POSUDEK VEDOUCÍHO ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Paralelní algoritmy pro operace s hustými maticemi**
Jméno autora:	**Alexandr Krastenov**
Typ práce:	bakalářská práce
Fakulta:	Fakulta jaderná a fyzikálně inženýrská (FJFI)
Katedra:	Katedra matematiky
Vedoucí práce:	Doc. Ing. Tomáš Oberhuber, Ph.D.
Pracoviště vedoucího práce:	KM FJFI ČVUT

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání a motivace k jeho vypsání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce a krátké průvodní slovo k motivaci pro zadání práce.

Cílem tohoto zadání bylo zoptimalizovat zejména násobení hustých matic na GPU v knihovně TNL . V této knihovně je pouze základní implementace této důležité operace, která nedosahuje optimálního výkonu. V rámci této práce měl student porovnat TNL s jinými knihovnami jako např. cuBLAS, prostudovat optimalizace popsané v odborných článcích, implementovat vybrané algoritmy a porovnat je mezi sebou. Ten nejvýkonnější by se pak měl zařadit do knihovny TNL. Podobně se měla optimalizovat transpozice hustých matic v TNL.

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.

Student bohužel implementoval jen poměrně triviální optimalizace popsané v jednom internetovém blogu, který popisoval základní optimalizační techniky pro implementaci násobení hustých matic. Pokročilejší kernel využívající registry multiprocesorů GPU není implementovaný správně minimálně z důvodu neefektivního načítání maticových dlaždic z globální do sdílené paměti, jak je vidět ve výpisu kódu na str. 37, řádek 17, kde se do pole AValues přistupuje s krokem matrixRows a tedy ne pomocí sloučených přístupů. Autor také bohužel nezvládl implementovat žádný kernel využívající tenzorová jádra GPU. Ve výsledku tak došlo jen k malému urychlení kódu již existujícího kódu v knihovně TNL. Výpočetní studie ukazuje, že jiné knihovny jako cuBLAS nebo Magma stále nabízejí znatelně vyšší výkon.

Aktivita a samostatnost při zpracování práce **podprůměrná**

Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven. Posuďte schopnost studenta samostatné tvůrčí práce.

Student docházel na pravidelné konzultace a práci na tématu se věnoval systematicky v průběhu celého akademického roku. Bohužel nemohu říci, že by byl silný v samostatném přístupu. Opakovaně, asi na pěti konzultacích , jsem ho upozorňoval např. neoptimální implementaci podmínek v kernelu 1.6. při načítání maticových dlaždic do sdílené paměti. Ač jde o poměrně jednoduchou optimalizaci, opakovaně nosil na konzultace špatně napsaný kód. Podobně jsem ho opakovaně upozorňoval na nesloučené přístupy do globální paměti také při načítání maticových dlaždic. Ani do odevzdání nebyl schopný toto opravit a to přesto, že stejnou techniku používá ve všech předchozích kernelech. Komunikace se studentem často připomínala dotazování generativních modelů, neboť mnohokrát vytvořil zcela nový kód, který ale vůbec neřešil problém, na který jsem ho upozorňoval. Často jsem měl pocit, že student jen různě zkouší měnit různé části kódu a pouze sleduje, zda proběhnou unit testy. Nad samotnými změnami se ale již příliš nezamýšlel. Také mě zarazilo, že na konci semestru, kdy student sepisoval třetí kapitolu popisující jednotlivé optimalizace, se ukázalo, že jednotlivým optimalizacím vlastně nerozumí. Část popisující jednotlivé kernely je stále velmi povrchní a vůbec nevysvětluje podstatu použitých optimalizací. Právě toto považuji za nejzásadnější nedostatek této bakalářské práce.

Odborná úroveň	**podprůměrná**
Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.	
Jak již bylo zmíněno, autor implementoval jen velmi triviální optimalizace a ty složitější již bohužel implementovat nezvládl. Přesto lze v této práci najít něco přínosného. Jednak jde o podrobné porovnání násobení hustých matic s jinými knihovnami a to i se zaměřením se na obdélníkové matice. Dále se autorovi podařilo významně urychlit operaci transpozice, kde se dosahuje dokonce lepšího výkonu i v porovnání s jinými knihovnami. Transpozice sice není kritická operace jako např. právě násobení matic, ale i tak je to potěšující výsledek. Dále bych vyzdvihl celkově pěkně provedenou výpočetní studii, kde jsou doloženy výstupy z profileru, je uveden roofline model a výpočty byly provedeny jak na dostupných GPU tak na profesionálních výpočetních GPU.	

Formální a jazyková úroveň	**průměrná**
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.	
Dále bych pochválil jazykovou úroveň překládané práce. Text je dobře strukturovaný a povětšinou dobře srozumitelný. Autor celkem dobře popisuje základy programování GPU s pomocí CUDA, základní principy programování v TNL a základní optimalizační techniky programování GPU. Bohužel výklad jednotlivých kernelů připomíná spíše jen přepis přiloženého kódu do angličtiny a nevysvětluje podstatu samotnou. Pokud jde o část s výpočetní studií, autor zde zcela nepochopitelně převrací orientaci svislé osy, takže dole jsou delší časy a nahoře kratší. Grafy pak působí naprosto zavádějícím dojmem. Studenta jsem na toto před odevzdáním upozorňoval, ale opravu neprovedl. Z textu mi také není jasné, jaký je význam grafů 4.6 a 4.12.	

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.	
Všechny použité zdroje jsou korektně citovány.	

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.
Bohužel se nepodařilo vytvořit GPU kernel pro násobení hustých matic, který by byl porovnatelný s jinými knihovnami. Pouze pro velmi malé matice, jsou implementované kernely rychlejší, ale to byl i původní kernel v knihovně TNL. Pro větší matice dosahují implementované kernely cca. výkonu na úrovni 20% až maximálně 70% výkonu ostatních knihoven. Kernel pro transpozici je sice znatelně rychlejší, ale to bohužel v reálných výpočtech není příliš důležité.

III. CELKOVÉ HODNOCENÍ A NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Případně uveďte otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

S výsledky předkládané práce jsem spíše nespokojený. Na studenta bych měl dotaz, zda by mohl vysvětlit jak funguje kernelu 1.6?

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

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK VEDOUCÍHO ZÁVĚREČNÉ PRÁCE

Datum: 9.8.2024

Podpis: