Avaliação do comportamento do Map-Reduce em ambientes Homogêneos e Heterogêneos

Alexandre K.S. Miyazaki[1] e Claudio R. Geyer[1]

Instituto de Informática - Universidade Federal do Rio Grande do Sul (UFRGS)[1]

MapReduce

- ✓ Fornece um mecanismo que permite abstrair a complexidade envolvida no paralelismo das aplicações [1]
- ✓ Baseia-se em linguagens funcionais de alto nível como LISP. **Proposta concebida para uso em grandes *clusters*, com recursos (processadores, memória e armazenamentos) homogêneos.**
- ✓ O *Hadoop* é a implementação *open source* mais popular do *MapReduce*, proposta pela *Apache Software Foundation* e usada por corporações como Yahoo, Amazon e FaceBook

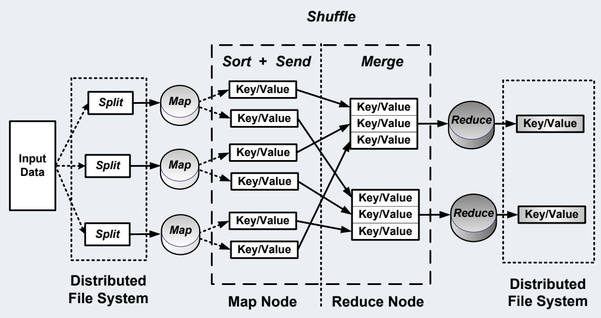

Figura: Estrutura do Modelo de Programa

O Simulador MRSG

- ✓ Desenvolvido sobre a plataforma SimGrid [2].
- ✓ Alternativa para testes em ambientes reais, onde é necessário uma grande quantidade de recursos.
- ✓ Fornece uma série de facilidades para implementação de algoritmos teóricos, mudanças na configuração do sistema a ser testado .
- ■ **Objetivo do Trabalho:** Analisar o comportamento do MapReduce em um ambiente real, executando testes na Grid'5000 [3], obtendo resultados para um comparativo com testes realizados no Simulador.

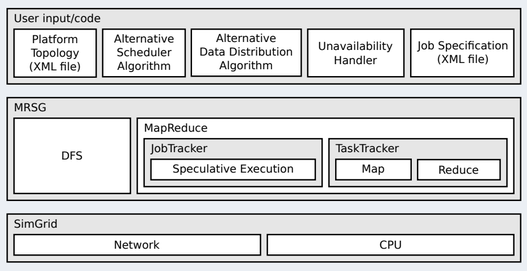

Figura: Arquitetura do simulador MRSG

Atividades

- ✓ Aprender a usar a Grid'5000, a grade computacional francesa, onde foram realizados os testes em ambiente real.
- ✓ Aprender a usar o Hadoop, que é uma implementação aberta do MapReduce.
- ✓ Pesquisar aplicações intensivas em dados para o Hadoop.
- ✓ Realizar testes na Grid'5000.

Configurações de testes

- ✓ A configuração do ambiente real variou em número de nós e tamanho de entrada linearmente.
- ✓ Foram feitos experimentos com 32, 64,128 e 200 nós, e o tamanho de entrada acompanhou linearmente o aumento em nós, fazendo com que cada máquina tivesse pelo menos 3 *chunks*, que são pedaços de 64Mb da entrada.
- ✓ Duas aplicações foram testadas, o TeraSort, aplicação de ordenação de dados, e um filtro de log, que recebe um arquivo contando a disponibilidade de máquinas, contendo uma linha para cada vez que a máquina esteve disponível e produz uma saída que tem uma linha por máquina.
- ✓ A configuração de testes foi a mesma para ambas aplicações.

Experimento - Filtro de log com 32 nós homogêneo

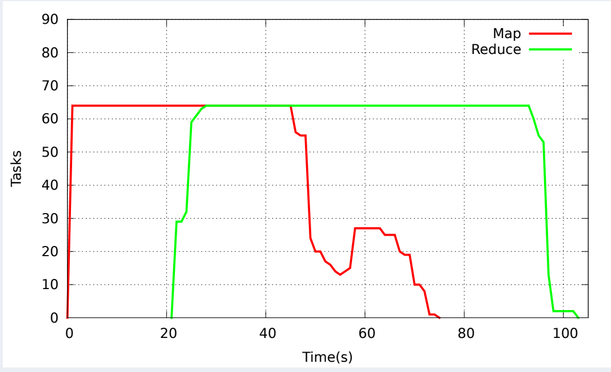

Figura: Filtro de log com 32 nós na Grid'5000

- ✓ Testes seguiram o comportamento esperado.
- ✓ "Lançamento especulativo" de tarefas.

Experimento - TeraSort com 32 nós Homogêneo

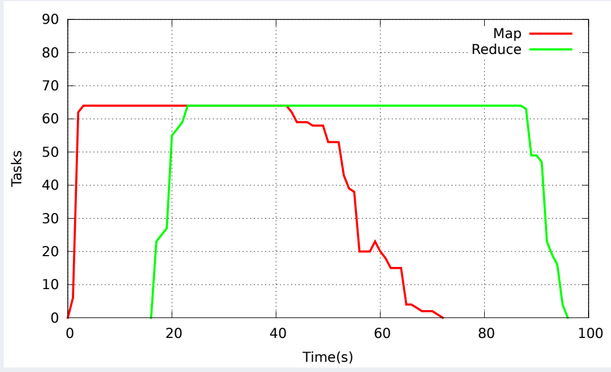

Figura: Terasort com 32 nós na Grid'5000

- ✓ Comportamento semelhante com o filtro de log por serem duas aplicações semelhantes, intensivas em dados.

Referências

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *OSDI*, 2004, pp. 137–150.

[2] SimGrid. Simgrid Homepage, Aug. 2011. Available on: <http://simgrid.gforge.inria.fr/>. Accessed on August 2011.

[3] Grid'5000. Grid5000 Homepage, Aug. 2011.Available on : <http://www.grid5000.fr/>. Acessed on August 2011.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Edital FAPERGS/CNPq n. 008/2009 Projeto de Pesquisa: **"GREEN-GRID: Computação de Alto Desempenho Sustentável"**.

{aksmiyazaki , geyer}@inf.ufrgs.br[1]