

Obsah

1	Architektura moderních procesorů	5
1.1	Hlavní součásti architektury počítačů založených na jednojádrových procesorech	6
1.2	Výpočty na CPU	10
1.2.1	Vektorové zpracování – instrukční sady	11
1.3	Optimalizační techniky	16
1.3.1	Obecné optimalizace	16
1.3.2	Optimalizace zaměřené na cykly	17
1.4	Paralelní systémy a mechanismy	19
1.4.1	Flynnova taxonomie paralelních architektur	20
1.4.2	Metody programování paralelních systémů	21
1.4.3	Hodnocení kvality paralelních programů	24
1.5	Architektura grafických procesorů	25
1.5.1	Programování GPU	29
2	Popis architektury NVIDIA CUDA	31
2.1	Programátorský model	32
2.2	Paměťový model	33
2.3	Spouštěcí model	33
2.4	Základy programování	36
2.4.1	Základní datové typy	38
2.4.2	Automaticky definované proměnné	38
2.4.3	Ošetření chyb	39
2.4.4	Funkce pro práci s globální pamětí	40
2.4.5	Zjištění dostupných zařízení	42
2.4.6	Události	43
2.4.7	Komunikace v rámci warpu	44
2.4.8	Podporované matematické funkce	45
3	Programování v NVIDIA CUDA	49
3.1	Typy pamětí a jejich použití	49
3.1.1	Globální paměť	50
3.1.2	Sdílená paměť	53
3.1.3	Lokální paměť a registry	54
3.1.4	Paměť pro konstanty	55
3.1.5	Paměť určená jen pro čtení	55
3.1.6	Specifikace umístění proměnných	56
3.1.7	Hlavní paměť (hostitele)	56
3.1.8	Paměť textur	58
3.1.9	<i>Surface Memory</i>	61
3.2	Synchronizace	62

3.2.1	Synchronizace v rámci bloku	63
3.2.2	Atomické operace	63
3.2.3	Synchronizace paměťových přístupů	65
3.3	Možnosti urychlení výpočtů	66
3.3.1	Využití více zařízení pro výpočty	66
3.3.2	Proudy (fronty operací)	68
3.3.3	Sjednocený adresní prostor	70
3.3.4	Vzájemná komunikace zařízení	71
3.4	Návrh programu pro GPU a optimalizace kódu	71
3.4.1	Využití obecných optimalizačních technik na GPU	72
3.4.2	Optimalizace CUDA kódu	73
3.4.3	Vytížení multiprocesorů	75
3.4.4	Sdružený přístup do globální paměti	77
3.4.5	Přístup do sdílené paměti	79
4	Ukázky řešení vybraných problémů	81
4.1	Výpočet histogramu	81
4.2	Paralelní redukce	83
4.3	Násobení matic	85
4.4	Kombinatorická úloha – problém batohu	88
5	Spolupráce s ostatními jazyky a nástroji	93
5.1	Překlad zdrojových kódů v CUDA (NVCC)	93
5.1.1	Nastavení CUDA kompilace	93
5.1.2	Práce se soubory	95
5.2	Spolupráce CUDA a OpenGL	96
5.3	Součásti CUDA SDK	97
5.3.1	Knihovna CURAND	97
5.3.2	CUBLAS	97
5.3.3	CUFFT	98
5.3.4	NPP	98
5.3.5	THRUST	98
5.3.6	CUSPARSE	100
5.3.7	Visual Profiler	100
6	OpenCL	101
6.1	Modely OpenCL	101
6.1.1	Model platformy (<i>platform model</i>)	101
6.1.2	Prováděcí (<i>execution</i>) model	102
6.1.3	Programový (<i>programming</i>) model	102
6.1.4	Paměťový (<i>memory</i>) model	103
6.1.5	Práce s pamětí	104
6.2	Základy programování	104
6.2.1	Synchronizace	104
6.2.2	Inicializace a spouštění kernel funkcí	104
6.3	Rozdíly mezi OpenCL a CUDA	112

A	Vnitřní architektura grafických karet	115
A.1	Architektura karet firmy NVIDIA	115
A.1.1	Architektura multiprocesorů	115
A.1.2	Přehled GPU	116
A.1.3	CUDA hardwarové parametry a limity	116
A.2	Vývoj architektury CUDA karet v závislosti na <i>CUDA capabilities</i>	117
A.2.1	Architektura a vlastnosti CUDA CC 1.X karet	117
A.2.2	Architektura a vlastnosti CUDA CC 2.0 karet	118
A.2.3	Architektura a vlastnosti CUDA CC 2.1 karet	119
A.2.4	Architektura a vlastnosti CUDA CC 3.0 karet	119
A.2.5	Architektura a vlastnosti CUDA CC 3.5 karet	120
A.3	Architektura karet firmy ATI/AMD	120
A.3.1	Architektura VLIW4 a VLIW5	120
A.3.2	Architektura GCN	121