

Univerzita Pavla Jozefa Šafárika v Košiciach
Prírodovedecká fakulta

ZRÝCHLENIE VÝPOČTU SPLAJN POVRCHOV

DIPLOMOVÁ PRÁCA

Študijný odbor:	Informatika
Školiace pracovisko:	Ústav informatiky
Vedúci záverečnej práce:	doc. RNDr. Csaba Török, CSc.
Konzultant:	RNDr. Lukáš Miňo

Košice 2016

Bc. Viliam Kačala

Podakovanie

Rád by som poďakoval vedúcemu záverečnej práce doc. RNDr. Csabovi Törökovi, CSc. za cenné pripomienky, odborné vedenie a obetavosť počas tvorby práce. Taktiež by som sa rád poďakoval RNDr. Lukášovi Miňovi za cenné rady a pomoc počas tvorby aplikačnej časti tejto práce.

Zadanie práce

Abstrakt

Splajny sú dôležitá súčasť počítačovej grafiky. Jedná sa o matematický model krivky a plochy slúžiaci na čo „najlepšie spojenie“ konečnej množiny bodov. Termín „najlepšie spojenie“ v našom prípade znamená hladkú, matematicky ľahko vyjadriteľnú plochu s čo najmenším zakrivením. Využitie splajnov v grafike je veľmi široké od rôznych CAD aplikácií, v štatistike, alebo v analýze dát. Splajny existujú v mnohých formách, či už vo forme krivky v rovine, rôznych trojrozmerných telies, atď.. Táto práca si dáva za cieľ navrhnúť, analyzovať a implementovať nový algoritmus pre bikubickú interpoláciu v trojrozmernom priestore.

Abstract

Splines are important part of computer graphics. It is a mathematical model of curve and surface for the "best connection" of any finite set of points. The term "best connection" in this case means smooth, easily calculable mathematical surface with minimal curvature. Use of splines in graphics varies from large variety of CAD applications, statistics or in data analysing. Splines exist in many forms, whether in the form of curves in the plane, a variety of three-dimensional bodies, etc.. This work aims to design, analyze and implement a new algorithm for counting and generating splines bicubic clamped interpolation in three-dimensional space.

Obsah

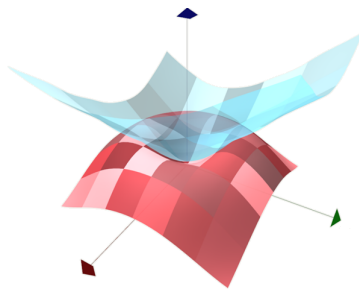
1	De Boorov model a redukovaný algoritmus	7
1.1	Polynómy	7
1.2	Splajny	8
1.2.1	Konštrukcia hermitovho splajnu	13
1.3	Trojdiagonálna LU dekompozícia	14
1.4	De Boorov výpočet derivácií	17
1.5	Počítanie derivácií redukovanou sústavou	19
2	Zrýchlenie	25
2.1	Procesorová architektúra	26
2.1.1	Kešovanie	26
2.1.2	Inštrukčný paralelizmus	28
2.1.3	Rýchlosť aritmetických operácií	31
2.2	Teoretické zrýchlenie	34
2.2.1	Cena plného algoritmu	37
2.2.2	Cena redukovaného algoritmu	39
2.2.3	Pamäťové nároky	42
2.2.4	Zhrnutie	43
2.3	Merané zrýchlenie	44

Úvod

Témou diplomovej práce sú priestorové splajn povrchy, pričom naším cieľom je preskúmať nové poznatky o splajnoch, na ktorých istý čas pracuje vedúci tejto práce doc. RNDr. Csaba Török, CSc.

Výsledok tejto práce je návrh novej metódy výpočtu derivácií splajnu v jeho uzloch, jej porovnanie s de Boorovým postupom z ktorého vychádzame, vysvetlenie zrýchlenia a implementácia aplikácií na vizuálne a výkonnostne porovnávanie oboch metód.

Prínos zrýchlenia výpočtu splajnov spočíva v lepších možnostiach modelovania ľubovoľných trojrozmerných útvarov v podobe polynomických funkcií. Akékoľvek zrýchlenie totiž znamená možnosť v reálnom čase modelovať zložitejšie objekty alebo fyzikálne dáta, čo je užitočné nielen vzhľadom na vizuálnu reprezentáciu, ale aj na skúmanie fyzikálnych vlastností ako napríklad aerodynamické vlastnosti lietadiel a podobne.



Technológia v ktorej tieto poznatky implementujeme je Microsoft Silverlight. Jedná sa o veľmi schopný nástroj na tvorbu webových aplikácií s plnou podporou hardvérovej akcelerácie užívateľského prostredia a možnosťou tvorby rýchlej trojrozmernej grafiky. Výhodou tohto frameworku je, keďže beží na platforme Microsoft .NET, možnosť jednoduchej portácie na desktopovú prípadne mobilnú aplikáciu. Na výkonnostné porovnanie nášho algoritmu s doterajším postupom však implementujeme samostatnú natívnu aplikáciu nad C++.

Štruktúra práce je nasledovná.

- **De Boorov model a redukovaný algoritmus**
 - **Polynómy**
Základné pojmy o polynómoch.
 - **Splajny**
Definícia krivkových a povrchových splajn.
 - **Trojdiagonálna LU dekompozícia**
Spôsob počítania trojdiagonálnych sústav.
 - **De Boorov výpočet derivácií**
V tejto časti si vysvetlíme splajn interpoláciu podľa Carla de Boora.
 - **Počítanie derivácií redukovanou sústavou**
Ukážka modifikovaného postupu pre kubické splajny a jeho rozšírenie pre bikubické splajny.
- **Zrýchlenie**
Očakávané zrýchlenie výpočtov novým algoritmom.
 - **Procesorová architektúra**
Tu si objasníme vplyv moderných inštrukčných sád na zrýchlenie redukovaného algoritmu.
 - **Teoretické zrýchlenie**
Spočítanie časovej zložitosti vysvetlenie vplyvu základných aritmetických a pamäťových operácií.
 - **Merané zrýchlenie**
Výsledky z našej implementácie.
- **Implementácia a užívateľská príručka**
Podrobnosti implementácie v MS Silverlight a C++. Stručný návod na použitie programov.

Formálne tézy diplomovej práce sú:

- Analýza modelov interpolačných splajnov.
- Redukovaný algoritmus výpočtu koeficientov splajn povrchov.
- Testovanie faktorov vplývajúcich na rýchlosť výpočtu splajn koeficientov

Kapitola 1

De Boorov model a redukovaný algoritmus

V tejto kapitole si postupne zdefinujeme pojem polynómu a tento postupne rozšírime na pojem splajnu. Následne si popíšeme dve metódy výpočtu derivácií splajnu a konštrukciu celej splajnovej plochy.

1.1 Polynómy

Pre úplnosť skôr než si povieme o splajnoch, predstavíme si pojem polynomickej funkcie ktorý je neoddeliteľnou súčasťou pri definícii splajnu.

Definícia 1.1 Nech n je z $\mathbb{N} \cup \{0\}$ a pre každé i z $\{0, \dots, n\}$ je a_i z \mathbb{R} , pričom $a_n \neq 0$. Funkciu $p: \mathbb{R} \rightarrow \mathbb{R}$ tvaru

$$p(x) = \sum_{i=0}^n a_i x^i$$

nazveme *polynomická funkcia jednej premennej stupňa n* .

Označenie 1.2 Polynomicke funkcie jednej premennej budeme skrátene nazývať *polynomicke funkcie*. Polynomicke funkcie stupňa 3 nazveme *kubické funkcie*.

V práci však budeme pracovať najmä s funkciami dvoch premenných. Analogicky si pre dve premenné zdefinujeme aj polynomicke funkcie.

Definícia 1.3 Nech n a m sú z $\mathbb{N} \cup \{0\}$ a pre každé i z $\{0, \dots, n\}$ a j z $\{0, \dots, m\}$ je a_{ij} z \mathbb{R} , pričom $a_{nm} \neq 0$. Funkciu $p: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ tvaru

$$p(x, y) = \sum_{i=0}^n \sum_{j=0}^m a_{ij} x^i y^j$$

nazveme *polynomicke funkcia dvoch premenných stupňov n a m*

Označenie 1.4 Polynomicke funkcie dvoch premenných budeme skrátene nazývať *bipolynomicke funkcie*. Bipolynomicke funkcie stupňov 3 a 3 nazveme *bikubické funkcie*.

1.2 Splajny

V našej práci pracujeme s hermitovskými splajnami [5], ktoré sú štandardne triedy C^1 , teda splajnami ktorých prvé derivácie v uzloch sa rovnajú.

Najbežnejšie kubické splajny triedy C^2 , teda tie pri ktorých v uzloch máme zaručenú rovnosť aj derivácií druhého rádu, sú *naturálne* a *clamped* splajny. Pre naše potreby budeme uvažovať iba druhé menované pričom ich budeme konštruovať použitím hermitovských bázových funkcií (pozri časť 1.2.1).

Keďže naším cieľom je zrýchlený algoritmus počítania povrchových splajnov, ktorý vznikol zovšeobecnením algoritmu pre krivkové splajny tak si zavedieme pojem splajnu pre krivky aj plochy. Iste by bolo možno elegantnejšie zdefinovať splajn všeobecne, ale rozdelenie definície špeciálne pre krivkové splajny v rovine a povrchové v priestore bude čitateľnejšie. Formálna definícia krivkového splajnu v rovine [1] je nasledovná.

Definícia 1.5 Nech $I \geq 0$ je prirodzené číslo, (u_0, \dots, u_{I-1}) je rastúca postupnosť. Nech pre každé i z $\{0, 1, \dots, I-2\}$ funkcie S_i sú polynomicke funkcie premennej x , ktoré spĺňajú

- $S_i(u_{i+1}) = S_{i+1}(u_{i+1})$,
- $\frac{dS_i}{dx}(u_{i+1}) = \frac{dS_{i+1}}{dx}(u_{i+1})$

Funkciu S z intervalu $[u_0, u_{I-1}]$ do \mathbb{R} pre ktorú platí:

$$S(x) = \begin{cases} S_0(x) & \text{pre } x \in [u_0, u_1], \\ S_1(x) & \text{pre } x \in [u_1, u_2], \\ \vdots & \\ S_{I-1}(x) & \text{pre } x \in [u_{I-2}, u_{I-1}], \end{cases} \quad (1.1)$$

nazveme *krivkový splajn na uzloch (u_0, \dots, u_{I-1})* .

Označenie 1.6 Pri označeniach z predchádzajúcej definície označme:

- Funkcie S_i nazveme *segmenty splajnu*.

- Hodnoty u_i nazveme *uzly*.
- Uzly u_0 a u_{I-1} nazveme *krajné uzly*.
- $z_i = S(u_i)$ nazveme *funkčné hodnoty splajnu v uzloch*.
- $d_i = \frac{dS(u_i)}{dx}$ nazveme *derivácie v uzloch*.

Na to aby sme zostrojili segmenty splajnu potrebujeme mať dané uzly, funkčné hodnoty v uzloch aj derivácie v uzloch. Je však možné zostrojiť splajn aj keď okrem uzlov a funkčných hodnôt máme známe len dve krajné derivácie d_0 a d_{I-1} . Zvyšné hodnoty derivácií totiž môžeme s malou odchýlkou vypočítať napríklad De Boorovou interpoláciou. Túto techniku, ale pre splajnové povrchy, si rozoberieme v nasledujúcich častiach kapitoly. Teraz rozšírime túto definíciu na povrchový splajn v priestore.

Definícia 1.7 Nech $I \geq 0$ a $J \geq 0$ sú prirodzené čísla, (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) sú rastúce postupnosti. Nech pre každé i z $\{0, 1, \dots, I-2\}$ a j z $\{0, 1, \dots, J-2\}$ funkcie $S_{i,j}$ sú bipolynomicke funkcie premenných x a y , ktoré spĺňajú

- $S_{i,j}(u_{i+1}, y) = S_{i+1,j}(u_{i+1}, y)$,
- $S_{i,j}(x, v_{j+1}) = S_{i,j+1}(x, v_{j+1})$,
- $\frac{\partial S_{i,j}}{\partial x}(u_{i+1}, y) = \frac{\partial S_{i+1,j}}{\partial x}(u_{i+1}, y)$
- $\frac{\partial S_{i,j}}{\partial y}(x, v_{j+1}) = \frac{\partial S_{i,j+1}}{\partial y}(x, v_{j+1})$,

Funkciu S z intervalu $[u_0, u_{I-1}] \times [v_0, v_{J-1}]$ do \mathbb{R} pre ktorú platí:

$$S(x, y) = \begin{cases} S_{0,0}(x, y) & \text{pre } (x, y) \in [u_0, u_1] \times [v_0, v_1], \\ S_{0,1}(x, y) & \text{pre } (x, y) \in [u_0, u_1] \times [v_1, v_2], \\ \vdots & \\ S_{0,J-1}(x, y) & \text{pre } (x, y) \in [u_0, u_1] \times [v_{J-2}, v_{J-1}], \\ S_{1,0}(x, y) & \text{pre } (x, y) \in [u_1, u_2] \times [v_0, v_1], \\ \vdots & \\ S_{1,J-1}(x, y) & \text{pre } (x, y) \in [u_1, u_2] \times [v_{J-2}, v_{J-1}], \\ \vdots & \\ S_{I-1,0}(x, y) & \text{pre } (x, y) \in [u_{I-2}, u_{I-1}] \times [v_0, v_1], \\ \vdots & \\ S_{I-1,J-1}(x, y) & \text{pre } (x, y) \in [u_{I-2}, u_{I-1}] \times [v_{J-2}, v_{J-1}], \end{cases} \quad (1.2)$$

nazveme *splajn na uzloch* (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) .

Označenie 1.8 Pre úplnosť analogicky ako pri krivkách označme:

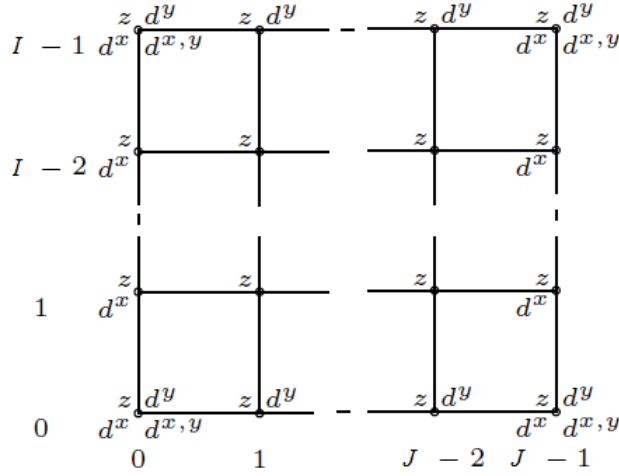
- Funkcie S_{ij} nazveme *segmenty splajnu*.
- Dvojice $\langle u_i, v_j \rangle$ nazveme *uzly*.
- Uzly $\langle u_0, v_0 \rangle, \langle u_{I-1}, v_0 \rangle, \langle u_0, v_{J-1} \rangle$ a $\langle u_{I-1}, v_{J-1} \rangle$ nazveme *rohové uzly*.
- $z_{i,j} = S(u_i, v_j)$ nazveme *funkčné hodnoty splajnu v uzloch*.
- $d_{i,j}^x = \frac{\partial S(u_i, v_j)}{\partial x}$ nazveme (*smerové*) *x-ové derivácie v uzloch*.
- $d_{i,j}^y = \frac{\partial S(u_i, v_j)}{\partial y}$ nazveme (*smerové*) *y-ové derivácie v uzloch*.
- $d_{i,j}^{xy} = \frac{\partial^2 S(u_i, v_j)}{\partial x \partial y}$ nazveme *zmiešané derivácie v uzloch*.

