Capítulo 2

Boas Práticas para a Implementação e Gerência de um Centro de Supercomputação Desassistido

Albino A. Aveleda, Ricardo P. Pareto, Alvaro L.G.A. Coutinho

Núcleo Avançado de Computação de Alto Desempenho (NACAD), COPPE/UFRJ
{albino, padilha, alvaro}@nacad.ufrj.br

Abstract

The High Performance Computing Center (NACAD) of COPPE / UFRJ is a laboratory specialized in the application of high performance computing to problems of engineering and science in general. NACAD also has extensive experience in the administration, management and tools development to support the Supercomputing Center and to develop and implement innovations in the machine management environment. The present mini-course proposes to share some of these best practices adopted by NACAD-COPPE/UFRJ made in the implementation of the Lobo Carneiro supercomputer.

Resumo

O Núcleo Avançado de Computação de Alto Desempenho (NACAD) da COPPE/UFRJ é um laboratório especializado na aplicação de computação de alto desempenho a problemas de engenharia e ciências em geral. O NACAD também possui grande experiência na administração, gerência e ferramentas de apoio ao Centro de Supercomputação e de desenvolver e implementar inovações no ambiente de administração e gerência da máquina. O presente minicurso propõe compartilhar algumas dessas melhores práticas adotadas pelo NACAD-COPPE/UFRJ feitas na implantação do supercomputador Lobo Carneiro.

Palavras-chave*: segurança, consumo de energia, portal de usuários, Internet das Coisas, indústria 4.0, aprendizado de máquina*

2.1. Introdução

Este trabalho tem a proposta de compartilhar algumas das melhores práticas para o gerenciamento e desenvolvimento de ferramentas de apoio ao Centro de Supercomputação desassistido que hospeda o supercomputador Lobo Carneiro. O foco do trabalho é mostrar o desenvolvimento e implementação das inovações no ambiente de administração e gerência do supercomputador.

Em um ambiente de computação de alto desempenho, normalmente a maior quantidade de ativos são os computadores. Eles foram e são uma das mais importantes ferramentas utilizadas que impulsionam o atual desenvolvimento tecnológico. Então, por que não usar esse potencial e desenvolver um ambiente desassistido que possa atuar de forma mais inteligente e automatizada? Dessa maneira torna-se possível que o sistema atue em caso de alguma falha no processo e se adapte para mitigar seus efeitos. A fim de viabilizar esse tipo de arquitetura foi desenvolvido um ambiente que integra várias tecnologias emergentes, tais como: internet das coisas, computação em nuvem, aplicativos (celulares e *tablets*) etc. A integração de várias dessas tecnologias também é conhecida como indústria 4.0.

O termo indústria 4.0 se originou a partir de um projeto de estratégias do governo alemão voltadas à tecnologia. "*Estamos a bordo de uma revolução tecnológica que transformará fundamentalmente a forma como vivemos, trabalhamos e nos relacionamos. Em sua escala, alcance e complexidade, a transformação será diferente de qualquer coisa que o ser humano tenha experimentado antes*", diz Klaus Schwab [Schwab]. Seu fundamento básico implica que conectando máquinas, sistemas e ativos, torna-se possível criar redes inteligentes ao longo de toda a cadeia de valor que podem controlar os módulos da produção de forma autônoma. Ou seja, as fábricas inteligentes terão a capacidade e autonomia para agendar manutenções, prever falhas nos processos e se adaptar aos requisitos e mudanças não planejadas na produção.

No nosso caso especifico, o uso dessa tecnologia é para preservar o investimento fazendo com que o supercomputador opere com segurança e baixo consumo de energia, dentro da faixa definida pelo fornecedor. Caso ocorra algum problema durante a operação desassistida, o sistema de controle irá interferir de forma autônoma para preservar a máquina. Podem-se citar alguns exemplos de problemas e ações de controle, ou seja:

- Falta de fornecimento de energia elétrica: apesar do nobreak manter o supercomputador funcionando, é necessário monitorar a temperatura e a carga das baterias do nobreak, a fim de permitir um desligamento correto das máquinas, principalmente no que se refere a manutenção da integridade dos arquivos contidos na área de armazenamento;
- Problemas na refrigeração: é necessário o monitoramento da temperatura do ambiente, independente da falta de energia, para evitar que caso ocorra algum problema na refrigeração o supercomputador não ultrapasse a temperatura máxima permitida de operação. A perda da garantia do fabricante do computador e de sistemas auxiliares pode ser uma consequência danosa desse tipo de falha;
- Problemas com o nobreak: é necessário o monitoramento das baterias e da carga do nobreak para verificar se o funcionamento está dentro dos limites de operação, a fim de evitar o desligamento precoce na falta de fornecimento de energia elétrica.

Sendo assim, de forma a conduzir esta exposição da forma mais clara possível, este texto está organizado da seguinte forma:

- A seção 2.2 apresenta as premissas do desenho da solução do Centro de Supercomputação;
- A seção 2.3 aborda um item cada vez mais importante nos centros de supercomputação, que é o alto consumo de energia elétrica;
- A seção 2.4 mostra o uma visão geral do desenho, desenvolvimento e implementação do Portal de Usuários e da Wiki de Suporte.
- A seção 2.5 discute algumas das soluções de segurança, tanto física como lógica, aplicadas ao supercomputador;
- A seção 2.6 introduz o uso de Internet das Coisas aplicadas ao ambiente do supercomputador;
- A seção 2.7 faz as considerações finais.

2.2. Desenho da Solução do Centro de Supercomputação

O Núcleo Avançado de Computação de Alto Desempenho (NACAD) da COPPE/UFRJ possui grande experiência na administração, gerência e desenvolvimento de ferramentas de apoio ao Centro de Supercomputação. Aproveitando-se da experiência adquirida durante a implementação do cluster Galileu, que fez parte da lista do TOP500 [TOP500] entre 11/2009 a 06/2012, e em 2010 foi a maior supercomputador da América Latina, foi feito o desenho da solução no supercomputador atual do NACAD, o Lobo Carneiro.

O supercomputador Lobo Carneiro possui a seguinte especificação: total de nós de processamento: 252; 504 CPUs Intel Xeon E5-2670v3 (Haswell): 6048 Cores; cores/nó processamento: 24; threads/nó processamento: 48; memória por nó de processamento: 64 GBytes; total de memória RAM: 16 TBytes (distribuída); sistema de arquivo paralelo: Intel Lustre (500 TBytes); armazenamento em disco: 60 TBytes; rede: Infiniband FDR - 56 Gbs topologia hipercubo. O desempenho de pico do Lobo Carneiro no HPL [Dongarra] é de 191TFlop/s.

