Eötvös Loránd Tudományegyetem
Informatikai Kar
Komputeralgebra Tanszék

Cache optimalizált lineáris szita párhuzamos megvalósítása

Dr Vatai Emil
adjunktus

Huszti Péter
Programtervező Informatikus MSc

Budapest, 2017

Tartalomjegyzék

1. Bevezetés

1.1. A dolgozat célja, motiváció

Az alkalmazott matematikában nagyon fontos szerepet játszanak a prím számok, elég csak például a nyílt kulcsos titkosítási módszerekre gondolnunk. Ebből kifolyólag az évek során nagyon sok módszert fejlesztettek ki prímszázmok keresésére, például a Fermat-teszt, vagy a Miller-Rabin teszt. Ezen módszerekkel elég gyorsan el lehet dönteni egy darab számról, hogy prím-e vagy sem, és így ezért az ilyen algoritmusokkal nagyon nagy prímeket is meg lehet találni viszonylag gyorsan.

Előfordulhat azonban probléma, hogy egy adott intervallumban szeretnénk megtalálni az ott lévő összes prím számot. Az ilyen feladatok megoldására a leghatékonyabb módszerek a szitáló módszerek. Ezek úgy működnek, hogy kiválasztunk bizonyos számokat, és azokkal "végigszitálunk" a vizsgált intervallumon, és az érintetlenül hagyott számok lesznek a nekünk megfelelő, jelen esetben prím számok. Ezek közül is a legegyszerűbb egészen az ókorig nyúlik vissza, az ún. Erathosztenészi-szita. Ez az algoritmus ahhoz képest, hogy milyen rég óta ismert, meglepően hatékonyan működik. De természetesen vannak hátrányai, például hiába tűnik nagyon gyorsnak komplexitás szempontjából, elég nagy intervallumokra nagyon meg fog növekedni a memória olvasások száma, ami mint köztudott nagyságrendekkel lassabb, mint a processzorok utasítás végrehajtó képessége már a memóriában lévő adatokon. Nem meglepő módon sok féle képpen fel lehet javítani az algoritmus teljesítményét mai modern eszközökkel.

A dolgozat célja az, hogy a fent említett "naív" szitáló algoritmus teljesítményét növeljük, miközben a program skálázható is maradjon, és így a valós gyakorlati életben is alkalmazni lehessen. A dolgozat során két féle módon próbáltam meg javítani a teljesítményen.

Az első és talán legkézenfekvőbb módszer a program párhuzamosítása. Ezt viszonylag egyszerűen meg lehet tenni, mivel az algoritmust könnyedén fel lehet darabolni kisebb, egymással ekvivalens részfeladatokra, amiket szét lehet osztani a processzor szálai közt. Ráadásul a konkurens programokra jellenző veszélyek nem állnak fent, így sok nehézségtől meg tudjuk kímélni magunkat, ami a teljesítményt is javítja.

A másik megközelítés, hogy a memóriaműveletek számát próbáljuk meg minimalizálni. Erre egy hatékony megvalósítása az ún. COLS - cache optimalizált lineáris szita - algoritmus, aminek az megvalósítása is része a dolgozatnak.

Végül a fentiekből magától értetődik egy újabb gyorsítási lehetőség, hogy a COLS algoritmust is meg lehet valósítani párhuzamos szálakon, ami mint majd később látjuk megint csak nagy teljesítménynövekedéssel járhat.

A dolgozat során implementálva lett a fent említett négy algoritmus, és az volt vizsgálva, hogy milyen esetekben (a probléma mérete, hardver specifikációja) mennyire tudják, ha egyáltalán lehetséges, és mint kiderült nagyon is az, felgyorsítani a fenit módszerek a probléma megoldását. Az algoritmusokat C++ nyelvben valósítottam meg, csak és kizárólag a standard C++11 szabvány által kínált lehetőségeket felhasználva.

Az elészített programmal könnyen és gyorsan lehet egy megadott intervallumon megkeresni a prím számokat, továbbá megfelelően paraméterezhető a rendelkezésre álló processzor(ok) tulajdonságainak ismeretében. A program jól skálázható, és elméletben akár nagyobb, sok központi számítóegységből álló konfigurációk, plédául szuperszámítógépeken is lehet használni, és így nagy, valós problémák megoldására is lehetőséget nyújt.

