POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Paralelní algoritmy pro operace s řídkými maticemi
Jméno autora:	Vít Novotný
Typ práce:	bakalářská práce
Fakulta:	Fakulta jaderná a fyzikálně inženýrská (FJFI)
Katedra:	Katedra matematiky
Oponent práce:	Ing. Jakub Klinkovský, Ph.D.
Pracoviště oponenta práce:	KSI FJFI ČVUT v Praze

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Zadání práce se týká relativně jednoduchých paralelních algoritmů pro sčítání a násobení řídké a husté matice, ale jejich implementace v rámci knihovny TNL probíhala "od nuly". Zadání má také 5 bodů a klade důraz na kvalitní a udržitelnou implementaci, což zvyšuje náročnost zadání.

Splnění zadání **splněno s výhradami**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

První dva body zadání považuji za splněné, protože práce prokazuje dostatečné porozumění programování GPU pomocí CUDA a TNL s využitím existujících datových struktur pro reprezentaci řídkých matic. V implementační části vzniklo několik funkcí pro uvedené algoritmy, několik unit testů a benchmark pro porovnání vlastních implementací, takže další dva body zadání jsou také splněny. V posledním bodě se nějak zapomnělo na dokumentaci, která není zmíněná v práci a chybí také v samotné implementaci, ale jistě nebude těžké ji před obhajobou doplnit.

Zvolený postup řešení **vhodný s výhradami**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postup řešení spočívá v tradičních krocích studia teorie a dostupných nástrojů, následující implementace a vyhodnocení výsledků. Implementace začala jednodušším algoritmem pro sčítání řídké a husté matice, poté vznikly dvě varianty komplikovanějšího paralelního algoritmu pro násobení řídké a husté matice. K použité metodice pro porovnávání výkonu jednotlivých funkcí mám ale výhrady, které uvedu dále v posudku.

Odborná úroveň **podprůměrná**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Algoritmus pro sčítání řídké a husté matice je popsaný a implementovaný správně. Jak je řečeno v práci, tento algoritmus se nepoužívá příliš často, a proto další optimalizace není důležité zkoumat.
Teoretický popis algoritmu pro násobení řídké a husté matice je také správně, ale shrnutí pomocí "pseudo-kódu" mohlo být lépe formalizované, aby bylo jasné, kde probíhá paralelizace. Celkově je mezi popisem teorie a paralelní implementace v práci obrovská propast a bez pečlivého studia kompletního kódu se nedá posoudit, jestli je paralelizace opravdu v pořádku. K problémům s paralelizací se dostanu ještě na konci posudku.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Formální a jazyková úroveň **průměrná**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Jazyková úroveň je v pořádku, práce je psána spisovnou angličtinou a obsahuje jen minimum jazykových a typografických chyb. Práce má vhodnou strukturu a členění, text ale obsahuje jen málo doplňujících prvků (jen 1 tabulka a 3 obrázky/grafy). Některé vzorce a výrazy by měly být matematicky korektnější (např. zápis prvků matice z nějaké množiny v sekci 1.3).

Výběr zdrojů, korektnost citací **průměrné**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Citace v práci jsou použité korektně a v souladu s citačními zvyklostmi, ale jejich výběr mohl být podrobnější. Práce např. cituje jen 3 z 8 publikací v seznamu doporučené literatury v zadání, z nichž dvě jsou CUDA C++ Programming Guide a TNL Users' Guide. V teoretické/rešeršní části by bylo vhodné popsat více různých přístupů z jiných zdrojů, i když by potom třeba nebyly implementovány, ale na druhou stranu chápu snahu soustředit se na jednu část a tu dotáhnout do konce.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Po přečtení práce i kódu jsem v implementaci našel několik problémů:

Paralelizace

- Popis implementace všech algoritmů neobsahuje určení parametrů pro spuštění paralelního výpočtu, jako je velikost bloku a velikost gridu.
- Ve variantě *Multiplication kernel 1* je provedena paralelizace jen přes řádky výsledné matice, přičemž každé vlákno musí spočítat postupně hodnoty na celém řádku. Výsledek je však hustá matice a všechny výpočty jejích prvků jsou na sobě nezávislé, proto by bylo možné snadno provést paralelizaci i ve druhém směru, tedy přes sloupce výsledné matice. Vlákna v 2D bloku by se navíc namapovat jedním nebo druhým způsobem podle toho, jestli jsou husté matice uložené v paměti po řádcích nebo po sloupcích, aby se v obou případech dosáhlo sloučení přístupů paměti při čtení dat z husté vstupní matice a při zápisu do výsledné matice.
- Druhá varianta *Multiplication kernel 2* trpí stejným problémem, jako první verze. Používá sdílenou paměť pro načtení dat z řídké matice na základě publikace [7] citované v práci, ale neuvažuje případ, kdy je ve vstupní řídké matici více hodnot, než je staticky alokovaná velikost sdílené paměti kernelu. V kódu se navíc pro přístup k proměnným "sm_val" a "sm_col", které jsou alokované ve sdílené paměti, používá globální index "row_id", což selže v případě, že má výstupní matice velký počet řádků.

Unit testy
V textu se píše, že byla vytvořena rozmanitá množina testovacích případů, ale v kódu se ve skutečnosti testují jen 2 případy pro každou operaci. Navíc by bylo dobré otestovat přinejmenším případy, kdy rozměry matic neodpovídají požadavkům daných operací - v takovém případě by mělo dojít např. k vyvolání výjimky.

Benchmarky
- Práce nezmiňuje, jestli jsou použité husté matice uložené po řádcích nebo po sloupcích. V benchmarku by bylo vhodné otestovat a srovnat oba případy.
- Také není zmíněno, jaký je počet a rozmístění nenulových prvků v případě řídkých matic. V kódu je vidět, že zvolené řídké matice jsou jen diagonální (1 prvek na každém řádku), což je velmi speciální případ.
- Celkově by metodika zvolená pro tento benchmark měla být obsáhlejší, např. pro řídké matice by bylo vhodné najít nějaké reprezentanty praktických úloh (existují databáze jako třeba SuiteSparse Matrix Collection) nebo náhodným způsobem generovat počet a rozmístění nenulových prvků.

K **prezentaci výsledků** mám další připomínky:
- Grafy mají netypicky invertovanou vertikální osu zobrazující čas výpočtu - hodnoty blízké nule jsou nahoře a velké hodnoty jsou dole.
- Výsledky pro sčítání jsou opačně, než bych čekal - GPU by mělo být pomalejší než CPU pro malé rozměry, ale rychlejší pro velké rozměry. Čím je to způsobeno?
- Celkově je těžké dělat závěry z grafů, které zobrazují přímo délku výpočtu. Lepší by bylo spočítat nějakou veličinu, kterou je možné porovnat s jinými algoritmy nebo použitým hardware, např. paměťovou propustnost v GB/s (objem dat potřebný pro spočítání daného výsledku vydělený dobou výpočtu). V ideálním případě by pak práce bývala mohla obsahovat podrobnější analýzu implementace, např. pomocí CUDA profileru nebo jiných nástrojů.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.
Ačkoli má předložená práce řadu problémů, zadání považuji za splněné a po vyřešení zmíněných problémů to bude dobrý základ pro další rozvoj knihovny TNL.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **E - dostatečně.**

Datum: 14.8.2024 Podpis: