

Politechnika Poznańska
Wydział Informatyki i Zarządzania
Instytut Informatyki

Praca dyplomowa magisterska

**RÓWNOLEGŁE ODKRYWANIE REGUŁ ASOCJACYJNYH
ZAIMPLEMENTOWANE NA PROCESORY GRAFICZNE**

inż. Tomasz Kujawa

Promotor
dr inż. Witold Andrzejwski

Poznań, 2011

Tutaj przychodzi karta pracy dyplomowej;
oryginał wstawiamy do wersji dla archiwum PP, w pozostałych kopiach wstawiamy ksero.

Spis treści

1	Wstęp	1
1.1	Cel i zakres pracy	3
2	Podstawy teoretyczne	4
2.1	Definicje	4
2.1.1	Model teoretyczny	4
2.2	Aktualna wiedza	6
2.2.1	Algorytm Apriori	6
	Generowanie zbiorów częstych	7
	Generowanie reguł asocjacyjnych	9
2.2.2	Algorytm FP-growth	10
	Przykład budowy FP-drzewa	11
	Kompresja bazy danych	13
	Konstrukcja FP-drzewa	13
	Eksploracja FP-drzewa	14
	Przykład eksploracji FP-drzewa	15
3	Uniwersalna architektura procesorów wielordzeniowych	18
3.1	Czasy przetwarzania równoległego	18
3.1.1	Procesor	19
3.2	Wpływ procesorów graficznych na procesy obliczeniowe	21
3.2.1	Krótką historia kart graficznych	21
3.2.2	Wczesne obliczenia na kartach graficznych	23
3.3	CUDA	24
3.3.1	Czym jest architektura CUDA?	25
3.3.2	Od wyświetlania grafiki do obliczeń ogólnego przeznaczenia	25
3.3.3	Skalowalny model programistyczny	26
3.3.4	Używanie architektury CUDA	28
3.4	Kwantyfikator parallel czyli CUDA	28
4	Technologie	30
4.1	Język C# i platforma .NET	30

4.1.1	Platforma .NET	31
	Język, środowisko uruchomieniowe i biblioteki	31
4.1.2	Kompilacja i język MSIL	32
4.1.3	Język programowania C#	33
4.1.4	Zapytania do obiektów czyli LINQ	34
4.1.5	Kwantyfikator as parallel czyli Parallel LINQ	35
4.1.6	.NET Framework 4.0	36
4.2	CUDA	36
4.2.1	CUDA na platformie .NET	36
4.3	Środowisko programistyczne	37
4.3.1	Visual Studio 2010	37
4.3.2	System kontroli wersji	38
4.3.3	Budowanie projektu	38
4.4	Testowanie	39
4.4.1	Usprawnienie procesu testowania	39
	Code Contracts	39
	Pex i Moles	40
5	Algorytm równoległy	42
5.1	Algorytm ParallelApriori	42
5.1.1	Generowanie zbiorów częstych	42
	Procedura pAprioriGen	43
5.1.2	Generowanie reguł asocjacyjnych	44
5.2	Algorytm CudaApriori	45
5.2.1	Przygotowanie danych wejściowych do algorytmu	45
5.2.2	Ograniczenie bazy danych	46
5.2.3	Generowanie zbiorów częstych	47
5.2.4	Generowanie reguł asocjacyjnych	49
6	Eksperymenty	50
6.1	Parametry wejściowe oraz maszyna testowa	50
6.2	Wyniki eksperymentów	50
6.2.1	Wyniki pomiarów poszczególnych implementacji algorytmów	50
	Wyniki dla minsup = 0.2, minconf = 0.1	50
	Wyniki dla minsup = 0.2, minconf = 0.4	51
	Wyniki dla minsup = 0.05, minconf = 0.1	51
	Wyniki dla minsup = 0.05, minconf = 0.4	51
6.2.2	Wyniki pomiarów dla algorytmu CudaApriori	51
6.3	Podsumowanie eksperymentów i wnioski	51

7 Podsumowanie i wnioski	61
Literatura	62

Rozdział 1

Wstęp

Proces informatyzacji przedsiębiorstw, rozpoczęty kilka dekad temu, wprowadził światową gospodarkę na nowe, dotąd nieznane tory rozwoju. Skrócenie procesu produkcyjnego, wprowadzenie kontroli komputerowych, czy też skomputeryzowanych maszyn skróciło i ułatwiło produkcję, a także zarządzanie procesami w firmach i przedsiębiorstwach. Przed ludźmi stanęły możliwości, ale także wyzwania, z którymi nigdy wcześniej nikt nie musiał sobie radzić. Zmiany, jakie nastąpiły przez ostatnie trzy dekady są nieodwracalne i zmuszają programistów do tworzenia nowych aplikacji, które będą w stanie sprostać stawianym im wymaganiom.

Informatyzacja firm, instytucji oraz innych jednostek organizacyjnych powinna realizować dwa podstawowe cele. Z jednej strony powinna ona usprawniać pracę pojedynczego pracownika poprzez automatyzację realizowanych przez niego rutynowych zadań. Dzięki wykorzystaniu możliwości komputerów działania te powinny być wykonywane szybciej i w sposób bardziej niezawodny. Z drugiej strony celem informatyzacji jest wpływanie na działanie całych firm w wyniku wspomagania decyzji kadry zarządzającej przedsiębiorstwami. Szybka analiza bazująca na pełnej i aktualnej informacji o stanie firmy może ułatwić kadrze zarządzającej podejmowanie trafnych i szybkich decyzji o strategicznym znaczeniu dla rozwoju danego przedsiębiorstwa.

Wprowadzenie komputerów do właściwie każdej przestrzeni ludzkiego życia wpłynęło na wyprodukowanie olbrzymich ilości danych. Reprezentowane są one w sposób umożliwiający ich składowanie i przetwarzanie komputerowe przez aplikacje analityczne. W chwili obecnej ludzkość jest świadkiem eksplozji w produkcji danych produkowanych przez różnego rodzaju systemy komputerowe. Analiza tych danych przynieść może wymierne korzyści nie tylko w kwestiach finansowych, ale również poznawczych. Dzięki analizie zebranych w przeszłości informacji możliwe jest lepsze dopasowanie planów w przyszłości - na tej podstawie planowane mogą być np. akcje marketingowe, czy też promocje w supermarketach spożywczych. Wykorzystanie wiedzy uzyskanej w ten sposób jest niezwykle szerokie i może być użyte w każdym obszarze działalności firmy.

Odkrycie zależności pomiędzy zgromadzonymi danymi bez zastosowania narzędzi informatycznych jest procesem bardzo skomplikowanym i wymagającym do realizacji dużo czasu. Przy obecnej złożoności większości systemów oraz rozmiarom danych produkowanych przez te systemy, koszt czasowy jest na tyle duży, że ręczna analiza tych danych stała się niemożliwa. Dlatego też tworzone

są narzędzia umożliwiające odkrywanie prawidłowości w dużych zbiorach danych, by człowiek na tej podstawie mógł podejmować decyzje i wyciągać wnioski.

Dział informatyki, który zajmuje się odkrywaniem ukrytych dla człowieka prawidłowości i reguł w danych nazywa się eksploracją danych (ang. *Data Mining*, w literaturze spotkać można również określenie drażenie danych, ekstrakcja danych, pozyskiwanie wiedzy, czy też wydobywanie danych [EN05]), który jest jednym z etapów procesu *odkrywania wiedzy z baz danych* (*KDD*, ang. *Knowledge Discovery in Databases*). Proces odkrywania wiedzy w bazach danych obejmują zwykle działania bardziej złożone niż tylko eksploracja danych. Są to między innymi selekcja danych, transformacja lub kodowanie danych, czy też raportowanie i prezentowanie odkrytych informacji [EN05]. Eksploracja danych to proces odkrywania wiedzy w postaci nowych, użytecznych, poprawnych i zrozumiałych wzorców w bardzo dużych wolumenach danych [FPSSU96]. Możliwość stosowania technik eksploracji danych w praktyce, wymagają efektywnych metod przeszukiwania ogromnych plików lub baz danych. Warto przy tym wspomnieć, że tego typu technologie nie są w chwili obecnej dobrze zintegrowane z systemami zarządzania bazami danych.

Eksploracja danych odbywa się najczęściej w środowisku baz lub hurtowni danych, które stanowią doskonałe źródła danych do analizy - głównie ze względu na łatwość dostępu oraz usystematyzowaną strukturę przechowywanych informacji. Ponieważ liczba odkrytych wzorców w wielu przypadkach może być bardzo duża, odkryte wzorce bardzo często zapisuje się w osobnych relacjach bazy lub hurtowni danych. Pozwala to na ich dalsze przetwarzanie w trybie off-line przez użytkowników końcowych. Pojęcie eksploracji zyskuje coraz większą popularność (również w wymiarze marketingowym) i jest wykorzystywane w wielu dziedzinach ludzkiego życia.

Jednym z najczęściej wykorzystywanych modeli wiedzy w eksploracji danych są reguły asocjacyjne. Reguła asocjacyjna ma postać $X \Rightarrow Y$, gdzie X oraz Y są wzajemnie rozłącznymi zbiorami elementów. Przykładem reguły, która mogła zostać odkryta w bazie danych sklepu komputerowego, może być reguła postaci *komputer* \wedge *myszka* \Rightarrow *monitor*. Prezentuje ona fakt, że klienci kupujący komputer oraz myszkę z dużym prawdopodobieństwem kupią również monitor. W [AIS93] po raz pierwszy sformułowany został problem odkrywania reguł asocjacyjnych wraz z algorytmem Apriori, który jest podstawą wielu algorytmów znajdujących reguły asocjacyjne. Algorytm ten został następnie rozszerzony w pracy [Agr94].

W ostatnich latach pojawiły się nowe możliwości wykorzystania współczesnych komputerów. W roku 2007 firma NVIDIA udostępniła programistom uniwersalną architekturę obliczeniową CUDA (ang. *Compute Unified Device Architecture*), który umożliwia wykorzystanie mocy obliczeniowej procesorów graficznych (*GPU*, ang. *Graphics Processing Unit*), bądź innych procesorów wielordzeniowych, do rozwiązywania ogólnych problemów obliczeniowych w sposób znacząco wydajniejszy niż w przypadku tradycyjnych, sekwencyjnych procesorów [Cor07a]. Choć w grach komputerowych moc obliczeniową jednostek graficznych można wykorzystać do obliczeń fizyki, to CUDA idzie jeszcze dalej, umożliwiając przyspieszenie obliczeń w takich dziedzinach, jak biologia, fizyka, kryptografia, bioinformatyka oraz innych naukach. Specjalnie dla potrzeb tego segmentu NVidia opracowała kartę graficzną o nazwie *Tesla* [Cor07b]. Układy te są pierwszymi układami

produkowanymi na masową skalę, które przeznaczone zostały do pracy *obliczeniach ogólnego przeznaczenia na układach GPU (GPGPU, ang. General-Purpose Computing on Graphics Processing Units)*, czyli segmencie do tej pory zarezerwowanym dla klasycznych procesorów obliczeniowych.

Do tej pory bardzo małe jest zainteresowanie wykorzystaniem tej technologii w procesie odkrywania wiedzy, a w szczególności znajdowania reguł asocjacyjnych. Wyniki przeprowadzonych eksperymentów pozwalają przypuszczać, że algorytm wykorzystujący możliwości procesorów wielordzeniowych, a w szczególności GPU, będzie wyraźnie szybszy od klasycznych algorytmów eksploracji danych zaimplementowany na tradycyjnych procesorach.

1.1 Cel i zakres pracy

Celem pracy jest zaprojektowanie i zaimplementowanie algorytmu odkrywającego reguły asocjacyjne, który będzie wykorzystywał możliwości współczesnych kart graficznych dzięki wykorzystaniu technologii CUDA oraz porównanie zaprojektowanego i zaimplementowanego algorytmu do innych, podstawowych algorytmów odkrywania reguł asocjacyjnych. W ramach pracy dokonane zostanie również zebranie wiedzy dotyczącej algorytmów eksploracji reguł asocjacyjnych.

Niniejszy rozdział jest wprowadzeniem do pracy, w której opracowany i zaimplementowany został równoległy algorytm odkrywający reguły asocjacyjne. W rozdziale 2 zebrano aktualną wiedzę na temat algorytmów oraz zagadnienia związane z odkrywaniem reguł asocjacyjnych - w tym dwa podstawowe algorytmy, jakimi są Apriori oraz FPGrowth. Kolejna ważna część pracy, to technologia CUDA, którą opisano w rozdziale 3. Pozostałe użyte w pracy technologie przedstawiono w rozdziale 4. Rozdział 5 zawiera opis równoległego algorytmu korzystającego z możliwości kart graficznych. Wyniki z wykonanych eksperymentów zostały opracowane i zebrane w rozdziale 6. Podsumowanie i wnioski do pracy

Rozdział 2

Podstawy teoretyczne

W rozdziale tym przedstawiony zostanie przegląd literatury, który stanowi podstawy wiedzy na temat eksploracji danych, a w szczególności problemu odkrywania reguł asocjacyjnych w dużych zbiorach danych. Zebrana wiedza posłużyła autorowi do opracowania algorytmu wykorzystującego możliwości współczesnych kart graficznych.

2.1 Definicje

2.1.1 Model teoretyczny

Niech $I = \{i_1, i_2, \dots, i_m\}$ będzie *zbiorem elementów* o liczności $|I| = m$. *Transakcją* T nazwano dowolny, niepusty podzbiór $X \subseteq I$ zbioru elementów. Bazą danych DB nazwano dowolny zbiór par (id, X) , gdzie X jest transakcją, a id jest dowolną wartością unikalną w ramach bazy danych nazywaną *identyfikatorem transakcji*. Bez utraty ogólności założono iż $id \in \mathbb{N}$.