A dolgozat során elkészített programot fel lehet használni többek között Cunningham-láncok keresésére is. Ezek a láncok egymáshoz közel elhelyezkedő prímekből állnak és felhasználják például az ún. Primecoin digitális fizetőeszköz bányászására.

A COLS algoritmust, és a mögötte húzódó elméletet nem csak az Erathosztenészi-szita felgyorsítására lehet felhasználni, hanem többek között például az ún. SIQS algoritmus, vagyis az öninicializáló kvadratikus szita javítására is, ami egy elég hatékony faktorizáló algoritmus.

2. Matematikai háttér

2.1. Prím számok, faktorizáció

2.1. Definíció. Egy p természetes számot prímnek nevezünk, ha $\forall a, b$-re amire $p|a \cdot b \rightarrow (p|a \vee p|b)$.

Természetes számok körében ez a definíció ekvivalens azzal, hogy egy prím számnak kettő, és csak kettő osztója van, 1 és önmaga.

A prím számok kitüntetett szerepet játszanak a matematikában. Többek közt felhasználják őket hasítótáblákhoz, pszeudovéletlen számok generálásához vagy nyílt kulcsú titkosításokhoz. Utóbbiak széles körben elterjedtek, valószínűleg sokan ismerik például az RSA kódolást, az SSH-t vagy a HTTPS-t. Ezek mind fontos részét képezik a modern kornak. A nyílt kulcsú kódolások olyan matematikai problémákon alapulnak, amelyeket megoldani nehéz, vagyis a mai eszközeinkkel valós időben nem lehetséges, viszont ellenőrizni egy lehetséges megoldást gyors és egyszerű. A leggyakrabban használt ilyen probléma a prím faktorizáció.

Számelmélet alaptétele: minden pozitív szám felírható egyértelműen prímszámok szorzatára.

Viszont, ennek a felbontásnak a megkeresése NP-nehéz probléma, vagyis nem tudunk jelenleg sokkal jobb módszert annál, mint hogy kipróbáljuk az összes lehetséges prím számot, hogy oszthója-e a felbontani kívánt számnak.

Tehát jól látszik, hogy a prím számok megtalálása kiemelten fontos feladat. Rengeteg módszer létezik arra, hogy prímeket keressünk. A dolgozat az ún. szitáló módszerekkel foglalkozik, konkrétabban ezek felhasználása prímszámok keresésére. Ezeknek a módszereknek megvan az az előnye, hogy egy adott intervallumban megtalálják az összes ott előforduló prímet, viszont ha konkrétan csak egy darab számról akarjuk eldönteni, hogy prím-e, akkor ezeknél a módszereknél léteznek sokkal hatékonyabbak is.

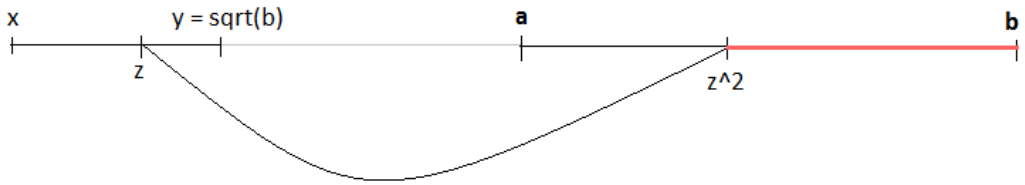
2.2. Eratoszthenészi-szita

Az Eratoszthenészi-szita, mint a nevéből is látszik már egy nagyon régen ismert algoritmus. Ennek ellenére, bármilyen meglepő is, ha gyorsan meg akarjuk keresni egy intervallumban az összes prím számot, akkor ehhez az algoritmushoz kell visszanyúlnunk. Ez egy egyszerű kizárásos algoritmus. A számelmélet alaptétele szerint az intervallumunkban minden szám, amelyik nem prím, osztható nálánál kisebb prím szám(ok)kal. Tehát, ha lenne egy listánk a kisebb prímekről (ezt a táblát nevezzük szitatáblának), akkor azoknak meg tudnánk találni az intervallumunkban lévő többszöröseit, és amely szám egyik kis prímnek sem többszöröse, az prím szám. Ez az alap ötlet.