Para aumentar a eficiência da refrigeração o supercomputador possui um isolamento entre a entrada de ar (corredor frio) e a saída de ar (corredor quente).

O supercomputador Lobo Carneiro está instalado fisicamente a uma distância de aproximadamente 4 (quatro) Km do NACAD, no parque tecnológico da UFRJ, em ambiente especialmente projetado para esse fim e próximo das empresas sediadas neste local. Em função da distância e de uma equipe reduzida, foi necessário desenvolver um ambiente autônomo, pois por exemplo, caso houvesse algum problema de falta no fornecimento de energia, não teríamos acesso a rede para poder interagir com o sistema a fim de monitorá-lo ou desligá-lo.

Desta forma foram definidas as principais premissas de operação do supercomputador Lobo Carneiro:

- Desenvolver uma melhor gerência sobre o supercomputador, levando em consideração restrições financeiras, de forma a automatizar o máximo possível a operação e manutenção do sistema.
- Autonomia de operação 24x7 sem a necessidade de intervenção humana para o caso de algum tipo de falha do sistema. Em uma universidade pública é muito difícil, e as vezes complicado, manter uma estrutura 24x7. Seria necessária uma equipe bem maior para poder montar uma escala de trabalho 24x7, aumentando significativamente o custo de pessoal.
- Monitoramento e gerenciamento das temperaturas do corredor frio, corredor quente, sala do nobreak e sala de operação. Isto permite um melhor controle da temperatura de todos os ambientes e possibilitando que o supercomputador opere dentro das especificações de fábrica, mantendo assim a garantia de todos os equipamentos.
- Maior controle sobre o consumo de energia, que atualmente é um dos fatores críticos no custo total de centros de supercomputação.
- Controle sobre todos os equipamentos ligados/desligados no site. O controle deve permitir acesso remoto para ligar e desligar quaisquer equipamentos do supercomputador, incluindo até os sistemas de armazenamento que não possuem botão de liga e desliga.
- Controle do sistema de UPS que inclui: carga do UPS, carga das baterias, tensões das fases de entrada etc.
- Desenvolvimento de um Portal de Usuários para permitir um ambiente de interação com os usuários do sistema. O portal deve prover:
    - Gerenciamento dos usuários
    - Abertura de chamados de suporte
    - Integração com o supercomputador
        - Automação na abertura de contas
        - Segurança nas comunicações entre o Portal e o supercomputador
- Segurança de acesso tanto físico como lógico.

O perfil de operação do supercomputador deve permitir jobs longos de até 1.000 horas de *walltime* e jobs com múltiplos nós de processamento. Limitamos ao máximo de 40 nós o que equivale a 960 cores reais ou 1.920 threads para que a máquina possa ser usada por mais de um usuário simultaneamente. A Figura 2.1 mostra uma listagem parcial dos jobs, executando no dia 21/08/2018, onde as informações dos usuários foram removidas. Pode-se ver jobs sendo executados a mais de 912 horas, o que equivale 38 dias de execução.

O sistema também deve permitir o controle dos recursos utilizados para evitar que um ou mais grupos/projetos venham a monopolizar os recursos do supercomputador através de dezenas de jobs sendo alocados para processar ao mesmo tempo. Sendo assim, é necessário definir alguns limites para usuários/grupo na máquina, tais como: o número de job simultâneos, o número máximo de nós alocados, etc.

```
user@service1:~> qstat -a
service1:
                                                            Req'd  Req'd   Elap
Job ID          Username Queue  Jobname    SessID NDS TSK Memory Time  S Time
--------------- -------- ------ ---------- ------ --- --- ------ ----- - -----
97589.service1  xxxxxx   workq  job31_bh2   43909   2  96    --  999:0 R 918:4
97949.service1  yyyyyy   workq  job_02_1    15840   2  96    --  672:0 R 278:3
97950.service1  yyyyyy   workq  job_02_2    18190   2  96    --  672:0 R 263:5
102580.service1 xxxxxx   workq  job34_bh2   12856   2  96    --  999:0 R 765:5
102581.service1 xxxxxx   workq  job35_bh2   16957   2  96    --  999:0 R 796:3
102989.service1 xxxxxx   workq  job02b_bh1  48539   2  96    --  999:0 R 789:0
102990.service1 xxxxxx   workq  job02_bh12  42434   2  96    --  999:0 R 658:2
103217.service1 zzzzzz   workq  job_run     43872   1  48    --  1000: R 737:3
104076.service1 zzzzzz   workq  job_run      1040   2  96    --  1000: R 427:0
105986.service1 wwwwww   workq  job_F       37483   1  48    --  999:0 R 286:4
. . .
```

Figura 2.1. Listagem parcial dos Jobs em execução no dia 21/08/2018.

2.3. Consumo de Energia

O consumo de energia ainda é um dos grandes obstáculos para atingir a computação exascala e diversas pesquisas estão sendo realizadas com a finalidade de diminuir o consumo por FLOPS, isto é, operações de ponto flutuante por segundo (FLoating-point Operations Per Second).

Atualmente uma das maiores preocupações de um centro de supercomputação é o consumo de energia. Isso se deve a grande capacidade de condensação dos equipamentos por rack, atingindo algumas vezes o consumo maior do que 40KW/rack. O maior consumo de energia pelos computadores gera um maior calor. Com isso, torna-se necessária uma infraestrutura de refrigeração para controlar a temperatura e mantê-la na faixa de operação. Isto contribui para aumentar ainda mais o consumo elétrico e, por conseguinte, a um maior custo financeiro mensal.

Dependendo das condições das tarifas em um determinado país, pode-se gastar em energia o equivalente ao custo do supercomputador em poucos anos [Martin]. Para efetuar este controle de forma integrada são desenvolvidas ações no sistema de filas e na operação do supercomputador. Algumas destas medidas são discutidas em seguida.

2.3.1. Integração com o PBS-Pro

O hardware do supercomputador Lobo Carneiro (SGI ICE-X) permite um maior controle sobre o consumo energético. Este recurso viabilizou a integração com o sistema de fila (PBS-Pro) para permitir uma coleta de informações de consumo de energia durante a execução de um job. Esta coleta é feita por *resources in hooks* escritos em Python que se

integram ao PBS-Pro durante a execução dos jobs. Desta forma é possível definir perfis de consumo energético por job e/ou fila. Além de poder coletar o consumo energético por job, isto é, pelos nós de processamento usados. A Figura 2.2 ilustra a arquitetura de medição do consumo energético. Entretanto, nem todos os equipamentos da SGI possuem esses medidores. A família SGI ICE-X possui racks de computação personalizados. Dentro de cada rack há dois ou quatro subsistemas de energia de 12VDC. Uma interface proprietária da SGI é usada para ler a energia de cada rack que permite coletar as informações dos nós computacionais (lâminas ou *blades* em Inglês).

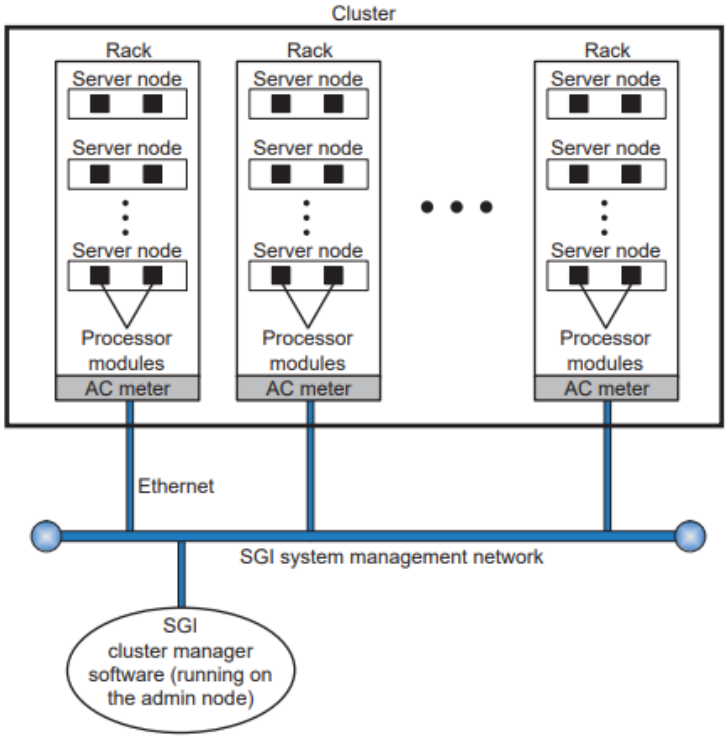


Figura 2.2. Arquitetura do Sistema de Medição de Energia da SGI [PMG].

Apesar do sistema do NACAD-COPPE/UFRJ monitorar todo o consumo energético, não há como calcular o consumo individual das áreas compartilhadas. Entretanto, como o maior consumo se dá nos nós de processamento, essa informação retorna um valor próximo do consumo real do usuário. O usuário pode facilmente obter as informações do PBS-Pro sobre o consumo dos nós de processamento usados durante a execução do seu job. Basta para isso incluir no script do job as informações referentes ao seu e-mail, como mostrado na Figura 2.3.

Esse recurso permitiu com que fosse feito uma comparação [Canesin] entre consumo energético entre nós com CPUs Intel Xeon (SGI ICE-X) e nós com CPUs ARM através do protótipo da máquina do projeto MontBlanc [Rajovic et al] instalado no Barcelona Supercomputing Center. O projeto MontBlanc possui equipamentos de medição externos para monitorar o consumo energético.

```
#!/bin/bash
#PBS -l select=2:ncpus=48:mpiprocs=24
#PBS -l walltime=400:00:00
#PBS -j oe
#PBS -V
#PBS -N mpi-intel
#PBS -m ea
#PBS -M user@domain.br

# load modules
module load intel
# change directory
cd ${PBS_O_WORKDIR}
# run
mpirun ./mpitest
```

Figura 2.3. Exemplo de job com solicitação de e-mail ao término do job.

```
From: root <loboc@nacad.ufrj.br>
Date: 2018-06-05 8:51 GMT-03:00
Subject: PBS JOB 93269.service1
To: user@domain.br

PBS Job Id: 93269.service1
Job Name:   xxxxxxx
Execution terminated
Exit_status=0
resources_used.cpupercent=2700
resources_used.cput=8871:19:46
resources_used.energy=92.9448
resources_used.mem=3403776kb
resources_used.ncpus=48
resources_used.vmem=5038460kb
resources_used.walltime=380:19:14
```

Figura 2.4. Exemplo de e-mail com informação de consumo de energia por job.

A Figura 2.4 ilustra um e-mail com a saída do consumo de energia do job, observado no item "*resources_used.energy*" em KWh.

```
service:~ # qmgr -c "set server power_provisioning=True"
service:~ # qmgr -c "set node @default power_enable=True"

service:~ # pbsnodes -a | grep resources_available.eoe | uniq
   resources_available.eoe = 100W,150W,200W,250W,300W,350W,400W,
   450W,500W,NONE
```

Figura 2.5. Configuração do PBS-Pro.

Depois do script *hook* pronto, a configuração do PBS-Pro se torna muito simples como pode ser visto na Figura 2.5. Basicamente é habilitado o *power_provisioning* no PBS-Pro e depois habilitado o atributo *power_enable* em todos os nós computacionais.

2.3.2. Consumo de energia do supercomputador

No Brasil, durante o "horário de ponta", que corresponde ao intervalo das 17:30hs às 20:30hs nos dias úteis, o preço da energia oscila, dependendo da distribuidora, em torno de 5 a 6 vezes o preço em uma hora normal. Esse é um dos motivos da implantação do horário de verão no Brasil, para tentar diminuir o consumo durante esse período.

Como o supercomputador consome muitos KWh de energia, se for possível diminuir o consumo do mesmo durante o horário de ponta, isso acarretaria numa grande redução do custo de energia. Entretanto, como mostrado na Figura 2.1 alguns dos nossos usuários tem jobs que duram mais de 1 (um) mês sendo executados, o que dificulta esse tipo de controle. Não obstante, o Lobo Carneiro possui algumas ferramentas que podem contribuir para diminuir o consumo de energia.

O controle de energia pode ser feito em diversos níveis: sistema, rack ou nó. A Figura 2.6 mostra como é feito essa coleta. O primeiro exemplo, Figura 2.6(a), corresponde a todo o sistema, o segundo, Figura 2.6(b), de apenas um (1) rack e o último, Figura 2.6(c), o de um nó em modo de execução. O consumo das lâminas (blades) da SGI ICE-X oscilam entre 50-400W e os demais nós oscilam entre 150-600W. Dessa maneira, pode-se limitar o consumo por sistema, rack ou nó de acordo com a necessidade.

Para exemplificar, suponha que se deseje limitar o consumo do nó r1i3n9 para o máximo de 200W, bastaria utilizar o comando mostrado na Figura 2.7. Logo, para o nó em questão o consumo diminuiria em torno de 23%.