W tabeli 2.1 zaprezentowany został przykładowy zbiór elementów oraz zestaw transakcji. Na podstawie tych danych obliczane będą wartości wprowadzanych kolejno definicji.

Wsparciem (ang. *support*) $sup(X)$ transakcji X , w bazie danych nazwano częstość wystąpień transakcji w bazie danych. Formalnie przedstawia to wzór 2.1.

$$sup(X) = \frac{|\{id : (id, Y) \in DB \wedge X \subseteq Y\}|}{|DB|} \quad (2.1)$$

Łatwo zauważyć, że jeśli poziom ten jest niski, to oznacza to, że elementy zbioru X w transakcjach rzadko występują razem.

I	{ beer, bread, butter, diapers, jam, juice, milk, water }
Transakcje	1: { bread, milk, butter, beer } 2: { bread, butter, water, jam, beer } 3: { beer, diapers, bread, butter, jam } 4: { butter, milk, juice } 5: { diapers, beer, juice, water }

TABLICA 2.1: Przykładowe dane

Dla przykładowego zbioru $X = \{\text{milk}\}$ wartość $\text{sup}(X)$ została obliczona w przykładzie 2.2, ponieważ zbiór X jest podzbiorem dwóch transakcji - $t_1 = \{\text{bread, milk, butter, beer}\}$ oraz $t_4 = \{\text{butter, milk, juice}\}$.

$$\begin{aligned} \text{sup}(X) = \text{sup}(\{\text{milk}\}) &= \frac{|\{1, 4\}|}{|DB|} = \\ &= \frac{2}{5} \end{aligned} \quad (2.2)$$

Definicja 1. Niech będą dane dwie transakcje X i Y takie, że $X \cap Y = \emptyset$ oraz $Y \neq \emptyset$. Regułą asocjacyjną R nazwano implikację postaci $X \Rightarrow Y$.

Poziom ufności (ang. *confidence*) jest miarą określającą jakość reguły asocjacyjnej [EN05].

Definicja 2. Poziom ufności (*conf*) reguły asocjacyjnej $R: X \Rightarrow Y$ jest równy

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)} \quad (2.3)$$

Z definicji 2 wynika, że poziom ufności może być interpretowany, jako estymacja prawdopodobieństwa w transakcji zbioru Y pod warunkiem wystąpienia w niej również zbioru X - co oznaczone jest poprzez $P(Y|X)$. Formalnie zapisane jest to za pomocą wzoru 2.4 [Mue95].

$$\text{conf}(X \Rightarrow Y) = p(Y \subseteq T | X \subseteq T) = \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)} \quad (2.4)$$

Reguły asocjacyjne zazwyczaj powinny spełniać pewne wymagania zdefiniowane przez użytkownika - minimalne wsparcie oraz minimalny poziom ufności, oznaczane odpowiednio *minsup* oraz *minconf*. Wyznaczają one dla aplikacji progi, jakie powinny spełniać zbiory oraz reguły, aby były brane pod uwagę w trakcie analizy. Motywacją za tymi minimalnymi wartościami jest fakt, by do analizy brane były tylko te zbiory, które pojawiają się w DB wystarczającą liczbę razy.

Przykładową regułą asocjacyjną znaną dla danych przedstawionych w tabeli 2.1 może być reguła $\{\text{butter}\} \Rightarrow \{\text{bread, beer}\}$, która zostałaby znaleziona dla *minsup* = 0,6. Taka reguła posiada współczynnik pewności równy $\frac{3}{4} = 75\%$, obliczony w przykładzie 2.5.

$$\begin{aligned} \text{conf}(X \Rightarrow Y) &= \text{conf}(\{\text{butter}\} \Rightarrow \{\text{bread, beer}\}) = \\ &= \frac{\text{sup}(\{\text{butter}\} \cup \{\text{bread, beer}\})}{\text{sup}(\{\text{butter}\})} = \\ &= \frac{\text{sup}(\{\text{butter, bread, beer}\})}{\text{sup}(\{\text{butter}\})} = \\ &= \frac{3}{4} = 75\% \end{aligned} \quad (2.5)$$

Definicja 3. Zbiorem częstym $X \subseteq I$ nazywamy taki zbiór, który spełnia zależność $\text{sup}(X) \geq \text{minsup}$.

Generowanie reguł asocjacyjnych zazwyczaj sprowadza się do dwóch, niezależnych kroków:

1. Minimalne wsparcie jest używane do odnalezienia wszystkich zbiorów częstych w bazie danych DB .

2. Znalezione zbiory częste oraz minimalny poziom ufności są używane do wygenerowania reguł asocjacyjnych.

W naiwnym podejściu znalezienie wszystkich zbiorów częstych w bazie danych DB jest zadaniem wymagającym przeszukania wszystkich możliwych kombinacji bez powtórzeń ze zbioru I . Zbiór X nazywamy k -zbiorem, jeśli $|X| = k$, tzn. zbiór X ma k elementów. Zbiór możliwych zbiorów elementów ma licznosc równą $2^n - 1$ (wszystkie zbiory, poza zbiorem pustym, który nie jest w tym wypadku zbiorem sensownym w znaczeniu poddania go analizie), czyli zbiór ten jest $(2^n - 1)$ -zbiorem.

Warto zauważyć, że dla każdego zbioru częstego Y , każdy jego podzbiór X jest również zbiorem częstym [AIS93]. Korzystając z tej właściwości wsparcia możliwe jest w sposób efektywny znalezienie wszystkich zbiorów częstych w zadanej bazie danych - z tej zależności korzysta algorytm apriori opisany w rozdziale 2.2.1. Dodatkowo, wszystkie reguły zbudowane na podstawie zbioru częstego Y muszą spełniać warunek minimalnego wsparcia, ponieważ spełnia ten warunek zbiór Y , a suma zbiorów reguły jest zbiorem wyjściowym Y .

2.2 Aktualna wiedza

W rozdziale tym zebrana została oraz opracowana dotychczasowa wiedza (ang. *state-of-the-art*) na temat algorytmów odkrywania reguł asocjacyjnych. Przedstawione zostaną dwa podstawowe algorytmy wykorzystywane w tym procesie: Apriori oraz FP-growth. W chwili obecnej te dwa algorytmy stanowią podstawę, na której budowane są nowe algorytmy, wykorzystujące możliwości współczesnych algorytmów - bazujących na FPGrowth: [LQ05, ZY08] lub na algorytmie Apriori: [PCY95, YC06].

2.2.1 Algorytm Apriori

Pierwszy algorytm odkrywający reguły asocjacyjne został przedstawiony w roku 1994 w pracy [AS94], który jest rozszerzeniem algorytmu zaprezentowanego przez autorów w [AIS93]. Niżej przedstawione zostaną szczegóły działania tego algorytmu nazwanego algorytmem *Apriori*.

Zagadnienie odkrywania reguł asocjacyjnych można podzielić na dwa etapy [AIS93]:

1. Odkrywanie zbiorów częstych, których wartość wsparcia jest wyższa od wartości *minsup*.
2. Generowanie reguł asocjacyjnych na podstawie znalezionych zbiorów częstych. Reguła $X \Rightarrow Y$ jest wynikiem działania algorytmu dla zbioru $Z = X \cup Y$, jeżeli spełnia ona nierówność $conf(X \Rightarrow Y) \geq minconf$. Ponieważ zbiór $Z = X \cup Y$ jest zbiorem częstym, to reguła spełnia również warunek przekraczania minimalnego wsparcia.

Na tym etapie możliwe jest tworzenie reguł, w których w zbiorze *poprzedników* (X z oznaczeń z definicji 1) jest wiele elementów oraz jeden w *następniku* (zbiór Y z definicji 1) [AIS93] lub dopuszczana jest możliwość wielu elementów również w następniku [AS94]. W niniejszej

L_k	Zbiór zawierający k -zbiory. Każdy zbiór zawarty w L_k zawiera dwa pola: i) zbiór oraz ii) wartość <i>support</i> .
C_k	Zbiór k -zbiorów kandydatów (potencjalnych zbiorów częstych). Każdy zbiór zawarty w C_k zawiera dwa pola: i) zbiór oraz ii) wartość <i>support</i> .

TABLICA 2.2: Oznaczenie w opisach algorytmów

pracy analizowany jest sposób generowania reguł, w którym oba zbiory mogą być zbiorami wieloelementowymi.

W kolejnych podrozdziałach przedstawione zostaną etapy tworzące razem algorytmy Apriori.

Generowanie zbiorów częstych

W celu wyznaczenia zbiorów częstych algorytm dokonuje analizy bazy danych DB , by w kolejnych iteracjach generować rodziny coraz to liczniejszych zbiorów, będących zbiorami częstymi dla zadanej wartości *minsup*. Algorytm zaczyna od znalezienia wszystkich zbiorów jednoelementowych, które są zbiorami częstymi. W każdym kolejnym kroku generowane są zbiory częste na podstawie zbiorów wygenerowanych w kroku poprzednim. Proces ten jest kontynuowany do momentu aż nie zostaną znalezione żadne zbiory częste.

Algorytm generuje zbiory kandydatów jedynie na podstawie zbiorów częstych odkrytych w kroku poprzednim - co ważne generowanie ich odbywa się bez wielokrotnego przeglądania bazy danych transakcji. Intuicja podpowiada, że każdy podzbiór zbioru częstego jest zbiorem częstym. Zatem, każdy zbiór częsty zawierający k elementów może być wygenerowany na podstawie połączenia dwóch zbiorów posiadających $k - 1$ elementów, a na koniec kasując te zbiory, których jakikolwiek podzbiór nie jest częsty [AS94].

Tabela 2.2 zawiera spis oznaczeń używanych w opisie algorytmu.

Procedura APRIORI FREQUENT SET GENERATION przedstawia pseudokod realizujący opisywany w tym rozdziale algorytm generowania zbiorów częstych.

APRIORI FREQUENT SET GENERATION

```

1   $L_1 \leftarrow \{1\text{-zbiory częste} \}$ 
2  for ( $k = 2; L_{k-1} \neq \emptyset; k++$ )
3      do  $C_k \leftarrow \text{aprioriGen}(L_{k-1})$ 
4          for each transakcja  $t \in DB$ 
5              do  $C_t \leftarrow \text{subset}(C_k, t)$ 
6                  for each kandydat  $c \in C_t$ 
7                      do  $c.\text{count}++$ 
8           $L_k \leftarrow \{c \in C_k | c.\text{count} \geq \text{minsup}\}$ 
9   $\text{Answer} \leftarrow \bigcup_k L_k$ 

```

Procedura aprioriGen Procedura *aprioriGen* reprezentuje proces tworzenia zbiorów k -elementowych kandydatów na podstawie zbiorów wejściowych $(k - 1)$ -elementowych. Procedura ta jest podzielona na dwa etapy: łączenia oraz przycinania.

Jak łatwo zauważyć wynikiem działania JOIN STEP są zbiory k -elementowe, które powstały na podstawie zbiorów wejściowych L_{k-1} , a ich zawartość różni się tylko jednym elementem - ostatnim. Ważnym faktem jest to, iż elementy w zbiorach są uporządkowane leksykograficznie, co wykorzystywane jest w tej procedurze.

JOIN STEP

```

1  insert into  $C_k$ 
2  select p.item1, p.item2, ..., p.item $k-1$ , q.item $k-1$ 
3  from  $L_{k-1}$  p,  $L_{k-1}$  q
4  where p.item1 = q.item1, ..., p.item $k-2$  = q.item $k-2$ , p.item $k-1$  < q.item $k-1$ 

```

Warto zauważyć, że JOIN STEP jest ekwiwalentem rozszerzania zbioru L_{k-1} (p) każdym elementem zbioru elementów I , a następnie kasowania tych (k) -zbiorów otrzymanych przez dodanie (k) -tego elementu, który nie był w L_{k-1} (q).

Warunek p.item _{$k-1$} < q.item _{$k-1$} zapewnia, że nie będą generowane duplikaty. Dlatego też po etapie łączenia zachodzi zależność $C_k \supseteq L_k$.

Następnym krokiem jest PRUNE STEP, w którym usuwane są wszystkie elementy $c \in C_k$, którego jakiegokolwiek podzbiór $(k - 1)$ -elementowy zbioru c nie należy do L_{k-1} .

PRUNE STEP

```

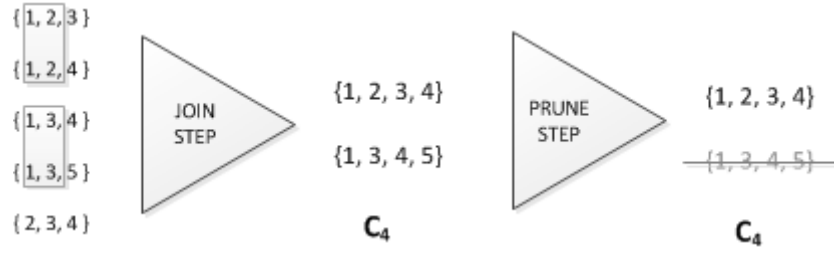
1  for each zbiór  $c \in C_k$ 
2      do
3          for each  $(k - 1)$ -podzbiór  $s$  zbioru  $c$ 
4              do if  $s \notin L_{k-1}$ 
5                  then delete  $c$  z  $C_k$ 

```

Celem operacji przycinania (ang. *prune*) jest ograniczenie rozmiaru zbioru C_k przed sprawdzeniem wsparcia dla kandydatów w bazie danych DB . W tym celu wykorzystywana jest właściwość, z której wynika, że jeśli jakiś $(k - 1)$ -podzbiór danego kandydata ($c \in C_k$) nie występuje w L_{k-1} , to kandydat c nie jest zbiorem częstym i powinien być usunięty z C_k .

Przykład działania procedury aprioriGen Niech zbiór L_3 będzie równy $\{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}\}$. Po działaniu procedury JOIN STEP wartość zbioru C_4 będzie zatem równa $\{\{1, 2, 3, 4\}, \{1, 3, 4, 5\}\}$. W następnym kroku, czyli wewnątrz procedury PRUNE STEP usunięty zostanie z C_4 element $\{1, 3, 4, 5\}$ (patrz linia 5), ponieważ element $\{1, 4, 5\}$ nie należy do L_3 (patrz linia 4). Dlatego też $C_4 = \{\{1, 2, 3, 4\}\}$.

Opisane działanie zostało przedstawione schematycznie na rysunku 2.1, na którym zaznaczono poszczególne etapy wraz z zaznaczeniem części wspólnych poszczególnych zbiorów wejściowych (patrz procedura JOIN STEP).



RYSUNEK 2.1: Zilustrowanie działania procedury aprioriGen

Generowanie reguł asocjacyjnych

Po zakończeniu pierwszego etapu algorytm przystępuje do drugiego, czyli do budowania reguł asocjacyjnych na podstawie odkrytych zbiorów. Podobnie, jak w [AS94] algorytm będący przedmiotem analizy niniejszej pracy, generuje wszystkie możliwe reguły asocjacyjne dla zadanego zbioru. Mniej ogólny sposób generowania reguł został przedstawiony w pracy [AIS93], jednakże podjęto decyzję, że jest to sposób zbyt mało użyteczny w środowisku produkcyjnym.

Aby wygenerować reguły, dla każdego zbioru częstego l znajdowane są niepuste podzbiory - podzbiór taki oznaczony jest jako a . Dla takich oznaczeń wygenerowana zostanie reguła $a \Rightarrow (l - a)$, jeżeli spełniona jest nierówność $\frac{\text{support}(l)}{\text{support}(a)} \geq \text{minconf}$. Warto zauważyć, że dla każdego zbioru częstego generowane są wszystkie możliwe niepuste podzbiory - zapewnia to, że odkryte zostaną wszystkie możliwe reguły.

Istnieje możliwość poprawienia tego podejścia poprzez generowanie podzbiorów dużego zbioru w sposób rekurencyjnego wywołania *najpierw w głąb* (ang. *depth-first fashion*). Na przykład, dla zbioru $\{A, B, C, D\}$, najpierw rozważany jest podzbiór $\{A, B, C\}$, potem $\{A, B\}$ itd. Wtedy, jeśli podzbiór a zbioru l nie jest źródłem reguły asocjacyjnej, podzbiór a nie musi być rozważany w generowaniu reguł ze zbioru l . Dla przykładu, jeśli reguła $\{A, B, C\} \Rightarrow \{D\}$ nie posiada wystarczającego współczynnika conf , wówczas nie ma potrzeby sprawdzać reguły $\{A, B\} \Rightarrow \{C, D\}$. Nie jest pominięta żadna możliwa reguła, ponieważ wartość sup dowolnego $\tilde{a} \in a$ nie może być większe niż wartość sup dla zbioru a . Dlatego też, wartość conf dla reguły $\tilde{a} \Rightarrow (l - \tilde{a})$ nie może być większa niż poziom ufności $a \Rightarrow (l - a)$. Stąd, jeżeli a nie zwrócił reguły zawierającej wszystkich elementów z l w a , jako poprzednika, nie dokona tego również \tilde{a} [AS94].

Procedura GENERATE FREQUENT ITEMSETS prezentuje generowanie reguł asocjacyjnych na podstawie odkrytych k -zbiorów częstych l_k będących elementami zbioru L_k ($l_k \in L_k$).

GENERATE FREQUENT ITEMSETS

```

1  for each zbiór częsty  $l_k, k \geq 2$ 
2      do call genrules( $l_k, l_k$ )

```

W powyższym algorytmie wykorzystana została funkcja GENRULES, która na podstawie dwóch zbiorów generuje reguły asocjacyjne. Zapis pseudokodu tej funkcji przedstawiony jest poniżej.

TID	Lista elementów	Posortowana lista elementów częstych
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

TABLICA 2.3: Przykładowy zbiór transakcji

GENRULES(l_k : k -zbiór częsty, a_m : m -zbiór częsty)

```

1   $A \leftarrow \{(m-1)\text{-zbiór } a_{m-1} | a_{m-1} \subset a_m\}$ 
2  for  $a_{m-1} \in A$ 
3      do  $conf \leftarrow \frac{support(l_k)}{support(a_{m-1})}$ 
4      if  $conf \geq minconf$ 
5          then output reguła  $a_{m-1} \Rightarrow (l_k - a_{m-1})$  //ufność =  $conf$  oraz wsparcie=  $support(l_k)$ 
6          if  $m - 1 > 1$ 
7              then call genrules( $l_k, a_{m-1}$ ) //generowanie reguł podzbiorów zbioru  $a_{m-1}$ 
```

2.2.2 Algorytm FP-growth

Podstawową wadą algorytmu Apriori jest wysoki koszt przetwarzania dużych zbiorów danych. Przykładowo, dla 10^4 1-zbiorów częstych, algorytm Apriori wygeneruje około 10^7 2-zbiorów kandydatów, które następnie poddane zostaną weryfikacji, czy są zbiorami częstymi. Poza tym algorytm ten wymaga wielokrotnego odczytywania zawartości bazy danych - w każdym kroku algorytmu należy odczytać całą bazę danych w celu obliczenia wsparcia zbiorów kandydujących.

Wymienione wyżej wady algorytmu Apriori nie występują w algorytmie *FP-growth* przedstawionym w pełni w pracy [HPYM04]. Algorytm ten pozwala wyeliminować konieczność generowania tak dużej liczby kandydujących zbiorów elementów oraz ogranicza liczbę dostępu do bazy danych do absolutnego minimum. Co więcej algorytm ten charakteryzuje się kompletnością, co oznacza, że znajdowane są wszystkie wzorce o określonej częstości.

Algorytm FP-growth można podzielić na trzy podstawowe kroki.

1. W kroku pierwszym generowana jest skompresowana wersja bazy danych DB , mająca postać drzewa częstych wzorców - jako posortowane listy elementów częstych w każdej z transakcji.
2. Drugim krokiem jest transformacja tak skonstruowanego drzewa do postaci *FP-drzewa* (patrz definicja 4).
3. Trzeci krok polega na analizie FP-drzewa celem odnalezienia reguł asocjacyjnych. W kroku tym stosowana jest metoda dziel i zwyciężaj (ang. *divide-and-conquer*) zamiast podejścia Apriori, czyli generowania na każdym poziomie zbioru kandydatów na zbiory częste, a następnie odcinaniu kandydatów nie spełniających kryteriów akceptacji. Takie podejście przekształca problem znajdowania długich reguł w problem szukania krótszych, a następnie konkatencji wyników.

Przykład budowy FP-drzewa

Niech zbiór transakcji DB będzie zdefiniowany przez tablicę 2.3. Dla przykładu przyjęto, że $minsup = 3$.

1. Ponieważ w budowie FP-drzewa (patrz definicja 4) będą uczestniczyć jedynie elementy będące elementami częstymi, to pierwszym etapem algorytmu jest jednokrotne przejście za wartości DB oraz zidentyfikowanie zbioru elementów częstych (z wartością f_{count} dla każdego elementu, przechowujący liczbę jego wystąpień).
2. Jeśli zbiory elementów częstych, dla każdej z transakcji, mogą być przechowywane w relatywnie małej strukturze - warto jest zachować wynik skanowania, by uniknąć w kolejnych etapach kolejnych, niepotrzebnych i wyraźnie obniżających wydajność algorytmu, skanów bazy danych.
3. Jeśli kilka transakcji posiada te same zbiory elementów częstych, możliwe jest przechowywanie tych zbiorów jako jednego, z licznikiem w ilu transakcjach on występuje ($count$). Czy dwa zbiory są identyczne można w łatwy sposób zweryfikować, jeśli w każdym ze zbiorów są posortowane zgodnie z przyjętym dla wszystkich porządkiem.
4. Możliwe jest również częściowe współzapamiętywanie zbiorów, jeśli prefiksy (czyli części zbiorów posortowanych) są identyczne dla dwóch transakcji. Te same części mogą być scalone do jednego prefiksu, jeśli wartość $count$ jest właściwie obliczona. Jeśli elementy częste są posortowane w porządku malejącej liczby wystąpień, istnieje większe prawdopodobieństwo znalezienia większej liczby wspólnych prefiksów.

Dla danych zawartych w tabeli 2.3 przedstawione zostanie teraz działanie algorytmu.

Na początek, skan bazy danych DB tworzy listę elementów częstych: $\langle f : 4, c : 4, a : 3, b : 3, m : 3, p : 3 \rangle$ (gdzie para $a : 0 \dots n$ reprezentuje odpowiednio - element a oraz liczbę jego wystąpień), w której elementy posortowane są zgodnie z malejącą liczbą wystąpień. Sposób sortowania jest istotny, ponieważ proces budowy drzewa będzie właśnie oparty na tym porządku. Dla wygody analizy oraz łatwiejszego zrozumienia, pierwszą kolumnę z prawej w tablicy 2.3 tworzą elementy częste z danej transakcji, posortowane właśnie w ten sposób.

Co więcej, korzeń drzewa jest tworzony i oznaczony jako $null$, czyli wartość pusta. FP-drzewo jest tworzone poprzez skanowanie raz jeszcze bazy danych DB i działania opisane przez kolejne kroki poniżej.

1. Skan pierwszej transakcji ($TID = 100$) prowadzi do wprowadzenia w drzewie pierwszej gałęzi: $\langle f : 1, c : 1, a : 1, m : 1, p : 1 \rangle$. Warto zauważyć, że elementy częste z danej transakcji są wylistowane zgodnie z porządkiem w liście elementów częstych.
2. Analizując drugą transakcję ($TID = 200$), dla której lista elementów częstych $\{f, c, a, b, m\}$ współdzieli ten sam prefiks $\{f, c, a\}$ z istniejącą już ścieżką $\{f, c, a, m, p\}$, dlatego też liczba

wystąpień na tej ścieżce inkrementowana jest o 1, a także tworzony jest nowy element $b : 1$ oraz dodawany jako potomek elementu $a : 2$, a następny nowy element $m : 1$ jest dodawany, jako potomek elementu $b : 1$.

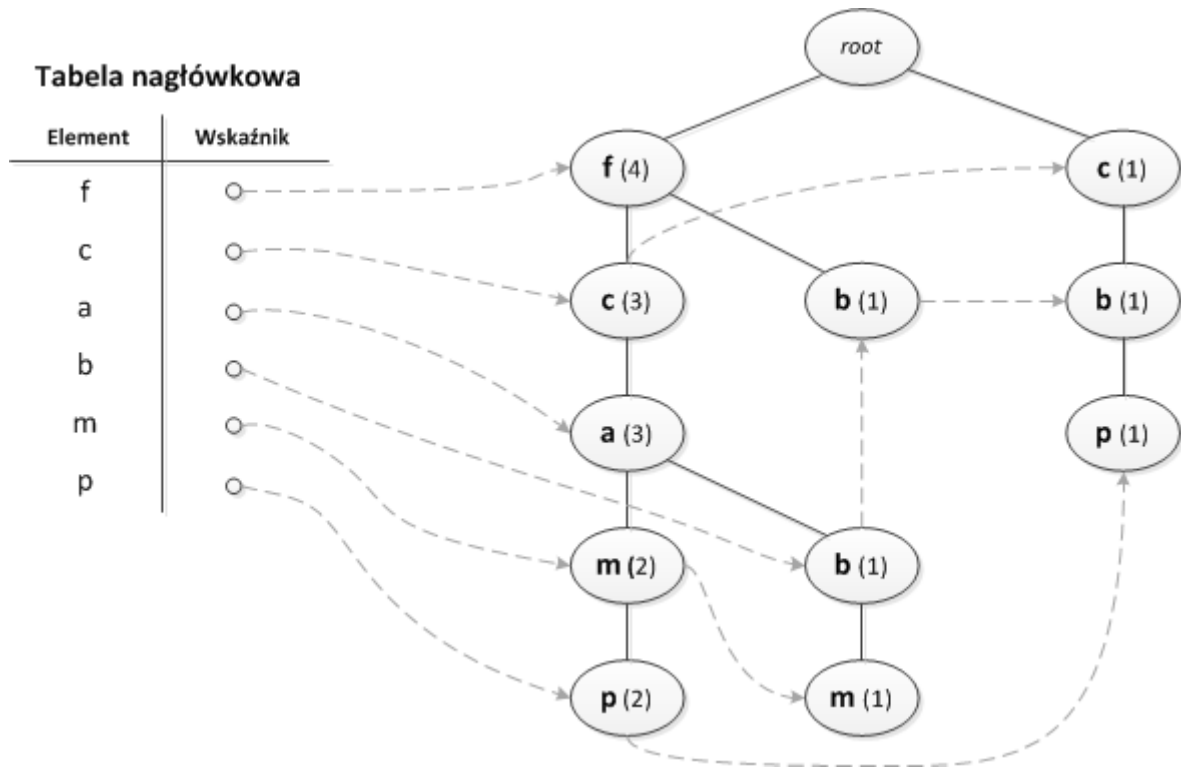
3. Trzecia transakcja ($TID = 300$) $\{f, b\}$ współdzieli jedynie pierwszy element $\{f\}$ z f -poddrezwem prefikсовym, wartość *count* dla elementu $f : 2$ jest inkrementowana o 1, a nowy element $b : 1$ jest dodawany, jako bezpośredni potomek elementu $f : 3$.
4. Skan czwartej transakcji ($TID = 400$) prowadzi do utworzenia nowej gałęzi drzewa $\langle c : 1, b : 1, p : 1 \rangle$, ponieważ transakcja ta nie współdzieli żadnego fragmentu z istniejącą już ścieżką w FP-drzewie.
5. Ostatnia transakcja ($TID = 500$) posiada identyczną listę elementów częstych, jak transakcja pierwsza ($TID = 100$), dlatego też elementy na tej ścieżce są inkrementowane odpowiednio o 1.

By ułatwić *przechodzenie drzewa* (ang. *tree traversal*) stworzona została również tablica nagłówkowa, w której każdy element posiada wskaźnik do elementu drzewa. Wierzchołki reprezentujące ten sam element, są oprócz tego połączone bezpośrednio między sobą. Po przeskanowaniu wszystkich transakcji, drzewo razem z połączeniami pomiędzy wierzchołkami zostało zaprezentowane na rysunku 2.2.

Definicja 4. FP-drzewo (ang. *frequent-pattern tree*) jest to ukorzeniony, etykietowany w wierzchołkach graf acykliczny spełniający poniższe cechy.

1. Korzeniem drzewa jest jeden element *null*, zbiór poddrzew prefiksowanych elementami (jako dzieci elementu *null*) oraz tablicy nagłówkowej zawierającej wpisy *element* \rightarrow *wskaźnik na element drzewa*.
2. Każdy wierzchołek poddrzewa składa się z trzech elementów: nazwy elementu (ang. *item name*), licznika (ang. *count*) oraz wskaźnika na inny wierzchołek. Nazwa elementu (*itemName*) w sposób jednoznaczny identyfikuje element ze zbioru elementów I , licznik przechowuje liczbę transakcji reprezentowanych przez ścieżkę od *null* do tego elementu, natomiast wskaźnik wskazuje na kolejny wierzchołek w FP-drzewie, którego nazwa jest identyczna do danego.
3. Każdy wpis w *tablicy nagłówkowej* (ang. *frequent-item-header table*) składa się z dwóch elementów: nazwy elementu oraz wskaźnika na pierwszy element w drzewie posiadający identyczną nazwę.

Rysunek 2.2 przedstawia przykładowe FP-drzewo wraz z tablicą nagłówkową reprezentujący dane z przykładu wprowadzonego i analizowanego wcześniej.



RYSUNEK 2.2: Ilustracja przykładowego FP-drzewa dla zbioru danych z tablicy 2.3

Kompresja bazy danych

Pierwszy etap polega na znalezieniu wszystkich 1-zbiorów częstych występujących w bazie danych DB . Po ich odnalezieniu (F_1 - zbiór znalezionych 1-zbiorów częstych) z każdej transakcji T usuwany jest ten element, który nie należy do F_1 .

W wyniku usunięcia elementów nie tworzących jednoelementowych zbiorów częstych, baza ma zazwyczaj znacznie mniejszy rozmiar niż wyjściowa baza danych. Dodatkowo w tym kroku elementy w każdej transakcji zostają posortowane według malejącej wartości ich wsparcia.

Konstrukcja FP-drzewa

Algorytm 1. Konstrukcja FP-drzewa.

Input: Baza danych transakcji DB oraz minimalne wsparcie ($minsup$).

Output: FP-drzewo utworzone na podstawie zawartości DB

Metoda: Poniżej zostanie opisany proces konstrukcji FP-drzewa.

1. Przeskanowanie bazy danych transakcji DB odbywa się jednokrotnie. Utworzony na tej podstawie zostanie zbiór F , zawierający 1-zbiory częste. Posortowany malejąco zbiór F na podstawie wartości $support$ dla każdego elementu tworzy listę $FList$, czyli listę wszystkich elementów tworzących jednoelementowe zbiory częste.
2. Tworzony jest pierwszy element drzewa - korzeniem zostaje (zgodnie z definicją 4) element z etykietą $null$. Dla każdej transakcji w bazie danych DB wykonywane jest, co następuje.

Wybierane są elementy częste z transakcji, a następnie sortowane zgodnie z kolejnością w $FList$. Niech taka posortowana lista elementów częstych ma postać $[p|P]$, gdzie p jest pierwszym elementem, a P jest pozostałą częścią listy. Następnie wywoływana jest funkcja $insertTree([p|P], T)$, gdzie T jest FP-drzewem.

Funkcja insertTree Jeśli drzewo T ma dziecko N takie, że $N.itemName = p.itemName$, zwiększana jest wartość $N.count$ o wartość 1; w przeciwnym wypadku tworzony jest nowy element N z wartością $count = 1$, a wskaźnik rodzica ustawiany jest na T oraz wskaźnik sąsiedztwa elementu ustawiany jest na element z takim samym $itemName$. Jeśli lista P była niepusta, to wywoływana jest funkcja $insertTree(P, T)$ rekurencyjnie, w przeciwnym wypadku kończony jest działanie funkcji.

Eksploracja FP-drzewa

Po utworzeniu FP-drzewa przeprowadzana jest jego analiza w celu znalezienia wszystkich zbiorów częstych. Eksploracja bazuje na obserwacji, że dla każdego 1-zbioru częstego α wszystkie częste nadzbiory tego zbioru są reprezentowane w FP-drzewie przez ścieżki zawierające wierzchołek (bądź wierzchołki) α .

Analiza rozpoczyna się od znalezienia dla każdego 1-zbioru częstego α wszystkich ścieżek w FP-drzewie, których końcowym wierzchołkiem jest wierzchołek odpowiadający zbiorowi α . Pojedyncza ścieżka, na której końcu znajduje się wierzchołek α w dalszej analizie będzie nazywana *ścieżką prefiksową wzorca α* .

Poniżej zaprezentowany zostanie pseudokod algorytmu przeszukiwania FP-drzewa celem odnalezienia reguł asocjacyjnych.

Algorytm 2. *FP-growth: Przeszukiwanie FP-drzewa celem odnalezienia reguł asocjacyjnych.*

Input: Baza danych transakcji DB reprezentowana przez FP-drzewo zwrócone przez algorytm 1 oraz minimalne wsparcie ($minsup$).

Output: Kompletny zbiór reguł asocjacyjnych.

Metoda: Wywołanie $FP-GROWTH(FP - drzewo, null)$.

FP-GROWTH($Tree, \alpha$)

```

1  if  $Tree$  zawiera jedną ścieżkę prefiksową
2      then  $P \leftarrow$  ścieżka prefiksowa drzewa  $Tree$ 
3           $Q \leftarrow$  wieloczęściowa ścieżka z najwyższym elementem zastąpionym przez korzeń  $null$ 
4      for each kombinacja ( $\beta$ ) elementów z  $P$ 
5          do generuj regułę  $\beta \cup \alpha$  z  $support =$  minimalna wartość  $support$  elementów w  $\beta$ 
6              niech  $freqPatternSet(P)$  będzie zbiorem wygenerowanych do tej pory reguł
7      else  $Q \leftarrow Tree$ 
8      for each element  $a_i \in Q$ 
9          do generuj regułę  $\beta \leftarrow a_i \cup \alpha$  z  $support = a_i.support$ 
10             stwórz warunkową bazę wzorców z  $\beta$  oraz FP-drzewo ( $Tree_\beta$ ) dla  $\beta$ 
11             if  $Tree_\beta \neq \emptyset$ 
12                 then call FP-GROWTH( $Tree_\beta, \beta$ )
13             niech  $freqPatternSet(Q)$  będzie zbiorem wygenerowanych do tej pory reguł
14 return  $freqPatternSet(P) \cup freqPatternSet(Q) \cup (freqPatternSet(P) \times freqPatternSet(Q))$ 

```

Przykład eksploracji FP-drzewa

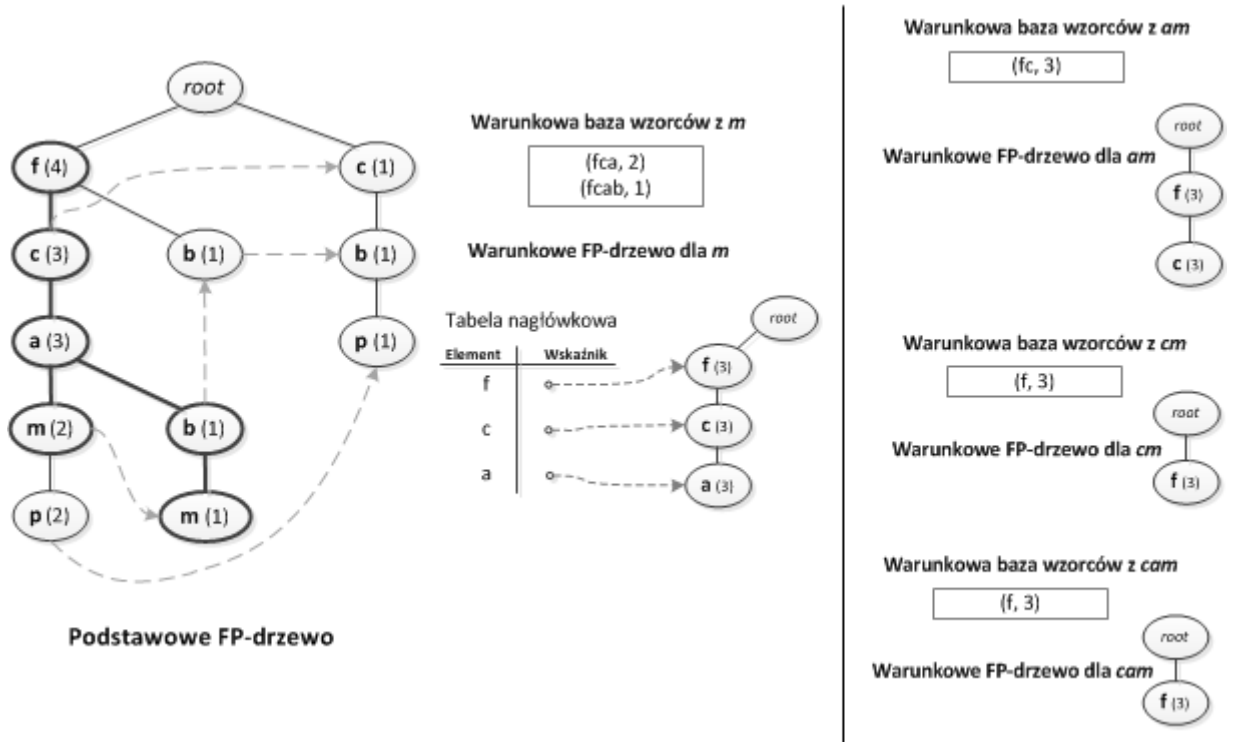
W rozdziale pt. *Przykład budowy FP-drzewa* zaprezentowano sposób budowy FP-drzewa. Opierając się na przykładzie wprowadzonym w tablicy 2.3 przedstawiony zostanie przykład działania eksploracji zbudowanego na tej podstawie FP-drzewa, które zostało zaprezentowane na rysunku 2.2.

Przykład należy rozpocząć od spostrzeżenia, że wszystkie reguły asocjacyjne, zawierające element częsty a_i mogą być zebrane poprzez wyjście ze wskaźnika z tablicy nagłówkowej FP-drzewa, a następnie podążając zgodnie z kolejnymi wskaźnikami na wystąpienia danego elementu w drzewie. W przykładzie analiza zostanie przeprowadzona od końca - tj. podążając od elementu najmniej liczego spośród elementów częstych.

Dla elementu p wzorcem częstym jest $p : 3$, która posiada dwie ścieżki w drzewie: $\langle f : 4, c : 3, a : 3, m : 2, p : 2 \rangle$ oraz $\langle c : 1, b : 1, p : 1 \rangle$. Pierwsza ścieżka oznacza, że wartość " (f, c, a, m, p) " występuje w bazie danych DB dwukrotnie. Warto zauważyć, że pierwsza ścieżka oznacza również, że ciąg " (f, c, a) " występuje w bazie dwukrotnie, a " (f) " nawet czterokrotnie. Jednakże występują one jednocześnie z p jedynie dwukrotnie. Dlatego też, do analizy wartości występujących razem z p tylko ścieżka prefiksowana $\langle f : 2, c : 2, a : 2, m : 2 \rangle$ (lub zapisane krócej: $\langle fcam : 2 \rangle$) jest brana pod uwagę. Podobnie w przypadku drugiej ścieżki - wynika z niej, że ciąg " (c, b, p) " występuje tylko raz w transakcjach w bazie danych DB , lub też prefiksowana ścieżka dla p ma postać $\langle cb : 1 \rangle$. Te dwie ścieżki prefiksowane elementu p (czyli $\{(fcam : 2), (cb : 1)\}$) tworzą bazę wzorca elementu p (ang. *subpattern-base*), która jest nazywana *warunkową bazą wzorców z p* (ang. *conditional pattern base*) - to znaczy bazę wzorca pod warunkiem występowania elementu p [HPYM04]. Konstrukcja FP-drzewa na tej warunkowej bazie wzorców (które nazywane jest *warunkowym FP-drzewem elementu p* ang. *conditional FP-tree*) prowadzi tylko do jednej gałęzi - $(c : 3)$. Stąd też tylko jeden

wzorzec jest dziedziczony - ($cp : 4$). (Warto zauważyć, że wzorec jest zbiorem elementów, który jest reprezentowany przez ciąg.) Poszukiwanie częstych wzorców związanych z elementem p kończy działanie.

Dla wierzchołka m jego bezpośrednim wzorcem częstym jest ($m : 3$), posiada on również dwie ścieżki: $\langle f : 4, c : 3, a : 3, m : 2 \rangle$ oraz $\langle f : 4, c : 3, a : 3, b : 1, m : 1 \rangle$. W obu tych ścieżkach można zauważyć, że m występuje razem z p , jednakże p w tym miejscu nie musi być brane pod uwagę, ponieważ wszystkie wzorce częste dla tego elementu zostały znalezione wcześniej (opisano to akapit wcześniej). Wykonując działania podobnie, jak w poprzednim akapicie wyznaczona zostanie warunkowa baza wzorców o postaci $\{(fca : 2), (fcab : 1)\}$. Skonstruowane na tej podstawie FP-drzewo, czyli warunkowe FP-drzewo elementu m ($\langle f : 3, c : 3, a : 3 \rangle$) posiada jedną ścieżkę częstą, co można zaobserwować na rysunku 2.3. Następnie przeprowadzona jest rekurencyjna eksploracja tego warunkowego drzewa poprzez wywołanie $mine(\langle f : 3, c : 3, a : 3 \rangle | m)$.



RYSUNEK 2.3: Eksploracja warunkowego FP-drzewa dla elementu m (FP-tree $|m$) [HPYM04]

Rysunek 2.3 ilustruje, że wywołanie $mine(\langle f : 3, c : 3, a : 3 \rangle | m)$ poddaje przetwarzaniu trzy elementy: $(a), (c), (f)$ sekwencyjnie. Przetwarzanie pierwszego tworzy częsty wzorec o postaci $(am : 3)$, warunkową bazę wzorców $\{(fc : 3)\}$, a następnie wywoływana jest funkcja $mine(\langle f : 3, c : 3 \rangle | am)$; kolejny tworzy częsty wzorec o postaci $(cm : 3)$, warunkową bazę wzorców $\{(f : 3)\}$, a następnie wywoływana jest funkcja $mine(\langle f : 3 \rangle | cm)$; trzecie natomiast tworzy jedynie częsty wzorec $(fm : 3)$. Dalsze wywołanie metody $mine(\langle f : 3, c : 3 \rangle | am)$ tworzy dwa wzorce $(cam : 3)$ oraz $(fam : 3)$, warunkową bazę wzorców $\{(f : 3)\}$, która prowadzi do wywołania $mine(\langle f : 3 \rangle | cam)$, czego wynikiem jest najdłuższy wzorec częsty $(fcam : 3)$. Podobnie, wywołanie $mine(\langle f : 3 \rangle | cm)$

Element	Warunkowa baza wzorców	Warunkowe FP-drzewo
p	$\{(fcam : 2), (cb : 1)\}$	$\{(c : 3)\} p$
m	$\{(fca : 2), (fcab : 1)\}$	$\{(f : 3, c : 3, a : 3)\} m$
b	$\{(fca : 2), (f : 1), (c : 1)\}$	\emptyset
a	$\{(fc : 3)\}$	$\{(f : 3, c : 3)\} a$
c	$\{(f : 3)\}$	$\{(f : 3)\} c$
f	\emptyset	\emptyset

TABLICA 2.4: Wyniki eksploracji FP-drzewa dla przykładu opartego na danych w tablicy 2.3

tworzy jeden wzorec $(fcm : 3)$.

Z przeprowadzonej analizy wynika, że zbiór wszystkich wzorców częstych zawierający m ma postać $\{(m : 3), (am : 3), (cm : 3), (fm : 3), (cam : 3), (fam : 3), (fcam : 3), (fcm : 3)\}$. Wynika z tego, że jedna ścieżka w FP-drzewie może być przeanalizowana poprzez wygenerowanie wszystkich kombinacji (bez powtórzeń) elementów na tej ścieżce.

Podobnie, b tworzy $(b : 3)$, posiada również trzy ścieżki: $\langle f : 4, c : 3, a : 3, b : 1 \rangle$, $\langle f : 4, b : 1 \rangle$ oraz $\langle c : 1, b : 1 \rangle$. Ponieważ warunkowa baza wzorca tego elementu $\{(fca : 1), (f : 1), (c : 1)\}$ nie generuje żadnych elementów częstych, eksploracja po elemencie b kończy działanie.

Element a tworzy jeden element częsty $(a : 3)$ oraz bazę wzorca $\{(fc : 3)\}$, czyli FP-drzewo posiadające jedną ścieżkę. Wynika z tego, że zbiór częstych wzorców może być wygenerowany poprzez wygenerowanie wszystkich ich kombinacji. Dokonując konkatenacji z $(a : 3)$, otrzymany zostanie zbiór $\{(fa : 3), (ca : 3), (fca : 3)\}$. Wierzchołek c generuje $(c : 4)$, a także bazę wzorca $\{(f : 3)\}$, a zbiór częstych wzorców skojarzonych z $(c : 3)$ ma zawartość $\{(fc : 3)\}$. Analiza elementu f tworzy jedynie $(f : 4)$, lecz bez warunkowych baz wzorców.

W tablicy 2.4 zebrane zostały bazy wzorców oraz warunkowe FP-drzewa dla poszczególnych elementów.

Rozdział 3

Uniwersalna architektura procesorów wielordzeniowych

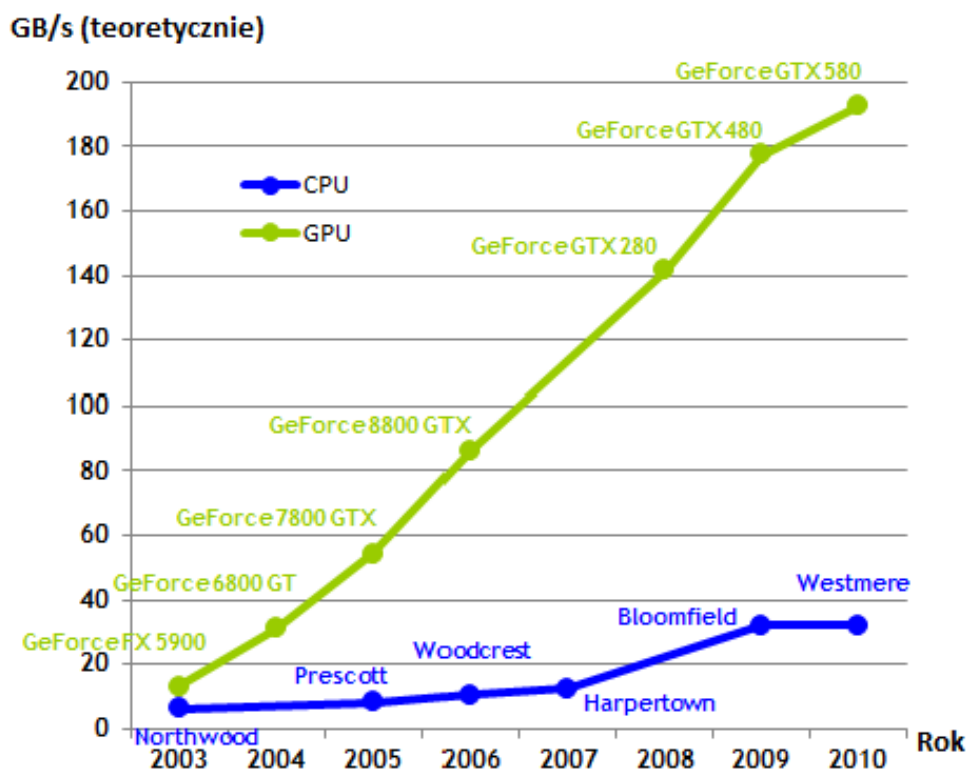
Rozdział ten stanowi wprowadzenie to zagadnienia obliczeń wielordzeniowych. Przedstawiona została w nim pokrótce historia rozwoju oprogramowania oraz sprzętu komputerowego, który podążał za co raz to wyższymi wymaganiami stawianymi przez użytkowników. Opisana zostanie również historia rozwoju *procesora* (*CPU*, ang. *Central Processing Unit*) - głównej jednostki obliczeniowej w praktycznie każdym urządzeniu elektronicznym. Później opisany zostanie rozwój karty graficznej oraz zmiana jej zastosowania na przestrzeni lat.

Ostatnim, lecz nie najmniej ważnym elementem tego rozdziału jest opis uniwersalnej architektury obliczeniowej *CUDA* (ang. *Compute Unified Device Architecture*) wprowadzonej przez firmę *nVidia* w roku 2007, która umożliwia wykorzystanie mocy obliczeniowej *procesorów graficznych* (*GPU*, ang. *Graphics Processing Unit*), bądź innych procesorów wielordzeniowych, do rozwiązywania ogólnych problemów obliczeniowych w sposób znacząco wydajniejszy niż w przypadku tradycyjnych, sekwencyjnych procesorów [Cor07a].

Na rysunku 3.1 przedstawiony został przyrost w przepustowości pamięci - odpowiednio dla GPU oraz CPU na przestrzeni lat. Łatwo zauważyć, że w ciągu 7 lat procesor graficzny zyskał około 10-krotną przewagę nad CPU. Dlatego też co raz większe jest zainteresowanie wykorzystaniem kart graficznych w dziedzinach innych niż tylko renderowanie grafiki.

3.1 Czasy przetwarzania równoległego

W poprzednich latach dokonał się znaczący postęp w przechodzeniu przemysłu komputerowego na obliczenia wykonywane równolegle. W roku 2010 większość komputerów konsumenckich była dostarczana do odbiorcy z procesorem zawierającym więcej niż jeden *rdzeń* (ang. *core*). Począwszy od procesorów dwurdzeniowych w laptopach do 8- czy 16-rdzeniowych stacji roboczych - od pewnego czasu obliczenia równoległe nie są już tylko domeną superkomputerów lub mainframe'ów (ang. *main* – główny, *frame* – struktura). Co więcej, urządzenia elektroniczne takie jak telefony komórkowe czy też przenośne odtwarzacze muzyki wyposażane są w procesory wielordzeniowe, co



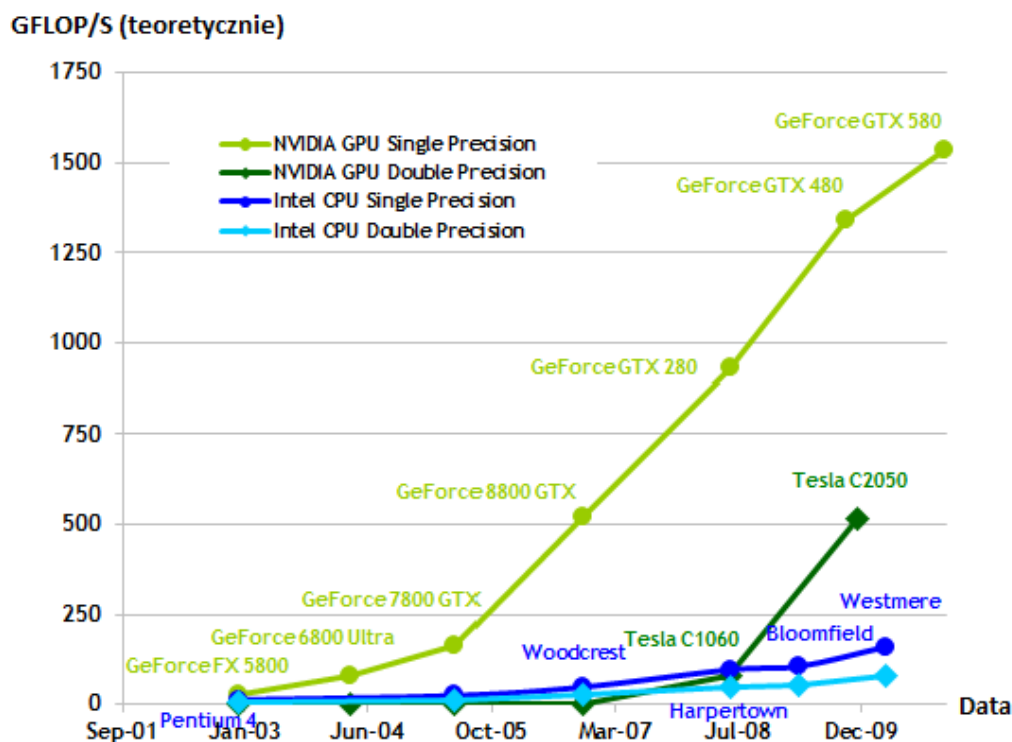
RYSUNEK 3.1: Przepustowość pamięci na CPU oraz GPU [Cor11b]

zapewnia im możliwości dalece przekraczające te dostępne dla ich poprzedników.

Wynika z tego, że coraz to więcej programistów będzie musiało radzić sobie z implementacją oprogramowania przeznaczonego na jednostki równoległe, wykorzystywać nowe technologie, które będą pozwalały na dostarczenie nowatorskich rozwiązań dla co raz bardziej wymagającej rzeszy użytkowników. Wiersze poleceń to przeżytek - od dawna komputerem steruje się za pomocą skomplikowanych interfejsów graficznych. To samo tyczy się telefonów komórkowych - w chwili obecnej telefon to tylko jedna z wielu funkcji, jakich może dostarczyć współczesny aparat telefoniczny. Teraz telefony mogą jednocześnie grać muzykę, dostarczać informacji o obecnej lokalizacji przy użyciu modułu nawigacji satelitarnej (*GPS*, ang. *Global Positioning System*) i jednocześnie wyświetlać zdjęcia.

3.1.1 Procesor

Przez około 30 lat jedną z ważniejszych metod udoskonalania centralnych jednostek obliczeniowych, a przez to zwiększania komfortu korzystania z urządzenia przez użytkownika, było zwiększanie prędkości z jaką operował zegar procesora. W latach 80 XX wieku procesor przeznaczony na rynek konsumencki operował z prędkością oscylującą w okolicach 1MHz. Około 30 lat później, w czasach współczesnych, większość komputerów osobistych wyposażona jest w procesory o prędkościach od 1 do 4GHz, czyli obecne jednostki są około 1000 szybsze od wczesnych procesorów. Zwiększanie prędkości, z jaką operuje zegar procesora jest niezawodnym źródłem zwiększania



RYSUNEK 3.2: Operacje zmiennopozycyjne na sekundę na GPU oraz CPU [Cor11b]

szybkości, należy jednak podkreślić, że nie jest to jedyna metoda na zwiększanie jego wydajności.

Jednakże ograniczenia technologiczne wyznaczają pewne granice, w jakich może wzrastać prędkość zegara procesora. Dlatego też poszukuje się innych, równie niezawodnych źródeł zwiększenia wydajności. Nie można już dłużej polegać jedynie na zwiększaniu prędkości. Z powodu restrykcji na mocy oraz wydzielanym cieple oraz docieraniu do granicy rozmiaru tranzystora, naukowcy oraz producenci rozpoczęli poszukiwanie nowych źródeł i sposobów na zwiększenie możliwości procesorów.

Poza światem konsumentów, czyli w świecie tzw. superkomputerów przez dekady osiągnęto niezwykle wielkie przyrosty mocy w bardzo podobny sposób. Moc procesora używanego w tych komputerach rosła tak samo szybko, jak w przypadku przyrostów procesorów desktopowych. Jednakże, poza wielkimi przyrostami mocy obliczeniowej na jednej jednostce, producenci superkomputerów tworzyli komputery, w których solidne przyrosty w wydajności osiągnęto dzięki zwiększaniu liczby używanych procesorów. Nie jest niczym niezwykłym, że pojedynczy superkomputer składa się z dziesiątek lub setek tysięcy procesorów działających równolegle.

W poszukiwaniu dodatkowych możliwości dla komputerów osobistych, poprawa wydajności w przypadku superkomputerów rodzi pytanie: Dlaczego zamiast zwiększać wydajność pojedynczej jednostki, nie umieścić w komputerze osobistym więcej rdzeni? W wypadku zwiększania liczby rdzeni rozwój jednostek obliczeniowych nie byłby ograniczony przez te same niedogodności, co w przypadku ciągłego zwiększania prędkości zegara procesora.

W roku 2005 wiodący producenci procesorów zaczęli oferować jednostki z dwoma, zamiast z

jednym rdzeniem. W latach następnych kontynuowano tę praktykę, tworząc jednostki trzy-, cztero-, sześćo- oraz ośmio-rdzeniowe. Czasami nazywa się ten okres *rewolucją wielordzeniową* [SK10], ponieważ zmiana podejścia do zwiększania wydajności jednostek w znaczący sposób wpłynęła na ewolucję konsumenckiego rynku komputerów.

W chwili obecnej praktycznie każdy komputer osobisty jest wyposażony w procesor dwurdzeniowy. Nawet na rynku niskobudżetowych komputerów z bardzo niskim zapotrzebowaniem na moc, dokonała się rewolucja wielordzeniowa - już nawet netbooki będą wyposażone w dwa rdzenie [Cor10].

3.2 Wpływ procesorów graficznych na procesy obliczeniowe

Wydawać by się mogło, że użycie procesora graficznego, jako jednostki obliczeniowej dla problemów nie związanych bezpośrednio z przetwarzaniem grafiki jest podejściem zupełnie nowym. Porównując to do klasycznych obliczeń na procesorach jest to w istocie koncepcja nowa. Jednakże obliczenia na jednostkach graficznych nie są tak nowe, jak mogłoby się wydawać na pierwszy rzut oka.

3.2.1 Krótka historia kart graficznych

W rozdziale 3.1.1 przedstawiony został rozwój procesora w dwóch płaszczyznach - prędkości oraz liczbie rdzeni. W międzyczasie karty graficzne przeżywały rewolucję rozwojową. Na końcu lat 80 oraz początku 90, wzrost popularności *graficznych interfejsów użytkownika* (*GUI*, ang. *Graphical User Interface*), a w szczególności systemów operacyjnych takich jak Microsoft Windows, wymusił na producentach sprzętu stworzenie nowego typu procesora. We wczesnych latach 90 użytkownicy zaczęli kupować karty graficzne 2D dla swoich komputerów osobistych. Te akceleratory grafiki oferowały wsparcie sprzętowe do operacji bitmapowych by umożliwić wykorzystanie graficznego interfejsu użytkownika.

Mniej więcej w tym samym czasie, w świecie profesjonalnych komputerów, firma o nazwie *Silicon Graphics* popularyzowała w latach 80-tych użycie grafiki trójwymiarowej w wielu dziedzinach, m.in. aplikacje rządowe oraz obronne, aplikacje wspierające naukę oraz wizualizację wyników przeprowadzanych badań naukowych, a także w tworzeniu trójwymiarowych efektów filmowych, czyli rzeczy do tej pory niedostępnych na rynku. W roku 1992 Silicon Graphics stworzyło interfejs programistyczny do swojego sprzętu poprzez wydanie biblioteki IRIS GL, która następnie ewoluowała do OpenGL (ang. *Open Graphics Library*), czyli specyfikacji uniwersalnego API do generowania grafiki. Firma ta wypuszczając tę bibliotekę chciała, by była ona ustandaryzowaną, niezależną od platformy biblioteką do tworzenia aplikacji trójwymiarowych. Zupełnie, jak w przypadku licznych rdzeni na procesorach (patrz rozdział 3.1.1) było tylko kwestią czasu, aż aplikacje 3D znajdą się w domach użytkowników na ich prywatnych komputerach.

Dwa znaczące powody w połowie lat 90-tych spowodowały, że pojawiła się nagle potrzeba stworzenia procesora graficznego potrafiącego tworzyć trójwymiarowe efekty (*3D*, ang. *three-dimensional*).

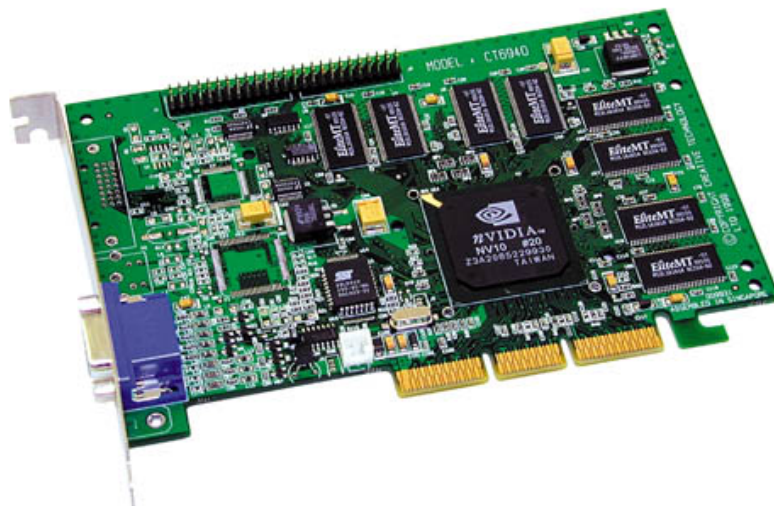


RYSUNEK 3.3: Zrzut ekranu z gry DOOM - klasycznej gry z gatunku FPS

Pierwszym było pojawienie się wciągających gier, w których gracz ogląda świat gry z perspektywy pierwszej osoby, czyli oczami bohatera (*FPP*, ang. *First Person Perspective*), takich jak Doom (patrz rysunek 3.3), Duke Nukem 3D oraz Quake. Przyczyniło się to do rozpoczęcia tworzenia bardziej realistycznych środowisk trójwymiarowych, w których możliwe byłoby tworzenie gier komputerowych. Pomimo tego, że grafika 3D mogła być wykorzystana w praktycznie wszystkich grach komputerowych, to wyjątkowa popularność gier z gatunku FPS (ang. *First Person Shooter*) spowodowała, że większy nacisk położono na prywatne jednostki komputerowe. W tym samym czasie firmy takie jak NVIDIA, ATI Technologies oraz 3dfx Interactive zaczęły produkować akceleratory graficzne, które były w stanie sprostać ciągle rosnącym wymaganiom klientów. Te strategiczne decyzje spowodowały, że grafika trójwymiarowa na dobre zadomowiła się na komputerach stacjonarnych i w nadchodzących latach to ona miała wieść prym wśród technologii komputerowych.

Wypuszczenie w drugiej połowie roku 1999 na rynek karty graficznej NVIDIA GeForce 256 przesunęło możliwości komputerów osobistych jeszcze dalej [Cor99] - wygląd tej karty przedstawiony został na rysunku 3.4. Po raz pierwszy obliczenia świetlne oraz transformacje obliczeniowe możliwe były do wykonania bezpośrednio na procesorze graficznym, co dawało jeszcze większe możliwości tworzenia aplikacji atrakcyjnych wizualnie. Ponieważ operacje te były już integralną częścią *potoku wywołań* (ang. *pipeline*) biblioteki OpenGL, karta GeForce 256 wyznaczyła początek progresywnego zwiększania operacji dostępnych w bibliotece OpenGL, które były implementowane na procesorze graficznym.

Uważa się, że wypuszczona w roku 2001 seria kart graficznych GeForce 3 jest jednym z większych przełomów w świecie technologii GPU [Cor01, SK10]. Seria ta była pierwszą w świecie, której chip był kompatybilny z nowym w ówczesnym czasie standardem firmy Microsoft o nazwie DirectX 8.0,



RYSUNEK 3.4: Karta graficzna NVIDIA GeForce 256

który wymagał by hardware zawierało *programowalne wierzchołki* (ang. *programmable vertex*) oraz *programowalne cieniowanie* (ang. *programmable pixel shading*) w kolejnych fazach przetwarzania. Po raz pierwszy programiści mogli mieć jakąkolwiek kontrolę nad obliczeniami dokonywanymi bezpośrednio na karcie graficznej.

3.2.2 Wczesne obliczenia na kartach graficznych

Wyprodukowanie kart, które posiadały następujące po sobie programowalne fazy (potok wywołań) spowodowały, że wielu naukowców zaczęło wykorzystywać ich możliwości nie tylko poprzez używanie OpenGL czy też DirectX do standardowych zadań. Takie podejście do obliczeń na kartach graficznych we wczesnych latach obliczeń na GPU było niezwykle skomplikowane. Ponieważ standardowe API graficzne takie jak OpenGL, czy też DirectX były jedynymi metodami do interakcji z GPU, to programowanie obliczeń na kartach graficznych nadal sprowadzało się do implementowania przetwarzania graficznego poprzez dotępne metody interfejsu programistycznego karty. Z tego też powodu, wielu naukowców wykonywało swoje obliczenia poprzez wspomniane wcześniej API w taki sposób, by ich problemy sprowadzone zostały do renderowania grafiki, a następnie odpowiedniego przetworzenia.

Zasadniczo każda karta graficzna we wczesnym roku 2000 zaprojektowana była w taki sposób, by produkować kolor dla każdego piksela na ekranie używając programowalnych jednostek arytmetycznych nazywanych *pixel shader*. W ogólności jednostka ta wykorzystuje pozycję (jako parę (x, y)) piksela oraz pewne dodatkowe informacje by obliczyć kolor danego piksela na ekranie. Tymi dodatkowymi informacjami mogą być kolory, wymiary tekstur oraz parametry, które mogą zostać podane, bądź obliczone w trakcie interakcji z użytkownikiem, bądź otoczeniem piksela. Ponieważ obliczenia wykonywane na wejściowych kolorach oraz teksturach były kontrolowane przez programistę, zauważono, że te wejściowe „kolory” mogą w rzeczywistości reprezentować dowolną daną.

Zatem jeśli wejście programu było dowolną daną reprezentowaną jako wartość reprezentującą coś zupełnie innego niż wartość, to programiści mieli możliwość wykorzystywania pixel shaderów, czyli krótkiego programu komputerowego w specjalnym języku, do wykonania na tych danych porządkanych obliczeń oraz przekształceń. Rezultatem tych obliczeń był „kolor”, który w danym kontekście oznaczał coś zupełnie innego niż tylko kolor na ekranie w dosłownym tego słowa znaczeniu. Można więc nazwać to „oszukiwaniem” karty graficznej poprzez poddawanie danych wejściowych przetwarzaniu graficznemu, jakby były to zwyczajne dane potrzebne do wyrenderowania obrazu. Takie podejście odznaczało się wyjątkowym poziomem pomysłowości, ale niestety wykonanie takich obliczeń było wyjątkowo zagmatwane i skomplikowane.

Z powodu wysokiej przepustowości obliczeń arytmetycznych na GPU, początkowe rezultaty takich eksperymentów obiecywały świetlaną przyszłość obliczeń na jednostkach graficznych. Jednakże model programistyczny stosowany do implementacji takich obliczeń był zbyt restrykcyjny dla rzeszy developerów by mógł być wykorzystywany na szeroką skalę. Możliwe było jedynie wykorzystywanie „kolorów” oraz podmian tekstur, co w wielu przypadkach stanowiło duże ograniczenie przy bardziej skomplikowanych obliczeniach. Było również sporo ograniczeń do tego jak i gdzie programista mógł zapisywać wyniki do pamięci karty, tak więc algorytmy wykorzystujące rozproszone lokacje nie były możliwe do implementacji na GPU. Co więcej, bardzo trudno było implementować algorytmy korzystające z obliczeń zmiennopozycyjnych (ang. *floating-point*), ponieważ nie można było przewidzieć, jak karta graficzna, jeśli w ogóle, wykona te obliczenia [SK10]. Ostatnim z zasadniczych ograniczeń był fakt braku jakiegokolwiek metody do debugowania kodu uruchamianego na karcie graficznej - na przykład w momencie wyraźnie błędnych obliczeń, zawieszania się komputera czy też w momencie nie zakończenia się programu.