Aby sme zostrojili splajnové segmenty potrebujeme mať dané všetky uzly aj funkčné hodnoty v uzloch a derivácie v uzloch. Analogicky ako pri krivkách, ak nemáme známe všetky derivácie, vieme napriek tomu zvyšné derivácie dopočítať De Boorovou interpoláciou s malou odchýlkou aby splajn bol spojitý a hladký. Na zostrojenie splajnu pomocou De Boorovho modelu interpolácie potrebujeme mať známe

- postupnosti uzlov (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) ,
- funkčné hodnoty $\{z_{0,0}, \dots, z_{I-1,0}, \dots, z_{0,J-1}, \dots, z_{I-1,J-1}\}$,
- smerové derivácie $\{d_{0,0}^x, \dots, d_{I-1,0}^x, d_{0,J-1}^x, \dots, d_{I-1,J-1}^x\}$,
- smerové derivácie $\{d_{0,0}^y, \dots, d_{0,J-1}^y, d_{I-1,0}^y, \dots, d_{I-1,J-1}^y\}$,
- zmiešané derivácie $\{d_{0,0}^{xy}, d_{I-1,0}^{xy}, d_{0,J-1}^{xy}, d_{I-1,J-1}^{xy}\}$.

Kým v prípade splajnových kriviek bolo evidentné hneď od začiatku ich výskumu a aplikácie, že okrem uzlov a funkčných hodnôt (z_0, \dots, z_{I-1}) v uzloch je potrebné zadať ešte dve podmienky, v našom prípade sú to dve hodnoty d_0 a d_{I-1} . V prípade interpolačných splajnových povrchov nebolo zrejmé ktoré smerové a zmiešané derivácie je potrebné zadať. De Boor navrhol vhodný model, ktorý okrem hodnôt $z_{i,j}$ na mriežke veľkosti $I \times J$ vyžaduje zadanie smerových derivácií na okrajoch mriežky uzlov a štyri zmiešané derivácie v rohoch.

Prirodzene vidieť, že takto definovaný krivkový aj povrchový splajn je triedy C^1 . Pri vhodne zvolených hodnotách derivácií v uzloch vieme ale zaručiť rovnosť aj druhých derivácií, čím dostaneme splajn z triedy C^2 čo nám zaručí hladkosť *spojenia* jednotlivých segmentov. Splajny sa taktiež dajú použiť na interpoláciu matematických funkcií.



Obr. 1.1: De Boorov model vstupných hodnôt pre splajnový povrch.

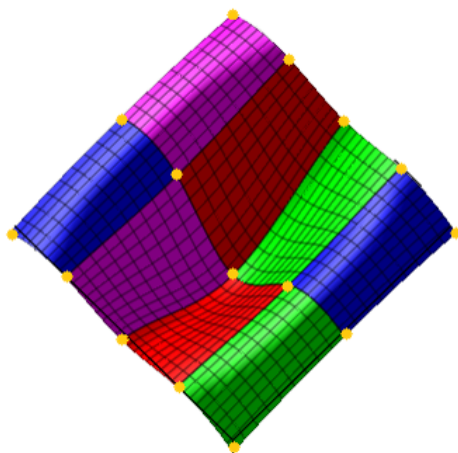
Definícia 1.9 Nech S je splajn na uzloch (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) a $f : [u_0, v_0] \times u_{I-1}, v_{J-1} \rightarrow \mathbb{R}$ je spojitá funkcia. Hovoríme, že splajn S interpoluje funkciu f ak platia tieto podmienky

- $z_{i,j} = f(u_i, v_j)$, pre každé i z $\{0, \dots, I-1\}$ a pre každé j z $\{0, \dots, J-1\}$.
- $d_{i,j}^x = \frac{\partial f(u_i, v_j)}{\partial x}$, pre každé i z $\{0, \dots, I-1\}$ a pre každé j z $\{0, J-1\}$.
- $d_{i,j}^y = \frac{\partial f(u_i, v_j)}{\partial y}$, pre každé i z $\{0, I-1\}$ a pre každé j z $\{0, \dots, J-1\}$.
- $d_{i,j}^{xy} = \frac{\partial^2 f(u_i, v_j)}{\partial x \partial y}$, pre každé i z $\{0, I-1\}$ a pre každé j z $\{0, J-1\}$.

Na splajn interpolujúcu funkciu f môžeme aplikovať De Boorovu interpoláciu a zo získaných derivácií už vieme zostrojiť jeho segmenty. Takto získaný splajn bude danú funkciu interpolovať, teda ju *napodobní*. To je žiadané napríklad v grafickom modelovaní, pretože pre počítač je jednoduchšie a rýchlejšie pracovať s polynómom interpolujúcim napríklad goniometrickú funkciu ako priamo s ňou.

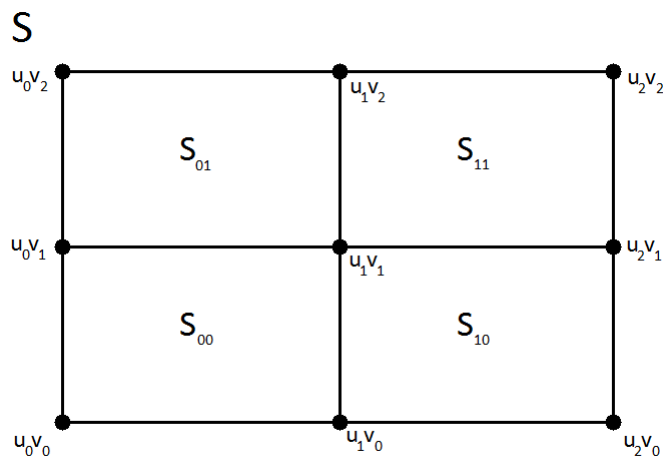
Obecne splajny minimalizujú integrál druhej derivácie funkcie (napr. zakrivenie, energie, ...). Poskytujú pružný nástroj na modelovanie reálnych situácií na lokálne požiadavky, pričom ich výpočet je rýchly a stabilný. Navyše sa dá ukázať, že interpolácia funkcie kubickým splajnom je jednoznačná. Témou tejto práce sú práve metódy výpočtu prvých derivácií, ktorými je možné dosiahnuť práve spomenutý cieľ.

Základná téma našej práce je úloha na základe vstupných uzlov u_0, \dots, u_{I-1} , v_0, \dots, v_{J-1} a funkčných hodnôt $z_{0,0}, \dots, z_{I,J}$ nájsť *hladkú*, po častiach



Obr. 1.2: Funkcia $\sin(\sqrt{x^2 + y^2})$ interpolovaná bikubickým splajnom.

definovanú funkciu $S : [u_0, u_{I+1}] \times [v_0, v_{J+1}] \rightarrow \mathbb{R}$ so spojitými deriváciami prvého aj druhého rádu takú, že pre každé $i \in 0, \dots, I-1$ a $j \in 0, \dots, J-1$ platí $z_{i,j} = S(u_i, v_j)$. Funkciu S nazývame splajn (konkrétne povrchový splajn), pričom jednotlivé časti nazveme *segmenty*.



Obr. 1.3: Ukážka uzlov pre štvorsegmentový splajn.

1.2.1 Konštrukcia hermitovho splajnu

Teraz si ukážeme ako môžeme podľa zadaných uzlov, ich funkčných hodnôt a derivácií zostrojiť segmenty splajnu ako funkcie dvoch premenných v priestore. Predpokladajme $I \geq 3$ a $J \geq 3$ z $\mathbb{N} \cup \{0\}$, rovnomerné uzly u_0, \dots, u_{I-1} a v_0, \dots, v_{J-1} . Úlohou je zostrojiť splajn S triedy C^2 interpolujúci funkciu f . Položme

- funkčné hodnoty splajnu $z_{0,0}, \dots, z_{I-1,0}, \dots, z_{0,J-1}, \dots, z_{I-1,J-1}$,
- x -ové derivácie splajnu $d_{0,0}^x, \dots, d_{I-1,0}^x, \dots, d_{0,J-1}^x, \dots, d_{I-1,J-1}^x$,
- y -ové derivácie splajnu $d_{0,0}^y, \dots, d_{I-1,0}^y, \dots, d_{0,J-1}^y, \dots, d_{I-1,J-1}^y$,
- zmiešané derivácie splajnu $d_{0,0}^{xy}, \dots, d_{I-1,0}^{xy}, \dots, d_{0,J-1}^{xy}, \dots, d_{I-1,J-1}^{xy}$.

Máme všetky potrebné hodnoty aby sme mohli definovať segmenty splajnu S podľa definície 1.7. Pre každé i z $\{0, \dots, I-2\}$ a j z $\{0, \dots, J-2\}$ položme segment $S_{i,j} : [u_i, v_j] \times [u_{i+1}, v_{j+1}] \rightarrow \mathbb{R}$ vzťahom:

$$S_{i,j}(x, y) = \lambda^T(x, u_i, u_{i+1}) \cdot \Phi(u_i, u_{i+1}, v_j, v_{j+1}) \cdot \lambda(y, v_j, v_{j+1}), \quad (1.3)$$

kde

$$\lambda(t, t_0, t_1) = \left(\frac{(t-t_1)^2(1-2\frac{t-t_0}{t_0-t_1})}{(t_0-t_1)^2}, \frac{(t-t_0)^2(1-2\frac{t-t_1}{t_1-t_0})}{(t_1-t_0)^2}, \frac{(t-t_1)^2(t-t_0)}{(t_0-t_1)^2}, \frac{(t-t_0)^2(t-t_1)}{(t_1-t_0)^2} \right)^T, \quad (1.4)$$

$$\Phi(t_0, t_1, s_0, s_1) = \left(\frac{(t-t_1)^2(1-2\frac{t-t_0}{t_0-t_1})}{(t_0-t_1)^2}, \frac{(t-t_0)^2(1-2\frac{t-t_1}{t_1-t_0})}{(t_1-t_0)^2}, \frac{(t-t_1)^2(t-t_0)}{(t_0-t_1)^2}, \frac{(t-t_0)^2(t-t_1)}{(t_1-t_0)^2} \right)^T, \quad (1.5)$$

Ak i je z $\{0, \dots, I-2\}$ a j je z $\{0, \dots, J-2\}$, tak štvorice uzlov (u_i, v_j) , (u_i, v_{j+1}) , (u_{i+1}, v_j) a (u_{i+1}, v_{j+1}) tvoria obdĺžnikový úsek nad ktorým sa nachádza splajnový segment. Každý segment $S_{i,j}$ je bikubická funkcia z $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$ do \mathbb{R} . Výsledná funkcia S teda vznikne zjednotením segmentov $S_{i,j}$. Na vypočítanie každého segmentu potrebujeme štyri uzly, a pre každý uzol príslušne hodnoty $z_{i,j}$, $d_{i,j}^x$, $d_{i,j}^y$ a $d_{i,j}^{xy}$.

Poznámka 1.10 Hodnoty $z_{i,j}$ musíme mať vždy vopred dané. Derivácie $d_{i,j}^x$, $d_{i,j}^y$ a $d_{i,j}^{xy}$ buď môžu byť dané všetky, alebo iba niektoré a z nich vieme s malou odchýlkou vypočítať hodnoty ostatných. Jeden z týchto spôsobov spočíva

v De Boorovej interpolácii ktorú si v ďalších dvoch častiach objasníme. Potom predstavíme modifikovaný spôsob interpolácie špeciálne pre splajny s rovnomerne rastúcimi uzlami.

1.3 Trojdiagonálna LU dekompozícia

Základ De Boorovej interpolácie tkvie v opakovanom počítaní systémov trojdiagonálnych lineárnych rovníc.

Definícia 1.11 Nech $n \geq 3$ je z $\mathbb{N} \cup \{0\}$. Sústavu rovníc tvaru

$$\begin{pmatrix} b_0 & c_0 & 0 & \cdots & 0 & 0 \\ a_0 & b_1 & c_1 & \cdots & 0 & 0 \\ 0 & a_1 & b & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b & c_{n-2} \\ 0 & 0 & 0 & \cdots & a_{n-2} & b_{n-1} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{n-1} \\ r_n \end{pmatrix} \quad (1.6)$$

nazveme *trojdiagonálna sústava lineárnych rovníc*.

Jeden z efektívnych spôsobov riešenia týchto rovníc spočíva v LU dekompozícii $A\mathbf{x} = L \underbrace{U\mathbf{x}}_{\mathbf{y}} = \mathbf{r}$, kde maticu A rozložíme na súčin matíc L a U v

tvaru

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ l_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & l_2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & l_{n-1} & 1 \end{pmatrix} \cdot \begin{pmatrix} u_0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & u_1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & u_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & v_{n-2} & 0 \\ 0 & 0 & 0 & \cdots & 0 & v_{n-1} \end{pmatrix}. \quad (1.7)$$

Pre k z $\{1, \dots, n\}$ sú hodnoty v_k a λ_k určené takto:

$$v_i = b, \left\{ \lambda_i = \frac{1}{v_{i-1}}, v = b - \lambda_i \right\}, i \in \{2, \dots, n\}. \quad (1.8)$$

Pre priamy a spätný chod máme

$$\text{Priamy: } L\mathbf{y} = \mathbf{r}, y_1 = r_1, \{y_i = r_i - \lambda_i\}, i \in \{2, \dots, n\}, \quad (1.9)$$

$$\text{Spätný: } U\mathbf{d} = \mathbf{y}, d_i = \frac{y_i}{u_i}, \left\{ d_i = \frac{1}{u_i}(y_i - d_{i+1}) \right\}, i \in \{n-1, \dots, 1\}. \quad (1.10)$$

LU dekompozíciou sa rieši ako de Boorova sústava (1.15), tak aj naša redukovaná (1.26). Nižšie si ukážeme optimalizovaný pseudokód tohto algoritmu prevzatý z [7]. Procedúra algoritmu je priestorovo optimalizovaná kde LU rozklad robíme priamo na vstupných vektorech predstavujúce diagonály a výsledok ukladáme do vektora obsahujúci pravú stranu.

Algoritmus 1 Všeobecná LU dekompozícia

```

1: procedúra VYRIEŠLU
2:   vstup:
3:      $a[0..n-2]$  spodná diagonála
4:      $b[0..n-1]$  hlavná diagonála
5:      $c[0..n-2]$  horná diagonála, po skončení sa nezachová
6:      $r[0..n-1]$  pravá strana
7:   výstup:  $r[0..n-1]$  pravá strana obsahujúca výsledok
8:
9:      $c[0] \leftarrow c[0]/b[0]$ 
10:     $r[0] \leftarrow r[0]/b[0]$  ▷ Priamy prechod
11:    pre  $i$  od 1 až  $n-1$ 
12:       $m \leftarrow 1/b[i] - a[i] \cdot c[i-1]$ 
13:       $c[i] \leftarrow m \cdot c[i]$ 
14:       $r[i] \leftarrow m \cdot (r[i] - a[i] \cdot r[i-1])$  ▷ Spätný prechod
15:    pre  $i$  od  $n-2$  až 0
16:       $r[i] \leftarrow r[i] - c[i] \cdot r[i+1]$ 

```

Lemma 1.12 *Nech n je počet rovníc trojdiagonálnej lineárnej sústavy. Algoritmus 1 nájde riešenie v lineárnom čase, pričom vyžaduje $5n$ pamäte.*

V prípade počítania derivácií je sústava rovníc v špeciálnom tvare podľa 1.15 pre plný alebo 1.26 pre redukovaný algoritmus v nasledujúcich častiach. V oboch prípadoch všetky čísla na hornej aj spodnej diagonále majú hodnotu 1. Na hlavnej diagonále sú to hodnoty 4 pri plnom algoritme a v prípade redukovaného postupu sú to hodnoty -14 . Posledný prvok diagonály tiež môže obsahovať hodnotu -15 ak počet rovníc je párny. Teda pri implementácii si nemusíme pamätať vektory, stačí nám zapamätať si iba štyri hodnoty. Upravený algoritmus, ktorý rieši LU dekompozíciu špeciálne pre plný a redukovaný algoritmus je nižšie.

