Comparação de desempenho do processamento paralelo de consultas de banco de dados em CPUs multi-core e GPUs

Simone Dominico[1], **Marco Antonio Zanata Alves**[1], **Eduardo Cunha de Almeida**[1]

[1]Programa de Pós graduação em Informática – Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

{sdominico,mazalves,eduardo}@inf.ufpr.br

Resumo. *O processamento paralelo é uma solução para melhorar o desempenho de consultas de banco de dados, reduzindo o tempo de resposta, e aumentando a vazão no processamento de consultas. Com a evolução de* hardware *surgiram novas tecnologias para o paralelismo. Uma delas é o uso de GPUs (Graphics Processing Units) para processamento de propósito geral. A GPU é uma unidade de processamento massivamente paralela com um número maior de núcleos executando em uma frequência menor comparado a CPU (Central Processing Unit). Neste contexto, este artigo apresenta um estudo comparativo do processamento de uma operação de consulta na GPU e CPU.*

1. Introdução

Ao longo dos anos, o volume de dados vem crescendo exponencialmente com estimativas globais na ordem de 180 zettabytes [Holst 2021] para o ano 2025. Esses dados são gerenciados por diferentes bancos de dados, enfatizando a necessidade de processamento rápido de consultas. Atualmente, processadores possuem múltiplos núcleos e diferentes níveis de compartilhamento na hierarquia de memória, os quais afetam a performance de consultas. Portanto, espera-se aumento de desempenho em execuções paralelas beneficiadas pelos múltiplos núcleos de processamento.

Em paralelo, com a evolução do *hardware*, servidores modernos estão adotando cada vez mais as GPUs, visando melhorar a capacidade energética e computacional do sistema. Os recursos de computação das GPUs modernas atendem à demanda de processamento de grandes quantidades de dados. Com isso, uma arquitetura massivamente paralela está disponível para realizar processamento de consultas [Volk et al. 2010]. Desta forma, este trabalho apresenta uma comparação do desempenho do processamento paralelo de consulta entre uma CPU multi-core e GPUs.

2. Metodologia experimental e Resultados

Os experimentos foram realizados comparando a execução de uma consulta simples que realiza uma agregação de duas colunas. A consulta é escrita na linguagem C utilizando OpenMP para o paralelismo e adaptada para a execução em uma GPU. No código implementado para a GPU foi utilizada a diretiva do OpenMP *teams distribute* para construir equipes de *threads*, permitindo a melhor distribuição do trabalho e adequação à organização hierárquica das unidades de processamento das GPUs. Foram criadas 15 equipes com 20 *threads*. Na implementação para a CPU o código foi paralelizado definindo a quantidade de 20 *threads*.

Nos experimentos, foi utilizada uma máquina com dois soquetes, onde cada um possui um processador *Intel Xeon Silver* 4114 executando o sistema operacional Ubuntu,

versão 18:04:01 LTS. A máquina possui duas placas de vídeo GeForce GTX TITAN Black com memória de 6 GB cada. Para a obtenção dos tempos de execução, os experimentos foram executados 20 vezes, calculando o tempo através da média dos resultados. O coeficiente de variação máximo encontrado para os resultados com CPU foi de 0,12 e GPU ficou em 0,20. Esses valores indicam baixa dispersão dos dados: dados homogêneos.

A Figura 1 mostra o tempo necessário para concluir a operação da agregação incluindo CPU e GPU para diferentes quantidades de tuplas, variando do menor ao maior sendo possível para armazenar na memória da GPU. Com quantidades menores que 80 mil tuplas, a operação processada na GPU não apresenta um bom desempenho. No entanto, quando a quantidade de tuplas é maior, aumentando também o cálculo realizado na agregação, o desempenho da CPU corresponde ao da GPU. Já para a quantidade de tuplas máxima analisada a transferência de dados entre CPU e GPU torna-se irrelevante e a execução na GPU reduz em 16% o tempo de processamento.

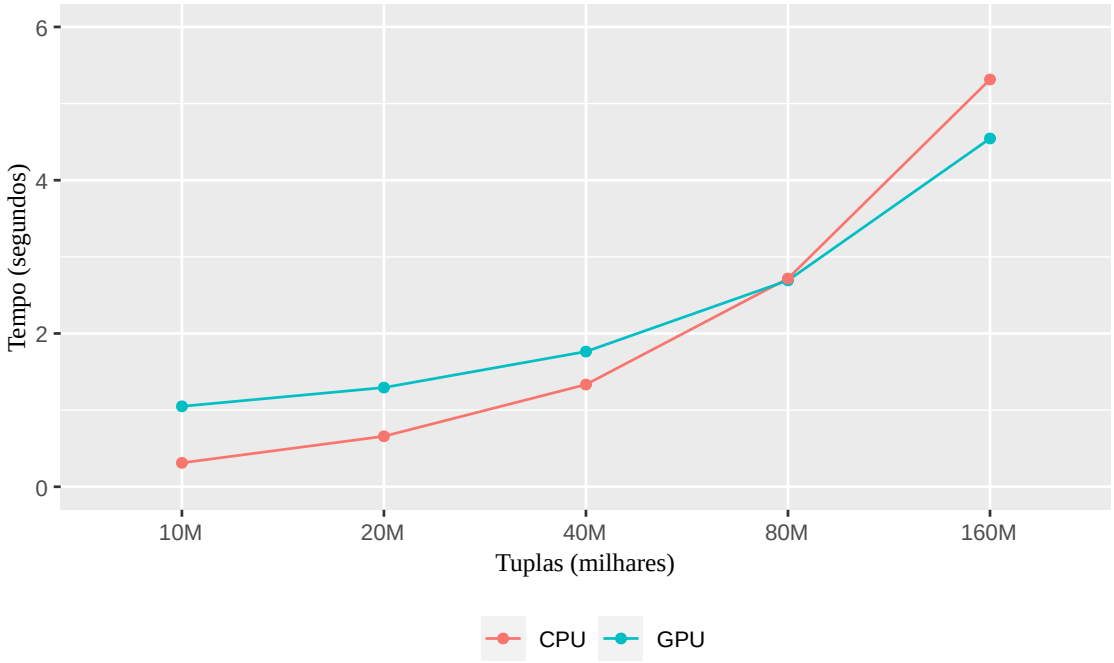

Figura 1. Comparação do tempo de execução entre CPU e GPU de uma operação de agregação.

3. Conclusão

A análise apresentada nesse artigo mostra que a execução de uma operação de consulta na GPU apresenta um bom desempenho quando o tempo de computação se sobrepõe a transferência de dados. Como trabalhos futuros, novas operações de consultas analíticas serão analisadas e avaliadas na GPU.

References

Holst, A. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025.

Volk, P. B., Habich, D., and Lehner, W. (2010). Gpu-based speculative query processing for database operations. In *ADMS@ VLDB*, pages 51–60.