Az algoritumus:

1. Készítünk egy listát a kisebb prímekről, amelyekkel majd ki fogjuk szitálni a vizsgált intervallumot. Ezt a listát magát is el lehet készíteni egy kisebb Erathosztenészi szitával.

   De mit is jelent az, hogy kis prímek? Jelöljük az intervallumunkat, ahol keressük a prímeket $[a, b]$-vel. A első ötlet természetesen, hogy vizsgálunk minden 1-nél nagyobb de b-nél kisebb prímszámot. De kicsit jobban belegondolva erre egyáltalán nincs szükség.



1. ábra.

   Elég $\sqrt{b}$-ig felmennünk a szitatábla felső korlátjával. Ahhoz, hogy ezt belássuk, vegyük észre, hogy minden a szitatáblában lévő, prímmel csak annak négyzetétől kell kezdenünk szitálni.

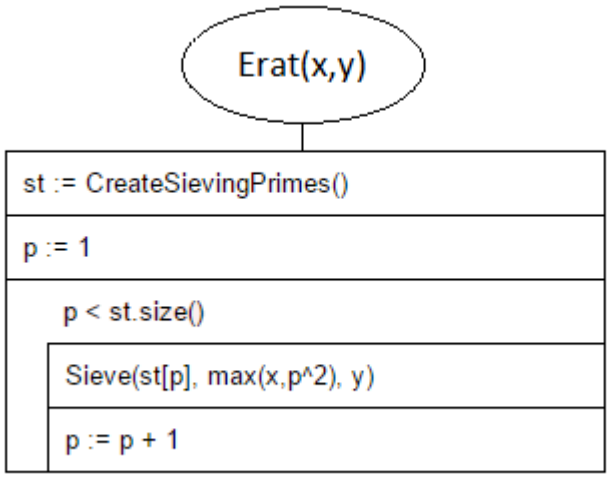
   Jelöljük a szitatáblát [x,y]-nal. Ekkor legyen $x < z < y$. Ekkorra már kiszitáltunk minden x és z közé eső prímmel. Vagyis azok a számok, amelyek oszthatóak valamely z-nél kisebb prímmel, már ki vannak szitálva. Vagyis az első olyan szám, amit vizsgálnunk kell, az a z^2.

   Ebből már jól látszódik, egrészt az, hogy a fenti jelölést használva $y = \sqrt{b}$, másrészt ahogy haladunk előre a szitálóprímekkel, a vizsgált intervallum egyre kisebb lesz. Ez az egyszerűsítés rengeteget javít a program teljesítményén. Tfh. hogy $[x, y] = [10^6, 4 \cdot 10^6]$. Ekkor a naív $4 \cdot 10^6$ szitatábla méret helyett elég mindössze $\sqrt{4 \cdot 10^6} = 2000$ méretű tábla.

2. Legyen p a szitatábla első eleme, a legkisebb prím: 2.

3. Jelöljük meg p összes többszörösét a vizsgált intervallumban, kedzve p^2-től, ha $a < p^2$.

4. Legyen p a szitatábla következő eleme, a következő szitáló prím. Ha nincs ilyen akkor vége az algoritmusnak. Ha van ilyen, akkor folytassuk a 3. ponttal.



2. ábra. Erathosztenészi-szita

Az iteráció végén a jelöletlenül maradt számok lesznek a keresett prím számok. Ha egy ilyen szám összetett lenne, akkor már biztosan meg lett volna jelölve, hiszen a számelmélet alaptétele szerint felírható kisebb prímszámok szorzataként, vagyis egy kisebb prímszámnak a többszöröse, amikkel pedig már szitáltunk.

2.2.1. Hátrányok

Az algoritmus kisebb intervallumokra nagyon jól használható. A dolgozat során elvégzett mérések szerint körülbelül 2^17-es nagyságrendig tartotta a lépést a később szóba kerülő COLS algoritmussal. De igazából algoritmikus módszerekkel nehéz tovább gyorsítani ezt a módszert. A nem prím számok kiszitálását nem tudjuk megúszni.

Viszont közismert tény, hogy a memória műveletek nagyságrendekkel lassabbak, mint ahogy egy processzor képes végrehajtani műveleteket a már meglévő, a cache memóriában tárolt adatokon. És itt jön ki igazából a nagy hátránya az Erathosztenészi-szitának. Mi történik akkor, ha akkora intervallumot akarunk vizsgálni, amekkora nem fér be a memóriába?

TODO: insert ábra

Ahogy az ábrán is látszódik, ilyen esetben nem tudunk végigszitálni egy adott kis prímmel az egész intervallumon. Amint elérjük a cache "végét", nem fogja találni a keresett következő számot a program, ezt hívják **cache miss**nek. Ilyenkor be kell kérni a hiányzó adatot jó esetben a RAM memóriából, rosszabb esetben a

háttértárról. Ráadásul, amint kiszitáltunk egy prímmel, kezdhetjük beolvasni előről az összes adatot, jó esély van rá, hogy a processzor már kidobda az intervallum elején kiszitált számokat. Vagyis ez azt jelenti, hogy legrosszabb esetben az intervallum összes számát újra és újra be kell olvasni, egészen pontosan $\sqrt{b}$-szer. Mondani sem kell, hogy ez mennyire lelassítja az program teljesítményét. Sejteni lehet, hogy a vizsgált intervallum növekedésével nem egyenesen arányos a program futásának ideje. Ezt később be is látjuk majd, lásd: 5.3

2.3. Szegmensenkénit szitálás

A 2.2.1 fejezetben látott hátrányt mindenképp ki kell küszöbölni, ha valós környezetben is alkalmazható alkalmazást implementálni. Ugyanis a naív Erathosztenészi-szita nagyobb számok illetve intervallumok esetében az idő nagy részében a memóriából fog olvasni. Ez a probléma csak akkor fog jól látszódni, ha már akkora intervallumot szitálunk, ami nem fér bele a cache memóriába. Ez azért fordul elő, mert egyesével vesszük ki a prímeket a szitatáblából, és utána egyesével dolgozunk velük. Viszont ha megfordítanánk az algoritmust, és az intervallumból vennénk ki egy darabot, és ezzel dolgoznánk a szitatáblán, akkor megoldódna minden memória problémánk. Ugyanis a szitatábla mindössze $\sqrt{b}$ méretű, ami valószínűleg könnyedén elfér a lokális memóriában.

Az előbb említett darabot, amit kiveszünk az intervallumból nevezzük szegmensnek, vagy angolul *chunk*-nak. Ötlet: daraboljuk fel az egész intervallumot ilyen szegmensekre, és ezeket egyesével szitáljuk ki teljesen, majd írjuk vissza a memóriába. Így elérhetjük, hogy minden vizsgált szám mindössze egyszer kerüljön beolvasásra, nagyságrendekkel redukálva így a memória olvasások és írások számát.

Így tehát az algoritmus:

1. Készítsük el a kisebb prímekből alló szitatáblát a naív algoritmushoz hasonló módon.

2. Daraboljuk fel a vizsgált intervallumot alkalmasan sok szegmensre. Ara, hogy pontosan mennyire finom felosztás a legjobb választás nem lehet általánosan jó választ adni. Függ a használt számítógép specifikációjától. Valamint természetesen egy számítógép közben sok más dologra is használja a cache memóriáját, így nem lehet előre egy legjobb felosztást mondani. Ráadásul ha túl finom felosztást választunk, akkor pedig lehet, hogy csak felesleges overheadet okozunk a programnak.

   Viszont kisebb, de már elég nagy példákon kísérletezve be lehet lőni azt, hogy körülbelül mi lenne a legmegfelelőbb felosztás.

