Posudek oponenta diplomové práce

předložené na Matematicko-fyzikální fakultě Univerzity Karlovy

Autor:	Bc. Josef Matějka
Název práce:	Efficient sorting algorithms for memory hierarchies
Stud. program:	teoretická informatika
Rok odevzdání:	2024
Jméno a tituly oponenta:	Mgr. Martin Mareš, Ph.D.
Pracoviště:	Katedra aplikované matematiky
Kontaktní e-mail:	mares@kam.mff.cuni.cz

Popis práce

Práce se zabývá třídicími algoritmy v cache-oblivious výpočetním modelu. To je model počítače s hierarchickou pamětí, který docela dobře popisuje cacheování na současném hardwaru.

Algoritmů tohoto druhu už bylo známých několik, zejména Funnelsort a Distribution sort. Oba mají asymptoticky optimální jak časovou složitost, tak I/O složitost, ale pro praktické použití nejsou vhodné kvůli vysokým multiplikativním konstantám. Předložená práce navrhuje nový randomizovaný třídicí algoritmus Squaresort inspirovaný jednak Quicksortem, jednak paralelním algoritmem Columnsort.

První část práce popisuje známé třídicí algoritmy, cache-oblivious výpočetní model a algoritmus Columnsort. Druha část vybuduje nový algoritmus Squaresort a analyzuje jeho worst-case i průměrnou složitost. Závěrečná část se věnuje experimentům na reálném hardwaru.

Hodnocení

Student navrhl zajímavý třídicí algoritmus, který do světa cache-oblivious třídění přenáší myšlenky dělení na přihrádky pomocí pivotů volených náhodně. To je myšlenka, která se už osvědčila ve světe klasického třídění a vedla k jedněm z prakticky nejlepších algoritmů.

Student prokázal schopnost provést netriviální analýzu randomizovaného algoritmu a dokazat jeho optimalitu ve střední hodnotě, a to jak pro časovou, tak pro I/O složitost.

Implementace algoritmu v C++ je příjemně jednoduchá a experimenty jasně ukazují, že algoritmus je mnohem efektivnější než jak Funnelsort, tak Introsort použitý ve standardní knihovně jazyka.

K práci mám nicméně řadu výhrad:

- Popis známých algoritmů v úvodu se soustřeďuje na triviální algoritmy typu Bubblesort. Nepopisuje nejen žádný z cache-oblivious algoritmů, ani Introsort, ačkoliv je o něm v práci několikrát řeč a experimenty ho používají jako referenční.
- Popis Distribution sortu postrádám o to víc, že je mu nový algoritmus v mnoha ohledech podobný, zejména principem rekurzivního rozdělování setříděných bloků

vstupu do přihrádek. Ačkoliv vysvětlení Squaresortu vychází primárně z Columnsortu, přijde mi přímočařejší dívat se na Squaresort jako na Distribution sort, ve kterém jsou přihrádky místo postupného štěpení (jako u B-stromu) vymezeny pivoty.

- Analýza časové složitosti pomíjí volbu pivotů. Ta ve skutečnosti není triviální, mimo jiné proto, že v použitém modelu výpočtu je k dispozici pouze generátor náhodných bitů, nikoliv rovnoměrně náhodných čísel.

- Speciálně se volba pivotů může nezastavit (hned dvojím způsobem: buď při generování indexů, nebo restarty po nalezení duplicity). Tudíž navzdory tvrzení oddílu 5.2 nemůže být worst-case složitost konečná. Toho se v uvedeném výpočetním modelu není snadné zbavit.

- V analýze složitosti jsou četné překlepy ve formulích a velké skoky v argumentaci. Finální výsledky jsou nicméně správně.

- V oddílu 4.2.4 se pracně dokazuje, že velikosti jednotlivých přihrádek mají stejné střední hodnoty. V oddílu 5.1 se ale bez důkazu použije, že mají dokonce stejné distribuce. Buď jsme to ochotni považovat za zřejmé, a pak je oddíl 4.2.4 zbytečný, nebo nejsme, a pak nám výsledek z 4.2.4 nestačí.

- Experimenty testují implementaci, která neodpovídá popisu algoritmu – pivoty místo náhodného výběru ze vstupu bez opakování náhodně vybírá z univerza s opakováním. To nejspíš dost ovlivní výsledky těch testů, které třídí hodnoty s malým rozsahem.

- Experimenty s tříděním ve vnější paměti je těžké posoudit, protože popis je velmi skoupý na detaily. Pravděpodobně spoléhají na diskovou cache v jádru Linuxu, o které se ví, že je výrazně neoptimální.

- K práci není přiložena implementace experimentů, takže výsledky nejsou snadno reprodukovatelné.

Text je psaný příjemně čtivou angličtinou. Dojem kazí občasné překlepy a nepořádné formulace. Použité zdroje jsou korektně citovány.

Přes všechny výhrady považuji jádro práce za kvalitní a výsledky za zajímavé a užitečné. Práci proto doporučuji uznat jako diplomovou.

V Praze dne 29. srpna 2024
Martin Mareš