Se durante a hora de ponta fosse adotado esse procedimento em todo o supercomputador, o impacto na conta de energia seria grande, pois o custo durante esse período é muito alto. O NACAD ainda está fazendo a análise de desempenho dos jobs para identificar o impacto do consumo em sua execução.

```
# mpower system get_power
System Power Stats:
Instant                           : 57384.71
Minimum during sampling period    : 202.61
Maximum during sampling period    : 16217.39
Average during sampling period    : 12461.97
KwH during sampling period        : 340691.45
```

(a)

```
# mpower rack get_power rack1
r1lead:
Instant                           : 38145.33
Minimum during sampling period    : 216.78
Maximum during sampling period    : 16217.39
Average during sampling period    : 6694.35
KwH during sampling period        : 215816.91
Sampling period                   : 29014806.75
```

(b)

```
# mpower node get_power r1i3n9
r1i3n9        262W
```

(c)

Figura 2.6. Coleta de informações de consumo.

```
# mpower node set_limit r1i3n9 200
# mpower node get_power r1i3n9
r1i3n9        200W
```

Figura 2.7. Definindo um limite de consumo.

2.4. Desenvolvimento do Portal de Usuários e de Suporte

2.4.1. Portal de Usuários

O desenvolvimento de um portal para o Lobo Carneiro surgiu em função da necessidade de facilitar a troca de informações com os usuários. Esses tipos de Portais também estão presentes em várias instalações de centros de supercomputação ao redor do mundo, tais como no TACC [TACC] e no XSEDE [XSEDE]. Este portal deve permitir a integração com o supercomputador e a inclusão de scripts para facilitar a implantação de novas funcionalidades. Estas premissas foram importantes para a escolha do framework. O framework escolhido foi o Django [Django], que é desenvolvido em Python. As principais caraterísticas do portal devem ser:

- Atualização on-line de várias informações coletadas do supercomputador, por exemplo, o status do supercomputador, número de jobs em execução, número de jobs na fila etc.
- Informações das contas para os usuários, como o uso de espaço de armazenamento e o SU (system units) que é calculado em função do tempo de processamento por core alocado, isto é, *número de cores alocado* x *tempo de processamento* (*walltime*).
- Sumário dos dados dos usuários para os coordenadores.
- Notícias de modo geral, como instalação de softwares e bibliotecas, informes sobre o sistema etc.
- Permitir a abertura de chamados de suporte e seu respectivo acompanhamento.
- Permitir que os menus sejam sensíveis a conta, isto é, o menu disponibilizado para um usuário comum é diferente do aparece para o coordenador, que por sua vez também é diferente do que é disponibilizado para os administradores.
- Documentação geral do sistema.
- Informações de administração etc.

A seguir será descrita uma visão geral dos principais menus do Portal, tomando como referência o menu de administração. Alguns dessas opções podem não estar disponíveis para certos tipos de usuários.

- Perfil: mostra o perfil do usuário, suas estatísticas de uso e permite a abertura de chamado de suporte. Com base nos chamados feitos pelos usuários foi possível criar um FAQ, perguntas mais frequentes, do supercomputador Lobo Carneiro. Permitindo assim, uma melhor integração do usuário com o Portal.
- Coordenação: mostra a solicitação de registro, pois nesse portal o coordenador tem independência de controlar a sua equipe. Também mostra a equipe do projeto e sua estatística de uso.
- Recursos: mostra informações sobre os equipamentos disponíveis no centro de supercomputação. Mostra uma visão geral dos equipamentos, status dos servidores e nós computacionais, informações sobre a fila e lista de IPs que foram bloqueados durante a autenticação. Este bloqueio será explicado melhor em um item posterior.
- Administração: mostra todos os chamados de suporte incluindo o status deles, todos os projetos e usuários, permite postagem de notícias e acesso aos logs.
- Documentação: informações de acesso ao sistema, Guia do Usuário, documentação dos compilares Intel e PGI.
- Sobre: informações sobre registro e renovação de contas.

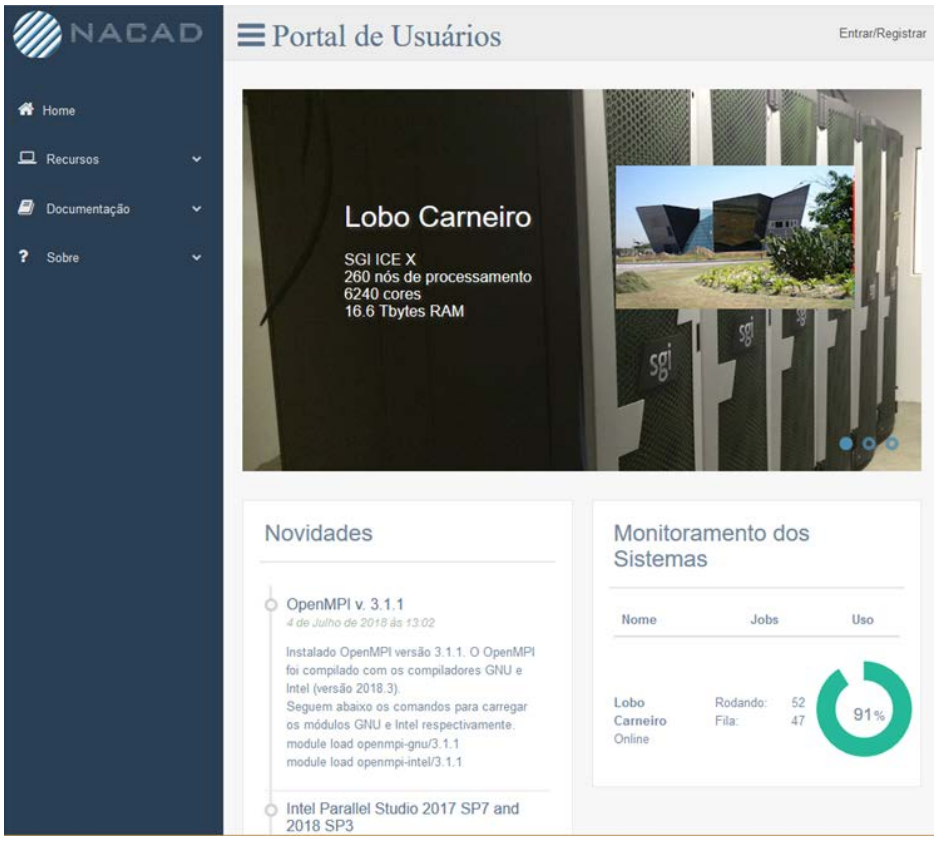


Figura 2.8: Portal de Usuários.

A Figura 2.8 mostra a página de entrada do Portal de Usuários. No canto superior direito é possível entrar na conta e com isso ter acesso a outras informações.

