Aritmetika s vysokou přesností na GPU

školitel: Ing. Tomáš Oberhuber, Ph.D.

Matěj Novotný

6. února 2012

Motivace

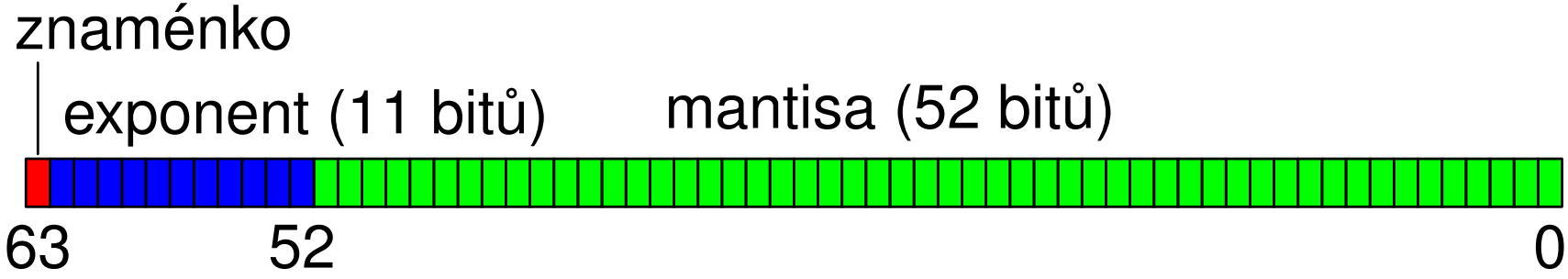
- Počítačová aritmetika s vysokou přesností
- Využití grafické karty (GPU) za účelem urychlení

Čísla s plovoucí desetinnou čárkou

Norma IEEE 754

- Definuje uložení desetinných čísel v počítači
- Definuje aritmetické operace a jejich vlastnosti
- Většina dnešních procesorů normu splňuje
- Binární formáty čísla: 16-bitové, **32-bitové** (single precision), **64-bitové** (double precision), 128-bitové
- Dekadické formáty čísla: 32-bitové, 64-bitové, 128-bitové

Double precision



- 64 bitů
- Hardwarová podpora = rychlé výpočty
- znaménko $\in \{0, 1\}$, exponent $\in \mathbb{N}_0$, cifry $m_i \in \{0, 1\}$
- $$(-1)^{\text{znaménko}}(1, m_{51}m_{50}\ldots m_0)_2 \times 2^{\text{exponent}-1023}$$
- 53 bitů mantisy $\sim$ 16 dekadických cifer

Ukázka

Číselná posloupnost

- $$a_n = \frac{34}{11} a_{n-1} - \frac{3}{11} a_{n-2}$$
$$a_0 = 1,\ a_1 = \frac{1}{11}$$
- $a_{50} = ?$
- double precision $a_{50} \doteq 2\,388\,034$
- skutečná hodnota $a_{50} \doteq 0,851 \times 10^{-52}$

Vyšší přesnost

- softwarové výpočty
- oproti hardwarově podporvaným typů nižší výkon

Multi-term

- číslo reprezentované sumou desetinných čísel (např double)
- omezená přesnost
- využití výpočetních jednotek

Multi-digit

- číslo uloženo po cifrách se společným exponentem
- volitelná přesnost
- pomalé

Moje práce

- Nastudování a následný popis aritmetiky **Quad-Double**
- Vysvětlení algoritmů a dokázání jejich korektnosti
- Implementace aritmetiky na CPU a GPU
- Otestování správnosti výpočtů a porovnání výkonu

Quad-double

- výpočty typu multi-term
- číslo uloženo pomocí 4 double
-

$$a_0 + a_1 + a_2 + a_3$$

$$a_0 > a_1 > a_2 > a_3$$

Implementace

Aritmetické operace

- sčítání, sčítání přesné, odčítání
- násobení, násobení rychlé
- dělení

CPU

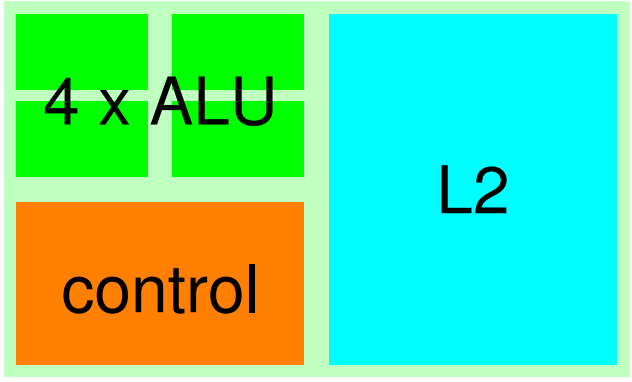
- naivní implementace v jazyce C

GPU

- CUDA

CPU vs GPU

CPU



- Optimalizace na nízké latence
- Složitá řídící logika
- Velká nezávislá vlákna

GPU

- Optimalizace na datový paralelizmus a datovou propustnost
- Jednoduchá řídící logika
- Mnoho výpočetních jednotek

Testy

- open source knihovna GNU MP
  - aritmetika typu multi-digit
- testy na konkrétních číslech – ověření správnosti implementace
- testy na náhodných číslech
  - 1 920 000 čísel
  - postupné sčítání, násobení, dělení čísel mezi sebou
  - porovnání výsledků z CPU, GPU a z knihovny GNU MP
  - měření času výpočtu na CPU a GPU

Dosažené časy

operace	čas na CPU	čas na GPU	urychlení
sčítání	0,136541 s	0,005856 s	**23** ×
sčítání přesné	0,354374 s	0,011432 s	**31** ×
násobení	0,320313 s	0,014787 s	**21** ×
násobení rychlé	0,156445 s	0,008782 s	**17** ×
dělení	1,249169 s	0,051122 s	**24** ×

double aritmetika na CPU

operace	čas na CPU
sčítání	0,002400 s
násobení	0,002400 s
dělení	0,002406 s

Budoucnost

- implementace v C++ s použití šablon
- vyšší přesnost (8, 16 členů)
- obecné optimalizace
  - rozvedení polí do jednotlivých proměnných
- optimalizace kódu pro GPU na konkrétní použití
  - rovnice s citlivou vazbou na počáteční podmínky
  - úlohy LA se špatně podmíněnou maticí

Děkuji za pozornost.