Uma Arquitetura P2P de Distribuição de Atividades para Execução Paralela de *Workflows* Científicos

Vítor Silva[1], Jonas Dias[1], Daniel de Oliveira[2], Eduardo Ogasawara[3], Marta Mattoso[1]

[1]COPPE/Universidade Federal do Rio de Janeiro (UFRJ), Brasil

[2]Universidade Federal Fluminense (UFF), Brasil

[3]CEFET/RJ, Brasil

{silva, jonasdias, marta}@cos.ufrj.br,
danielcmo@ic.uff.br, eogasawara@cefet-rj.br

Abstract. *Scientific workflows are composed of activities that model scientific experiments. Many Scientific Workflow Management Systems use High Performance Computing environments to parallelize the execution of these activities in large-scale workflows. Data distribution, control, and optimizing the parallel execution of these activities can be a complex task due to scalability of involved resources. This paper presents DEW, a data and activity distribution mechanism for a parallel workflow execution engine. DEW is based on a hierarchical P2P network that enables distributed control in workflow execution using distributed disk and in the presence of high occurrence of churn events.*

Resumo. *Workflows científicos são compostos de atividades que modelam experimentos científicos. Vários Sistemas de Gerência de Workflows Científicos fazem uso de ambientes de Processamento de Alto Desempenho para paralelizar a execução de atividades do workflow sobre grandes fluxos de dados. Gerenciar a distribuição dos dados, controlar e otimizar a execução paralela dessas atividades pode ser complexo devido à escala dos recursos envolvidos. Esse artigo apresenta o DEW, um mecanismo de distribuição de dados e atividades para execução paralela. O DEW é baseado em uma rede P2P hierárquica que permite o controle distribuído na execução de workflows, com disco distribuído e alta volatilidade de recursos computacionais.*

1. Introdução

Os experimentos científicos baseados em simulação computacional são caracterizados pela execução de uma sequência de atividades (programas) que possuem dependências de dados entre si. Para facilitar a gerência da execução do experimento, cientistas utilizam *workflows* científicos na modelagem de seus experimentos (Mattoso et al. 2008), que podem ser gerenciados por Sistemas de Gerência de *Workflows* Científicos (SGWfC). Além de apoiar a modelagem, os SGWfC permitem a execução, monitoramento e visualização dos resultados dos *workflows*.

Devido à necessidade do processamento de grande volume de dados pelos *workflows* científicos, as técnicas de paralelismo aplicadas em ambientes de

Processamento de Alto Desempenho (PAD) são utilizadas para reduzir o tempo de execução total. Originalmente, os ambientes de PAD são classificados como homogêneos ou heterogêneos. Ambientes homogêneos, como clusters, possuem recursos computacionais homogêneos normalmente conectados em uma rede dedicada de alta velocidade e baixa latência, o que facilita a manutenção do equipamento, a paralelização de aplicações científicas por meio de bibliotecas especializadas (e.g. MPI ou OpenMP) (Chandra 2001, Gabriel et al. 2004) e a obtenção de alto desempenho, mas ao mesmo tempo, limita o número de recursos computacionais. Os ambientes heterogêneos, e.g. grades computacionais e nuvens híbridas, são caracterizados por apresentar os recursos heterogêneos do ponto de vista de poder de processamento, topologia de redes e software, o que amplia a possibilidade de envolver mais recursos computacionais no mesmo ambiente. Porém, esse ambiente traz dificuldades na manutenção do equipamento e na paralelização de aplicações científicas, o que implica na complexa gerência distribuída da execução paralela do *workflow*.

A dificuldade de manutenção do equipamento e a gerência distribuída dos dados e da rede são algumas características que propiciam a ocorrência de falhas em ambientes de PAD de grande escala. Mesmo ambientes tradicionalmente classificados como homogêneos podem apresentar alta volatilidade durante sua execução paralela, caso possuam um grande número de nós com armazenamento distribuído e componentes heterogêneos (Raicu et al. 2008). A heterogeneidade também dificulta a coleta de dados de proveniência (Freire et al. 2008) da execução do *workflow*. A proveniência é o registro da informação relacionada à execução do *workflow*. Portanto, é necessário manter um registro sólido do recurso computacional responsável pela execução de cada atividade e a movimentação de dados envolvida na execução do *workflow*. Tais informações são essenciais para proporcionar confiabilidade e reprodutibilidade dos resultados obtidos.

