Armazenamento distribuído de dados e checkpointing de aplicações paralelas em grades oportunistas

Raphael Yokoingawa de Camargo

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE DOUTOR
EM
CIÊNCIAS

Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Fabio Kon

Durante a elaboração deste trabalho o autor recebeu o auxílio financeiro do CNPq

São Paulo, 30 de Agosto de 2007

Armazenamento distribuído de dados e checkpointing de aplicações paralelas em grades oportunistas

Este exemplar corresponde à redação
final da tese devidamente corrigida
e defendida por Raphael Yokoingawa de Camargo
e aprovada pela Comissão Julgadora.

São Paulo, 30 de agosto de 2007

Banca Examinadora:

- Prof. Dr. Fabio Kon (orientador) - IME-USP.
- Prof. Dr. Marco Dimas Gubitoso - IME-USP
- Prof. Dr. Luiz Eduardo Buzato IC - Unicamp
- Prof. Dr. Markus Endler - PUC-Rio
- Prof. Dr. Francisco Brasileiro - UFCG

Dedico este trabalho à Selma, meu grande amor,
cuja compreensão, dedicação e carinho trouxeram equilíbrio e grandes alegrias à minha vida.

Agradecimentos

Agradeço de modo especial ao professor Fabio Kon que, com sua competente orientação, constantemente me motivou a superar os desafios da carreira acadêmica. Ao CNPq, pelo seu apoio financeiro durante todos estes anos.

A todos meus amigos do LCPD, das listas `aabalados` e `psycohunters` e da ETFSP. Dos simples almoços no bandejão às longas baladas noturnas, sua companhia foi essencial para tornar minha vida mais alegre. E um agradecimento especial àqueles que ainda contribuiram com meu trabalho de doutorado.

Aos meus pais, por todo o apoio e carinho que possibilitaram o meu amadurecimento e a conclusão deste doutorado, e às minhas irmãs e meus familiares, por sua constante presença a meu lado.

À Selma, por me apoiar nos momentos mais difíceis. Seu afeto e compreensão me deram forças para superar este grande desafio que é o doutorado.

RESUMO

Grades computacionais oportunistas utilizam recursos ociosos de máquinas compartilhadas para executar aplicações que necessitam de um alto poder computacional e/ou trabalham com grandes quantidades de dados. Mas a execução de aplicações paralelas computacionalmente intensivas em ambientes dinâmicos e heterogêneos, como grades computacionais oportunistas, é uma tarefa difícil. Máquinas podem falhar, ficar inacessíveis ou passar de ociosas para ocupadas inesperadamente, comprometendo a execução de aplicações. Um mecanismo de tolerância a falhas que dê suporte a arquiteturas heterogêneas é um importante requisito para estes sistemas. Neste trabalho, analisamos, implementamos e avaliamos um mecanismo de tolerância a falhas baseado em checkpointing para aplicações paralelas em grades computacionais oportunistas. Este mecanismo permite o monitoramento de execuções e a migração de aplicações entre nós heterogêneos da grade.

Mas além da execução, é preciso gerenciar e armazenar os dados gerados e utilizados por estas aplicações. Desejamos uma infra-estrutura de armazenamento de dados de baixo custo e que utilize o espaço livre em disco de máquinas compartilhadas da grade. Devemos utilizar somente os ciclos ociosos destas máquinas para armazenar e recuperar dados, de modo que um sistema de armazenamento distribuído que as utilize deve ser redundante e tolerante a falhas.

Para resolver o problema do armazenamento de dados em grades oportunistas, projetamos, implementamos e avaliamos o middleware OppStore. Este middleware provê armazenamento distribuído e confiável de dados, que podem ser acessados de qualquer máquina da grade. As máquinas são organizadas em aglomerados, que são conectados por uma rede *peer-to-peer* auto-organizável e tolerante a falhas. Dados são codificados em fragmentos redundantes antes de serem armazenados, de modo que arquivos podem ser reconstruídos utilizando apenas um subconjunto destes fragmentos. Finalmente, para lidar com a heterogeneidade dos recursos, desenvolvemos uma extensão ao protocolo de roteamento em redes *peer-to-peer* Pastry. Esta extensão adiciona balanceamento de carga e suporte à heterogeneidade de máquinas ao protocolo Pastry.

ABSTRACT

Opportunistic computational grids use idle resources from shared machines to execute applications that need large amounts of computational power and/or deal with large amounts of data. But executing computationally intensive parallel applications in dynamic and heterogeneous environments, such as opportunistic grids, is a daunting task. Machines may fail, become inaccessible, or change from idle to occupied unexpectedly, compromising the application execution. A fault tolerance mechanism that supports heterogeneous architectures is an important requisite for such systems. In this work, we analyze, implement and evaluate a checkpointing-based fault tolerance mechanism for parallel applications running on opportunistic grids. The mechanism monitors application execution and allows the migration of applications between heterogeneous nodes of the grid.

But besides application execution, it is necessary to manage data generated and used by those applications. We want a low cost data storage infrastructure that utilizes the unused disk space of grid shared machines. The system should use the machines to store and recover data only during their idle periods, requiring the system to be redundant and fault-tolerant.

To solve the data storage problem in opportunistic grids, we designed, implemented and evaluated the OppStore middleware. This middleware provides reliable distributed storage for application data, which can be accessed from any machine in the grid. The machines are organized in clusters, connected by a self-organizing and fault-tolerant peer-to-peer network. During storage, data is codified into redundant fragments, allowing the reconstruction of the original file using only a subset of those fragments. Finally, to deal with resource heterogeneity, we developed an extension to the Pastry peer-to-peer routing substrate, enabling heterogeneity-aware load-balancing message routing.

Sumário