comunidade de usuários e, principalmente, por fornecer uma maneira simples de se exibir na web os dados coletados.

### 3.2. Instalação e configuração do Globus Toolkit

Objetivando automatizar a instalação e configuração do Globus Toolkit no cluster Enterprise, foi desenvolvido em *bash* um script que realiza o download, instalação e configuração básica do mesmo, em sua versão 2.4.3. O script permite a instalação de outras versões, bastando alterar as variáveis referentes à versão. No entanto, para manter a conformidade com a versão do Globus utilizada pelo grid Sinergia, optou-se por usar a 2.4.3. A versão 2.4.3 foi escolhida por ser a mesma utilizada pelos sites formadores do grid e por ser a versão recomendada pela equipe de suporte do grid Sinergia em seu guia de instalação.

Após a configuração, instalação e verificação de um grid de teste, foi realizada a instalação do Globus Toolkit no cluster Enterprise de acordo com o padrão estabelecido entre os parceiros do grid Sinergia.

O cluster Enterprise já possuía uma política de submissão de jobs a ser respeitada quando um job fosse submetido através do grid. Visando obedecer a essa política, instalamos um gerenciador de jobs (*jobmanager*) do Globus que o integra ao SGE, traduzindo as requisições feitas ao Globus para requisições do SGE, de modo que para o usuário a utilização e manutenção da fila tornam-se transparentes.

Outra integração necessária foi a entre o MDS do Globus e o Ganglia [13] para o cluster Enterprise, visto que o sistema de monitoramento padrão do grid é o próprio MDS do Globus e removendo o overhead de haver dois sistemas de monitoramento executando no cluster.

### 3.3. Configuração do Firewall

A política de segurança do DI da UFES proíbe que máquinas pertencentes ao cluster possuam IP "válido", impedindo que o cluster seja acessível através da internet. Inicialmente, portanto, para que a submissão de jobs para o cluster fosse realizada, era necessário que um usuário do cluster antes realizasse login remoto em uma máquina da rede do LCAD, para depois realizar login no front-end do cluster e, por fim, submeter seu job para o cluster.

Para não entrar em conflito com a política de segurança do DI, optou-se por utilizar técnicas de NAT no LCAD.

A máquina front-end do cluster (lcad10) já estava configurada para utilizar lcad1 como gateway. Utilizamos, portanto, essa configuração para possibilitar o acesso dos serviços do Globus ao cluster criando regras de NAT em lcad1, transformando-a realmente no gateway de lcad10. Assim, publicamos a informação de que lcad1 deve ser a máquina utilizada para acessar o cluster Enterprise através do grid e a máquina lcad1 se encarrega de repassar os pacotes enviados pelos serviços do Globus para o front-end do cluster.

A configuração do firewall, então, foi realizada da seguinte maneira:

- Para as redes internas (DI e LCAD), todos os pacotes que entram e saem de lcad1 são liberados, como acontecia antes da implantação do firewall;
- É habilitado, para lcad10 o acesso a toda a rede interna;
- Para os serviços de lcad1 (ssh, cvs etc.), o acesso é permitido, como antes da implantação do firewall;

- Às portas alocadas dinamicamente pelo Globus (definidas entre o intervalo de 50000 a 51000 no cluster Enterprise), é permitida a saída de pacotes;
- Às portas alocadas para os serviços do Globus (2119 para o GRAM, 2135 para o MDS e 2811 para o GridFTP) a entrada de pacotes é liberada e redirecionada para o cluster.

Foram adicionadas regras ao firewall do DI da UFES para permitir o tráfego de dados através das portas utilizadas pelo Globus em lcad1.lcad.inf.ufes.br e adicionada uma entrada no servidor de nomes (DNS) do DI que especifica que lcad10.lcad.inf.ufes.br é um apelido para lcad1.lcad.inf.ufes.br. Essa modificação é necessária, pois se há alguma diferença entre o nome do host utilizado como argumento para submissão do job e o nome do host que efetivamente trata dessa submissão, a execução é abortada pelo Globus.