As arquiteturas ponto-a-ponto (do inglês *peer-to-peer*, ou simplesmente P2P), são frequentemente utilizadas em ambientes heterogêneos de larga-escala devido ao seu controle de execução descentralizado, proporcionando escalabilidade e resiliência. Ambientes P2P assumem a ocorrência de falhas nos nós, acomodando ingressos e saídas frequentes de nós na rede (*churn*) *(*Wu et al. 2008). Chamamos de volatilidade de nós os "*churns*" das redes P2P. Precaver-se de falhas durante a execução paralela é sempre importante e é fundamental em ambientes heterogêneos. A resiliência e a alta escalabilidade de ambientes P2P podem beneficiar a execução de *workflows* científicos em larga-escala, proporcionando tolerância a falhas, paralelização e execução paralela do *workflow* envolvendo um grande número de nós com coleta de proveniência.

Esse artigo apresenta o DEW (*Distributed Execution control for Workflow engines*), um mecanismo de distribuição de dados e atividades com controle de execução descentralizado para a execução paralela de *workflows* científicos em ambientes heterogêneos. O DEW implementa uma rede P2P hierárquica em ambientes de *cluster* ou nuvens de computadores, onde cada recurso computacional do ambiente é considerado um nó da rede P2P. Sendo assim, considerando um cluster, cada nó computacional será considerado um *peer* na rede P2P. A arquitetura do DEW visa a apoiar a alta escalabilidade e a resiliência, assumindo ambientes sem disco compartilhado e com alta volatilidade. Foram realizados experimentos com a distribuição de atividades de *workflows* no DEW que foram analisados quanto à eficiência, variando-se o número de nós e a volatilidade.

Além desta introdução, esse artigo está organizado em cinco seções. A Seção 2 apresenta os trabalhos relacionados à proposta do DEW. A Seção 3 descreve a tecnologia P2P, considerando os tipos existentes quanto à sua centralização. Já a Seção 4, apresenta a arquitetura do DEW com os mecanismos de distribuição. A Seção 5 mostra as avaliações realizadas com a arquitetura e a Seção 6 conclui.

2. Trabalhos Relacionados

Tendo em vista a limitação do controle centralizado na paralelização de *workflows* científicos, algumas propostas descentralizadas vêm surgindo, ainda que de modo ainda incipiente. Rahman et al. (2010a) propõem um sistema descentralizado para compartilhamento de recursos de processamento e de armazenamento em ambiente de grades por meio da sobreposição de uma rede P2P. Os diferentes recursos são organizados e controlados por meio de uma tabela *hash* distribuída (DHT). Enquanto isso, um espaço de coordenação realiza a alocação das atividades do *workflow* nos recursos utilizando a DHT. A vantagem principal está associada ao mecanismo de distribuição de recursos computacionais, que apresenta a descoberta determinística de recursos e o controle no número de mensagens geradas, diferentemente de técnicas de *broadcast* existentes em alguns SGWfCs, como o Triana (Taylor et al. 2007), que inundam o tráfego de mensagens do ambiente. Da mesma forma, o ambiente Sunflower (Papuzzo and Spezzano 2011) monitora e controla a execução de *workflows* científicos em ambientes computacionais de grade e nuvem. O Sunflower faz uso de uma rede P2P, tendo como contribuições o balanceamento das atividades dos *workflows* por meio de recursos e a recuperação de falhas.

A abordagem Heracles (Dias et al. 2010a) é uma proposta que motivou o desenvolvimento do DEW. Heracles tem como contribuição um mecanismo para paralelizar a execução de atividades de *workflows* em clusters com um número elevado de nós. Usando uma rede P2P hierárquica, o Heracles realiza a distribuição das atividades tendo em vista o dinamismo da rede e a tolerância a falhas. Os resultados apresentados pelo Heracles, por meio de simulações do ambiente de execução, evidenciam os benefícios da abordagem P2P para execução de atividades de *workflows* em ambientes com volatilidade.

O Hadoop (Apache Software Foundation 2009) é outra proposta nos moldes do paradigma de MapReduce que executa aplicações em ambientes de PAD e permite a instanciação de uma grande quantidade de nós. Outro ponto positivo é o seu comportamento dinâmico (permite a entrada e saída de nós), o que contempla a recuperação de falhas. Contudo, o controle do Hadoop é centralizado, o que implica que todos os nós respondam ao nó principal. Portanto, caso haja a falha do nó principal, a aplicação precisa ser reinicializada. Além disso, o Hadoop é voltado para a execução de uma atividade e não para a gerência do fluxo de dados das atividades de um *workflow*.