Jeśli wcześniej wymienione ograniczenia nie były przeszkodą, ktokolwiek kto chciał wykorzystać moc obliczeniową karty graficznej do wykonywania obliczeń nie związanych z przetwarzaniem grafiki musiał w dalszym ciągu nauczyć się OpenGL bądź DirectX, gdyż to one pozostawały jedyne metody do interakcji z GPU. Nie tylko oznaczało to przechowywanie wyników w teksturach graficznych oraz wykonywanie obliczeń poprzez wywoływanie funkcji OpenGL lub DirectX, ale dodatkowo oznaczało to pisanie obliczeń w specjalnych językach programowania graficznego (ang. *shading languages*). Wymaganie by naukowcy dokonywali obliczeń na ograniczonych zasobach oraz specjalnych językach programowania oraz nauczania się specyfiki jednostek graficznych przed przystąpieniem do wykorzystania możliwości GPU do obliczeń spowodowało, że metoda ta nie przyjęła się, jako wiodąca w świecie naukowców.

3.3 CUDA

Dopiero po około pięciu latach od wydania serii 3 GeForce obliczenia na jednostkach graficznych były gotowe rzeczywiście być implementowane przez programistów z pełnym wykorzystaniem możliwości kart. W listopadzie 2006 roku NVIDIA wypuściła na rynek pierwszą kartę graficzną wspierającą DirectX 10 - GeForce 8800 GTX [Cor06b]. Karta ta była również pierwszą jednostką

zbudowaną na architekturze CUDA. Architektura ta zawierała nowe komponenty zaprojektowane specjalnie pod obliczenia wykonywane na GPU oraz eliminowała wiele ograniczeń, które powstrzymywały programistów przed wykorzystywaniem możliwości karty do obliczeń nie związanych z przetwarzaniem grafiki.

3.3.1 Czym jest architektura CUDA?

W przeciwieństwie do poprzednich generacji kart graficznych, w których jednostki były podzielone na pixel i vertex shader'y, w kartach zgodnych z architekturą CUDA zawarta jest zunifikowana linia shaderów (ang. *shader pipeline*). Pozwala to każdej jednostce arytmetyczno-logicznej (*ALU*, ang. *Arithmetic Logic Unit* lub *Arithmetic and Logical Unit*) na czipie być wykorzystaną przez program wykonujący podstawowe operacje obliczeniowe.

Ponieważ firma NVIDIA zamierzała tę nową rodzinę kart graficznych była wykorzystywana w innych rejonach niż tylko wyświetlanie grafiki, wspomniane wcześniej jednostki ALU są zbudowane w oparciu o wymagania IEEE (ang. *Institute of Electrical and Electronics Engineers*) co do arytmetyki zmiennopozycyjnej. Zostały one również tak zaprojektowane, by zbiór dostępnych operacji był dostosowany do ogólnego użycia niż tylko do wyświetlania grafiki.

Co więcej, jednostki wykonawcze na GPU zostały wyposażone w możliwość czytania oraz zapisu do pamięci, a także do *pamięci podręcznej* (ang. *cache*) programu, znanej również jako *pamięć współdzielona* (ang. *shared memory*).

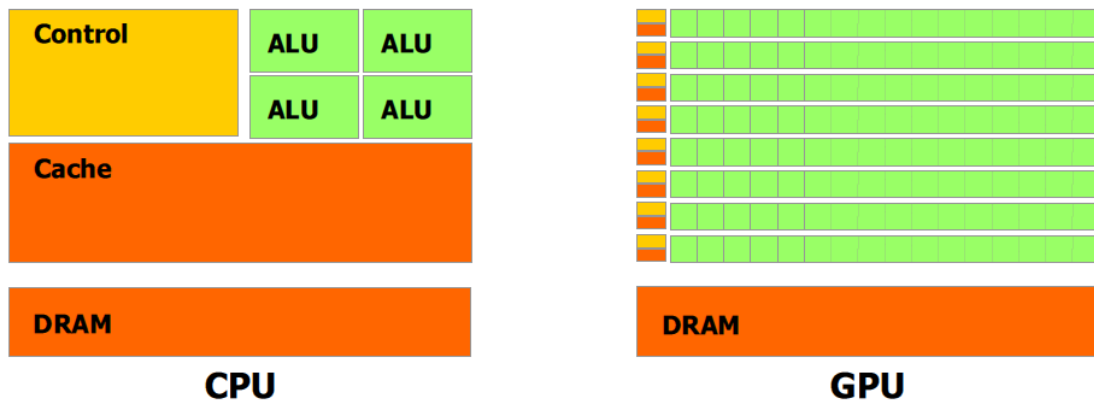
Wszystkie te usprawnienia wprowadzone w architekturze CUDA celem rozszerzenia możliwości kart graficznych poza jedynie wyświetlanie grafiki [SK10].

3.3.2 Od wyświetlania grafiki do obliczeń ogólnego przeznaczenia

Jak przedstawiono na rysunkach 3.1 oraz 3.2 procesory graficzne rozwijały się w kierunku przetwarzania równoległego - głównie ze względu wymagań stawianych przez konsumentów na tworzenie realistyczny grafik 3D w czasie rzeczywistym. Z tego też powodu GPU stały się wyspecjalizowanymi jednostkami o wysokim stopniu zrównoleglenia zadań, wielordzeniowymi oraz przetwarzającymi informacje w wielu wątkach.

Powodem tak dużej rozbieżności w możliwościach przetwarzania równoległego (zauważalne na rysunku 3.2) pomiędzy CPU a GPU jest fakt wysokiej specjalizacji na intensywne obliczenia w sposób równoległy (czyli to, co jest głównym składnikiem renderowania grafiki) - dlatego też w jednostkach kart grafiki więcej tranzystorów przeznaczonych jest do przetwarzania danych zamiast na cache danych oraz przepływ sterowania - schematycznie zilustrowane to zostało na rysunku 3.5.

Opisując bardziej szczegółowo, to GPU jest tak zaprojektowane by spełniać wymagania problemów, które mogą być zapisane jako obliczenia równoległe - ten sam program (metoda) jest wykonywany na każdym elemencie danych równolegle. Ponieważ ten sam program jest uruchomiony na każdym elemencie danych, to istnieje mniejsze wymaganie w kontroli skomplikowanego przepływu danych. Dodatkowo z powodu wysokiej intensywności arytmetycznej, to opóźnienie dostępu do



RYSUNEK 3.5: Schematyczne przedstawienie budowy CPU oraz GPU [Cor11b]

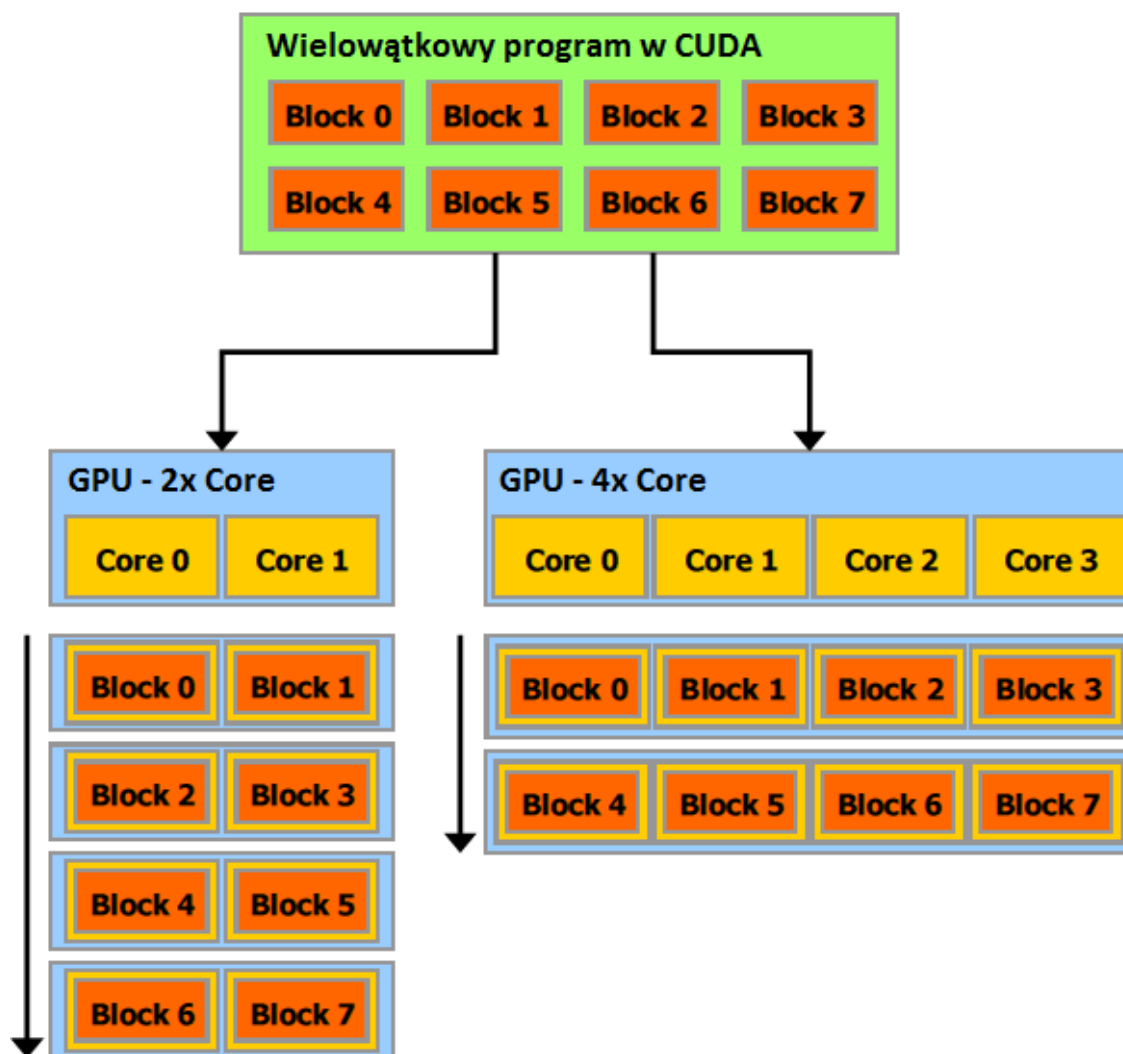
pamięci może być ukryte przy pomocy obliczeń zamiast dużych części danych przechowywanych w cache jednostki.

Równoległe przetwarzanie danych mapuje elementy danych na równoległe wątki przetwarzające te dane. Wiele aplikacji, które przetwarzają duże zbiory danych może używać podejścia równoległego przetwarzania danych do przyspieszenia obliczeń. W renderowaniu grafiki 3D duże zbiory pikseli oraz wektory są mapowane na równoległe wątki. Podobnie w aplikacjach przetwarzających obraz oraz media, takie jak post-przetwarzanie wyrenderowanych obrazów, kodowanie i dekodowaniu video, skalowanie obrazów, czy też rozpoznawanie wzorców może mapować bloki obrazu oraz pikseli na równoległe wątki. W rzeczywistości, wiele algorytmów poza dziedziną przetwarzania obrazów są przyspieszane poprzez równoległe przetwarzanie danych - począwszy od przetwarzania sygnałów czy też symulacji fizycznych do obliczeń finansowych, czy też biologii obliczeniowej.

3.3.3 Skalowalny model programistyczny

Nadejście wielordzeniowych procesorów oraz multirdzeniowych GPU oznacza, że obliczenia na procesorach będą odbywać się równoległe. Co więcej równoległość kontynuuje skalowalność zgodnie z prawem Moora - czyli zgodnie z empirycznym prawem, które polega na obserwacji, że ekonomicznie optymalna liczba tranzystorów w układzie scalonym w kolejnych latach posiada trend wykładniczy (podwaja się w niemal równych odcinkach czasu). Wyzwaniem jest stworzenie takiego oprogramowania, które w niezauważalny sposób będzie skalowało wykorzystywanie równoległości do zmieniającego się środowiska, w którym wzrastać będzie liczba rdzeni procesora, zupełnie jak aplikacje graficzne, które transparentnie skalują swój paralelizm do wielu rdzeni GPU, niezależnie od tego na jakiej karcie graficznej są one uruchamiane, co wiąże się bezpośrednio z liczbą dostępnych jednostek na procesorze.

Równoległy model programistyczny CUDA jest stworzony by przezwyciężyć wyzwanie, o którym wspomniano wyżej, z jednoczesnym ograniczeniem czasu potrzebnego programistom na nauczanie się nowego podejścia poprzez wykorzystywanie do tworzenia aplikacji standardowych języków programowania, takiego jak na przykład język C.



RYSUNEK 3.6: Zilustrowanie automatycznej skalowalności [Cor11b]

Na architekturę CUDA składają się trzy kluczowe elementy: hierarchię grup wątków, współdzieloną pamięć oraz barierę synchronizacyjną. Elementy te są udostępnione programiście poprzez ograniczony zbiór rozszerzeń języka programowania.

Te abstrakcyjne elementy dostarczają drobnoziarnistą równoległość na poziomie danych oraz wątków, zagnieżdżone w gruboziarnistej równoległości danych oraz *zadań* (ang. *task*). Takie podejście prowadzi programistę do podziału problemu na większe podproblemy, które mogą być rozwiązywane niezależnie przez bloki wątków, a każdy podproblem może być podzielony na mniejsze, które mogą być rozwiązywane wspólnie przez wątki wewnątrz jednego bloku.