Algoritmus 2 Špeciálna LU dekompozícia

```
1: procedúra VYRIEŠLU
2:   vstup:
3:
4:    $b$  hodnota prvkov hlavnej diagonály
5:    $b_l$  hodnota posledného prvku hlavnej diagonály
6:    $r[0..n-1]$  pravá strana
7:   výstup:  $r[0..n-1]$  pravá strana obsahujúca výsledok
8:
9:    $p \leftarrow [0..n-2]$   $\triangleright$  Pomocný vektor nahradzujúci hornú diagonálu
10:   $m \leftarrow 1/b$ 
11:   $p[0] \leftarrow q$ 
12:   $r[0] \leftarrow q \cdot r[0]$ 
13:  pre  $i$  od 1 až  $n-2$ 
14:     $q \leftarrow 1/b - p[i-1]$ 
15:     $p[i] \leftarrow m$ 
16:     $r[i] \leftarrow m \cdot (r[i] - r[i-1])$ 
17:   $m \leftarrow 1/b_l - p[n-1]$ 
18:   $p[i] \leftarrow m$ 
19:   $r[i] \leftarrow m \cdot (r[n-1] - r[n-2])$ 
20:  pre  $i$  od  $n-2$  až 0
21:     $r[i] \leftarrow r[i] - p[i] \cdot r[i+1]$ 
```

Lemma 1.13 *Nech n je počet rovníc trojdiagonálnej lineárnej sústavy, všetky prvky na spodnej a hornej diagonále majú hodnotu 1 a pre hodnoty b_0, b_1, \dots, b_{n-1} na hlavnej diagonále platí $b_0 = b_1 = \dots b_{n-2}$. Algoritmus 2 nájde riešenie v lineárnom čase, pričom vyžaduje $2n$ pamäte.*

Tento modifikovaný algoritmus má menej než polovičné pamäťové nároky než v prípade všeobecných trojdiagonálnych sústav, čo nám v praxi umožňuje riešiť väčšie úlohy. Navyše pri reálnej implementácii nemusíme vektory r a p inicializovať pri každom volaní procedúry *VyriešLU* ale namiesto toho môžeme použiť prednačítané vyrovňavacie vektory, tzv. *buffery*. To nám ušetrí mnoho strojových cyklov, ktoré procesor strávi neustálym alokovaním a zmazaním potrebnej operačnej pamäte.

Máme už potrebný základ aby sme mohli prejsť k cieľu práce a tým je práve De Boorov výpočet derivácií splajnu a z neho odvodený efektívnejší algoritmus.

1.4 De Boorov výpočet derivácií

Skôr než začneme s popisom De Boorvho modelu výpočtu derivácií v uzloch si položíme dve označenia s ktorými budeme pracovať.

Označenie 1.14 Postupnosť (a_0, a_1, \dots) nazveme *rovnomere rastúcou* ak pre ľubovoľné i z $\{0, 1, \dots\}$ platí $a_{i+1} > a_i$ a pre ľubovoľné i, j z $\{0, 1, \dots\}$ platí $a_{j+1} - a_j = a_{i+1} - a_i$.

Označenie 1.15 Pre rovnomerne rastúcu postupnosť (a_0, a_1, \dots) označme hodnotu $h_a = a_1 - a_0$.

Nech sú dané rovnomerne rastúce postupnosti uzlov (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) , kde I, J sú z $\mathbb{N} \cup \{0\}$, pričom chceme interpolovať funkčné hodnoty $\{z_{0,0}, \dots, z_{0,J-1}, \dots, z_{I-1,0}, \dots, z_{I-1,J-1}\}$. Výsledný splajn bude teda tvorený $(I-1) \cdot (J-1)$ segmentami. Ako bolo spomenuté, každý segment splajnu potrebuje na svoje zostrojenie štyri uzly a hodnoty z , d^x , d^y a d^{xy} . Získanie derivácií však môže byť niekedy nákladné. Príkladom môže byť v prípade, keď funkčné hodnoty $z_{i,j}$ sú získané vyhodnotením nejakej funkcie f . To môže byť v praxi na počítači značne pomalé najmä v prípade, ak pracujeme so symbolicky zapísanou funkciou vo forme textového reťazca, ktorú je potrebné dynamicky interpretovať počas behu programu.

Algoritmus nájdený Carlom de Boorom [4] umožňuje s malou odchýlkou vypočítať hodnoty derivácií v uzloch na základe nasledujúcich vstupných hodnôt, ktoré máme dané:

- $z_{i,j}$ pre $i \in \{0, \dots, I-1\}$, $j \in \{0, \dots, J-1\}$.
- $d_{i,j}^x$ pre $i \in \{0, I-1\}$, $j \in \{0, \dots, J-1\}$.
- $d_{i,j}^y$ pre $i \in \{0, \dots, I-1\}$, $j \in \{0, J-1\}$.
- $d_{i,j}^{xy}$ pre $i \in \{0, I-1\}$, $j \in \{0, J-1\}$.

Poznámka 1.16 De Boorova interpolácia vo všeobecnosti nepredpokladá len rovnomerne rastúce postupnosti uzlov (u_0, \dots, u_I) a (v_0, \dots, v_J) . Náš postup v ďalšej časti článku ale funguje len s takýmito postupnosťami. Preto v tejto časti budeme uvažovať De Boorov postup špeciálne pre rovnomerne rastúce postupnosti uzlov.

Príklad 1.17 Na príklade uzlov z obrázka 1.3 potrebujeme poznať hodnoty následovne:

- $z_{0,2}, z_{1,2}, z_{2,2},$
 $z_{0,1}, z_{1,1}, z_{2,1},$
 $z_{0,0}, z_{1,0}, z_{2,0},$

- $d_{0,2}^x, \quad , d_{2,2}^x,$
 $d_{0,1}^x, \quad , d_{2,1}^x,$
 $d_{0,0}^x, \quad , d_{2,0}^x,$

- $d_{0,2}^y, d_{1,2}^y, d_{2,2}^y,$
 $\quad , \quad , \quad ,$
 $d_{0,0}^y, d_{1,0}^y, d_{2,0}^y,$

- $d_{0,2}^{xy}, \quad , d_{2,2}^{xy},$
 $\quad , \quad , \quad ,$
 $d_{0,0}^{xy}, \quad , d_{2,0}^{xy},$

Zvyšné derivácie d^x , d^y a d^{xy} vieme jednoznačne vypočítať pomocou $2(I) + J + 5$ lineárnych sústav s celkovo $3IJ + I + J + 2$ rovnicami. Nižšie uvádzame modelové rovnice, pomocou ktorých sú zostrojené tieto sústavy lineárnych rovníc.

Pre $j \in \{0, \dots, J-1\}$, teda pre každý stĺpec j vypočítame parciálne derivácie d^x

$$d_{i+1,j}^x + 4d_{i,j}^x + d_{i-1,j}^x = \frac{3}{h_u}(z_{i+1,j} - z_{i-1,j}), \quad (1.11)$$

$$i \in \{1, \dots, I-2\}$$

Pre $j \in \{0, J-1\}$, teda pre prvý a posledný stĺpec vypočítame parciálne derivácie $d^{x,y}$

$$d_{i+1,j}^{xy} + 4d_{i,j}^{xy} + d_{i-1,j}^{xy} = \frac{3}{h_u}(d_{i+1,j}^y - d_{i-1,j}^y), \quad (1.12)$$

$$i \in \{1, \dots, I-2\}$$

Pre $i \in \{0, \dots, I-1\}$, teda pre každý riadok i vypočítame parciálne derivácie d^y

$$d_{i,j+1}^y + 4d_{i,j}^y + d_{i,j-1}^y = \frac{3}{h_v}(z_{i,j+1} - z_{i,j-1}), \quad (1.13)$$

$$j \in \{1, \dots, J-2\}$$

Pre $i \in \{0, \dots, I-1\}$, teda pre každý riadok j dopočítame parciálne derivácie $d^{x,y}$

$$d_{i,j+1}^{xy} + 4d_{i,j}^{xy} + d_{i,j-1}^{xy} = \frac{3}{h_v}(d_{i,j+1}^x - d_{i,j-1}^x), \quad (1.14)$$

$$j \in \{1, \dots, J-2\}$$

Každá z týchto sústav má takýto maticový tvar:

$$\begin{pmatrix} 4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 4 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 4 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_{N-3} \\ D_{N-2} \end{pmatrix} = \begin{pmatrix} \frac{3}{h}(Y_2 - Y_0) - D_0 \\ \frac{3}{h}(Y_3 - Y_1) \\ \frac{3}{h}(Y_4 - Y_2) \\ \vdots \\ \frac{3}{h}(Y_{N-2} - Y_{N-4}) \\ \frac{3}{h}(Y_{N-1} - Y_{N-3}) - D_{N-1} \end{pmatrix}, \quad (1.15)$$

kde podľa toho o ktorú z modelových rovníc sa jedná, hodnoty N , D a Y zadávame následovne. Nech k z $1, \dots, K-1$. Potom

- $N = I$, $h = h_u$, $D_k = d_{k,j}^x$ a $Y_k = z_{k,j}$, pre rovnicu 1.11,
- $N = I$, $h = h_u$, $D_k = d_{k,j}^{xy}$ a $Y_k = d_{k,j}^y$, pre rovnicu 1.12,
- $N = J$, $h = h_v$, $D_k = d_{i,k}^y$ a $Y_k = z_{i,k}$, pre rovnicu 1.13,
- $N = J$, $h = h_v$, $D_k = d_{i,k}^{xy}$ a $Y_k = d_{i,k}^x$, pre rovnicu 1.13.

Po vypočítaní všetkých derivácií môžeme funkčné hodnoty jednoznačne interpolovať splajnom.

1.5 Počítanie derivácií redukovanou sústavou

V rámci svojej bakalárskej práce som popisoval nový postup pre kubické splajnové krivky triedy C^2 , teda splajny, kde interpolovaná funkcia f je typu $\mathbb{R} \rightarrow \mathbb{R}$. Cieľom tejto práce je analyzovať a vylepšiť De Boorov algoritmus pre bikubické splajny z predchádzajúcej časti na základe výskumu [2] doc. Töröka a RNDr. Szaba o vzťahu medzi bikvartickými polynómami a bikubickými splajnami a odvodenie sústav na základe tohto výsledku [3].

Označenie 1.18 Teraz popíšeme tento nový algoritmus, ktorý pracovne označíme ako *redukovaný algoritmus*. Ďalej budeme pôvodný de Boorov postup, označovať pojmom *plný algoritmus*.

Vstupné hodnoty sú identické ako pri pôvodnom algoritme. Teda máme dané (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) , kde I, J sú z $\mathbb{N} \cup \{0\}$, pričom chceme zostrojiť splajn S , ktorý interpoluje hodnoty $\{z_{0,0}, \dots, z_{I-1,0}, \dots, z_{0,J-1}, \dots, z_{I-1,J-1}\}$ aby pre každé $i \in \{0, \dots, I-1\}$, $j \in \{0, \dots, J-1\}$ platilo $z_{i,j} = S(u_i, v_j)$. Pre pripomenutie potrebujeme ešte poznať tieto hodnoty.

- $d_{i,j}^x$ pre $i \in \{0, I-1\}$, $j \in \{0, \dots, J-1\}$.
- $d_{i,j}^y$ pre $i \in \{0, \dots, I-1\}$, $j \in \{0, J-1\}$.
- $d_{i,j}^{xy}$ pre $i \in \{0, I-1\}$, $j \in \{0, J-1\}$.

Zvyšné derivácie d^x , d^y a d^{xy} vieme jednoznačne vypočítať pomocou $2(I) + J + 5$ lineárnych sústav s celkovo $3IJ + I + J + 2$ rovnicami: Nižšie uvádzame modelové rovnice, pomocou ktorých sú zostrojené tieto sústavy lineárnych rovníc. Označme I_l a J_l indexy po ktoré budeme iterovať. Platí:

$$I_l = \begin{cases} I-2 & \text{ak } I \text{ je nepárne,} \\ I-3 & \text{ak } I \text{ je párne,} \end{cases}$$

$$J_l = \begin{cases} J-2 & \text{ak } J \text{ je nepárne,} \\ J-3 & \text{ak } J \text{ je párne,} \end{cases}$$

Pre $j \in \{0, \dots, J-1\}$, teda pre každý stĺpec j vypočítame parciálne derivácie d^x

$$d_{i+2,j}^x - 14d_{i,j}^x + d_{i-2,j}^x = \frac{3}{h_u}(z_{i+2,j} - z_{i-2,j}) - \frac{12}{h_u}(z_{i+1,j} - z_{i-1,j}), \quad (1.16)$$

$$i \in \{2, 4, \dots, I_l\}$$

Rovnica je podobná ako plnom algoritme 1.11. Všimnime si, že sústavu rovníc teraz budujeme len pre párne indexy i , teda vyriešením tejto sústavy získame iba polovicu žiadaných hodnôt d^x . Pre $i \in \{1, 3, \dots, I_l\}$ a $j \in \{0, \dots, J-1\}$ zvyšné derivácie d^x vypočítame ako

$$d_{i,j}^x = \frac{3}{4h_u}(z_{i+1,j} - z_{i-1,j}) - \frac{1}{4}(d_{i+1,j}^x - d_{i-1,j}^x) \quad (1.17)$$

Pre $i \in \{0, \dots, I-1\}$, teda pre každý riadok i analogicky vypočítame parciálne derivácie d^y

$$d_{i,j+2}^y - 14d_{i,j}^y + d_{i,j-2}^y = \frac{3}{h_v}(z_{i,j+2} - z_{i,j-2}) - \frac{12}{h_v}(z_{i,j+1} - z_{i,j-1}), \quad (1.18)$$

$$i \in \{2, 4, \dots, I_l\}$$

Následne analogicky pre $i \in \{1, 2, \dots, I-1\}$ a $j \in \{1, 3, \dots, J_l\}$ zvyšné derivácie d^y vypočítame ako

$$d_{i,j}^y = \frac{3}{4h_v}(z_{i,j+1} - z_{i,j-1}) - \frac{1}{4}(d_{i,j+1}^y - d_{i,j-1}^y) \quad (1.19)$$

Pre $j \in \{0, J-1\}$, teda pre prvý a posledný stĺpec vypočítame parciálne derivácie $d^{x,y}$ rovnako ako pri plnom algoritme.

$$d_{i+1,j}^{xy} + 4d_{i,j}^{xy} + d_{i-1,j}^{xy} = \frac{3}{h_u}(d_{i+1,j}^y - d_{i-1,j}^y), \quad (1.20)$$

$$i \in \{1, \dots, I-2\}$$

Pre $i \in \{0, I-1\}$, teda pre prvý a posledný riadok analogicky vypočítame parciálne derivácie $d^{x,y}$

$$d_{i,j+1}^{xy} + 4d_{i,j}^{xy} + d_{i,j-1}^{xy} = \frac{3}{h_v}(d_{i,j+1}^x - d_{i,j-1}^x), \quad (1.21)$$

$$j \in \{1, \dots, J-2\}$$

Pre $i \in \{2, 4, \dots, I_l\}$, teda pre každý stĺpec i dopočítame parciálne derivácie $d^{x,y}$

$$\begin{aligned} d_{i,j+2}^{xy} + 4d_{i,j}^{xy} + d_{i,j-2}^{xy} = & \\ & \frac{1}{7}(d_{i-2,j+2}^{xy} - d_{i-2,j-2}^{xy}) - 2d_{i-2,j}^{xy} \\ & + \frac{3}{7h_u}(d_{i-2,j+2}^y - d_{i-2,j-2}^y) + \frac{3}{7h_v}(-d_{i-2,j+2}^x - d_{i-2,j-2}^x) \\ & + \frac{9}{7h_u}(d_{i,j+2}^y - d_{i,j-2}^y) + \frac{9}{7h_u h_v}(-z_{i-2,j+2} + z_{i-2,j-2}) \\ & + \frac{12}{7h_u}(-d_{i-1,j+2}^y - d_{i-1,j-2}^y) + \frac{12}{7h_v}(d_{i-2,j+1}^x - d_{i-2,j-1}^x) \\ & + \frac{3}{7h_v}(d_{i,j+2}^x - d_{i,j-2}^x) + \frac{27}{7h_u h_v}(-z_{i,j+2} + z_{i,j-2}) \\ & + \frac{36}{7h_u h_v}(z_{i-1,j+2} - z_{i-1,j-2} + z_{i-2,j+1} - z_{i-2,j-1}) \\ & - \frac{6}{h_u}d_{i-2,j}^y + \frac{144}{7h_u h_v}(-z_{i-1,j+1} + z_{i-1,j-1}) + \frac{24}{h_u}d_{i-1,j}^y, \end{aligned} \quad (1.22)$$

$$j \in \{4, 6, \dots, J_l-2\}$$

Následne vypočítame zvyšné derivácie $d^{x,y}$. Najprv pre $i \in \{1, 3, \dots, I_l\}$ a

$j \in \{1, 3, \dots, J_l\}$ platí

$$\begin{aligned}
d_{i,j}^{xy} = & \frac{1}{16}(d_{i+1,j+1}^{xy} + d_{i+1,j-1}^{xy} + d_{i-1,j+1}^{xy} + d_{i-1,j-1}^{xy}) \\
& - \frac{3}{16h_v}(d_{i+1,j+1}^x - d_{i+1,j-1}^x + d_{i-1,j+1}^x - d_{i-1,j-1}^x) \\
& - \frac{3}{16h_u}(d_{i+1,j+1}^y + d_{i+1,j-1}^y - d_{i-1,j+1}^y - d_{i-1,j-1}^y) \\
& + \frac{9}{16h_u h_v}(z_{i+1,j+1} - z_{i+1,j-1} - z_{i-1,j+1} + z_{i-1,j-1}).
\end{aligned} \tag{1.23}$$

Nakoniec pre $i \in \{1, 3, \dots, I_l + 1\}$ a $j \in \{2, 4, \dots, J_l\}$

$$d_{i,j}^{xy} = \frac{3}{4h_v}(d_{i,j+1}^x - z d_{i,j-1}^x) - \frac{1}{4}(d_{i,j+1}^{xy} - d_{i,j-1}^{xy}) \tag{1.24}$$

a pre $i \in \{2, 4, \dots, I_l\}$ a $j \in \{1, 3, \dots, J_l + 1\}$

$$d_{i,j}^{xy} = \frac{3}{4h_v}(d_{i,j+1}^x - z d_{i,j-1}^x) - \frac{1}{4}(d_{i,j+1}^{xy} - d_{i,j-1}^{xy}) \tag{1.25}$$

Označenie 1.19 Zaveďme dve označenia.

- Rovnice 1.17, 1.19 budeme súhrne označovať pojmom *resty*.
- Rovnice 1.23, 1.24, 1.25 budeme súhrne označovať pojmom *zmiešané resty*.

Modelové sústavy rovníc 1.16 a 1.18 majú takýto maticový tvar

$$\begin{pmatrix} -14 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -14 & 1 & \cdots & 0 & 0 \\ 0 & 1 & -14 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -14 & 1 \\ 0 & 0 & 0 & \cdots & 1 & \mu \end{pmatrix} \cdot \begin{pmatrix} D_2 \\ D_4 \\ D_6 \\ \vdots \\ D_{v-2} \\ D_v \end{pmatrix} = \begin{pmatrix} \frac{3}{h}(Y_4 - Y_0) - \frac{12}{h}(Y_3 - Y_1) - D_0 \\ \frac{3}{h}(Y_6 - Y_2) - \frac{12}{h}(Y_5 - Y_3) \\ \frac{3}{h}(Y_8 - Y_4) - \frac{12}{h}(Y_7 - Y_5) \\ \vdots \\ \frac{3}{h}(Y_\nu - Y_{\nu-4}) - \frac{12}{h}(Y_{\nu-3} - Y_{\nu-5}) \\ \frac{3}{h}(Y_{\nu+\tau} - Y_{\nu-2}) - \frac{12}{h}(Y_{\nu-1} - Y_{\nu-3} - \theta D_{K+1}) \end{pmatrix}, \tag{1.26}$$

kde

$$\begin{aligned}
\mu = -15, \tau = 0, \theta = -4, \text{ a } \nu = N, & \quad \text{ak } K \text{ je párne,} \\
\mu = -14, \tau = 2, \theta = 1, \text{ a } \nu = N - 1, & \quad \text{ak } N \text{ je nepárne,}
\end{aligned} \tag{1.27}$$

a podľa toho o ktorú z modelových rovníc sa jedná, hodnoty N , D a Y zadávame následovne. Nech k z $1, \dots, K-1$. Potom

- $N = I$, $h = h_u$, $D_k = d_{k,j}^x$ a $Y_k = z_{k,j}$, pre rovnicu 1.16..
- $N = J$, $h = h_v$, $D_k = d_{i,k}^y$ a $Y_k = z_{i,k}$, pre rovnicu 1.18.

Analogicky vieme zostrojiť maticový tvar aj pre modelovú sústavu pre derivácie d^{xy} podľa 1.22.

Označenie 1.20 Špeciálne zavedme označenia pre tieto hodnoty:

- Parciálne derivácie d^x a d^y počítané rovnicami 1.17 a 1.19 budeme nazývať *zostatkové derivácie*.
- Zmiešané parciálne derivácie d^{xy} počítané rovnicami 1.23, 1.24 a 1.25 budeme nazývať *zmiešané zostatkové derivácie*.

Následujúca lema formálne zhrnie fakt, že splajn s vypočítanými deriváciami podľa plného alebo redukovaného algoritmu je interpolačný.

Lemma 1.21 (Miño-Török [3]) Nech sú zadané nasledujúce hodnoty:

- $I \geq 3$ a $J \geq 3$,
- (u_0, \dots, u_{I-1}) a (v_0, \dots, v_{J-1}) sú rovnomerne rastúce postupnosti uzlov,
- pre každé i z $\{0, \dots, I-1\}$ a j z $\{0, \dots, J-1\}$ sú z_{ij} funkčné hodnoty v uzloch
- pre každé i z $\{0, \dots, I-1\}$ a pre každé j z $\{0, J-1\}$ sú $d_{i,j}^x$ smerové derivácie na osi x ,
- pre každé i z $\{0, I-1\}$ a pre každé j z $\{0, \dots, J-1\}$ sú $d_{i,j}^y$ smerové derivácie na osi y ,
- pre každé i z $\{0, I-1\}$ a pre každé j z $\{0, J-1\}$ sú $d_{i,j}^{xy}$ zmiešané derivácie na osiach x a y .

Ďalej nech následovné hodnoty sú vypočítané podľa plného alebo redukovaného algoritmu:

- $d_{i,j}^x$, pre každé i z $\{0, \dots, I-1\}$ a pre každé j z $\{1, \dots, J-2\}$.
- $d_{i,j}^y$, pre každé i z $\{1, \dots, I-2\}$ a pre každé j z $\{0, \dots, J-1\}$.
- $d_{i,j}^{xy}$, pre každé i z $\{0, \dots, I-1\}$ a pre každé j z $\{0, \dots, J-1\}$ okrem $d_{0,0}^{xy}$, $d_{I-1,0}^{xy}$, $d_{0,J-1}^{xy}$ a $d_{I-1,J-1}^{xy}$.

Potom splajn S s vypočítanými segmentami podľa časti 1.2.1 s vyššie uvedenými hodnotami interpoluje funkčné hodnoty $\{z_{0,0}, \dots, z_{I-1,J-1}\}$.

Paralelizácia Všimnime si, že v prípade oboch algoritmov derivácie podľa jednotlivých premenných počítame v vždy jednom smere. Uvažujme napríklad výpočet smerových derivácií po osi x . Jedna LU dekompozícia vypočíta derivácie práve pre jeden riadok matice uzlov, pričom tieto výpočty sú vzájomne nezávislé. To umožňuje x -ové derivácie na jednotlivých riadkoch matice počítateľ paralelne. Analogická úvaha platí aj pre y -ové aj zmiešané derivácie. Pri zmiešaných zostatkových deriváciách u redukovaného algoritmu sa situácia trochu komplikuje pretože ich zatiaľ nevieme počítateľ nezávisle na sebe. To má negatívny vplyv na rýchlosť paralelnej verzie redukovaného algoritmu a preto bude zrýchlenie oproti paralelnému plnému algoritmu menšie ako v prípade sériových verzií.

Kapitola 2

Zrýchlenie

V predchádzajúcej kapitole sme popísali dva postupy výpočtu neznámych derivácií pre splajnové povrchy. Časová zložitosť oboch algoritmov je rovnaká a síce $O(I \cdot J)$. Pri určení asymptotickej časovej zložitosti zvyčajne zanedbávame rýchlosti jednotlivých krokov algoritmu ako sú napríklad aritmetické operácie, porovnávanie veľkosti čísel, kopírovanie a podobne. Keď porovnáваме rýchlosť asymptoticky rovnako rýchlych postupoch musíme brať do úvahy vplyvy jednotlivých týchto elementárnych krokov. Pochopiteľne tieto operácie nemusia byť rovnako rýchle a môže nastať situácia kedy algoritmus s väčším počtom krokov bude rýchlejší ako iný s menším počtom krokov, ak tieto kroky bude procesor počítača schopný vykonávať rýchlejšie.

To je práve aj náš prípad. Ak u oboch algoritmov spočítame všetky aritmetické operácie, tak výsledok bude väčší počet operácií práve u redukovaného algoritmu. Pri reálnej implementácii na počítači sa ale ukázalo, že napriek väčšiemu počtu aritmetických operácií ako sú sčítanie, násobenie a delenie, je redukovaný algoritmus rýchlejší čo si ukážeme v časti 2.2. V nasledujúcej časti si najprv popíšeme technické parametre procesorov v dôsledku ktorých dosahujeme zrýchlenie a potom spočítame počty aritmetických operácií oboch algoritmov. Následne budeme schopní k jednotlivým aritmetickým operáciám v jednotlivých krokoch priradiť *cenу*, ktoré určia váhu týchto krokov a z nich vypočítame celkové rýchlosti algoritmov a následne zrýchlenie redukovaného spôsobu.

Na moderných procesoroch v prípade matematických operácií s plávajúcou desatinnou čiarkou platí, že sčítanie a násobenie sú podobne rýchle, pričom delenie je niekoľkonásobne pomalšie. Pamäťové operácie sú pritom v závislosti od konkrétneho typu operačnej pamäte približne dvadsaťnásobne pomalšie ako sčítanie, resp. násobenie.¹

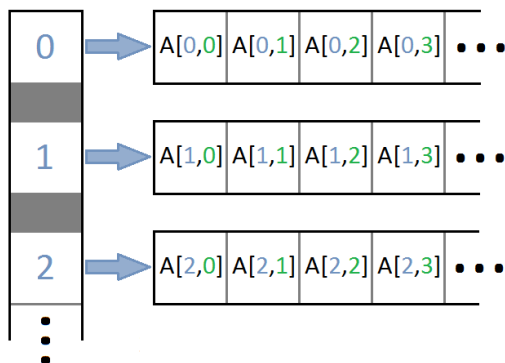
¹V praxi sa aj výpočty rádovo tisícov uzlov často zmestia do vyrovnávacej pamäte

2.1 Procesorová architektúra

Predtým ako začneme s počítaním operácií je nutné objasniť si ako moderné procesory narábajú s dátami a ako funguje aplikovanie výpočtov. Architektúry CPU prešli za posledné desaťročia značným vývojom. V dnešnej dobe už nemožno zvyšovať výpočtovú rýchlosť hrubou silou zvyšovaním frekvencie. Moderné procesorové mikroarchitektúry často používajú rôzne optimalizačné „triky a finty“ ako dosiahnuť zlepšenie výkonu a ktoré majú značný vplyv na reálne rýchlosti programov a algoritmov. Dvomi takými optimalizáciami sú takzvané kešovanie a inštrukčný paralelizmus.

2.1.1 Kešovanie

Pri návrhu algoritmu je vhodné si uvedomiť tvar dátových štruktúr do ktorých pri reálnej implementácii budeme ukladať funkčné hodnoty, uzly a derivácie. Naša implementácia v prípade kriviek používa klasické vektory, teda polia. Pri povrchoch sa hodnoty reprezentujú maticami. Maticu je možné interpretovať buď ako „pole polí“ respektíve „vektor vektorov“ (anglicky „jagged array“), alebo ako jedno spojité pole, kde prvok na i -tom riadku a j -tom stĺpci má v takejto reprezentácii index $n \cdot i + j$, kde n je počet stĺpcov. My budeme v meraniach a výpočtoch uvažovať práve prvú reprezentáciu znázorненú na obrázku nižšie.



Obr. 2.4: Vizuálna reprezentácia dátovej štruktúry „pole polí“.

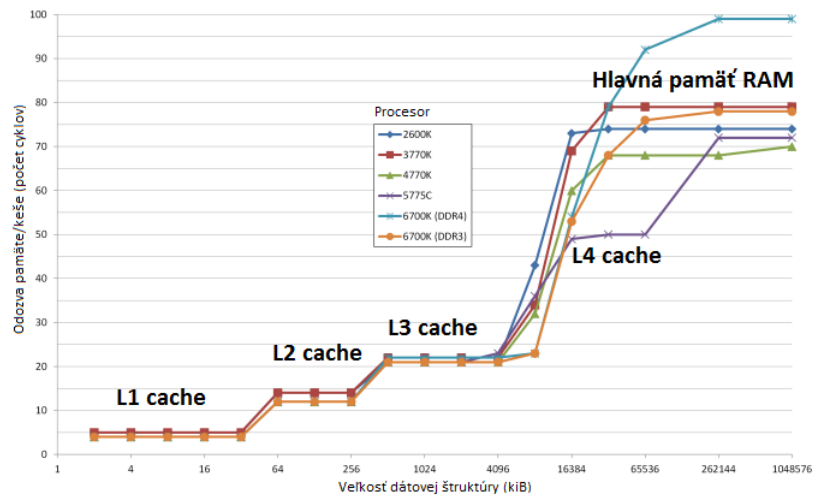
Operačná pamäť v dnešných počítačoch je rádovo pomalšia ako procesor, ktorý musí čakať desiatky až stovky strojových cyklov ak chce načítať hodnotu z pamäte. Moderné procesory preto používajú systém malých, ale procesora (*cache*). V tom prípade sú pamäťové operácie len $1 \times 10 \times$ pomalšie ako sčítanie.

rýchlych vyrovnávacích pamätí tzv. *cache* do ktorých sa prednačítavajú dáta a strojové inštrukcie programu z operačnej pamäte, v ideálnom prípade skôr ako ich procesor bude potrebovať. Toto kešovanie je plne automatický proces riadený samotným CPU. Programátor má len nepriame možnosti ovplyvnenia tohto procesu, napríklad vhodným výberom dátových štruktúr a podobne.

Uvažujme maticu $m \times n$ typu pole polí ako z obrázka vyššie. Ak z matice načítame prvok $a_{0,0}$ tak je vysoko pravdepodobné, že procesor následne bude mať nakešované aj prvky $a_{0,1}$, $a_{0,2}$ pretože sú z jedného poľa a teda sú v pamäti „vedľa seba“ (CPU kešuje pamäť po spojitých blokoch), ale už nemusí mať kešovaný prvok $a_{1,0}$ lebo tento je už z iného poľa a teda z úplne inej oblasti pamäte. Z toho v praxi vyplýva, že výraz $a_{0,0} + a_{0,1}$ procesor vyhodnotí rýchlejšie ako výraz $a_{0,0} + a_{1,0}$ pretože v prvom prípade má obe hodnoty vo vyrovnávacej pamäti.

Pozorný čitateľ si môže všimnúť, že v prípade použitia druhej reprezentácie matice, teda jedného veľkého spojitého poľa dĺžky $m \cdot n$ by sa zvýšila pravdepodobnosť kešovania aj prvku $a_{1,0}$. V tomto prípade sa totiž bude nachádzať v tom istom bloku pamäte ako prvok $a_{0,0}$ iba o n prvkov (t.j. jeden celý stĺpec) ďalej. Toto je pravda a všeobecne sa druhá menovaná reprezentácia považuje za efektívnejšiu vzhľadom na procesorový čas a pamäť počítača, ale to platí len pre malé matice najmä s malým počtom stĺpcov n .

V aplikácii sme totiž vyskúšali obe varianty reprezentácie a pre matice s rádovo 1000×1000 prvkami sa ako efektívnejšia ukazuje práve prvá reprezentácia, teda „pole polí“. Totiž ak jeden prvok matice má 8 bajtov (veľkosť číselného typu *double* vo väčšine jazykoch), tak matica zaberá v pamäti približne 7,63 mebibajtov. Pri matici typu pole polí máme 1000 polí každé zabierajúce práve 7,81 kibibajtov ($1 \text{ KiB} = 2^{10} \text{ B} = 1024 \text{ B}$, $1 \text{ MiB} = 2^{10} \text{ KiB}$, atď.). Pre CPU je kešovanie takýchto polí podstatne jednoduchšie ako v prípade matice reprezentovanej jedným poľom, ktoré samotné má veľkosť práve 7,63 MiB, čo presahuje kapacitu väčšiny vyrovnávacích pamätí. Obrázok 2.5 nižšie názorne ukazuje vzťah medzi veľkosťou dátovej štruktúry v KiB a počtom strojových cyklov, ktoré procesor musí čakať pre prístup k jej prvkom. Ako vidieť do pri dosiahnutí veľkosti 32 KiB, dáta sú namiesto veľmi rýchlej vyrovnávacej pamäte L1 kešované do trojnásobne pomalšej vyrovnávacej pamäte L2 atď..



Obr. 2.5: Graf rýchlosti prístupu k dátam rôznych veľkostí na moderných CPU[8].

2.1.2 Inštrukčný paralelizmus

V posledných rokoch sa v počítačovej vede stále viac spomína pojem paralelizmu. Dnešné procesory architektúry x86 využívajú až štyri úrovne paralelizácie výpočtov. Na najvyššej úrovni hovoríme o návrhu kedy je procesor zložený z niekoľkých autonómne pracujúcich „podprocesorov“ nazývaných *jadrá*. Tie zdieľajú systémové zbernice a pamäť pričom majú väčšinou vlastnú L1 alebo L2 cache. Každé jadro môže spracovávať na sebe nezávislé procesy prípadne jeden proces môže byť rozdelený do takzvaných vlákien, kde každé môže byť spracovávané iným jadrom. V tomto prípade hovoríme o takzvanom *vláknovom paralelizme* s ktorým sa stretávame najmä pri programovaní vo vyšších jazykoch ako sú C, Java, Haskell a iné.

Poznámka 2.1 V našej implementácii testujeme vláknovo paralelizované aj klasické sériové verzie algoritmov. Princíp paralelizovanej verzie sme stručne popísali na konci časti 1.5.

Ďalšie úrovne paralelizácie majú spoločné označenie *inštrukčný paralelizmus*. Pod týmto pojmom rozumieme *superskalárnosť*, *pipelining* a *vektORIZÁciu*. Pri tejto úrovni paralelizmu má programátor len obmedzené možnosti jeho ovplyvnenia, všetku „ťažkú“ prácu obstará prekladač a pri behu aplikácie zasa inštrukčný plánovač v samotnom procesorovom jadre.

Hlavný vplyv na výkon plného a najmä redukovaného algoritmu má superskalárnosť. Vysvetlíme čo tento pojem znamená. Každé z procesorové jadier je tvorené niekoľkými až desiatkami špecializovanými jednotkami ako sú aritmeticko logické jednotky (ALU), bitové posuvníky (Shift), numerické koprocessory (FPU) a podobne. Narozdiel od celých jadier, tieto jednotky vo všeobecnosti nedokážu fungovať samostatne a simultánne spracovávať viacero vlákien alebo procesov². Vedia iba v rámci jedného vlákna, za splnenia určitých podmienok, spracovávať niekoľko inštrukcií jediného procesu (vlákna) naraz. Nutnou podmienkou je napríklad vzájomná nezávislosť niekoľkých po sebe idúcich inštrukcií, teda keď výsledok jednej inštrukcie nezávisí na výsledku predchádzajúcej.

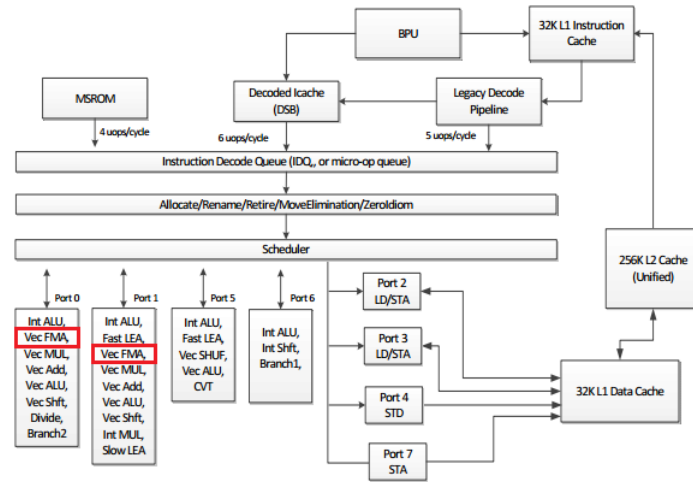
Tu ale paralelizácia nekončí. Aj samotné jednotky totiž dokážu simultánne spracovávať viac než jeden dátový vstup. Na tejto úrovni rozlišujeme medzi vektorizáciou, a pipeliningom. Prvá menovaná technika umožňuje jednu inštrukciu aplikovať na celé vektory, resp. polia. Vektorizáciu vieme použiť ak vykonávame operácie na vektore tak, že výpočet i -teho prvku nezávisí na výsledku výpočtu (napr.) $(i-1)$ -teho prvku. V algoritme 2 pre LU dekompozíciu môžeme vidieť, že toto neplatí. Vektorizácia nás teda nemusí zaujímať.

Pipelining je založený na myšlienke princípu fungovania výpočtových jednotiek podobne ako výrobná linka vo fabrike, kde nový výrobok na linku vstúpi skôr ako je predchádzajúci dokončený [11]. Teda aj väčšina výpočtových jednotiek dokáže s novým výpočtom začať ešte pred dokončením práve prebiehajúcej operácie. Táto technika sa prejavuje najviac ak máme veľký počet operácií rovnakého typu, ktoré sú navyše na sebe nezávislé podobne ako pri vektorizácii. To opäť nie je prípad našich algoritmov takže ho zanedbáme.

Poznámka 2.2 Aby som neuviedol čitateľa v omyl, svoje závery ohľadom vektorizácie a pipeliningu upresním. Procesor pri vykonávaní algoritmu robí mnoho operácií nevýpočtového charakteru, ktoré sú programátorovi skryté a kde sa môžu (nielen) tieto techniky prejaviť. Ale aj po zanedbaní dostávame korešpondujúce výsledky meraní a vypočítaného zrýchlenia, ktoré sa líšia najviac o jednu desatinu. Uvažovaním vplyvu týchto techník pri počítaní zrýchlenia by sa dosiahnutý cieľ a síce vysvetlenie príčin urýchlenia redukovaného algoritmu nezmenil. Teda vzhľadom na množstvo potenciálne investovaného času by zanedbanie týchto techník bolo kontraproduktívne.

Spomenuli sme majoritný vplyv superskalárnosti. Na obrázku 2.6 vidíme schému jedného jadra procesorovej mikroarchitektúry Intel Skylake, ktorá je ku dňu písania práce najmodernejšou bežne dostupnou procesorovou generá-

²Existujú technológie ako napríklad Hyper-threading umožňujúce za určitých podmienok uplatniť vlákňový paralelizmus aj v rámci jediného superskalárneho jadra.



Obr. 2.6: Schéma výpočtového jadra mikroarchitektúry Intel Skylake [9].

ciou a na ktorej testujeme zrýchlenie redukovaného algoritmu. Jadrá (nielen) tejto architektúry sú vybavené dvomi sčítačkami čísiel s plávajúcou desatinnou čiarkou, ktoré sú na obrázku zvýraznené červenou farbou. Pointa superskalárnosti v tomto prípade spočíva v možnosti výpočty výrazov typu $a_0 \circ a_1 \circ \dots \circ a_n$, kde $\circ \in \{+, \cdot\}$ rozložiť medzi dve jednotky a dosiahnuť dvojnásobné zrýchlenie. Toto má značný vplyv na rýchlosť redukovaného algoritmu najmä pri počítaní pravých strán trojdiagonálnych rovníc, ktoré sú už na prvý pohľad zložitejšie ako v prípade plného algoritmu. Keďže procesor má ale iba jednu deličku, pre výrazy typu $a_0/a_1/\dots/a_n$ takýto trik fungovať nebude.

Procesorové jadro na obrázku vyššie má svojich 29 jednotiek prístupných cez osem zberníc (porty 0 až 7). To konkrétne v prípade procesorov Skylake znamená, že dokážu naraz využívať maximálne osem svojich výpočtových jednotiek za splnenia určitých podmienok. Málokedy ale proces alebo vlákno obsahuje vždy osmice vzájomne nezávislých inštrukcií. V situácií kedy jeden proces (vlákno) nemôže využiť všetky porty jadra, niektoré procesory umožňujú nevyužívané porty priradiť inému procesu (vláknu). Táto technika sa nazýva *Simultaneous Multi-Threading* (ďalej len SMT). V literatúre sa, v prípade procesoroch Intelu, často môžeme stretnúť s názvom *Hyper-Threading*. Procesory AMD majú implementovaný odlišný variant tejto techniky zvaný *Clustered Multi-Threading*. Tu namiesto klasických jadier je procesor vybavený tzv. modulmi, kde každý modul má dve samostatné „oklieštené“ jadrá

schopné spracovávať iba celočíselné operácie a jedinú FPU jednotku spoločnú pre oba jadrá. V praxi teda SMT predstavuje akýsi medzistupeň medzi vláknovým a inštrukčným paralelizmom, kedy procesor operačnému systému a bežiacim procesom hlási väčší počet jadier akým je v skutočnosti vybavený, pričom interne na úrovni mikroarchitektúry sa jedná iba o ďalšiu techniku superskalárnosti.

SMT má výrazne pozitívny vplyv na výkon procesora ak procesy (vlákna) zdieľajúce jedno jadro nepožadujú prístup k tým istým jednotkám. V opačnom prípade dochádza k javu kedy jeden proces (vlákno) musí čakať na uvoľnenie jednotky druhým procesom (vláknom) a vtedy môže dôjsť k zníženiu výkonu oproti identickému procesoru s neaktívnym SMT. Trocha predbehnime a priamo uveďme, že v praktických meraniach má SMT na vláknovo paralelnú verziu redukovaného algoritmu negatívny vplyv, čo si vysvetlíme v časti 2.3.

2.1.3 Rýchlosť aritmetických operácií

Procesory majú mnoho aritmetických jednotiek špecializovaných na určitý typ operácie. Je prirodzené predpokladať, že tieto jednotky budú pracovať navzájom rozličnými rýchlosťami. V prípade výpočtových algoritmov má vplyv na rýchlosť, pochopiteľne okrem výberu vhodných dátových štruktúr, najmä doba vykonania základných matematických operácií a síce sčítania, odčítania, násobenia a delenia.

Najlepším zdrojom ako zistiť rýchlosť týchto operácií je dokumentácia inštrukčných sád procesorov. Inštrukčná sada x86 sa od svojho prvotného uvedenia v roku 1978 dočkala mnohých rozšírení. Moderné procesory obsahujú niekoľko spôsobov ako napríklad vynásobiť dve čísla. Keďže nie je v silách otestovať všetky možné rozšírenia sady, zvolili sme si jedno konkrétne rozšírenie a síce Streaming SIMD Extensions (skrátene SSE), konkrétne vo verzii 4 (SSE4). Sady z rodiny SSE sú v čase písania práce najpoužívanejšími sadami (najmä verzia SSE2) ktoré sú podporované prakticky všetkými procesormi od roku 2003.

V súčasnosti už existuje modernejšia náhrada tejto sady zvaná Advanced Vector Extensions (skratka AVX), ktorej hlavný prínos spočíva vo vylepšených vektorových operáciach. Vektorové operácie však vyžadujú aby výpočty využívajúce jednotlivé prvky vektora boli navzájom nezávislé. To nie je prípad ani jedného z testovaných algoritmov. Nakonfigurovaním prekladača na generovanie týchto inštrukcií by sme iba znemožnili beh aplikácie na polovici z testovaných procesorov. Praktické zrýchlenie sme testovali na šiestich procesoroch architektúry x86, pričom sme pokryli väčšinu mikroarchitektúr v rozmedzí rokov 2007 až 2015.

V nasledujúcej tabuľke 2.1 uvidíme rýchlosti štyroch základných matematických operácií v rámci šiestich testovaných mikroarchitektúr. Tabuľka obašhuje tieto stĺpce.

- **Architektúra (rok)**

Testovaná mikroarchitektúra a rok jej uvedenia na trh. Architektúry sú zoradené abecedne podľa výrobcu a následne podľa roku vydania.

- **Odozva**

Počet strojových cyklov potrebných na vykonanie inštrukcie.

- **Inverzný prietok**

Počet strojových cyklov ktoré je nutné čakať kým je daná výpočtová jednotka schopná zopakovať inštrukciu. V prípade operácií sčítania, odčítania a násobenia je tento počet menší ako odozva. To znamená, že aritmetické sčítačky a násobičky sú schopné, vďaka technike pipelining-u, začať nový výpočet ešte pred dokončením aktuálneho výpočtu.

Podľa tabuľky vidno, že operácie sčítania a odčítania sú rovnako rýchle čo sa pochopiteľne dalo očakávať. Tieto dve operácie preto budeme spoločne označovať symbolom \pm . Od tejto chvíle ak spomenieme operáciu sčítania tak tým budeme súčasne myslieť aj operáciu odčítania. Ako vidieť zďaleka najpomalšie je práve delenie.

Architektúra (rok)	Odozva				Inverzný prietok			
	+	-	×	÷	+	-	×	÷
AMD Piledriver (2012)	5-6	5-6	5-6	9-27	0,5	0,5	0,5	5-10
Intel Penryn (2007)	3	3	5	6-21 ²	1	1	1	5-20 ²
Intel Nehalem (2008)	3	3	5	7-22	1	1	1	7-22
Intel Sandy Bridge (2011)	3	3	5	10-22	1	1	1	10-22
Intel Haswell (2013)	3	3	5	10-20	1	1	0,5	8-14
Intel Skylake (2015)	4	4	4	13-14	0,5	0,5	0,5	4

Tabuľka 2.1: Tabuľka aritmetických operácií na rôznych mikroarchitektúrach podľa [10]. Údaje predstavujú počet strojových cyklov.

Pre zaujímavosť si v tabuľke 2.2 nižšie ukážme praktické výsledky z testovacej aplikácie. Operácie sme merali o na 512 prvkovom poli, pričom aby sme dostali „rozumne“ dlhé časy výpočty boli opakované 500000 krát. Operácie

²Odozva je nadobúda menšie hodnoty ak je deliteľ celý.

boli v tvare $a[i] = a[i] \circ a[i-1]$, kde $\circ \in \{+, \cdot, \div\}$. Ďalej budeme merať aj pamäťovú operáciu kopírovania, pričom rozlišujeme medzi kopírovaním spojitých (\succ) a nespojitých (\therefore) dát. Kopírovanie v implementácii algoritmov používame pri zapísaní vypočítaných derivácií z LU dekompozície do výslednej matice, ktorá obsahuje derivácie prislúchajúce k uzlom.

Označenie 2.3 Označme dve triedy dátových štruktúr symbolmi:

- \succ označuje spojitú dátovú štruktúru ako napríklad vektory.
- \therefore označuje nespojitú dátovú štruktúru ako napríklad spájané zoznamy alebo vektory referencií na konkrétne dáta. Taktiež označuje aj spojitú vektory ak operácia kopírovania z jedného vektora do druhého nie je vykonávaná spojitou (napr. kopírujeme iba prvky na nepárnych indexoch).

V tabuľke 2.2 sú namiesto mikroarchitektúr uvedené konkrétne modely procesorov, pričom ich poradie zodpovedá poradiu z minulej tabuľky.

Procesor	\pm	\times	\div	\succ	\therefore
AMD FX-6300	247	236	445	10,3	204
Intel Core 2 Duo E8200	204	261	417	7,66	158
Intel Core i5 650	857	927	928	9,1	130,9
Intel Core i3 2350M	261	362	894	11,5	175
Intel Core i5 4440	175	243	573	5,41	83,8
Intel Core i7 6700K	77	80	147	3,5	68

Tabuľka 2.2: Rýchlosť aritmetických operácií na konkrétnych CPU. Údaje sú v milisekundách.

V praxi vidno, že sčítavanie a násobenie môžeme považovať za podobne rýchle operácie. Je nutné podotknúť, že pomery rýchlosti operácií sú závislé od povahy testovania. My sme doby trvania aritmetických operandov $\circ \in \{+, \cdot, \div\}$ merali na jednom vektore, pričom testy boli v tvare $a[i] = a[i] \circ a[i-1]$, teda výpočet i -teho prvku závisel od výsledku výpočtu predchádzajúceho $(i-1)$ -teho prvku. Týmto sme sa snažili povahu testu čo najviac napodobniť tvaru výpočtov v interpolačných algoritmoch.

Poznámka 2.4 V prípade, ak by test bol v tvare $a[i] = b[i] \circ c[i]$, tak výpočty pre jednotlivé i by boli na sebe nezávislé. To by procesoru umožnilo použiť techniku pipeliningu, pričom pomery časov násobenia a sčítania by

vyšli, vďaka väčšiemu prietoku týchto dvoch operácií oproti deleniu, násobne väčšie. Jednak by takýto test nesúhlasil s tvarom výpočtov v testovaných algoritmoch a taktiež by nekorešpondovalo teoretické zrýchlenie s meraným zrýchlením.

Ešte ostáva ukázať reálny vplyv inštrukčného paralelizmu. V nasledujúcej tabuľke sú uvedené namerané násobky časov trvania vyhodnotenia výrazu vzhľadom na počet operácií pre procesorovú architektúru Intel Skylake.

Procesor	2	3	4	5	6	7	8	9	10
\pm	1,61	2,15	2,59	3,03	3,43	3,85	4,25	4,68	5,11
\times	1,61	2,16	2,6	3,04	3,45	3,86	4,27	4,69	5,11
\div	2	3	4,01	5,01	6,02	7,02	8,03	9,03	10,04

Tabuľka 2.3: Rýchlosť aritmetických operácií na konkrétnych CPU. Údaje sú v milisekundách.

Podľa tabuľky 2.3 môžeme povedať, že ak t je doba vyhodnotenia výrazu s jedinou operáciou sčítania alebo násobenia tak vďaka inštrukčnému paralelizmu doba vyhodnotenia výrazu s n operáciami konverguje k $\frac{1}{2} \cdot tn$. To neplatí pre delenie, kde doba vyhodnotenia rastie lineárne s rastúcim počtom operácií.

Poznámka 2.5 Keďže každá testovaná mikroarchitektúra má v jadre práve po dva numerické koprocesory (FPU) tak tento jav platí aj pre ne. Napríklad pre mikroarchitektúru Intel Sandy Bridge je vyhodnotenie výrazu s desiatimi operáciami sčítania 5,1 násobne pomalšie ako vyhodnotenie výrazu s jednou operáciou. Analogicky pre násobenie vychádza tento pomer na 5,93 a pre delenie pomer vychádza na 9,99. V čase písania práce sme nestihli podrobne otestovať aj ostatné mikroarchitektúry.

Poznámka 2.6 V nasledujúcich častiach budeme pre pomer rýchlosti delenia a sčítania namiesto údajov z dokumentácie uvažovať nami namerané výsledky, ktoré presnejšie korešpondujú s meraným aj teoretickým zrýchlením redukovaného algoritmu.

2.2 Teoretické zrýchlenie

V tejto časti si spočítame počty operácií procedúr tvoriacich plný a redukovaný algoritmus. Na základe poznatkov z časti o procesorovej architektúre si

formálne zdefinujeme pojmy pre počty a ceny aritmetických operácií ktoré budeme v nasledujúcich častiach.

Označenie 2.7 Množinu všetkých matematických výrazov budeme označujeme symbolom \mathbb{V} .

Zdefinujeme si funkciu o ktorá vráti počet aritmetických operácií sčítania.

Definícia 2.8 Nech \mathcal{V} je matematický výraz obsahujúci p^\pm sčítaní a odčítaní, p^\times násobení, p^\div delení, p^\rightarrow spojitých kopírovaní a p^\cdot nespojitých kopírovaní. Definujeme funkciu $o : \mathbb{V} \rightarrow \mathbb{N}^5$ vzťahom

$$o(\mathcal{V}) = \langle p^\pm, p^\times, p^\div, p^\rightarrow, p^\cdot \rangle.$$

Funkciu o budeme nazývať *počet operácií výrazu \mathcal{V}* .

Označenie 2.9 Podľa časti 2.1.2 o inštrukčnom paralelizme položíme dve premenné, ktoré nám pomôžu definovať ceny aritmetických operácií.

- Hodnota β značí *faktor inštrukčného paralelizmu* operácií s plávajúcou desatinnou čiarkou. Inými slovami hodnota β značí počet jednotiek jadra procesora schopných počítať desatinné čísla.
- Hodnota γ^\div značí *pomer odozvy delenia a sčítania* pri operáciách s plávajúcou desatinnou čiarkou. Inými slovami hodnota γ^\div značí koľkokrát je delenie pomalšie oproti sčítaniu.
- Hodnota γ^\rightarrow značí *pomer rýchlosti kopírovania a násobenia* pri operáciách s plávajúcou desatinnou čiarkou a spojitých dátach.
- Hodnota γ^\cdot značí *pomer rýchlosti kopírovania a násobenia* pri operáciách s plávajúcou desatinnou čiarkou a nespojitých dátach.

Poznámka 2.10 Pre moderné procesory architektúry x86 budeme podľa tabuľky 2.3 uvažovať $\beta = 2$. Ďalej podľa tabuľky 2.2 budeme súhrne uvažovať $\gamma^\div = 3$, $\gamma^\rightarrow = 1/25$ a $\gamma^\cdot = 1$.

Poznámka 2.11 Faktor inštrukčného paralelizmu β budeme uvažovať iba pre operácie sčítania a násobenia. Žiadny bežne dostupný procesor totiž nedokáže inštrukčne paralelizovať delenie.

Teraz si definujeme funkciu c ktorá vráti cenu aritmetických operácií pre nejakú matematickú operáciu berúc do úvahy hodnoty β a γ^\div .

Definícia 2.12 Nech p^\pm je počet sčítaní a odčítaní, p^\times počet násobení, p^\div počet delení, $p^>$ počet spojitých kopírovaní a $p^{\dot{>}}$ počet nespojitých kopírovaní. Definujme funkcie cien operácií

- $c : \mathbb{N}^5 \rightarrow \mathbb{N}^5$ vzťahom

$$c(\langle p^\pm, p^\times, p^\div, p^>, p^{\dot{>}} \rangle) = \left\langle \left\lceil \frac{p^\pm}{\beta} \right\rceil, \left\lceil \frac{p^\times}{\beta} \right\rceil, \lceil \gamma^\div p^\div \rceil, \lceil \gamma^> p^> \rceil, \lceil \gamma^{\dot{>}} p^{\dot{>}} \rceil \right\rangle,$$

ktorú nazveme *cena operácií*.

- Nech V je matematický výraz a $o(V) = \langle p^\pm, p^\times, p^\div, p^>, p^{\dot{>}} \rangle$. Potom *cena operácií výrazu V* je funkcia $c : \mathbb{V} \rightarrow \mathbb{N}^5 \times \mathbb{N}$ v tvare

$$c(\mathcal{V}) = c(o(\mathcal{V})).$$

V kontexte počítania operácií a cien budeme z dôvodu zjednodušenia procedúry považovať za množiny matematických výrazov, pričom algoritmy budeme považovať za množiny obsahujúce procedúry a výrazy. Stručne doplníme počty aj pre procedúry, resp. algoritmy.

Označenie 2.13 Nech \mathcal{P} je procedúra. Funkciu

$$o(\mathcal{P}) = \sum_{v \in \mathcal{P}} o(v)$$

nazveme *počet operácií procedúry \mathcal{P}* .

Analogicky označme aj ceny procedúr, resp. algoritmov.

Označenie 2.14 Nech \mathcal{P} je procedúra. Funkciu

$$c(\mathcal{P}) = \sum_{v \in \mathcal{P}} c(v)$$

nazveme *cena operácií procedúry \mathcal{P}* .

Dôvod, prečo je náš redukovaný algoritmus rýchlejší je práve fakt, že pri ňom dochádza k značne menšiemu počtu delení, ktoré je oproti ostatným trom operáciám výrazne pomalšie. Navyše v prípade redukovaného algoritmu sa prejavuje superskalárnosť procesora (najmä) pri príprave pravých strán rovníc pre LU dekompozíciu ako podľa 1.16, 1.18, 1.20, 1.21, 1.22 v časti 1.5 o počítaní derivácií redukovaným spôsobom. V ďalšej sekcii sy spočítame jednotlivé operácie a vypočítame teoretické zrýchlenie redukovaného algoritmu. To budeme počítat tak, že si zadefinujeme „procedúry“ ktoré predstavujú jednotlivé časti algoritmu podľa častí 1.3 o trojdiagonálnej LU dekompozícii, 1.4 o plnom algoritme a 1.5 o redukovanom algoritme.

Najprv položíme procedúry predstavujúce LU dekompozíciu spoločné pre oba postupy.

- Procedúra *InicalizujLU* inicializuje hodnoty pravej strany sústavy rovníc r_0, \dots, r_{K-1} a hodnotu b z LU dekompozície podľa rovnice 1.15 v prípade plného algoritmu, respektíve podľa rovnice 1.26 v prípade redukovaného algoritmu.
- Procedúra *VyriešLU* vypočíta sústavu rovníc na základe hodnôt pravej strany poskytnutými *InicalizujLU*.

Každý algoritmus inicializuje pravé strany inak, takže pre oba si ceny uvedieme osobitne.

2.2.1 Cena plného algoritmu

Ceny základných aritmetických a pamäťových operácií pre vyššie uvedené procedúry môžeme zhrnúť do tabuľky, pričom uvažujeme počty v našej ukážkovej implementácii³. Riadky tabuľky predstavujú jednotlivé procedúry a stĺpce udávajú počet vykonaných operácií, kde K je počet neznámych. Podotýkam, že operáciu odčítania budeme uvažovať ako sčítanie. Operácia \succ predstavuje spojitú kopírovanie vypočítaných derivácií do výslednej matice uzlov. Pri meraní sa ukázalo, že kopírovanie spojitých dát vieme má cenu 0,04 teda kopírovanie je 25 násobne rýchlejšie ako sčítanie alebo násobenie. V tabuľkách budeme pre sčítanie, násobenie a delenie uvádzať ich ceny podľa definície 2.12.

Procedúra	\pm	\times	\div	\succ
<i>InicalizujLU</i>	K	K	0	0
<i>VyriešLU</i>	$3K$	$2K$	$\gamma^{\div} K$	$\gamma^{\succ} K$

Tabuľka 2.4: Ceny operácií LU dekompozície pre plný algoritmus vzhľadom na počet neznámych K .

Procedúry *VyriešLU* a *InicalizujLU* neobsahujú matematické výrazy s viac ako jedným sčítaním prípadne násobením. Superskalárnosť procesora sa teda v prípade plného algoritmu neprejaví. Uvažujme procedúry predstavujúce implementáciu plného algoritmu podľa časti 1.4:

- Procedúra *VypočítajDx* vypočíta parciálne derivácie d^x pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.11. Jedno

³Repozitár so zdrojovými kódmi k aplikáciám možno nájsť na adrese <https://github.com/vildibald/VKDiplom-master>

volanie *VyriešLU* vypočíta derivácie d^x pre jeden stĺpec. Teda procedúra musí počítat LU pre každý stĺpec, ktorých je J .

- Procedúra *VypočítajDxy* vypočíta parciálne derivácie d^{xy} pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.12. Jedno volanie *VyriešLU* vypočíta derivácie d^{xy} pre jeden stĺpec. Procedúra musí počítat LU pre prvý a posledný stĺpec.
- Procedúra *VypočítajDy* vypočíta parciálne derivácie d^y pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.13. Jedno volanie *VyriešLU* vypočíta derivácie d^y pre jeden riadok. Teda procedúra musí počítat LU pre každý riadok, ktorých je I .
- Procedúra *VypočítajDyx* vypočíta parciálne derivácie d^{yx} pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.14. Jedno volanie *VyriešLU* vypočíta derivácie d^{yx} pre jeden riadok. Teda procedúra musí počítat LU pre každý riadok, ktorých je I .
- Procedúra *VypočítajPlný* vypočíta na základe vstupných hodnôt pre de Boorovu interpoláciu 1.4 zavolaním procedúr *VypočítajDx*, *VypočítajDxy*, *VypočítajDy* a *VypočítajDyx*.

Nech I značí počet uzlov na osi x a J značí počet uzlov na osi y . Všetky ceny sú v tvare $a \cdot IJ + b \cdot I + c \cdot J + d$. Pre zjednodušenie budeme ceny, tam kde je koeficient a nenulový, uvádzať v tvare $a \cdot IJ$. Keďže ceny algoritmov rastú kvadraticky tak pre veľké I a J bude odchýlka zanedbateľná. Pre plný algoritmus teda dostaneme tieto počty

Procedúra	\pm	\times	\div	\succ
<i>VypočítajDx</i>	$4IJ$	$3IJ$	γ^+IJ	γ^-IJ
<i>VypočítajDxy</i>	$8I$	$6I$	$2\gamma^+I$	γ^-I
<i>VypočítajDy</i>	$4IJ$	$3IJ$	γ^+IJ	γ^-IJ
<i>VypočítajDyx</i>	$4IJ$	$3IJ$	γ^+IJ	γ^-IJ

Tabuľka 2.5: Ceny operácií plného algoritmu.

Sčítaním všetkých cien operácií v tabuľke dostaneme nasledujúci výsledok.

Lemma 2.15 *Nech I a J označujú počty uzlov na osiach x a y . Potom sumárna cena operácií plného algoritmu je*

- $12IJ$ sčítaní,
- $9IJ$ násobení,
- $3\gamma^\div IJ$ delení,
- $3\gamma^\cdot IJ$ kopírovaní.

Cena plného algoritmu teda je

$$21IJ + 3\gamma^\div IJ + 3\gamma^\cdot IJ.$$

2.2.2 Cena redukovaného algoritmu

Redukovaný spôsob na počítanie derivácií využíva iný tvar trojdiagonálnej sústavy rovníc. Zmeníme teda počty operácií procedúr *VyriešLU* a *InicializujLU*. Ďalej položíme dve nové pomocné procedúry *InicializujZmiešLU*, ktorá inicializuje hodnoty r_0, \dots, r_{K-1} a b z LU dekompozície 1.8 pre zmiešané derivácie d^{xy} a k nej príslušnú *VyriešZmiešLU*.

V prípade redukovaného algoritmu nie je kopírovanie výsledku LU dekompozície spojité. Konkrétne to znamená, že potrebujeme do vektora dĺžky napríklad I na každú nepárnu pozíciu uložiť hodnotu z vektora $I/2$, pretože LU nám nájde iba polovicu hľadaných derivácií. Cena kopírovania bude podľa reálneho merania v závislosti od konkrétneho systému približne 0,7. Aby sme takéto kopírovanie vizuálne odlíšili od kopírovania v predchádzajúcej časti, budeme túto operáciu označovať symbolom \div .

Procedúra	\pm	\times	\div	\div
<i>InicializujLU</i>	$3/\beta K$	$2/\beta K$	0	0
<i>VyriešLU</i>	$3K$	$2K$	$\gamma^\div K$	$\gamma^\cdot K$

Tabuľka 2.6: Ceny operácií LU dekompozície pre redukovaný algoritmus vzhľadom na počet neznámych K .

V prípade inicializovania pravej strany pre LU dekompozíciu v procedúre *InicializujLU* dochádza k viacerým sčítaniam a/alebo násobeniam v rámci jedného výrazu. Tu sa následne prejaví superskalárnosť výpočtových jednotiek procesorového jadra podľa časti 2.1.2. Moderné procesory architektúry x86 obsahujú práve dve jednotky pre výpočty s pohyblivou desatinnou čiarokou čo znamená faktor inštrukčného paralelizmu $\beta = 2$. To implikuje zaujímavý dôsledok, kedy algoritmus s väčším počtom matematických operácií

je v praxi rýchlejší ako algoritmus s menším celkovým počtom operácií, ale s viacerými na sebe závislými výrazmi (t.j. prípad, keď výraz b musí byť vyhodnotený až po výraze a).