Os trabalhos mencionados motivam a proposta do DEW quanto à característica de descentralização do controle da execução. Decisões relacionadas aos recursos e à distribuição de atividades passam a ser tomadas por diferentes nós computacionais da rede e não por apenas um nó de controle. Sendo assim, o DEW é um mecanismo com controle descentralizado para distribuir dados e atividades da execução paralela de *workflows* científicos. Com o intuito de tirar proveito de características presentes apenas no Chiron, o DEW foi acoplado ao mesmo. O Chiron (Ogasawara et al. 2013) é uma

máquina de execução de *workflows* que faz-se valer de uma álgebra de *workflow* como insumo para direcionar a otimização da execução do *workflow*. O Chiron permite a escolha da melhor estratégia de execução de um *workflow*, a realização de possíveis otimizações no *workflow* e a coleta de dados de proveniência de forma distribuída. Entretanto, a máquina de execução do Chiron é centralizada e dedica-se apenas a disco compartilhado. Logo, a integração do DEW com o Chiron vem a proporcionar um mecanismo descentralizado.

O controle distribuído do DEW traz novos desafios para o modelo de execução e o processo de otimização adotados no Chiron, assim como a gerência de arquivos e o acesso aos dados de proveniência. Em relação à proveniência, o DEW mantém a modelagem do Chiron que adota o modelo do PROV-Wf (Costa et al. 2013) baseado no W3C PROV. Outra característica do Chiron que pode beneficiar o DEW é a definição de atividades com restrição (Ogasawara et al. 2011) que permitem a execução de aplicações paralelas programadas em MPI, por exemplo, reservando um conjunto de processadores para cada instância da atividade.

3. A Tecnologia P2P

Os ambientes P2P podem ser classificados quanto à sua organização em: centralizado e descentralizado. Os sistemas centralizados são coordenados por um ou mais super-nós e os descentralizados todos os nós podem desempenhar as mesmas funcionalidades, inclusive as de controle. Uma outra organização é a das redes P2P hierárquicas, em que os nós da rede são organizados em grupos menores. Os nós de um determinado grupo apresentam maior conectividade dentro deste grupo, assim como garantem a característica de descentralização do controle da rede. Além disso, alguns dos nós do grupo se conectam a outros grupos existentes em níveis hierárquicos maiores.

A partir de estudos realizados em trabalhos anteriores, como o SciMule (Ogasawara et al. 2010) e o Heracles (Dias et al. 2010a), o DEW segue a abordagem hierárquica. As principais vantagens estão relacionadas à escalabilidade obtida com a dinamicidade dos nós, ao custo de manutenção reduzido, ao balanceamento de carga e à minimização dos efeitos de *volatilidade*. Outro ponto importante é a boa relação de custo e benefício entre os modelos centralizado e descentralizado para permitir a escalabilidade da rede. Assim, o custo de manutenção, o balanceamento de carga e a minimização das consequências da volatilidade se apresentam em um ponto de equilíbrio quando comparados às demais abordagens.

4. A Arquitetura do DEW (*Distributed Execution control for Workflow engines*)

O DEW é um mecanismo com controle distribuído responsável por gerenciar a execução das atividades de *workflows* científicos por meio de uma rede P2P hierárquica em *clusters* e nuvens de computadores. Seu objetivo é possibilitar a paralelização de *workflows* em ambientes heterogêneos e com sistemas de arquivos distribuídos, garantindo a escalabilidade dos nós e de recuperação da rede nas ocorrências de falhas.

A rede P2P criada pelo DEW apresenta nós com dois tipos de responsabilidade: o nó de execução e o nó de distribuição. Os nós de execução têm o objetivo de processar as ativações recebidas do seu nó de distribuição, assim como informar o estado da execução de uma ativação e fornecer os dados de proveniência da execução. Ativação é o menor conjunto de dados suficiente para executar uma instância de uma

atividade, incluindo o conjunto de parâmetros, informações a respeito dos dados de entrada, dados de saída e a própria aplicação, os quais são utilizados para possíveis transferências de dados e para a execução da atividade. Uma ativação é autocontida, de forma que o nó de execução, após recebê-la, tem condições de executar uma instância da atividade por completo. Os nós de distribuição, escalonam as ativações pendentes (ainda não processadas) para os nós de execução disponíveis. A configuração de um nó na rede acontece conforme novos recursos computacionais são identificadas e adicionados à rede P2P. A escolha de nós de distribuição utiliza a proporção entre o número de nós de distribuição e o número total de nós existentes. Caso esse valor seja inferior a um limite, o primeiro nó de execução a constatar essa condição é elencado como nó de distribuição.

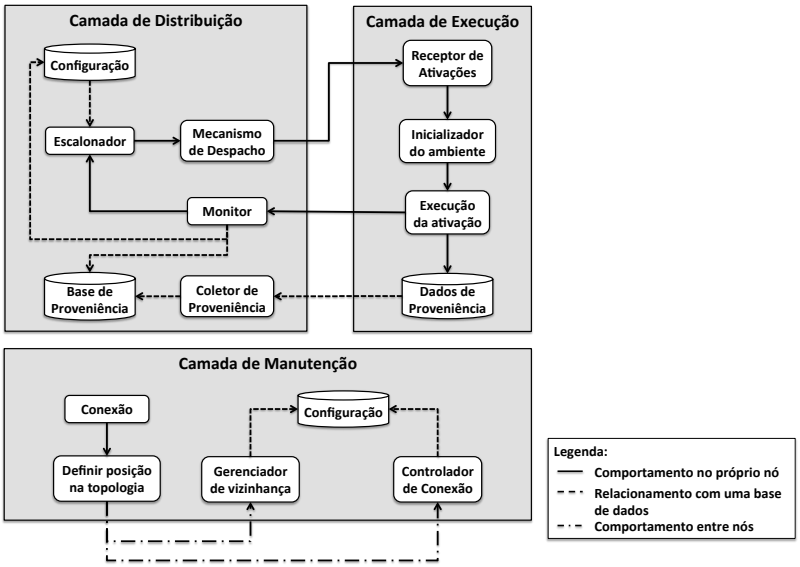

Figura 1. Arquitetura do DEW

A arquitetura do DEW (Figura 1) é composta de três camadas: Distribuição, responsável pelo escalonamento de ativações e pelo despacho das mesmas para os nós de execução; Execução, que lida com a recepção e execução das ativações recebidas; e Manutenção, que gerencia a configuração da rede P2P, a inserção de novos nós e a atualização da topologia.

A camada de distribuição é inicializada com as propriedades do nó de distribuição a partir da base de configuração. Após essa etapa, o escalonador é ativado e gerencia as ativações a serem distribuídas para os nós de execução usando o mecanismo de despacho. Vale ressaltar que os nós de distribuição consultam as informações relevantes da ativação (proveniência prospectiva) na base de proveniência. Já o monitor realiza um processo contínuo de análise do estado de execução das ativações, a partir de mensagens recebidas dos nós de execução. O elemento coletor de proveniência fornece as informações de proveniência para a base de proveniência. A camada de distribuição está presente apenas para os nós de distribuição.

A camada de execução é ativada pela recepção de uma ativação. De forma análoga, pode-se dizer que o mecanismo de despacho corresponde ao transmissor de ativação. Após obter uma ativação, o ambiente é configurado de acordo com as especificações obtidas da atividade do *workflow*. Em seguida, a ativação é executada.

Durante esse processo, dados de proveniência são encaminhados para os nós de distribuição, que realizam as devidas mudanças na base de proveniência.

Por último, a camada de manutenção apresenta três responsabilidades: o estabelecimento de conexões com novos nós, a definição do tipo de nó e seu posicionamento na topologia e a atualização das configurações da rede. Esta última função tem o objetivo de atualizar dados de nós vizinhos e diferentes mecanismos de troca de mensagens estabelecidos (anúncios, *pipes*, entre outros).

5. Avaliação Experimental

Foi realizada uma avaliação para comparar o desempenho do controle distribuído com o controle centralizado, simulando, respectivamente, o DEW e o Chiron. O mecanismo de distribuição de ativações original do Chiron é baseado na troca de mensagens MPI, por meio de um nó central (*i.e.* rank 0), que permite a submissão de ativações para outros nós de execução. DEW estabelece uma rede P2P hierárquica na qual diversos nós podem distribuir e executar ativações.

