BIGHYBRID: UM TOOLKIT PARA A SIMULAÇÃO DE APLICAÇÕES BIG DATA SOBRE INFRAESTRUTURAS DE NUVEM E COMPUTAÇÃO VOLUNTÁRIA

GRUPO DE PROCESSAMENTO PARALELO E DISTRIBUÍDO(GPPD)
AUTOR: VINÍCIUS PITTIGLIANI PEREGO
ORIENTADOR: CLÁUDIO GEYER.

INTRODUÇÃO

O alto consumo de dados por aplicações modernas requer alta disponibilidade de sistemas distribuídos. Nesse cenário, computação em nuvem (*Cloud Computing*) surge como uma solução devido sua disponibilidade, mas cobrando usuários pelos recursos utilizados. Alternativamente, *Desktop Grid* (DG) pode ser utilizado para executar essas aplicações através de poder computacional obtido por doação de usuários enquanto suas máquinas estão ociosas.

Hadoop [1] é um framework que implementa MapReduce [2], um modelo de programação que abstrai a paralelização e facilita a criação de programas distribuídos. Esse trabalho propõe alterações na implementação original do Hadoop visando a sua utilização em infraestruturas híbridas baseadas em computação em nuvem e DG. As mudanças foram realizadas em um *toolkit* implementado sobre o SimGrid [3], uma ferramenta para simulação de sistemas distribuídos que permite ao usuário especificar plataforma de execução, algoritmos de escalonamento e obter métricas como o tempo de execução.

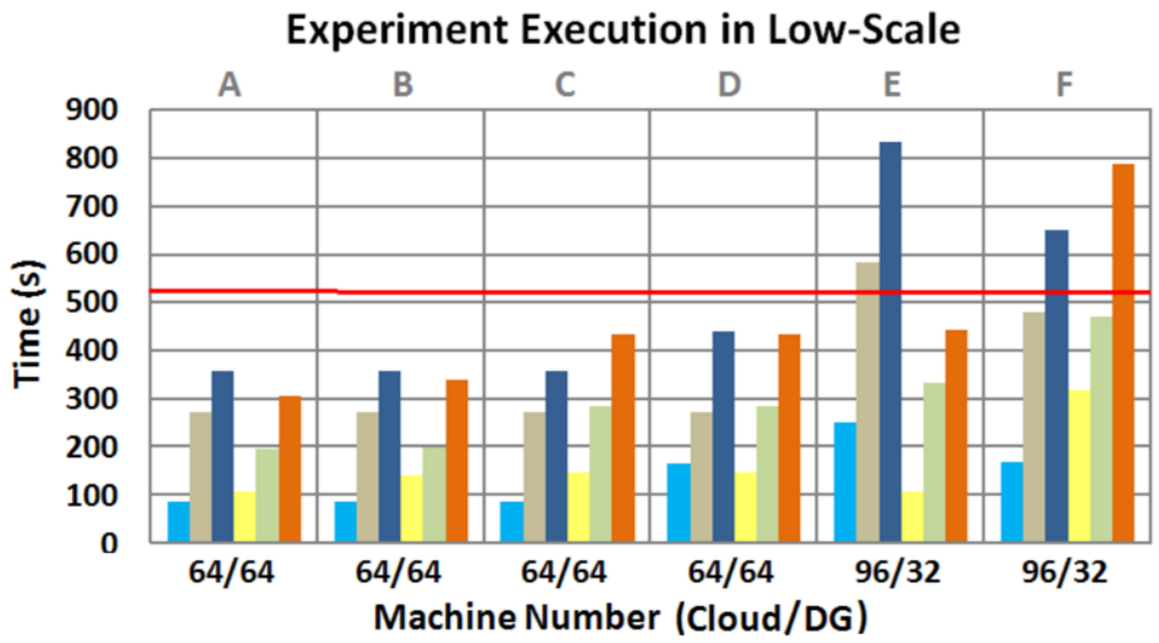
OBJETIVOS

- Simulação precisa de aplicações MapReduce, baseando-se em cenários reais.
- Análise de estratégias para ambientes híbridos, permitindo o monitoramento e obtenção de métricas da execução.
- Modularidade no projeto, visando futuras extensões.

MODELO DO TOOLKIT BIGHYBRID

O *toolkit* BigHybrid segue uma arquitetura em camadas, na qual a camada superior recebe um conjunto de dados com especificações sobre os ambientes de *Cloud* e DG, além da carga de trabalho. Ao processar essas informações, a segunda camada inicia a execução de dois diferentes *middlewares* com algoritmos específicos para o escalonamento em ambiente homogêneo (*Cloud*) e heterogêneo (DG). Na simulação de *Cloud*, são utilizados algoritmos semelhantes à implementação do Hadoop, enquanto o simulador de DG estende o simulador anterior com adaptações no escalonamento e um módulo de tolerância a falhas para tratamento de nós voluntários.

RESULTADOS



Item	A	B	C	D	E	F
Chunk size C/DG	64/16	64/32	64/64	128/64	64/64	64/64
Input Data C/DG	128/288	128/144	128/72	64/72	136/64	110/90

Cloud_MAP, Cloud_REDUCE, Cloud_JOB, DG_MAP, GD_REDUCE, DG_JOB

A imagem representa um experimento em um pequeno ambiente híbrido com entrada de dados de 128 GB e 128 máquinas. A linha vermelha representa uma única execução realizada somente na Cloud de 200 tarefas com chunks de 64 MB, com um tempo equivalente a 503 segundos. Os 5 itens abaixo são cenários híbridos que representam variações sobre o mesmo valor de entrada e número de máquinas distribuídos em Cloud (C) e DG. Os cenários A, B, C e D mostram um ganho de desempenho em relação ao cenário original. Nos cenários E e F são obtidos tempos maiores, devido a sobrecarga de chunks em C e poucas máquinas em DG.

CONCLUSÃO

Este trabalho apresentou uma ferramenta para avaliar uma solução alternativa ao processamento de *Big Data*. Os resultados obtidos do simulador apresentam boa acurácia, com aproximadamente 5% de erro em relação ao cenário real. Algumas extensões precisam ser feitas no trabalho proposto para expandir a capacidade de simulação, como a simulação de contenção de IO e a capacidade de simular outras aplicações Big Data além de MapReduce.

REFERÊNCIAS

[1] T. White. Hadoop: The definitive guide. ” O’Reilly Media, Inc.”, 2012.
[2] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008
[3] H. Casanova. Simgrid: A toolkit for the simulation of application scheduling. In Cluster computing and the grid, 2001. proceedings. first ieee/acm international symposium on, pages 430–437. IEEE, 2001.