O Portal de Usuários está instalado em um servidor localizado dentro do laboratório do NACAD e não junto com o supercomputador. Isso traz algumas vantagens, por exemplo, se houver algum problema no site do supercomputador os usuários poderão obter essa informação consultando o portal.

2.4.2. Wiki de Suporte

Além do Portal de Usuários foi configurado em outra máquina um outro Portal contendo uma Wiki. A Wiki é usada apenas pela equipe de suporte e possui informações internas e restritas ao NACAD. Portanto, tais informações não são compartilhadas. A finalidade dela é de prover um aprendizado contínuo para a equipe de suporte. Toda a configuração do sistema, ajustes e correção de problemas são documentados nesta Wiki. Dessa forma, quando o problema se repetir os demais membros da equipe consultar o conteúdo da Wiki e podem corrigir o problema seguindo as orientações lá contidas. Seguem abaixo alguns dos principais tópicos que fazem parte desta Wiki:

- Instruções de boot e shutdown do supercomputador;
- Dicas e correções de problemas do sistema de fila PBS-Pro;
- Configurações específicas para os diversos servidores e nós computacionais, incluindo informações especificas de cada servidor e suas respectivas funções;
- Configuração, controle de cota e correções de problemas para os servidores de armazenamento, incluindo o *filesystem* paralelo Lustre;
- Instruções de desenvolvimento para compilação de forma otimizada para diversas ferramentas;
- Instruções sobre atualização e ajustes nos compiladores Intel, PGI e GNU;
- Instruções para o desenvolvimento de *hooks* para o PBS-Pro;
- Instruções para configuração do Portal de Usuários;
- Instruções de configuração de segurança para ambientes Linux;
- Informações sobre uso dos scripts desenvolvidos pela equipe;
- Outros tópicos.

O conteúdo da Wiki é continuamente atualizado. Qualquer intervenção no sistema motivada por atualizações, adaptações solicitadas pelos usuários, etc, e imediatamente documentada e incorporada a Wiki.

2.5. Segurança

No mundo atual, a questão de segurança está se tornando cada vez mais importante. No caso de um centro de supercomputação a importância aumenta. Entretanto, como o site do Lobo Carneiro é desassistido, o item de segurança se torna ainda mais fundamental. A segurança do Centro de Supercomputação deve levar em consideração tanto a parte física como a parte lógica.

Para endereçar a parte física foram instaladas câmeras de vídeos, que podem ser acompanhadas tanto pelos seguranças do Parque Tecnológico, como por computador e/ou celular, como mostra a Figura 2.9. Além das câmeras foram instalados vários sensores de

IoT (Internet das Coisas). Como por exemplo, sensores de temperatura que foram distribuídos na sala de operação, sala do nobreak, corredor frio e corredor quente.

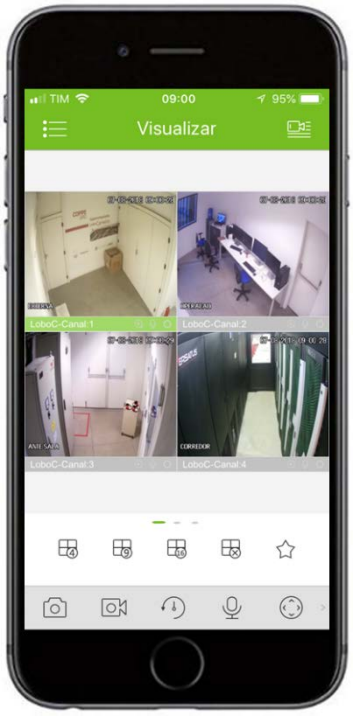


Figura 2.9: Monitoração das câmeras de vídeo através do celular.

2.5.1. Integração do supercomputador com o Portal

Para minimizar a superfície de ataque contra o supercomputador, toda a comunicação entre o Portal de Usuários e o supercomputador é feito por mão única, isto é, toda a comunicação é originada apenas do supercomputador. Ou seja, em nenhum momento o Portal envia dados para o supercomputador. Neste caso, se o Portal vier a ser comprometido, o invasor não terá acesso ao supercomputador. A Figura 2.10 ilustra de forma simplificada o sentido das comunicações entre os sistemas. O usuário pode acessar tanto o Portal como o supercomputador através do servidor de acesso. Este por sua vez, atualiza o Portal com os dados coletados do supercomputador. Vários dados coletados também são colocados na nuvem para poderem ser consumidos por outros dispositivos, tais como celular e *tablets*. Alguns desses itens serão descritos com mais detalhes posteriormente.

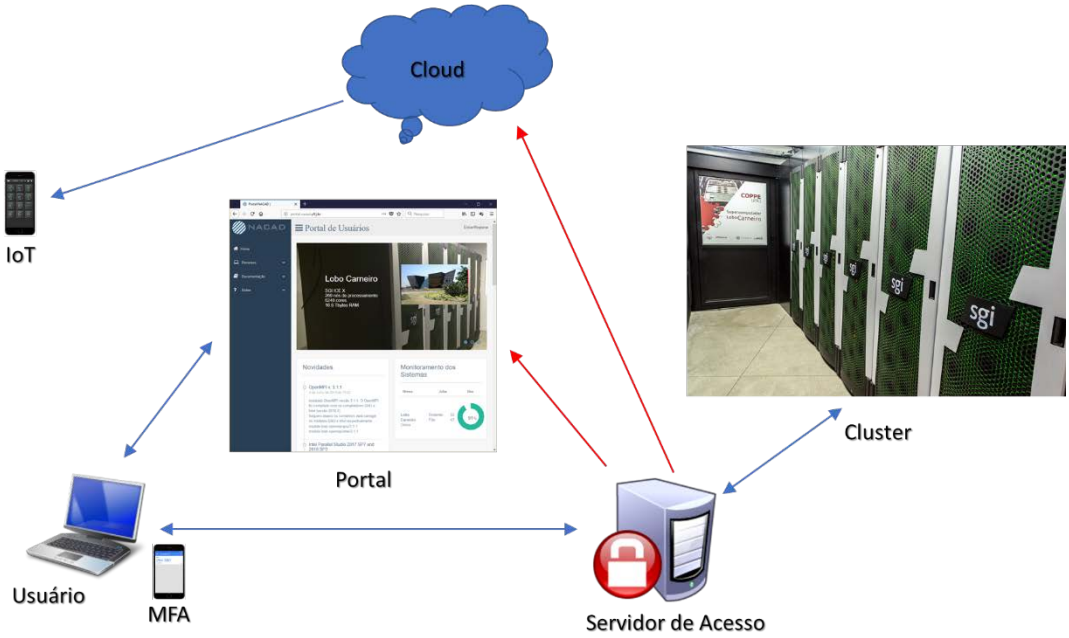
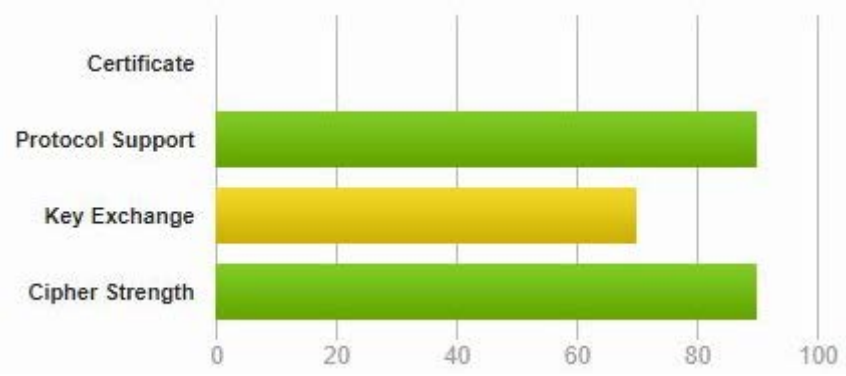


Figura 2.10. Infraestrutura lógica do supercomputador.

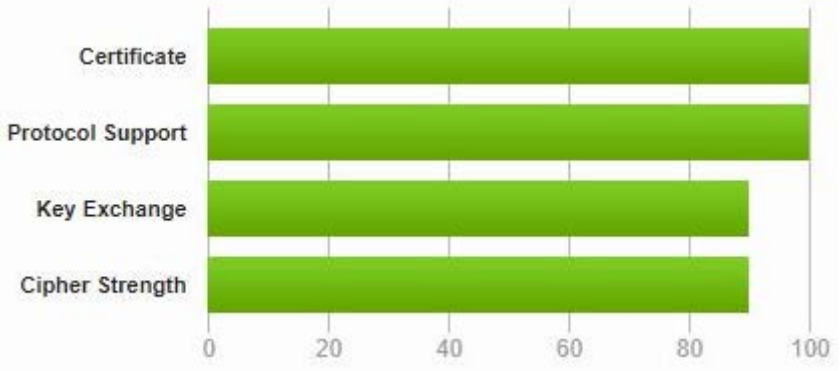
A fim de aumentar a segurança no Portal de Usuários foram implementados alguns ajustes no sistema. As configurações usuais não serão descritas neste item, apenas as mais importantes serão listadas e comentadas.

- Configuração do SELinux (*Security-Enhanced Linux*) que limita o escopo de possíveis danos que podem resultar da exploração de vulnerabilidades em aplicativos e serviços do sistema. O SELinux fornece um sistema flexível baseado no MAC (*Mandatory Access Control*), integrado ao kernel do Linux.
- Testes de penetração e outros testes. Para exemplificar, será mostrada a análise do item SSL do Portal. Durante a configuração, o Portal foi feito inicialmente usando a instalação padrão do servidor Web com um certificado auto assinado. O resultado do teste é mostrado na Figura 2.11(a). Com base nos resultados foram feitos ajustes no servidor Web, principalmente referentes aos protocolos suportados e chaves utilizadas, obtendo assim um aumento significativo na segurança do site, como mostrado na Figura 2.11(b).
- Configuração de um sistema de detecção de intrusão. Este sistema detecta qualquer alteração, inclusão ou remoção de arquivo do sistema. Ele utiliza uma base de dados para comparar com o sistema atual. Uma cópia dessa base de dados deve estar em outra máquina, pois no caso de uma invasão o invasor pode gerar outra base de dados diferente. A fim de mitigar esse problema foram criados alguns scripts que enviam diariamente um e-mail com o resultado da análise do sistema de intrusão e no assunto do e-mail vai junto o hash MD5 da base de dados em questão. Logo, se um invasor gerar uma nova base de dados, pode-se identificar rapidamente pelo cabeçalho do e-mail. A Figura 2.12 mostra o cabeçalho dos e-mails diários de duas máquinas, a do portal e a máquina de acesso com os respectivos hashes no campo assunto do e-mail. Na Figura 2.13 mostra

um trecho de uma mensagem padrão enviada por e-mail diariamente. Pode-se observar que este relatório é dividido em basicamente cinco partes, que são: sumário do relatório; arquivos que foram adicionados, removidos e modificados em relação ao banco de dados; e as informações com os detalhes das mudanças.



(a) Configuração padrão do servidor Web



(b) Configuração ajustada após análise do teste

Figura 2.11: Resultados dos testes.

Figura 2.12: E-mails automáticos do sistema de detecção de intrusão.

```
Summary:
  Total number of files:        73904
  Added files:                  34
  Removed files:                31
  Changed files:                22

---------------------------------------------------
Added files:
---------------------------------------------------
added: /var/log/btmp-20180701
added: /var/log/cron-20180624.gz
added: /var/log/httpd/access_log-20180611.gz
added: /var/log/httpd/error_log-20180624.gz
...

---------------------------------------------------
Removed files:
---------------------------------------------------
removed: /var/log/btmp-20170301
removed: /var/log/cron-20170226.gz
removed: /var/log/httpd/access_log-20170220.gz
removed: /var/log/httpd/error_log-20170220.gz
...

---------------------------------------------------
Changed files:
---------------------------------------------------
changed: /etc/sysconfig/network-scripts
...
changed: /var/log/lastlog
changed: /var/log/maillog
changed: /var/log/messages
...

---------------------------------------------------
Detailed information about changes:
---------------------------------------------------
Directory: /etc/sysconfig/network-scripts
 Mtime    : 2017-02-13 08:01:31              , 2018-06-11 17:03:05
 Ctime    : 2017-02-13 08:01:35              , 2018-06-11 17:03:08
...
File: /var/log/lastlog
 Mtime    : 2017-03-04 08:39:26              , 2018-06-25 09:58:31
 Ctime    : 2017-03-04 08:39:26              , 2018-06-25 09:58:31
 SHA256   : eghs7kR66UF58Nqe1EiV+P8UKtjiEzZ+,
Lqgbms9f4eSz4wzuw/EkP2nCblQ3RkZm
 SHA512   : QnP7b4I9EyG5kZp2i++mjM8Fb3V5xJsA,
BFEUOYzQqSHmyCzUQMjpaR1jLlIcjsQO
...
```

Figura 2.13: Relatório diário de alterações do sistema.