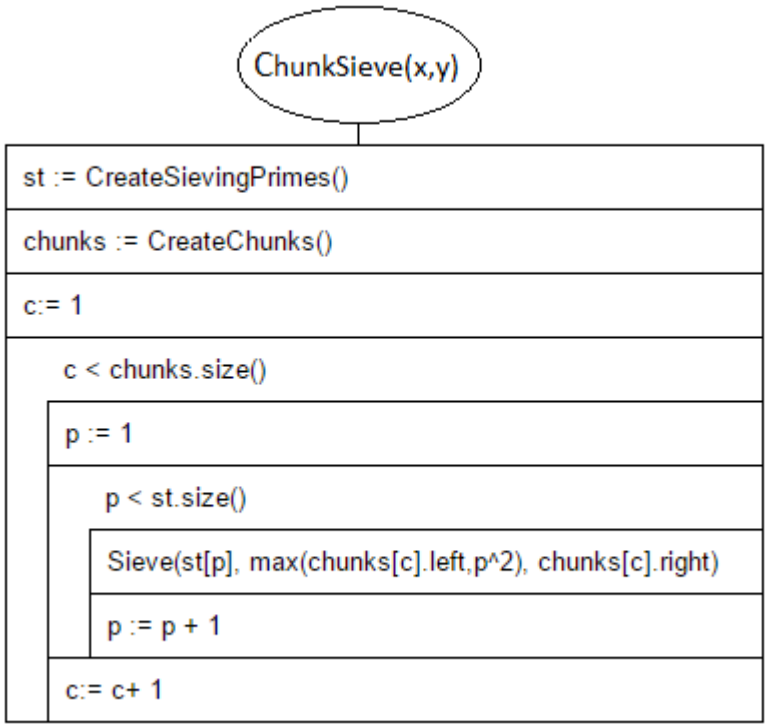
3. Olvassuk be az első szegmenst a memóriába.

4. Legyen p a szitatábla első eleme.

5. Jelöljük meg p összes többszörösét a beolvasott szegmensben, kezdve p^2-től, ha $a < p^2$.

6. Legyen p a szitatáblából a következő szitáló prím. Ha nincs ilyen, akkor kész vagyunk az aktuális szegmens kiszitálásával, vagyis már akár ebből a szegmensből ki is lehetne olvasni a benne lévő prím számokat. Ekkor olvassuk be a következő szegmenst, és folytassuk a 4-es ponttal.

   Ha viszont nincs több szegmens, akkor készen vagyunk, vége az algoritmusnak.



3. ábra. Szegmensenkéni szitálás

Amint később látni fogjuk, csak ezzel a módosítással volt, hogy több mint 3-szoros sebesség növekedést lehetett elérni. Lásd: 5.3

2.4. COLS

3. Felhasználói dokumentáció

3.1. A program használata

A program a **result/** mappába fogja elhelyezni a kiszámolt prímeket a result.txt fájlba, ha mást nem adtunk meg.

TODO: insert ábra

3.2. Rendszerkövetelmények

A program 64-bites Microsoft Windows operációs rendszer használatával íródott és lett tesztelve. (Windows 7 és 10) Így az ilyen operációs rendszereken biztosan helyesen működik.

Ahhoz, hogy UNIX alapú operációs rendszeren, illetve 32-bites rendszereken használni tudjuk, ahhoz újra kell fordítani a programot. Lásd: 4.2 Ezeken a rendszereken nem lett tesztelve a program, de nem használja ki a használt operációs rendszer által nyújtott lehetőségeket, ezért portolható. De természetesen ajánlott a tesztelő program használata is más rendszereken való használat előtt.

3.3. CD tartalma, telepítés

A *docs* mappában található a programhoz tartozó dokumentáció .pdf formátumban.

A *source* mappában található a futtatható fájl, ennek neve sieve.exe. Ezt a fájlt futtatva indíthatjuk el a programot. Ebben a mappában találhatóak továbba a program forrásfájljai is. A program konfigurálásához lásd: 3.1. A program módosításához, illetve újrafordításához lásd: 4.2

A program teszteléséhez használt fájlok a *test* mappában találhatóak. Bővebben lásd: 4.7

A *result* mappába helyezi el a program a kiszámított eredményeket. Természetesen a CD-re nem tud írni a program.

A *benchmark* mappában találhatóak a programban elkészített algoritmusok segítségéhez használt fájlok. Itt található még korábbi mérési eredmények benchmark_result_dátum.xlsx néven.

Telepítés:

A program használható közvetlenül a CD-ről is, bár ekkor nem tudjuk a kapott ererdményeket megteinteni. Ehhez a CD tartalmát másoljuk fel a használni kívánt számítógépre, és ezek után a program a fentebb említett *result* mappába fogja másolni az eredményeit.

