ACCELERATION OF BIOLOGICAL SEQUENCE COMPARISON ALGORITHMS USING FPGA

Patrik Beck
Bachelor Degree Programme (1), FIT BUT
E-mail: xbeckp00@stud.fit.vutbr.cz

Supervised by: Tomáš Martínek
E-mail: martinto@fit.vutbr.cz

ABSTRACT

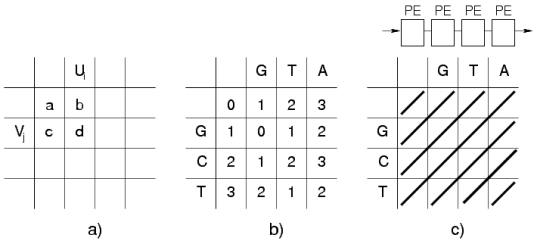
Hadware accelerators for approximate string matching (AM) play a significant role in biological algorithms. However, their wider use is often limited by the lack of flexibility and modularity. This paper presents a generic approach for designing of AM architectures that represents an essential step for their automated design. The proposed approach is evaluated on several biological tasks and the accelarator achieves speed-up 100-1000 in comparison with conventional processors.

1 ÚVOD

Bioinformatika vznikla spojením biológie s informatikou. Zaoberá sa náročnými výpočtami, napríklad pri analýze ľudského genómu, alebo pri simuláciách dejov v organizme. Veľa práci sa venuje urýchľovaniu často používaných algoritmov. Avšak väčšina implementácií je veľmi silne zameraná na konkrétnu úlohu [1] [3]. Využitie urýchlenia je potom limitované. Pri vytvorení generickej architektúry [2], je možné počítať väčšiu škálu úloh. Takto napisaný design je v podstate šablona, ktorú je možné prispôsobiť a zároveň optimalizovať na veľa rôznych úloh. Riešenie je vyskúšané na FPGA čipe Virtex II PRO - XC2VP50 na karte combo6x.

2 PRIBLIŽNÉ POROVNÁVANIE REŤAZCOV

Algoritmus Smith-Waterman reprezentuje špecifickú aplikáciu dynamického programovania (DP), ktorá rieši približnú zhodu reťazcov a bol prvý raz použitý v molekulárnej biológii na zarovnanie DNA sequencií. V skratke, algorimus rozdeluje úlohu na elementárne kroky, ktoré hľadajú lokálne približné zarovnania. Výpočet typicky prebieha v DP matici, ktorá po naplnení výsledkami všetkých elementárnych krokov definuje najlepšie zarovnanie(alebo zarovnania) a ich skóre.
Príklad porovnania reťazcov „GTA“ a „GCT“ s použitím algoritmu Smith-Waterman je zobrazený v obrázkoch 1a a 1b. Hodnota každej položky d v DP matici je vypočítaná z troch najbližších susedov a, b a c podľa nasledovných pravidiel:

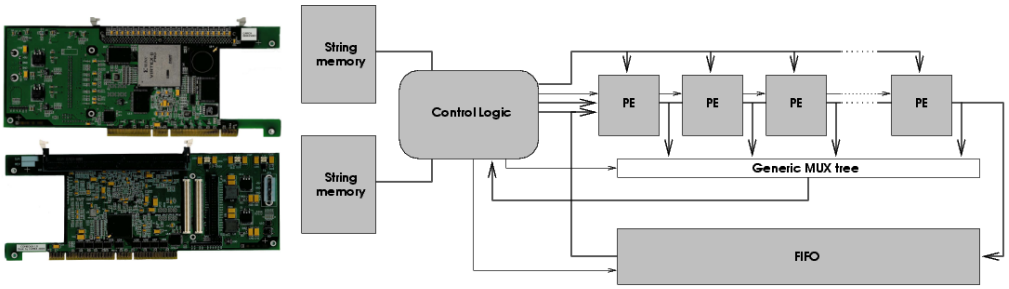


Obrázek 1: Algoritmus Smith-Waterman

$$d = min \begin{cases} a & if \quad U_i = V_j \\ a + sub & if \quad U_i \neq V_j \\ b + ins \\ c + del \end{cases}$$