V prípade zmiešaných derivácií používame odlišné pravé strany rovníc pre LU dekompozíciu. Osobitne si v tabuľke 2.7 spočítajme ceny aj pre tieto procedúry. Následne analogicky ako v predchádzajúcej sekcii položíme procedúry

Procedúra	\pm	\times	\div	$:\cdot$
<i>InicalizujZmiešLU</i>	$^{33}/_2K$	$^{17}/_2K$	0	0
<i>VyriešZmiešLU</i>	$3K$	$2K$	$\gamma^{\div}K$	$\gamma^{\cdot}K$

Tabuľka 2.7: Ceny operácií LU dekompozície pre zmiešané zostatkové derivácie.

predstavujúce implementáciu redukovaného algoritmu podľa časti 1.5::

- Procedúra *VypočítajDxResty* vypočíta zvyšné parciálne derivácie d^x podľa 1.17.
- Procedúra *VypočítajDx* vypočíta parciálne derivácie d^x pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.16. Jedno volanie *VyriešLU* vypočíta derivácie na párnych riadkoch d^x pre jeden stĺpec. Teda procedúra musí počítať LU pre každý stĺpec, ktorých je J . Zvyšné derivácie dopočítame procedúrou *VypočítajResty*.
- Procedúra *VypočítajDyResty* vypočíta zvyšné parciálne derivácie d^y podľa 1.19.
- Procedúra *VypočítajDy* vypočíta parciálne derivácie d^y pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.18. Jedno volanie *VyriešLU* vypočíta derivácie na párnych stĺpcoch d^y pre jeden riadok. Teda procedúra musí počítať LU pre každý riadok, ktorých je I . Zvyšné derivácie dopočítame procedúrou *VypočítajResty*.
- Procedúra *VypočítajDxy* vypočíta parciálne derivácie d^{xy} pomocou procedúr *InicalizujLU* a *VyriešLU*. Vstupné hodnoty vezmeme z 1.20 a 1.21. Jedno volanie *VyriešLU* vypočíta na párnych riadkoch(stĺpcoch) derivácie d^{xy} pre jeden stĺpec(riadok). Procedúra musí počítať LU pre prvý a posledný stĺpec a pre prvý a posledný riadok.
- Procedúra *VypočítajZmiešResty* vypočíta zvyšné parciálne derivácie d^{xy} podľa 1.23, 1.24 a 1.25.

- Procedúra *VypočítajDyx* vypočíta parciálne derivácie d^{xy} pomocou procedúr *InicalizujZmiešLU* a *VyriešZmiešLU*. Vstupné hodnoty vezmeme z 1.22, pričom na ich inicializovanie do LU dekompozície použijeme práve procedúru *InicalizujZmiešLU*. Jedno volanie *VyriešLU* vypočíta na párnych stĺpcoch d^{xy} pre jeden riadok. Procedúra musí počítať LU pre každý páry riadok, ktorých je I . Zvyšné derivácie dopočítame procedúrou *VypočítajZmiešResty*.
- Procedúra *VypočítajRedukovaný* vypočíta na základe vstupných hodnôt pre de Boorovu interpoláciu 1.4 postupným vykonaním predchádzajúcich siedmich procedúr.

Opäť pre zjednodušenie budeme počty v tvare $a \cdot IJ + b \cdot I + c \cdot J + d$ zanedbávať na $a \cdot IJ$. Ak hodnota I značí počet uzlov na osi x a hodnota J značí počet uzlov na osi y tak pre redukovaný algoritmus dostaneme počty v tabuľkách 2.8 a 2.9.

Procedúra	\pm	\times	\div	\div
<i>VypočítajDx</i>	$\frac{3}{2}IJ + \frac{3}{2\beta}IJ$	$1IJ + \frac{1}{\beta}IJ$	$\frac{1}{2}\gamma^{\dagger}IJ$	$\frac{1}{2}\gamma^{\dagger\dagger}IJ$
<i>VypočítajDy</i>	$\frac{3}{2}IJ + \frac{3}{2\beta}IJ$	$1IJ + \frac{1}{\beta}IJ$	$\frac{1}{2}\gamma^{\dagger}IJ$	$\frac{1}{2}\gamma^{\dagger\dagger}IJ$
<i>VypočítajDxy</i>	$8I + 8J$	$6I + 6J$	$2I + 2J$	$\gamma^{\dagger}I + \gamma^{\dagger}J$
<i>VypočítajDyx</i>	$\frac{3}{4}IJ + \frac{3}{4\beta}IJ$	$\frac{1}{2}IJ + \frac{1}{4\beta}IJ$	$\frac{1}{4}\gamma^{\dagger}IJ$	$\frac{1}{2}\gamma^{\dagger\dagger}$

Tabuľka 2.8: Ceny operácií pre derivácie počítané trojdiagonálnou sústavou.

Procedúra	\pm	\times	\div	$\dot{\cdot}$
<i>VypočítajDxResty</i>	$3/_{2\beta}IJ$	$1/_{\beta}IJ$	0	0
<i>VypočítajDyResty</i>	$3/_{2\beta}IJ$	$1/_{\beta}IJ$	0	0
<i>VypočítajZmiešResty</i>	$17/_{4\beta}IJ$	$7/_{4\beta}IJ$	0	0

Tabuľka 2.9: Ceny operácií pre zmiešané zostatkové derivácie.

Sčítaním operácií v predchádzajúcich dvoch tabuľkách dostaneme následujúci výsledok.

Lemma 2.16 *Nech I a J označujú počty uzlov na osiach x a y . Potom sumárna cena redukovaného algoritmu je*

- $15/_{4}IJ + 37/_{2\beta}IJ$ sčítaní,
- $5/_{2}IJ + 10/_{\beta}IJ$ násobení,
- $5/_{4}\gamma^{\div}IJ$ delení,
- $3/_{2}\gamma^{\dot{\cdot}}IJ$ kopírovaní.

Cena redukovaného algoritmu teda je

$$25/_{4}IJ + 57/_{2\beta}IJ + 5/_{4} \cdot \gamma^{\div}IJ + 3/_{2}\gamma^{\dot{\cdot}}IJ.$$

2.2.3 Pamäťové nároky

Redukovaný algoritmus by nemal byť lepší len v počte aritmetických operácií, ale taktiež priniesť menšie pamäťové nároky vyplývajúce z polovičnej veľkosti sústav rovníc riešených LU dekompozíciou. Uvažujme maticu uzlov veľkosti I a J . Potrebujeme si pamätať $I \cdot J$ funkčných hodnôt v uzloch, derivácií podľa x , y a zmiešaných derivácií. Samotné uzly si pamätať nemusíme lebo uvažujeme len uniformné splajny. Teda stačí si nám uložiť iba uzly $u_0, u_{I-1}, v_0, v_{J-1}$. Zvyšné uzly si vieme vypočítať vzorcom $u_i = u_0 + |u_{I-1} - u_0| \cdot \frac{i}{I}$ pre ľubovoľné i z $\{1, \dots, I-2\}$, respektíve $v_j = v_0 + |v_{J-1} - v_0| \cdot \frac{j}{J}$ pre ľubovoľné j z $\{1, \dots, J-2\}$. Potrebujeme si teda pamätať aspoň $4 \cdot I \cdot J + 4$ čísel pre funkčné hodnoty uzlov.

Počítanie všeobecnej LU dekompozície podľa algoritmu 1 potrebuje päť vektorov, kde tri predstavujú trojdiagonálu, jeden pravú stranu, v ktorom na konci bude výsledok. Nám ale stačí uvažovať iba procedúru pre náš špeciálny tvar trojdiagonálnej sústavy podľa algoritmu 2. Pri použití tejto priestorovo optimalizovanej varianty výpočtu LU dekompozície potrebujeme na vypočítanie $I \cdot J$ uzlov práve $I \cdot J + 2I + 2J$ pamäte.

2.2.4 Zhrnutie

Po spočítaní cien jednotlivých matematických a pamäťových operácií v častiach 2.2.1 a 2.2.2 pristúpime k formulácii zrýchlenia.

Veta 2.17 *Očakávané zrýchlenie redukovaného algoritmu je*

$$\frac{21 + 3\gamma^{\div} + 3\gamma^{\cdot}}{25/4 + 57/2\beta + 5/4 \cdot \gamma^{\div} + 3/2\gamma^{\cdot}}$$

Dôkaz.

Uvažujme mriežku uzlov veľkosti I a J . V lemach 2.15 a 2.16 dostávame ceny

- $21IJ + 3\gamma^{\div}IJ + 3\gamma^{\cdot}IJ$ pre plný algoritmus,
- $25/4IJ + 57/2\beta IJ + 5/4 \cdot \gamma^{\div}IJ + 3/2\gamma^{\cdot}IJ$ pre redukovaný algoritmus.

Zrýchlenie vyjadríme ako podiel cien oboch algoritmov.

$$\frac{21IJ + 3\gamma^{\div}IJ + 3\gamma^{\cdot}IJ}{25/4IJ + 57/2\beta IJ + 5/4 \cdot \gamma^{\div}IJ + 3/2\gamma^{\cdot}IJ}.$$

Ceny algoritmov sme počítali asymptoticky, t.j. cenu v tvare $a \cdot IJ + b \cdot I + c \cdot J + d$ sme redukovali na tvar $a \cdot IJ$. Po vykrátení IJ z výrazu vyššie dostaneme očakávané zrýchlenie redukovaného algoritmu.

□

Ukážme si dva príklady zrýchlenia aby sme názorne videli vplyv inštrukčného paralelizmu na rýchlosť redukovaného algoritmu. V prvom prípade predpokladajme porovnanie algoritmov na modernom procesore, pričom v druhom uvažujme primitívny procesor.

Príklad 2.18 Nech $\beta = 2$, teda procesorové jadro má práve dve jednotky pre výpočty s pohyblivou desatinnou čiarkou a $\gamma^{\div} = 3$. Ďalej nech $\gamma^{\cdot} = 1/25$ a $\gamma^{\div} = 3/5$. Analogicky ako v predchádzajúcom prípade dosadením do vzťahu dostaneme

$$\frac{30,12}{25,15} \approx 1,2.$$

V prípade inštrukčného paralelizmu s faktorom 2 je redukovaný algoritmus o 20% rýchlejší ako plný.

Príklad 2.19 Nech $\beta = 1$, teda procesorové jadro má iba jednu jednotku pre výpočty s pohyblivou desatinnou čiarkou a $\gamma^{\div} = 3$, to jest operácia delenia je trojnásobne pomalšia ako operácia sčítania podľa tabuliek 2.1 a 2.2

operácií v časti 2.1.3. Ďalej nech $\gamma^r = 1/25$ a $\gamma^i = 1$. Dosadením do vzťahu z predchádzajúcej vety dostaneme

$$\frac{30,12}{40} \approx 0,75.$$

Pomer je menší ako 1 čo znamená, že bez inštrukčného paralelizmu ($\beta = 1$) je redukovaný algoritmus pomalší ako plný.

Podľa príkladov vidíme, že rýchlosť redukovaného algoritmu je závislá na schopnosti hardvérovej architektúry procesora paralelizovať vyhodnotenia aritmetických výrazov s viacerými operandami. Je otázne nakoľko by s rastúcim faktorom inštrukčného paralelizmu β rástlo zrýchlenie. Keďže nemáme k dispozícii procesor s vhodnou hardvérovou výbavou, ktorý by mal viac ako dve jednotky pre výpočty s pohyblivou desatinnou čiarkou (ak vôbec taký existuje) je náročné predpokladať ako by si takýto stroj s redukovaným postupom poradil.

V nasledujúcej časti si ukážeme reálne výsledky a uvidíme či s nimi vypočítané zrýchlenie skutočne súhlasí.

2.3 Merané zrýchlenie

V kapitole Zrýchlenie sme určili teoretické zrýchlenie počítania uzlov, dosiahnuteľné zredukovaním veľkosti trojdiagonálnych sústav. V tejto časti si ukážeme reálne výsledky dosiahnuté v ukážkovej implementácii.

Ako bolo spomenuté, testovací program bol implementovaný v jazyku C++. Softvér obsahuje testy pre sekvenčné aj paralelné počítanie derivácií pre oba predmetné algoritmy. Použitý prekladač bol Intel C++ Compiler v 64 bitovej verzii nastavený na generovanie agresívne optimalizovaného binárneho kódu (-O2), pričom vláknovo paralelné verzie algoritmov boli implementované pomocou rozhrania OpenMP.

Testy boli vykonané na šiestich rôznych počítačových zostavách, všetky so systémom Windows 7 a 10. Testovacie stroje obsahujú rozličné multivláknové procesory od starého Penryn z roku 2007 až po najmodernejší Skylake z roku 2015 so vzájomne odlišnými architektúrami a hlavne spôsobmi vykonávania paralelizovaných procesov.

Stĺpec 1 obsahuje modely procesorov zoradených podľa mikroarchitektúry ako v tabuľke 2.1. Vedľa názvu modelu CPU sa nachádza údaj tvaru nC/mT , kde n je počet fyzických jadier procesora a m počet logických jadier. V prípade $m > n$ je procesor vybavený určitou formou SMT. Stĺpce 2 a 3 predstavujú časy behov sériovej verzie pre povrchové splajny. Ostávajúce stĺpce analogicky znázorňujú časy a pomer paralelných verzií implementovaných podľa myšlienky uvedenej na konci časti 1.5.

Procesor	Sériovo		Paralelne	
	Plný	Redukovaný	Plný	Redukovaný
FX-6300 _{3C/6T}	81	72	21	17
C2D E8200 _{2C/2T}	99	87	51	46
Ci5 650 _{2C/4T}	80	76	32	33
Ci3 2350M _{2C/4T}	93	75	41	43
Ci5 4440 _{4C/4T}	58	46	15	13
Ci7 6700K _{4C/8T}	36	30	7	8

Tabuľka 2.10: Reálne merania plného a redukovaného algoritmu na mriežke 1000×1000 uzlov. Údaje sú v milisekundách.

V tabuľke 2.11 budeme uvažovať merané rýchlosti aritmetických operácií z tabuľky 2.2 v časti 2.1.3. V stĺpcoch 2 až 4 máme pomery rýchlostí γ^+ , γ^* a $\gamma^{\cdot\cdot}$ aritmetických operácií získané podľa tabuľky 2.2 v časti 2.1.3. Stĺpec 5 predstavuje teoretické zrýchlenie redukovaného algoritmu vypočítané podľa vety 2.17 v predchádzajúcej časti. Posledný stĺpec 6 predstavuje merané zrýchlenie sériovej verzie redukovaného algoritmu podľa predchádzajúcej tabuľky.

Procesor	Pomery operácií			Zrýchlenie	
	γ^+	γ^*	$\gamma^{\cdot\cdot}$	teoretické	merané
FX-6300 _{3C/6T}	1,89	0,04	0,82	1,12	1,13
C2D E8200 _{2C/2T}	1,08	0,01	0,15	1,09	1,14
Ci5 650 _{2C/4T}	2,04	0,04	0,77	1,13	1,05
Ci3 2350M _{2C/4T}	3,43	0,04	0,67	1,21	1,24
Ci5 4440 _{4C/4T}	3,27	0,03	0,48	1,22	1,26
Ci7 6700K _{4C/8T}	2,98	0,02	0,4	1,21	1,2

Tabuľka 2.11: Porovnanie teoretického a meraného sériového zrýchlenia na mriežke 1000×1000 uzlov.

Merané zrýchlenie korešponduje s teoretickým zrýchlením v rámci malej odchýlky. Môžeme si všimnúť malé škálovanie výkonu paralelizovaného algoritmu najmä v prípade procesorov Intel obsahujúcich SMT kde je redukovaný algoritmus dokonca o niečo pomalší ako plný. Túto pomalosť sme nedokázali zatiaľ jednoznačne vysvetliť, no máme dve možné hypotézy:

- Prvou zmýšľanou príčinou môže byť fakt, že v plnom algoritme všetky

parciálne derivácie podľa nejakej premennej vypočítame trojdiagonálnymi sústavami naraz v jednom cykle. V redukovanom ale cez trojdiagonálne sústavy vyrátame len polovicu derivácií, pričom zvyšok dopočítame práve restami. Toto rozdelenie výpočtu na dve subprocedúry môže hypoteticky implikovať vysokú réžiu plánovača vlákien (tzv. *thread scheduler*) vrstvy OpenMP alebo operačného systému. V implementácií sa nám podarilo spojiť výpočet restov spolu s LU dekompozíciou do jedného kroku v prípade derivácií podľa premenných x a y a resty tak môžeme počítať súčasne s deriváciami počítanými cez LU. Táto optimalizácia však momentálne nie je možná pri zmiešaných deriváciách. Do budúca máme rozpracované zefektívnenie výpočtu zmiešaných restov, ktoré sa nám ale ešte nepodarilo do práce dokončiť.