4. Fejlesztői dokumentáció

4.1. Felhasznált technológiák

A program C++ programozási nyelv standard C++11-es verziójának használatával íródott. Semmilyen külső könyvtárat nem használ, így könnyen újrafordíthatja és portolhatja más rendszerre bárki.

A párhuzamos szálakon történő futtatáshoz a standard C++11 által kínált **thread** könyvtár van felhasználva.

A program fejlesztése során a verziókövetésre a GitHub online verziókövető rendszer volt használva. Egy program implementációja során kiemelten fontos egy hasonló szolgáltatás használata. Ennek segítségével könnyedén nyomon lehet követni a program fejlesztésének történetét, és egyszerűen lehet több számítógépről is folytatni a fejlesztést. A program megtalálható az alábbi helyen: https://github.com/peterhuszti/Thesis-MSc

A program tesztelése szintén standard C++ segítsgével lett megoldva. A dolgozat során ahol lehetett próbáltam a C++11-es szabvány által nyújtott lehetőségeket kihasználni. Erre jó példa a thread könyvtár használata, ami egy nagyon egyszerű API-t kínál szálak definiálására és általában konkurens programok készítésére. Továbba ki van hasnzálva a C++-ban újdonságnak számító lambda kifejezések ereje is. Ezeknek a lambdáknak, vagy más néven névtelen eljárásoknak a különlegessége, hogy nem tartozik hozzájuk azonosító. Ezeket felhasnzálva egyszerűen tudunk eljárásokat paraméterül adni magasabb rendű függvényeknek, például szálak viselkedésének a megadásakor.

A benchmark elkészítéséhez a Benchpress nevű frameworköt használtam. Ez egy nagyon egyszerűen és könnyen használható eszköz, amivel C++11 nyelven írt programok, vagy akár csak külön függvények és eljárások futási sebességét lehet mérni. REF

4.2. Fordítás

Ahhoz, hogy le tudjuk újra fordítani a programot, ahhoz egy C++11 kompatibilis fordítóra van szükség. A program eredetileg a g++ 5.2.0-ás verziójával lett lefordítva, de semmi akadálya annak, hogy más megfelelő fordítóprogramot használjunk.

Ha viszont g++-t használunk, akkor a következő javasolt g++ [-o sieve] -O3 -std=c++11 main.cpp. Az -o kapcsoló nem szükséges, ezzel csupán az elkészített futtatható fájl nevét tudjuk megadni. Az -std=c++11 paraméterrel tudjuk megadni, hogy a fordító a C++11-es szabvány szerint próbálja meg lefordítani a programot. Az -O3 paraméter egy optimalizációs paraméter. Ezekről bővebben lásd: 5.2

4.3. A program felépítése

4.4. Adatszerkezetek

4.4.1. Szitatábla

4.4.2. Szegmensek

4.4.3. Körök és edények

4.5. Párhuzamos megvalósítás

4.6. Skálázhatóság

4.7. Tesztelés

4.8. Továbbfejlesztési lehetőségek

5. Az algoritmusok összehasonlítása

5.1. Használt hardverek

A dolgozat során kettő konfiguráción volt lehetőség tesztelni a program és az algoritmusok sebességét és teljesítményét:

'A' konfiguráció	Intel Core i5-5300U @ 2.3GHz
	2 core, 4 thread
	3 MB cache
	Max. memória sávszélesség: 25.6 GB/s
	8 GB RAM
	64-bit Windows 7

'B' konfiguráció	Intel Core i7-4790 @ 3.6GHz
	4 core, 8 thread
	8 MB cache
	Max. memória sávszélesség: 25.6 GB/s
	16 GB RAM
	64-bit Windows 10

Érdemes megjegyezni, hogy az 'A' konfiguráció egy laptop, míg a 'B' egy asztali számítógép. Később jól fog látszódni a két számítógép közötti teljesítmény különbség.

Korábban volt szó a skálázhatóságról, lásd 4.6. Érdekes lenne kipróbálni egy nagyobb, több processzort tartalmazó konfiguráción is a programot, de erre sajnos a dolgozat készítése során nem volt lehetőség.

5.2. Fordítási paraméterek

5.3. Az eredmények

5.4. Konklúzió

6. Összegzés