Semestrální projekt MI-PAP 2010/2011:

Paralelní řadící algoritmy

Pavel Benáček
Tomáš Čejka
magisterské studijum, FIT ČVUT, Kolejní 550/2, 160 00 Praha 6

April 27, 2011

Contents

List of Tables

1 Definice problému a popis sekvenčního algoritmu

1.1 Zadání

Podle zadání semestrální práce je úkolem implementace aspoň tří z následujících řadících algoritmů: Shearsort, 3D sort, Sudo-lichý Mergesort a Bitonic sort.

Pro řešení jsme si vybrali algoritmy Shearsort, Sudo-lichý Mergesort a Bitonic sort.

Pro účely měření času běhu výsledných programů a porovnání paralelních algoritmů se sekvenčním řazením jsme si implementovali sekvenční algoritmus Mergesort a ten vzali jako referenční sekvenční řešení.

Semestrální práce spočívala v implementaci vybraných řadících algoritmů jako sekvenční řešení, paralelní řešení na počítači se sdílenou pamětí s použitím knihovny OpenMP a následně s využitím technologie CUDA společnosti NVIDIA®.

1.2 Řešení

Algoritmy byly ve své podstatě navrženy tak, aby co nejlépe seděli možnostem a architektuře CUDA. Bylo dbáno na několik základních vlastností. A to:

- co největší počet vláken v bloku
- co nejvíce nezávislých bloků
- minimalizace komunikace do globální paměti

Tyto části jsou společné pro všechny algoritmy pro CUDA. Zaručují jistou rychlost a škálovatelnost problému.

2 Popis paralelního algoritmu a jeho implementace v CUDA

2.1 Sudo-lichý mergesort

2.2 Bitonic sort

2.3 Shearsort

Vzhledem k možnosti HW v grafickém akcelerátoru jsme zvolili implementaci pomoci even odd transpozition sortu. Toto využívání je možné zejména díky schopnosti vytvořit velký počet vláken v zařízení (a to,že každý prvek v poli má své vlastní vlákno). Při samotné implementaci bylo dbáno na několik základních pravidel. A to pravidla v úvodní části **??**.

Po nastartování kernelu se provede nakopírování dat do sdílené paměti. Samotné řazení probíhá ve sdílené paměti a mimo bloková komunikace (výměna) probíhá přez globální paměť (kde komunikaci provádí POUZE prvky na přechodu mezi bloky).

Samotný shearsort byl implementován jako kernel, tak veškerá část algoritmu běží na GPU. Pro tuto příležitost byl implementován algoritmus synchronizace mezi bloky, který byl popsán v **??**. Tento algoritmus nám umožnil synchronizaci mezi bloky a tak se nemuselo z kernelu vystupovat a provádět synchronizaci bloků v rámci programu hosta.

Program si sám spočítá rozložení matice tak, aby se do globální paměti vlezl celý sloupec 2D matice.

3 Naměřené výsledky a vyhodnocení

3.1 Způsob měření

Měření bylo prováděno nad daty, které se generovalo pro každý každý problém znova a které samozřejmě do doby běhu algoritmu nebylo započítáno. Měření času bylo prováděno prostředky z CUDA SDK. Pro vyhodnocení jsme zvolili konstatní počet vláken na blok.

3.2 Naměřené časy

3.2.1 Sekvenční Mergesort

3.2.2 Sekvenční Even-Odd Mergesort

3.2.3 Sekvenční Bitonic sort

3.2.4 Shearsort

Množství dat (2^n)	Čas
6	0,00191242
7	0,00198106
8	0,0035513
9	0,00360128
10	0,00558307
11	0,0119386
12	0,0435032
13	0,249303
14	1,28781

Table 1: Shearsort

3.2.5 Even-Odd Mergesort

3.3 Zrychlení

3.3.1 Bitonic sort

3.3.2 Even-Odd Mergesort

3.4 Shearsort

3.5 Vyhodnocení

4 Literatura

1. Stránky předmětu MI-PAP
2. GPU synblocks

A Grafy závislosti času na počtu vláken