2.5.2. Acesso e autenticação dos usuários

Diferentemente do Portal de Usuários, onde o usuário utiliza apenas uma interface web e ele não possui conta na máquina, o supercomputador Lobo Carneiro precisa fornecer uma conta para permitir o uso dos seus recursos. O usuário poderá fazer acesso ao supercomputador de casa, do laboratório, do trabalho, do cybercafé, etc. Isso pode gerar um grande problema, pois não há como garantir a segurança do lado do usuário. Em outras palavras, se o laboratório/acesso dele for comprometido, o invasor pode, por exemplo, trocar o cliente SSH, que é usado para acessar o supercomputador, e com isso coletar as contas e senhas das outras máquinas. De posse dessas senhas, o invasor poderia entrar no supercomputador com a conta de um usuário. Dessa forma a superfície de ataque do invasor aumentaria significativamente, pois ele já teria acesso ao supercomputador.

Para contornar esse problema adota-se um sistema conhecido como autenticação de multi-fator (MFA - *Multi-Factor Authentication*), no caso do Lobo Carneiro, de dois fatores. Logo, para ter acesso ao supercomputador o usuário deverá informar seu nome de usuário e senha (o primeiro fator – o que ele sabe) e também um código de autenticação de seu dispositivo de MFA (o segundo fator – o que ele tem). Juntos, esses vários fatores fornecem maior segurança para suas configurações e recursos de conta.

Atualmente o celular é considerado um produto essencial e praticamente todos possuem um smartphone. Logo, é totalmente factível a implementação de um sistema MFA usando o smartphone como o segundo fator de autenticação. Existem vários softwares que rodam tanto em IOS (Apple) como no Android, são eles: Google Authenticator, Microsoft Authenticator, FreeOTP e outros. O único requisito para o sistema funcionar corretamente é que o smartphone esteja com a hora correta. Basta então, que o usuário sincronize a hora do seu celular com a operadora de telefonia.

Do lado do supercomputador foi compilado, instalado e configurado o libpam do Google Authenticator e feitos os ajustes necessários no serviço NTP (*Network Time Protocol*) para que a hora esteja sempre sincronizada.

Durante a abertura de conta de um usuário o script executa várias etapas, tais como: criar a conta; criar grupo do projeto, caso necessário; configurar as cotas em disco (área home e área scratch). Ele configura também o Google Authenticator para o usuário gerando um QR-code que é enviado para o Portal de Usuários. O usuário por sua vez ao entrar no Portal e com um dos aplicativos do MFA instalado em seu celular, que deve ter acesso a câmera, deve apontar a câmera para o QR-code gerado pelo sistema a fim de ler e gerar a sua configuração automaticamente. A Figura 2.14 mostra a tela que aparece no Portal. Para aumentar ainda mais a segurança este QR-code só é mostrado apenas uma vez. Desta maneira, tenta-se evitar que outra pessoa possa usar o mesmo QR-code do usuário no caso de deixar o Portal aberto na sua conta ou tenha sua senha violada.

Figura 2.14: QR-code do MFA.

De posse do QR-code, o usuário já pode se autenticar no supercomputador. O processo é muito simples, basta o usuário se conectar no supercomputador usando a sua conta e senha e consultar o código no celular. Este código é mudado a cada minuto e só pode ser usado apenas uma vez, também por motivos de segurança. As Figuras 2.15 e 2.16 ilustram o processo de *login* do usuário.

Figura 2.15: Google Authenticator no smartphone.

```
login as: user
Using keyboard-interactive authentication.
Password: ********
Using keyboard-interactive authentication.
Verification code: 294380
------------------------------------------------------------------
          Welcome to the Lobo Carneiro Supercomputer
                      NACAD-COPPE/UFRJ
------------------------------------------------------------------

     ** WARNING: Unauthorized use/access is prohibited. **

If you log on to this computer system, you acknowledge your
awareness of and concurrence with the NACAD Acceptable Use
Policy. By attempting connection without permission, you are
in violation of federal law.

NACAD Usage Policies:
http://www.nacad.ufrj.br/informacoes/politica-de-uso
------------------------------------------------------------------
user@service:~>
```

Figura 2.16: Processo de login no supercomputador.

Além dos procedimentos de segurança descritos anteriormente foram implementadas outras medidas, tais como:

- Firewall GeoIP: São regras de firewall baseadas na localização física do IP, isto é, pode-se impedir que um ou mais países tentem se conectar ao supercomputador. Este tipo de filtragem diminui sensivelmente alguns tipos de ataques.
- Bloqueio e desbloqueio automático de tentativa de acesso: Para evitar ataques de força bruta para o SSH, tem-se um sistema que bloqueia automaticamente o IP por 30 minutos no caso de três erros da senha e/ou código. Isto permite que caso o evento em questão não seja um ataque, isto é, um falso positivo, o IP será liberado automaticamente sem a necessidade de intervenção da equipe de suporte.
- E outros processos de menor importância.

2.6. IoT (Internet das Coisas)

O uso de sensores externos implementados na infraestrutura do supercomputador juntos com outros sensores que faziam parte da solução permitiu um melhor controle sobre todas variáveis de ambiente. Desta forma, mesmo com uma equipe reduzida, é possível gerenciar um supercomputador remotamente. Apenas para enumerar, alguns dos sensores que fazem parte da solução são:

- Sensor de medição de energia, descrito anteriormente e mostrado na Figura 2.2.

- Sensores de temperatura dentro de todos os nós computacionais e demais servidores, que podem ser acessados pelo comando *ipmitool*.
- Sensores do nobreak que informa carga do nobreak, corrente, etc.

Com os diversos sensores disponíveis foi desenvolvida a coleta e tratamento desses dados de maneira centralizada, agregando informações úteis do sistema como um todo. Desta maneira foi disponibilizado várias informações sobre o supercomputador na nuvem, conforme ilustrado na Figura 2.10. Isso é feito através do protocolo MQTT (*Message Queue Telemetry Transport*), que se tornou o padrão para comunicações de IoT. Essas informações podem ser disponibilizadas como públicas ou privadas.

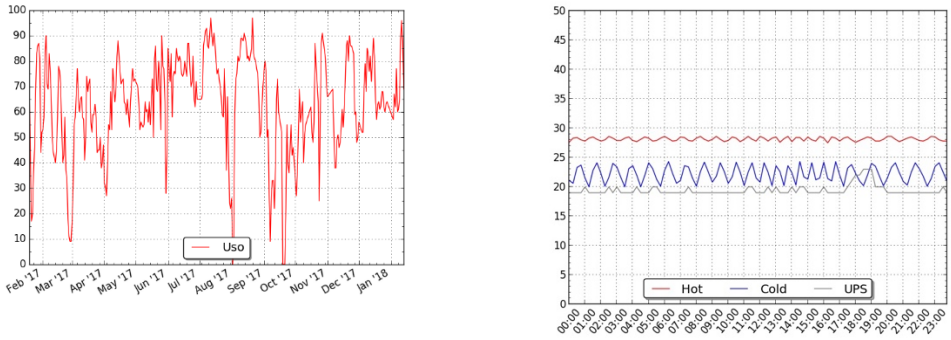
Este protocolo é oferecido em diversos provedores de serviços de nuvem pública, tais como: AWS (Amazon Web Services), Microsoft Azure, Google Cloud, etc. Também é possível, caso se queira, instalar um MQTT broker em um servidor próprio. Bastando para isso instalar o software de código aberto conhecido como Mosquitto [Mosquitto].