Nessa avaliação foi medida a eficiência da execução paralela do *workflow* considerando como parâmetros o número de nós existentes na rede e a volatilidade. A avaliação foi feita a partir do simulador SciMulator (Dias et al. 2010b) para alcançar resultados com um número elevado de nós. As execuções dos workflows tiveram duração média de quatro dias. Cada submissão de workflow envolveu um conjunto de dados de 256MB e aproximadamente 128 ativações em uma rede com número médio de vizinhos igual a 32 e conectividade máxima de 64 nós. O escalonamento considerando o custo da transferência de dados ainda não foi modelado no DEW. Por esse motivo, os experimentos consideraram que os dados estavam localizados geograficamente próximos aos nós.

A Figura 2(a) apresenta a eficiência considerando a variação no número de nós sem a ocorrência de volatilidade (*i.e.* nenhum nó entra ou sai da rede após a inicialização). Pelo gráfico é possível perceber que a abordagem do Chiron (representada como MPI) apresentou uma eficiência muito semelhante ao DEW em todas as configurações de nós utilizadas. Vale enfatizar, contudo, que a partir de 2048 nós, a eficiência do controle centralizado começa a se distanciar da eficiência obtida com o controle distribuído. Este resultado é decorrente da sobrecarga do nó central na troca de mensagens, gerando ociosidade nos nós de execução do Chiron com MPI.

A Figura 2(b) analisa a eficiência considerando uma volatilidade igual a 5%. Pelos resultados é possível perceber que a abordagem com controle centralizado apresentou resultado inferior ao DEW, uma vez que a estrutura da rede MPI é perdida na ocorrência de volatilidade em um determinado nó, afetando o desempenho do Chiron-MPI. O DEW, por conseguinte, apresentou resultado consistente se comparado à não ocorrência de volatilidade, pois a rede permite a redistribuição de ativações de nós que saem. Além disso, o desempenho da abordagem centralizada não se aproxima tanto da arquitetura DEW como na Figura 2(a), pois, para os maiores valores de nós, a probabilidade de que um nó sofra com a volatilidade (a falha de um nó implica a desestruturação da rede) é maior.

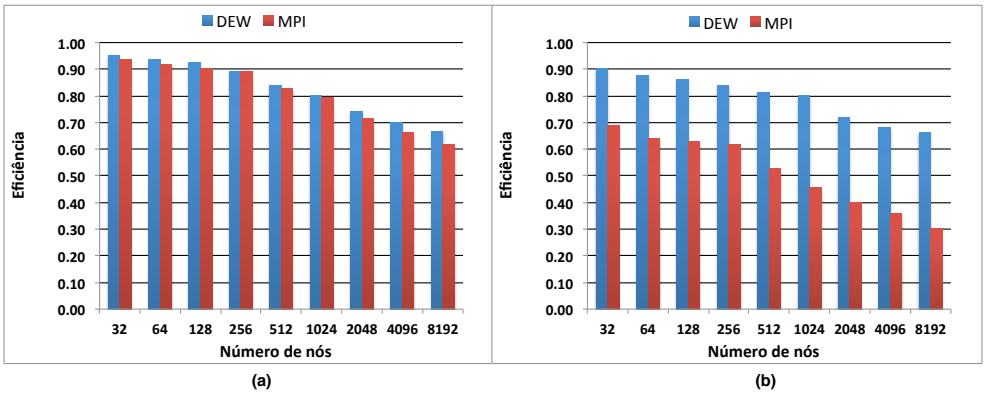

Figura 2. Eficiência sem volatilidade (a); Eficiência com volatilidade de 5% (b)

6. Conclusão

A execução de workflows científicos está cada dia mais associada ao uso em ambientes de PAD, o que contribui para a existência de diferentes mecanismos para a distribuição de atividades. O DEW é um mecanismo com controle descentralizado que pode ser acoplado a uma máquina de execução de *workflows*. O DEW apoia a execução paralela de *workflows* em ambientes heterogêneos e sem disco compartilhado por meio da organização de recursos em uma rede P2P hierárquica. A escolha da organização hierárquica está associada à escalabilidade da rede, uma vez que o aumento considerado no número de nós, em uma abordagem completamente descentralizada, implica o alto custo de manutenção da rede.

