Centro Federal de Educação Tecnológica de Minas Gerais

Departamento de Computação

Algoritmos paralelos de ordenação em Hadoop

Mariane Raquel Silva Gonçalves

Orientadora: Cristina Duarte Murta

Belo Horizonte
14 de março de 2012

Sumário

1 Proposta do Projeto

1.1 Tema do trabalho

A ordenação é o processo de reordenar uma sequência de entrada e produzir uma saída ordenada de acordo com um atributo. A ordenação paralela é o processo de uso de múltiplas unidades de processamento para ordenar em conjunto uma sequência. Na criação de algoritmos de ordenação paralela, é ponto fundamental ordenar coletivamente os dados de cada processo individual, de forma a utilizar todas as unidades de processamento e minimizar os custos de redistribuição de chaves entre os processadores. [Kale e Solomonik 2010]

Fatores como movimentação de dados, balanço de carga, latência de comunicação e distribuição inicial das chaves são considerados ingredientes chave para o bom desempenho da ordenação paralela, e variam de acordo com o algoritmo escolhido como solução. O presente projeto propõe a análise de desempenho, em termos do número de dados a serem ordenados e de máquinas utilizadas, de um ou mais algoritmos de ordenação paralela, implementados de acordo com o modelo MapReduce no ambiente Hadoop.

1.2 Relevância

A arquitetura paralela é um conceito conhecido há várias décadas

Com a recente mudança de foco no desenvolvimento de algoritmos sequenciais para paralelos, conhecer os modelos de programação paralela se tornou uma grande necessidade na computação. Os algoritmos paralelos ainda são um ramo pouco explo-

rado, devido a maior complexidade no desenvolvimento e recentes aplicações em sistemas *multicore*.

Uso crescente de computação paralela para sistemas computacionais gera a necessidade de algoritmos de ordenação inovadores, desenvolvidos para dar suporte a essas aplicações. O MapReduce[Dean e Ghemawat 2008] é um modelo de programação paralela criado pela Google para processamento e geração de grandes volumes de dados em *clusters*. Esse modelo propõe simplificar a computação paralela e ser de fácil uso, abstraindo conceitos complexos da paralelização - como tolerância a falhas, distribuição de dados e balanço de carga - e utilizando duas funções principais: map e reduce. A complexidade do algoritmo paralelo não é vista pelo desenvolvedor, que pode se ocupar em desenvolver a solução proposta. O Hadoop [White 2009] é uma implementação do MapReduce, um *framework* open source que provê o gerenciamento de computação distribuída.

1.3 Objetivos

O projeto propõe a

(i) Estudar a programação paralela, seus algoritmos e suas possibilidades de implementação em ambiente paralelo multicore; (ii) Implementar e avaliar o desempenho de um algoritmo de ordenação paralela; (iii) Estudar e implementar a solução no modelo MapReduce, com o software Hadoop; (iv) Comparar os resultados obtidos com os algoritmos de ordenação e analisar seu desempenho com relação à quantidade de dados a serem ordenados, variabilidade dos dados de entrada e número máquinas utilizadas.

1.4 Resultados esperados

Com a realização do trabalho, busca-se ampliar e consolidar conhecimentos adquiridos na área de computação paralela, assim como a capacidade de análise e desenvolvimento de algoritmos paralelos. Os resultados esperados incluem entendimento do modelo MapReduce, implementação de algoritmos de ordenação paralela em ambiente Hadoop e realização de experimentos para comparação de desempenho entre dois ou mais

algoritmos.

1.5 Metodologia

O início do projeto será destinado ao estudo da computação paralela, algoritmos de ordenação paralela, dos fatores que influenciam o desempenho de tais algoritmos, o modelo MapReduce e a plataforma Hadoop. O passo seguinte é conhecer detalhadamente o algoritmo paralelo a ser implementado e definir as estratégias para sua implentação ambiente Hadoop. O algoritmo implementado deve ser cuidadosamente avaliado para verificar um funcionamento adequado com diferentes entradas e número de máquinas. Em seguida, serão realizados experimentos para testes de desempenho dos algoritmos com relação à quantidade de máquinas, quantidade de dados e conjunto de dados. Os resultados obtidos serão analisados e permitirão comparar a performance dos algoritmos em cada situação.

1.6 Infraestrutura necessária

A infra estrutura necessária ao desenvolvimento do projeto será fornecida pelo Laboratório de Redes e Sistemas (LABORES) do Departamento de Computação (DECOM). Esse laboratório possui um cluster formado por cinco máquinas Dell Optiplex 380, que serão utilizadas na realização dos testes dos algoritmos. Os algoritmos serão desenvolvidos em linguagem Java, de acordo com o modelo MapReduce, no ambiente Hadoop.

O cluster a ser utilizado apresenta as seguintes características:

- 5 nodos
- Processador Intel Core 2 Duo de 3.0 GHz
- Disco rígido SATA de 500 GB 7200 RPM
- Memória RAM de 4 GB
- Placa de rede Gigabit Ethernet

- Sistema operacional Linux Ubunbu 10.04 32 bits (kernel 2.6.**XX**)
- Sun Java JDK 1.6.0 19.0-b09
- Hadoop 0.20.2

1.7 Cronograma de trabalho

Março	Pesquisa bibliográfica e escrita da proposta de projeto.
Abril	Pesquisa bibliográfica, contextualização de algoritmos de ordenação paralelos, familiarização com ambiente Hadoop e com o modelo MapReduce. Escrita dos itens introdução e referencial teórico do projeto.
Maio	Realização de experimentos com o algoritmo de ordenação paralela encontrado na literatura ou desenvolvido de acordo com a metodologia proposta. Descrição dos experimentos e da metodologia no texto do projeto.
Junho	Finalização e entrega do projeto.
Julho	Desenvolvimento algoritmos de ordenação
Agosto	Desenvolvimento e testes dos algoritmos desenvolvidos
Setembro	Análise dos resultados obtidos até o momento, em busca de pontos de melhorias no projeto
Outubro	Teste final dos algoritmos, análise e escrita dos resultados
Novembro	Finalização da escrita do projeto final, entrega e apresentação

Tabela 1: Cronograma proposto para o projeto

Citações:

[Kale e Solomonik 2010]

[Manferdelli, Govindaraju e Crall 2008]

[Dean e Ghemawat 2008]

[Asanovic et al. 2009]

Referências

[Asanovic et al. 2009]ASANOVIC, K. et al. A view of the parallel computing landscape. *Commun. ACM*, ACM, New York, NY, USA, v. 52, n. 10, p. 56–67, out. 2009. ISSN 0001-0782.

[Dean e Ghemawat 2008]DEAN, J.; GHEMAWAT, S. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, ACM, New York, NY, USA, v. 51, n. 1, p. 107–113, jan. 2008. ISSN 0001-0782.

[Kale e Solomonik 2010]KALE, V.; SOLOMONIK, E. Parallel sorting pattern. In: *Proceedings of the 2010 Workshop on Parallel Programming Patterns*. New York, NY, USA: ACM, 2010. (ParaPLoP '10), p. 10:1–10:12. ISBN 978-1-4503-0127-5.

[Manferdelli, Govindaraju e Crall 2008]MANFERDELLI, J. L.; GOVINDARAJU, N. K.; CRALL, C. Challenges and opportunities in Many-Core computing. *Proceedings of the IEEE*, v. 96, n. 5, p. 808–815, maio 2008. ISSN 0018-9219.

[White 2009]WHITE, T. *Hadoop: The Definitive Guide*. first edition. O'Reilly, 2009. Disponível em: <http://oreilly.com/catalog/9780596521981>.