De posse dessas informações, elas podem ser consumidas de várias formas. A Figura 2.8 mostra a carga do sistema e a quantidade de jobs sendo executados e na fila. Na Figura 2.17(a) mostra-se a carga do sistema no período de fevereiro de 2017 a janeiro de 2018. Vários dos picos inferiores foram durante a manutenção programada do sistema de refrigeração, cuja capacidade é diminuída pela metade para que a operação de manutenção seja possível sem o desligamento total do sistema. Em consequência, deve-se reduzir a quantidade de nós de processamentos em operação. Os outros picos são devidos a falhas no fornecimento de energia elétrica e, nesses casos, o sistema de proteção desliga todos os equipamentos de forma segura. Note que, por segurança, o religamento é manual.

Alguns dos equipamentos de armazenamento do supercomputador não possuem um botão liga/desliga ou uma maneira de desligar o hardware via software. Neste caso, foi implementado um *smart relay* para desligar de forma automática o disjuntor de energia, após o término do correto do processo de shutdown do serviço. Esta controladora também é usada para ligar o disjuntor. A Figura 2.17 mostra uma foto do *smart relay*.

Figura 2.17: Smart Relay.

A Figura 2.17(b) mostra o histórico diário das temperaturas coletadas nos sensores localizados na sala do UPS e nos corredores quente e frio em um dia típico de operação. Este gráfico é gerado diariamente e enviado por e-mail para a equipe de suporte.



(a) (b)

Figura 2.17: (a) Utilização do supercomputador; (b) Histórico da temperatura da UPS, e dos corredores quente e frio.

Como essas informações estão disponibilizadas na nuvem, elas podem ser consumidas onde for necessário. A Figura 2.18 mostra um exemplo de coleta de informações utilizando um celular.

Figura 2.18: Exemplo de coleta de informações através de IoT usando o celular.

Segue abaixo uma pequena descrição dos dados mostrados na Figura 2.18.

- Primeira linha informa o status do supercomputador se ele está ou não ligado e informa a temperatura do corredor frio e quente.
- Segunda linha informa a quantidade de jobs que estão sendo executados, os jobs que estão na fila e a carga do sistema.
- Terceira linha informa carga do UPS e carga da bateria; e informação da última atualização das informações na nuvem.
- Quarta linha mostra informações sobre as tensões de entrada das três fases do nobreak.

Para consultar estas informações basta abrir o aplicativo e visualizar os dados, pois a autenticação é feita de forma automática.

2.7. Considerações Finais

O presente trabalho teve por finalidade trazer informações sobre integração sobre diversos serviços, como: Portal de Usuários, segurança e uso de IoT em ambientes de supercomputação.

Foram abordadas algumas tecnologias que podem ser utilizadas para aumentar a segurança, como por exemplo o MFA, que permite uma autenticação com dois fatores, automação do sistema de detecção de intrusão, firewall GeoIP, etc.

O uso de algumas tecnologias emergentes, que fazem parte da indústria 4.0, como o IoT e o uso da nuvem foram integradas e trazidas para dentro de um centro de supercomputação. A integração dessas informações promoveu um salto de qualidade na gerência e operação desassistida do supercomputador Lobo Carneiro. Desde a sua instalação e até agosto de 2018 o supercomputador Lobo Carneiro já processou mais de 108.000 jobs, correspondendo a mais de 1.600.000 SUs, todos com segurança, com a máxima qualidade no serviço, já que foi mantida a integridade do sistema durante todo o período. Já foram atendidas, também até a presente data, mais de 500 requisições de usuários submetidas através do Portal.

Neste trabalho não foram considerados vários itens relativos a ambientes de HPC e centros de supercomputação. Entre estes destacamos:

- Ensinar bons hábitos aos usuários permitindo assim, que se crie uma nova cultura para lidar com a segurança da informação.
- Ajustes no supercomputador para maximizar o desempenho.
- Desenvolvimento de *hooks* para atuar junto com o sistema de fila PBS-Pro.
- Desenvolvimento de diversos scripts para customizar e automatizar tarefas importantes na operação do supercomputador.
- Implementações corriqueiras em sistemas computacionais, principalmente relativas à segurança da informação.
- Detalhes sobre ações de auditoria dos sistemas.
- Listagens de todas as ferramentas utilizadas.
- Detalhes sobre a cibersegurança do supercomputador.

Estes itens, pertinentes, serão objeto de publicações e futuras para disseminarmos as boas práticas de operação e gerência de Centros de Supercomputação no Brasil.

Agradecimentos

Este trabalho contou com o apoio do CNPq, FAPERJ e da PETROBRAS.

Referências

Schwab, K., “A Quarta Revolução Industrial”, Editora Edipro, 2016.

TOP500: https://www.top500.org/system/176647, agosto, 2018.

Dongarra, Jack J., Piotr Luszczek, and Antoine Petitet. "The LINPACK benchmark: past, present and future." Concurrency and Computation: practice and experience 15.9 (2003): 803-820.

PMG: SGI Power Management Guide, version 002, 2016.

Canesin, F.C., “Algoritmos de Integração Temporal para Solução Adaptativa e Paralela das Equações de Navier-Stokes”, Tese de Mestrado, 2017, (http://www.coc.ufrj.br/pt/dissertacoes-de-mestrado/590-msc-pt-2017/8609-fabio-cesar-canesin).

Rajovic, N., et al, “The mont-blanc prototype: an alternative approach for HPC systems”, Proceeding SC '16 Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Article No. 38, Salt Lake City, Utah, November, 2016.

Martin, S., “Total Cost of Ownership and HPC System Procurement”, SC17: Birds of Feather, https://eehpcwg.llnl.gov/assets/sc17_bof_tco_procurement.pdf, 2017.

TACC User Portal, https://portal.tacc.utexas.edu/, agosto 2018.

XSEDE User Portal, https://portal.xsede.org/#/guest, agosto, 2018.

Django: https://www.djangoproject.com, agosto, 2018.

Mosquitto, https://mosquitto.org/, agosto, 2018.