- Druhá zmýšľaná príčina môže tkvieť v tom, že v prípade redukovaného algoritmu sa „bije“ súčasne SMT a superskalárnosť ako dve rôzne techniky inštrukčného paralelizmu. Dnešné procesorové mikroarchitektúry majú výpočtové jadrá vybavené práve dvomi matematickými jednotkami pre operácie s plávajúcou desatinnou čiarkou (označujú sa skratkou FPU). V tomto prípade SMT môže teoreticky každú jednotku prideliť inému vláknu. Pri redukovanom algoritme ale máme veľa operácií v tvare $a_0 \circ a_1 \circ \dots \circ a_n$, kde $\circ \in \{+, \cdot\}$. Optimalizujúci prekladač môže takýto výpočet rozložiť do viacerých strojových inštrukcií. V tomto prípade dôjde k využitiu oboch FPU jadra v rámci jediného vlákna ale napriek tomu SMT na toto jadro priradí aj druhé vlákno ktoré potrebuje počítať práve na FPU jednotke. Teda dôjde k situácii kedy dve vlákna zdieľajú spoločné zdroje, v tomto prípade dve FPU jednotky. Dôjde tak k zbytočnej réžii plánovača vlákien či už na úrovni OS alebo inštrukčného plánovača samotného procesora, čo spôsobí pokles výkonu. Ak je táto hypotéza korektná, tak aj v tomto prípade by bolo potrebné upraviť resty pri zmiešaných deriváciách, aby sa ich výpočet dal vykonať paralelne v jednom kroku spolu s LU dekompozíciou. Proti tejto hypotéze hovorí fakt, že na procesoroch mikroarchitektúry AMD Piledriver redukovaný algoritmus dosahuje zrýchlenie napriek tomu, že tento procesor má implementovanú SMT. Pravda, jedná sa o mierne odlišnú implementáciu oproti Intelu, no inštrukčný plánovač tohto CPU musí riešiť identický problém dvoch vlákien zdieľajúcich spoločné zdroje.

Táto kapitola sa venovala počítaniu zrýchlenia redukovaného algoritmu vzhľadom na plný algoritmus. Ako sme videli, teoretické zrýchlenie je v súlade s nameraným zrýchlením v závislosti od typu procesora.

Záver

Podarilo sa nám v prípade rovnomerne rozložených uzlov urýchliť sériový výpočet derivácií splajnov v uzloch a tiež zmenšiť pamäťovú náročnosť výpočtu na polovicu. Napriek tomu, že výsledný redukovaný algoritmus obsahuje viac aritmetických operácií je vďaka povahe mikroarchitektúr moderných procesorov a tiež vďaka povahe samotných operácií rýchlejší o približne 20%. Prioritou teraz ostáva jednak zovšeobecniť redukovaný algoritmus aj pre splajny s nerovnomernými uzlami a upraviť zmiešané zbytkové derivácie aby sme mohli docieľiť ďalšie zrýchlenie a najmä zefektívniť vláknovú paralelizáciu. Vláknovo paralelné počítanie uzlov redukovaným algoritmom nie je vo všeobecnosti v súčasnosti rýchlejšie ale závisí na spôsobe dosahovania vláknového paralelizmu na konkrétnej procesorovej mikroarchitektúre. Konkrétne podľa nameraných výsledkov vyplýva skutočnosť, že procesory Intelu s implementovaným SMT sú pri paralelnom počítaní uzlov horšie ako procesory bez tejto technológie. Výhodou ale ostávajú polovičné pamäťové nároky čo umožňuje riešiť väčšie úlohy a tak rozdeliť prácu medzi väčší počet logických procesorov. To by v budúcnosti umožnilo výpočty akcelerovať použitím masívne paralelizovaných čipov ako sú napríklad grafické karty.

Prílohy

Implementácia a užívateľská príručka

Jedným z našich cieľom je vytvorenie aplikácií na vizuálne a výkonnostné porovnávanie splajnov, ktorých uzly sú počítané De Boorovim a našim redukovaným algoritmom. Grafickú vizualizáciu sme implementovali v Microsoft Silverlight, čo je nástroj na tvorbu webových aplikácií spustiteľných priamo vo webovom prehliadači. Samotný vývoj aplikácii môže prebiehať v ľubovoľnom programovacom jazyku bežiacom pod Common Language Runtime (CLR). V našom prípade sme siahli po jazyku C#.

Silverlight je kompatibilný s väčšinou moderných webových prehliadačov a operačných systémov vrátane Microsoft Windows, Apple OS X a vďaka technológii Moonlight – open-source implementácii Silverlight-u aj na väčšine Linuxových distribúcií. Framework je možné použiť aj na vývoj off-line aplikácií v operačných systémoch Windows Phone, Windows 8 a Windows RT.

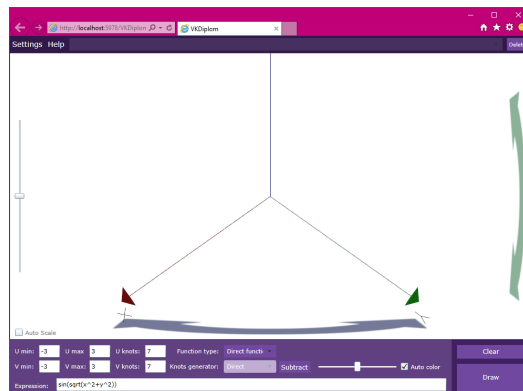
Poznámka 2.20 Naša aplikácia používa technológie, ktoré nie sú s Moonlight kompatibilné. V súčasnej dobe teda podporujeme iba MS Windows a Apple OS X. Prípadná portácia na GNU/Linux by musela byť vo forme off-line desktopovej aplikácie.

Pretože CLR znemožňuje korektne porovnávať procesorový čas a pamäťové nároky implementovaných algoritmov, rozhodli sme sa výkonnostný tester oddeliť do samostatnej aplikácie implementovanej v natívnom jazyku, konkrétne v C++.

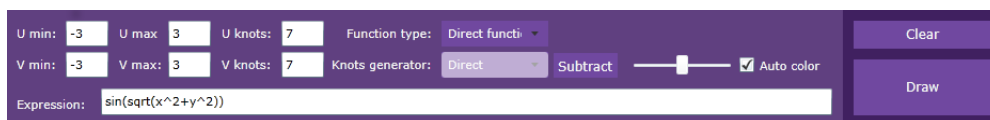
Vizualizácia

V tejto kapitole si ukážeme stručný návod na používanie a testovanie implementovaných aplikácií. Začnime aplikáciou na grafickú vizualizáciu.

Okno aplikácie sa skladá z troch hlavných častí, ktoré si trochu netradične prejdeme zdola nahor. Na obrázku 2.8 vidíme panel, pomocou ktorého mô-



Obr. 2.7: Aplikácia po spustení.



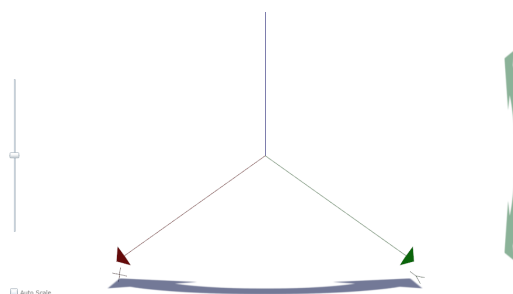
Obr. 2.8: Panel zadávania splajnu.

žeme na ľubovoľnom intervale a s ľubovoľným počtom uzlov (v rozumných medziach) interpolovať zapísanú funkciu dvoch premenných x a y . Popíšme si jednotlivé ovládacie prvky. Textové polia označené $U \min$, $U \max$ a $U \text{ knots}$ ($V \min$, $V \max$ a $V \text{ knots}$) značia interval a počty uzlov na osi x (osi y). Do spodného textového políčka označené *Expression* môžeme napísať takmer ľubovoľnú matematickú funkciu premenných x a y ktorú si želáme vykresliť. Vysúvateľný zoznam označený *Function type* umožňuje zvoliť typ interpolácie. Na výber máme ponúknuté tri možnosti akým spôsobom si želáme interpretovať matematický výraz v políčku *Expression*.

- *Direct function* vypočíta funkčné hodnoty potrebné pre grafickú vizualizáciu priamo z funkcie.
- *Bicubic* vypočíta z funkcie iba funkčné hodnoty v uzloch. Zvyšné potrebné body sú získané interpolačným hermitovým bikubickým splajnom.
- *Biquartic* funguje analogicky ako predchádzajúca položka. Použije sa ale bikvartický splajn.

Vysúvateľný zoznam označený *Knots generator* nám, v prípade výberu vykreslenia funkcie splajnom, umožňuje zvoliť plný (*De Boor*) alebo redukovaný

(*Reduced de Boor*) algoritmus. Tlačítko *Subtract* umožní vykresliť rozdiel medzi aktuálne zadanými hodnotami a nejakou už vykreslenou funkciou. Pomocou posuvníka označený *Auto color* môžeme namiesto automaticky vybranej farby vykresliť novú plochu vlastnou farbou. Nakoniec v pravej časti okna sa nachádzajú tlačidlá *Clear* a *Draw*, pričom prvé odstráni všetky vykreslené funkcie a druhé nám vypočíta a zobrazí funkciu podľa aktuálne zadaných hodnôt.



Obr. 2.9: Hlavné zobrazovacie okná.

Na obrázku 2.9 vidíme dominantnú časť aplikácie a síce zobrazovanie funkciových plôch. V ľavej časti okna môžeme nájsť posuvník na škálovanie osi z . Na spodnej a pravej strane hlavnej časti okna sú tlačidlá reprezentované modrou horizontálnou, respektíve zelenou vertikálnou šípkou. Pri ich držaní myšou je možné rotovať obrazom v danej osi.

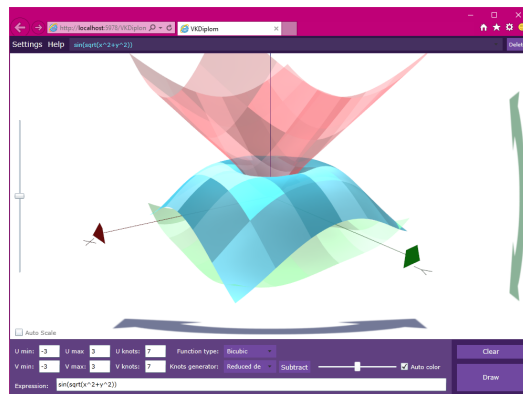


Obr. 2.10: Lišta nástrojov.

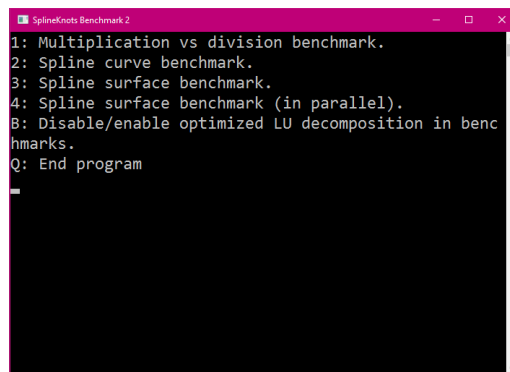
Položka *Settings* umožňuje prispôbiť grafické vykresľovanie splajnov. Tieto nastavenia je vhodné upraviť najmä pri starších počítačoch na dosiahnutie optimálnej plynulosti vykresľovania. V strednej časti panela máme k dispozícii vysúvateľný zoznam vykresľovaných plôch, ktorým ich môžeme zvýrazňovať.

Výkonnostný tester

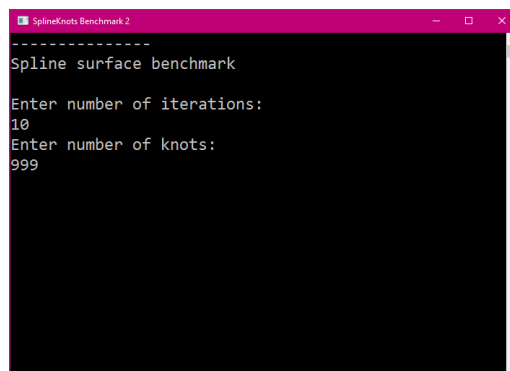
Teraz si popíšme druhú implementovanú aplikáciu zameranú na ukážku či skutočne je redukovaný algoritmus rýchlejší. Na obrázku 2.12 môžeme vidieť úvodné okno aplikácie, ktoré obsahuje šesť položiek. Ich význam si teraz vyjasníme.



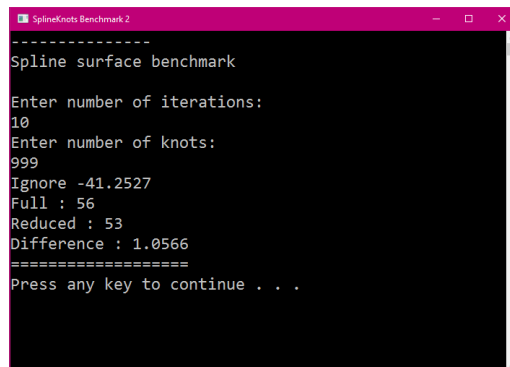
Obr. 2.11: Vykreslené tri splajnové povrchy.



Obr. 2.12: Výkonnostný tester.



Obr. 2.13: Zadanie údajov pre testovanie.



```
Spline surface benchmark
Enter number of iterations:
10
Enter number of knots:
999
Ignore -41.2527
Full : 56
Reduced : 53
Difference : 1.0566
=====
Press any key to continue . . .
```

Obr. 2.14: Výsledky testu.

1. *Multiplication vs division benchmark* spustí niekoľko variant testovania pomerov rýchlosti operácií sčítania, násobenia a delenia.
 2. *Spline curve benchmark* pomerá časy výpočtov derivácií pre krivkový splajn interpolujúci preddefinovanú funkciu plným a redukovaným algoritmom. Pred spustením aplikácie vyžaduje postupné zadanie dvoj údajov. Prvý udáva počet opakovaní testu pre každý algoritmus. Výsledný pomer vznikne vydelením aritmetických priemerov nameraných časov pre každý algoritmus. Čím je zadané číslo väčšie, tým je meranie menej náchylné na odchýlky. Odporúča sa zadať dvojcifernú hodnotu a potvrdiť klávesou **Enter**. Druhé zadané číslo znamená počet uzlov splajnu. Odporúča sa zadať rádovo 10^5 až 10^6 uzlov. Menší počet dokáže väčšina moderných počítačov vypočítať príliš rýchlo aby ten čas dokázala aplikácia rozumne zmerať.
 3. *Spline surface benchmark* analogicky ako predchádzajúci test testuje pomer časov výpočtu plného a redukovaného algoritmu pre plochy. Druhý zadany údaj tentoraz znamená počet uzlov pre jeden rozmer. Celkový počet uzlov je teda štvorec zadanej hodnoty. Pre väčšinu počítačov sa odporúča zadať rádovo 10^2 až 10^3 uzlov.
 4. *Spline surface benchmark (in parallel)* paralelná verzia testu pre plochy využívajúca na výpočet jednej matice uzlov všetky dostupné logické procesory.
- B. *Disable/enable optimized LU* prepne medzi naivnou a optimalizovanou variantou algoritmu pre počítanie LU dekompozície v testoch 2 a 3. Test 4 toto nastavenie neovplyvňuje z dôvodu značných pamäťových

nárokov naivnej verzie LU dekompozície (program na väčšine strojov zahltí celú operačnú pamäť).

- Q. *End program* ako názov napovedá, pri stlačení tlačidla **Q** dôjde k vypnutiu aplikácie.

Zoznam použitej literatúry

- [1] David Salomon, Curves and Surfaces for Computer Graphics, Springer, 2006
- [2] I. Szabó, L. Miño, C. Török, Biquartic polynomials in bicubic spline construction, Central European Journal of Computer Science, accepted for publication (2015)
- [3] L. Miño, C. Török, Fast algorithm for spline surfaces, Communication of the Joint Institute for Nuclear Research, Dubna, 2015, E11-2015-77
- [4] C. de Boor, Bicubic spline interpolation, Journal of Mathematics and Physics, 41(3),1962, 212-218.
- [5] E. Süli, D. Mayers, An Introduction to Numerical Analysis, Cambridge University Press, 2003, ISBN 0 521 00794 1
- [6] <https://github.com/vildibald/VKDiplom-master>, repozitár so zdrojovými kódmi k aplikáciám.
- [7] https://en.wikibooks.org/wiki/Algorithm_Implementation/Linear_Algebra/Tridiagonal_matrix_algorithm
- [8] <http://www.anandtech.com/show/9483/intel-skylake-review-6700k-6600k-ddr4-ddr3-ipc-6th-generation/9>
- [9] <http://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-optimization-manual.pdf>
- [10] http://www.agner.org/optimize/instruction_tables.pdf
- [11] <http://www.lighterra.com/papers/modernmicroprocessors/>