Premenné *sub*, *ins* a *del* reprezentujú penalty za nahradenie, vloženie a absenciu znaku a môžu byť nastaviteľné podľa rôznych požiadaviek porovnávania. Napríklad, ak výskyt redundantných znakov je menej prípustný ako líšiace sa znaky, penalta vkladania môže byť nastavená vyššie ako penalta nahradenia.
Viac výpočetných krokov algoritmu Smith-Waterman môže byť vykonaných nezávyslé. Táto vlastnosť je využitá k paralelizácii výpočtu. V DP matici to znamená, že diagonálne hodnoty môžu byť počitané zároveň, obrázok 1c.
Paralelny výpočet je obvykle uskutočnený množinou Procesných Elementov (PE) štrukturovaných do systolického pola. Každé PE v poly počíta jeden stĺpec a posiela vypočítané hodnoty susednému PE každý hodinový cyklus. Takže diagonálne hodnoty sú počítané paralelne.

3 HARDWAROVÁ IMPLEMENTÁCIA



Obrázek 2: Schéma designu

Najdôležitejším prvkom v designe je systolické pole, ktoré realizuje výpočet(skladá sa z niekoľkých PE). Ďalej dve pamäte pre uloženie reťazcov a v neposlednej rade FIFO, ktoré svojim prepojením posledného a prvého PE umožňuje porovnávať aj reťazce vačšie ako je počet PE v systolickom poly. Multiplexor naopak umožňuje zobrať výsledok z ktorého koľvek PE a tým umožnuje porovnanie reťazcov kratších ako je počet PE v systolickom poli.
Schéma na obrázku 3 vyjadruje jednotku, ktorá je schopná riešiť porovnanie dvoch reťazcov samostatne. Takýchto jednotiek môže byť v celom designe niekoľko(generický počet) v závislosti od zdrojov, ktoré FPGA čip poskytuje.

Tabulka 1: Výsledky s použitím čipu Virtex II xcv2p50-7

applikácia	SP	PE	Frekv.	BUps	Čas
VPCR+PRIMEX	81	10	241	195.2	353us
VPCR microarrays	1	40	241	9.6	12.4s
Oligo vs. Genóm	1	80	209	16.7	14.3s
Gén vs. Genóm	1	270	152	41.0	19.29m
Proteín+PRIMEX	79	8	177	111.6	5.57m
Proteín vs. Databáza	1	270	133	35.9	1.12m
Proteínové uhly	6	71	137	58.3	46s

4 DOSIAHNUTÉ VÝSLEDKY

Implementovaný design bol vyskúšaný na FPGA čipoch Virtex II - XC2V3000(na karte combo6) a Virtex II PRO - XC2VP50(na karte combo6x). Pre často používané aplikácie z oblasti bioinformatiky, ako napríklad porovnávanie jedného génu voči celému genómu, porovnánie proteínu s databázou proteínov, urýchlovanie algoritmu primex, boli vypočítané generické parametre. Dosiahnuté výsledky sú zhrnuté v tabulke 1. Sú prezentované nasledovné informácie: počet systolických polí na čipe (SP), počet PE v jenom systolickom poli (PE), frekvencia na ktorej implementácia beží (frekv.), výkon v miliardách update-ov za sekundu (BUps) a čas potrebný na riešenie úlohy. Pre porovnanie, BUps procesoru Intel Pentium 4 3,20 GHz je približne 0,05 BUps.

5 ZÁVER

Porovnávanie biologických sekvencií môže byť využité pri analýze genómu, rozpoznávaní dedičných chrôb, alebo pri hľadaní evolučného stromu. Vďaka generickému prístupu je možné riešiť veľké množstvo úloh a tak niekoľko násobne urýchliť bioinformatické aplikácie. Toto riešenie bolo vyskúšané na reálnom hardware a dosiahnuté zrýchlenie sa pohybuje v niekoľkých rádoch.

REFERENCE

[1] C. W. Yu, K. H. Kwong, K.-H. Lee, and P. H. W. Leong, "A smith-waterman systolic cell." in *Field Programmable Logic and Application (FPL 2003)*, Lisbon, Portugal, September 2003, pp. 375–384.

[2] T. V. Court and M. C. Herbordt, "Families of fpga-based algorithms for approximate string matching." in *IEEE International Conference on Application-Specific Systems, Architectures, and Processors (ASAP 2004)*, Galveston, TX, USA, September 2004, pp. 354–364.

[3] S. Guccione and E. Keller, "Gene Matching Using JBits." in *Field Programmable Logic and Application (FPL 2002)*, Montpellier, France, September 2002, pp. 1168–1171.