A avaliação experimental foi baseada na simulação e comparação da abordagem centralizada do Chiron, utilizando MPI, e o mecanismo de distribuição do DEW. Podemos perceber vantagens em relação à escalabilidade e à tolerância a falhas. Outro ponto importante foi a eficiência no caso de volatilidade igual a 5%, na qual o DEW apresentou melhor desempenho em relação à abordagem centralizada.

Como trabalhos futuros, pretendemos aprimorar o mecanismo de distribuição do DEW e sua integração com o mecanismo de execução do Chiron para realizarmos experimentos reais. Também desenvolveremos um modelo de custos para análise da transferência de dados.

Agradecimentos. Os autores gostariam de agradecer ao CNPq, Capes e FAPERJ pelo financiamento parcial deste trabalho.

Referências Bibliográficas

Apache Software Foundation, (2009), "Hadoop", *Apache Hadoop Website*

Chandra, R., (2001), *Parallel programming in OpenMP*. Morgan Kaufmann.

Costa, F., Silva, V., Oliveira, D., Ocaña, K., Dias, J., Ogasawara, E., Mattoso, M., (2013), "Capturing and Querying Workflow Runtime Provenance with PROV: a Practical Approach". In: *Proc. of the International Workshop on Managing and Querying Provenance Data at Scale (BigProv'13)*, Genova, Italy.

Dias, J., Ogasawara, E., de Oliveira, D., Pacitti, E., Mattoso, M., (2010a), "Improving Many-Task computing in scientific workflows using P2P techniques". In:

Proceedings of the 3rd IEEE Workshop on Many-Task Computing on Grids and Supercomputers, p. 1–10, New Orleans, Louisiana, USA.

Dias, J., Rodrigues, C., Ogasawara, E., Oliveira, D., Braganholo, V., Pacitti, E., Mattoso, M., (2010b), "SciMulator: Um Ambiente de Simulação de Workflows Científicos em Redes P2P". In: *Workshop P2P 2010*, p. 45–56, Gramado, Rio Grande do Sul - Brazil.

Freire, J., Koop, D., Santos, E., Silva, C. T., (2008), "Provenance for Computational Tasks: A Survey", *Computing in Science and Engineering, v.10*, n. 3, p. 11–21.

Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., et al., (2004), "Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation", *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, , p. 353–377.

Mattoso, A., Silva, F., Ruberg, N., Cruz, M., (2008), "Gerência de Workflows Científicos: Uma Análise Crítica no Contexto da Bioinformática", *COPPE/UFRJ*, n. Relatório técnico

Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P., Mattoso, M., (2011), "An Algebraic Approach for Data-Centric Scientific Workflows", *Proc. of VLDB Endowment*, v. 4, n. 12, p. 1328–1339.

Ogasawara, E., Dias, J., Oliveira, D., Rodrigues, C., Pivotto, C., Antas, R., Braganholo, V., Valduriez, P., Mattoso, M., (2010), "A P2P approach to many tasks computing for scientific workflows". In: *VECPAR'10*, p. 327–339, Berlin, Heidelberg.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., Oliveira, D., Porto, F., Valduriez, P., Mattoso, M., (2013), "Chiron: A Parallel Engine for Algebraic Scientific Workflows", *Concurrency and Computation*

Papuzzo, G., Spezzano, G., (2011), "Autonomic management of workflows on hybrid grid-cloud infrastructure", *CNSM '11 Proceedings of the 7th International Conference on Network and Services Management*

Rahman, M., Ranjan, R., Buyya, R., (2010), "Cooperative and decentralized workflow scheduling in global grids", *Future Generation Computer Systems*, v. 26, n. 5 (May.), p. 753–768.

Raicu, I., Foster, I. T., Yong Zhao, (2008), "Many-task computing for grids and supercomputers". In: *Proceedings of the Workshop on Many-Task Computing on Grids and Supercomputers*, p. 1–11, Austin, Texas, USA.

Taylor, I., Shields, M., Wang, I., Harrison, A., (2007), "The Triana Workflow Environment: Architecture and Applications", *Workflows for e-Science*, Springer, p. 320–339.

Wu, D., Tian, Y., Ng, K.-W., Datta, A., (2008), "Stochastic analysis of the interplay between object maintenance and churn", *Computer Communications*, v. 31, n. 2 (Feb.), p. 220–239.