### 3.4. Proteção de temperatura

Devido a falhas no sistema de refrigeração do cluster, eventualmente alguns nós de processamento sobre aqueciam e acabavam danificados devido ao excesso de temperatura.

Com esse problema em mente, foi desenvolvido um script, executado em cada nó de processamento de tempos em tempos (atualmente, de 5 em 5 minutos), que coleta a temperatura atual do processador e a compara com o limite de temperatura definido como aceitável. Caso esteja fora dos limites, então o script envia um e-mail informando aos administradores do laboratório que há problemas de temperatura no nó cuja temperatura passou do limite. Há, também, a opção de desligar o nó de processamento para evitar maiores danos.

As temperaturas coletadas também são repassadas para o Ganglia através do *gmetric*, possibilitando o armazenamento desses .

## 4. Discussão e Conclusões

Esse projeto de iniciação científica deve documentar as modificações realizadas no cluster em manuais de instalação e configuração para que o conteúdo gerado e o conhecimento obtido não se percam. Nesse sentido, um novo site para o LCAD foi criado em linguagem php, disponibilizando informações dinâmicas sobre o cluster, centralizando documentação estática e adicionando meios de contato com a administração do laboratório, permitindo maior interatividade entre administração e usuários.

Com a integração das informações sobre o status do cluster e a nova página web do LCAD, espera-se que toda a comunidade (administradores e usuários) envolvida possa ser beneficiada, pois aos administradores será permitido diagnosticar problemas mais rapidamente. Aos usuários será permitido conhecer a carga atual e o estado de cada nó independente do cluster, bem como ter acesso a um local com informações e documentação do laboratório centralizados.

## Referências


[1] I. Foster, C. Kesselman. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kauffman, 1998.

[2] Genome@home Web Page. http://www.stanford.edu/group/pandegroup/genome/

[3] SETI@home Web Page. http://www.seti.org/science/setiathome.html

[4] W. Cirne. Grids Computacionais: Arquiteturas, Tecnologias e Aplicações. Anais do Terceiro Workshop em Sistemas Computacionais de Alto Desempenho, Outubro 2002.

[5] I. Foster, C. Kesselman. Globus: A metacomputing infrastructure toolkit. International
Journal of Supercomputer Applications, vol. 11, no. 2, pg. 115–128. 1997.

[6] L. Ferreira, V. Berstis, J. Armstrong, M. Kendzierski, A. Neukoetter, M. Takagi, R. Bing-Wo, A. Amir, R. Murakawa, O. Hernandez, J. Magowan, N. Bieberstein. Introduction to Grid Computing with Globus, 2nd Edition. IBM Redbooks – http://ibm.com/redbooks, pg. 132 - 143. 2003.

[7] D. Wagner, B. Schneier. Analysis of the SSL 3.0 Protocol. The Second USENIX Workshop on Electronic Commerce Proceedings. USENIX Press, pg 29–40. 1996.

[8] A. Salomaa. Public-Key Cryptography. Springer. 1996.

[9] C. Adams, S. Lloyd. Understanding Pki. Addison-Wesley Professional. 2002.

[10] Sun Microsystems. Sun ONE Grid Engine Administration and User Guide - http://gridengine.sunsource.net/download/Manuals53/SGE53AdminUserDoc.pdf?contenttype=application/pdf

[11] M. Neves, T. Scheid, S. Charão. Monitoração de clusters com a ferramenta Ganglia: avaliação e adaptação. Anais do Sexto Workshop sobre Software Livre, pg. 271 - 276. 2005.

[12] O. Kirch, T. Dawson. Linux Network Administrator's Guide. O' Reilly. June 2000.

[13] Ganglia MDS Information Provider for MDS 2.4 - http://www.globus.org/toolkit/docs/2.4/mds/gangliaprovider.html