Taka dekompozycja zachowuje wszystkie dostępne funkcjonalności języka programowania umożliwiając wątkom kooperację w trakcie rozwiązywania każdego z podproblemów, a w tym samym czasie udostępnia automatyczną skalowalność programu. W rzeczywistości, każdy blok wątków może być zaplanowany na każdym z dostępnych rdzeni na procesorze, w dowolnym porządku, jednocześnie lub sekwencyjnie, tak że skompilowany program w CUDA może być wykonany na dowolnej liczbie rdzeni procesora. Na rysunku 3.6 zilustrowane zostało rozpropagowanie zadań

w dwóch przypadkach - na 2 rdzenie, a w drugim przypadku na 4 rdzenie procesora graficzne. Jedynie *system uruchomieniowy* (ang. *runtime system*) musi znać faktyczną liczbę jednostek procesora (liczbe rdzeni procesora). Wielowątkowy program jest podzielony na bloki wątków, które wykonywane są niezależnie jeden od drugiego, dlatego też GPU z większą liczbą rdzeni będzie automatycznie wykonywać program w krótszym czasie niż karta graficzna z mniejszą liczbą jednostek.

Ten skalowalny model programistyczny umożliwia architekturze CUDA bycie stosowaną w szerokim spektrum zastosowań: od wysokowydajnych kart graficznych GeForce czy też na profesjonalnych kartach Quadro na komputerach entuzjastów, kartach Tesla przeznaczonych tylko do obliczeń, do kart GeForce przeznaczonych na domowe komputery.

3.3.4 Używanie architektury CUDA

Wysiłek wprowadzenia nowej architektury przez firmę NVIDIA, która łączy możliwości klasycznego sposobu generacji grafiki z nowym podejściem do uniwersalnych obliczeń nie oznacza, że wprowadzono jedynie nową architekturę, bez odpowiedniego podejścia dla programowania. Bez względu na to, ile nowych funkcjonalności NVIDIA doda do swoich chipów, niemożliwe stało się używanie ich w dotychczasowy sposób - tj. poprzez OpenGL lub DirectX. Nie tylko wymagałoby to od użytkowników dalszego przekształcania swoich algorytmów do problemów renderowania grafiki (problem opisany w części 3.2.2), ale również w dalszym ciągu użytkownicy (programiści) musieli by programować swoje obliczenia w zorientowanych na renderowanie grafiki środowiskach (tzw. *shading languages*) takich jak GLSL z OpenGL'a czy też HLSL firmy Microsoft.

By osiągnąć możliwie wysoki poziom developerów korzystających z możliwości nowej architektury, NVIDIA rozszerzyła jeden z najbardziej znanych języków programowania - język C, dodając do niego relatywnie mały zbiór nowych słów kluczowych, by programiści mogli korzystać z nowych, specjalnych funkcji dostępnych jedynie na architekturze CUDA. Kilka miesięcy po wprowadzeniu karty GeForce 8800 GTX [Cor06b], NVIDIA wprowadziła na rynek darmowy kompilator dla języka CUDA C [Cor11c]. W tym samym momencie, CUDA C stał się pierwszym językiem zaprojektowanym przez firmę produkującą sprzęt graficzny, w celu wypełnienia luki w dziedzienie obliczeń ogólnego przeznaczenia realizowanych na kartach graficznych.

Oprócz języka, w którym tworzone mogą być programy korzystające z mocy GPU, NVIDIA dostarczyła również specjalne *sterowniki* (ang. *drivers*) by wykorzystać pełną moc architektury CUDA. Użytkownicy nie są już więcej zobowiązani do znania interfejsów programowania specyficznych dla procesorów graficznych (takich jak OpenGL czy też DirectX), a co więcej nie są już oni zmuszeni do sprowadzania swoich problemów do problemu renderowania grafiki.

3.4 Kwantyfikator *parallel* czyli CUDA

W pseudokodach algorytmów wykorzystujących technologię CUDA wprowadzone zostało słowo kluczowe **parallel**. Kwantyfikator ten oznacza, że dana pętla wykonywana jest w sposób równo-

legły na wielu wątkach - tzn. zakres dla pętli dzielony jest na poszczególne wątki w taki sposób, że jeden wątek wykonuje jedynie część iteracji - jest to zależne od liczby iteracji (*iteracji*) oraz liczby wątków (*watkow*). W przybliżeniu każdy wątek zostanie uruchomiony $\frac{\textit{iteracji}}{\textit{watkow}}$ razy.

Takie przedstawienie pseudokodu przy pomocy jednego dodatkowego słowa kluczowego w znaczny sposób ułatwia przedstawienie kodu. Szczegóły odnośnie użycia technologii CUDA zostaną przedstawione w miejscach, gdzie niniejszy opis będzie niewystarczający.

Rozdział 4

Technologie

Niniejszy rozdział zawiera przegląd oraz krótki opis technologii użytych w procesie realizacji projektu. Przedstawiono najważniejsze cechy każdej z nich, pomijając szczegóły, wprowadzając jedynie zarys informacji na temat poszczególnych komponentów, na których budowana była aplikacja. Dokonano również podstawowego wprowadzenia do każdej z wymienionych technologii, co jest potrzebne do pełnego przedstawienia sposobu ich wykorzystania.

4.1 Język C# i platforma .NET

Język C# oraz .NET zostały pierwszy raz przedstawione podczas międzynarodowej konferencji dla programistów o nazwie *Professional Developers Conderence (PDC)* w lipcu roku 2000 [Cor00] - choć środowisko programistów słyszało o planach firmy Microsoft znacznie wcześniej, jednakże pod różnymi nazwami (m.in. COOL, COM3, Lightning). Pomimo sporego w owym czasie szumu wokół niedoskonałych systemów operacyjnych (np. Windows Me), czy też sporych opóźnień w dostarczaniu owych rozwiązań, firma zdecydowała się na stworzenie własnej platformy programistycznej [Ske10a].

Należy również podkreślić, że firma Microsoft nie była jedyną, która wspierała tworzenie języka - oprócz niej w sponsoring tego języka włączył się firmy Hewlett-Packard oraz Intel. Język C# jest, podobnie, jak język Java, otwartym językiem, jednakże w zupełnie innym tego słowa znaczeniu. Microsoft sprzyja ścieżce standaryzacyjnej języka, gdy Sun (firma posiadająca prawa do platformy Java) stopniowo udostępnia kod źródłowy Javy na zasadach *wolnego oprogramowania* (ang. *open source*) i dopuszczając, a nawet zachęcając do tworzenia innych *środowisk uruchomieniowych* (ang. *runtime environments*) Javy. Istnieją alternatywne implementacje CLI oraz C# (najbardziej znanym jest Mono [Mon11]), jednakże należy podkreślić, że nie obejmują one swoją implementacją wszystkiego, co kryje się pod nazwą Microsoft .NET Framework. Ostatecznie język C# oraz .NET zostały udostępnione w roku 2002, razem z narzędziem programistycznym Visual Studio .NET 2002.

4.1.1 Platforma .NET

Kiedy wprowadzono termin ".NET", został on wcielony w szereg różnych technologii wychodzących spod sztandarów firmy Microsoft. Na przykład Windows Live ID był nazywany .NET Passport, pomimo tego, że nie istniało żadne bezpośrednie połączenie pomiędzy tą technologią, a tym co jest obecnie znane, jako .NET. Na całe szczęście z czasem zaprzestano używania nazwy .NET w innych usługach i produktach firmy z Redmond. W chwili obecnej nazwa ".NET" wiąże się jedynie ze światem programowania.

Należy podkreślić, że .NET zawiera wspólną specyfikację języka (*CLS*, ang. *Common Language Specification*), która udostępnia zestaw podstawowych reguł, niezbędnych do integracji języków na jednej platformie. CLS określa minimalne wymagania, jakie musi spełniać język z rodziny .NET. Kompilatory zgodne z CLS tworzą obiekty, które mogą ze sobą współpracować.

Poniżej przedstawione zostaną trzy składniki, z jakich zbudowana jest platforma .NET oraz łączące je zależności.

Język, środowisko uruchomieniowe i biblioteki

W skład platformy .NET wchodzi trzy podstawowe składniki: język programowania, biblioteki oraz środowisko uruchomieniowe. Wprawdzie rozróżnienie tych trzech nie zawsze jest możliwe, to ważna jest świadomość istnienia wyraźnych różnic pomiędzy tymi trzema składnikami.

Język Język C# jest zdefiniowany przez swoją specyfikację (patrz [Cor06a]), która opisuje format kodu programu w C#, obejmując jednocześnie składnię oraz zachowanie. Nie jest tam zawarta natomiast informacja o platformie, na jakiej zostanie on uruchomiony, poza kilkoma kilkoma aspektami współpracy kompilatora z platformą.

W teorii dowolna platforma, która wspiera wymagane funkcje (opisane w specyfikacji), może posiadać kompilator, który będzie budował oprogramowanie przeznaczone dla danej platformy. Na przykład, kompilator języka C# mógłby produkować na podstawie kodu źródłowego dowolny inny format niż tzw. *język pośredni* (*IL*, ang. *Intermediate Language*).

Środowisko uruchomieniowe Środowisko uruchomieniowe jest częścią .NET odpowiedzialną za dokonanie wszystkiego, by uruchomiony kod IL działał zgodnie ze specyfikacją języka. W implementacji firmy Microsoft środowisko uruchomieniowe nazywa się *Common Language Runtime*, czyli w skrócie *CLR*.

Biblioteki *Biblioteki* (ang. *libraries*) dostarczają kod, który może być wykorzystywany przez oprogramowanie. Większość bibliotek frameworka .NET są gotowymi produktami języka IL (patrz punkt 4.1.2) z *kodem natywnym* (ang. *native code*) wykorzystywanym tylko wtedy, gdy jest to potrzebne. Należy podkreślić, że kodu bibliotek jest znacznie więcej niż kodu środowiska uruchomieniowego. W taki sam sposób można spojrzeć na samochód (biblioteki), który jest znacznie bardziej skomplikowaną jednostką niż sam jego silnik (środowisko uruchomieniowe).

Warto podkreślić, że istnieje podział na biblioteki standardowe oraz pozostałe, które nie zostały ujęte w specyfikacji języka [Cor06a]. Pisząc program, który wykorzystuje tylko te pierwsze można mieć dużą pewność, że będzie on możliwy do uruchomienia na dowolnej implementacji platformy - czy to Mono, .NET, czy jakiegokolwiek innej.

Termin *.NET* odnosi się do kombinacji środowiska uruchomieniowego wraz z bibliotekami dostarczonymi przez Microsoft, a także kompilatory C# oraz VB.NET. .NET może być pojmowany, jako cała *platforma programistyczna* (ang. *development platform*) zbudowana nad systemem Windows.

4.1.2 Kompilacja i język MSIL

Programy pisane na platformie .NET nie są kompilowane do plików wykonywalnych, jak to ma miejsce w przypadku chociażby języka C++. Są one kompilowane do podzespółów, które składają się z instrukcji standardowego *języka pośredniego Microsoft* (*MSIL*, ang. *Microsoft Intermediate Language*). Środowisko CLR dokonuje przekształcenia instrukcji tego języka na kod maszynowy i wykonuje je. Pliki MSIL (często zamiennie używa się skrótowej nazwy IL) stworzone przez kompilator języka C# są niemalże takie same, jak pliki IL innych języków z rodziny .NET. Kluczową cechą środowiska CLR jest to, że jest ono wspólne, co oznacza, że programy napisane w języku VB.NET czy też C# są obsługiwane przez te samo środowisko uruchomieniowe.

Kompilacja kodu w języku C# do IL odbywa się w momencie budowania projektu. W wyniku czego kod IL jest zapisywany na dysku w postaci plików binarnych. W momencie uruchomienia programu zachodzi ponowna kompilacja - tym razem kodu IL. Jest to tak zwana kompilacja *JIT* (ang. *Just In Time*), w wyniku której powstaje kod maszynowy, wykonywany przez procesor maszyny, na której program został uruchomiony.

Kompilatory JIT, standardowo, są uruchamiane *na żądanie* (ang. *on demand*) - tzn. w momencie wywołania metody, kompilator JIT analizuje kod IL i tworzy bardzo wydajny (dzięki licznym optymalizacjom dokonywanym na tym etapie) kod maszynowy. Gdy aplikacja jest uruchomiona, kompilacja zachodzi jedynie wtedy, kiedy jest to potrzebne, a po kompilacji JIT kod znajduje się w pamięci podręcznej, przez co możliwe jest jego szybkie wykorzystywanie w późniejszych etapach przetwarzania. Wynika z tego, że aplikacja działa tym szybciej, im dłużej jest uruchomiona, ponieważ możliwe jest wykorzystanie większej ilości skompilowanego kodu.

Specyfikacja CLS wymusza podobieństwo kodu IL w przypadku każdego języka z rodziny .NET. Wynika z tego bezpośrednio, że obiekty utworzone w jednym języku mogą być wykorzystywane w programach napisanych w innych językach (z rodziny .NET oczywiście). Oznacza to np. fakt, że możliwe jest przygotowanie klasy bazowej w języku VB.NET oraz utworzenie od niej klas pochodnych w programie napisanej w języku C#.

Data	Framework	Visual Studio	C#	CLR
Luty 2002	1.0	2002	1.0	1.0
Kwiecień 2003	1.1	2003	1.2	1.1
Listopad 2005	2.0	2005	2.0	2.0
Listopad 2006	3.0	(rozszerzenia do wersji 2005)	[brak]	2.0
Listopad 2007	3.5	2008	3.0	2.0 SP1
Kwiecień 2010	4.0	2010	4.0	4.0

TABLICA 4.1: Poszczególne wersje .NET Framework oraz ich składniki [Ske10b]

4.1.3 Język programowania C#

Język C# jest niezwykle wydajnym narzędziem do implementacji współczesnych technik programistycznych. Jest on językiem obsługującym strukturalne, oparte na komponentach i obiektowe programowanie, czego należy oczekiwać od współczesnego języka, zbudowanego na doświadczeniach z C++ oraz Java.

Programiści mówiący o nowej wersji .NET, mają zazwyczaj na myśli ważne wydania framework'a. W większości przypadków wydanie nowej wersji platformy jest połączone z wydaniem Visual Studio, czyli zintegrowanego środowiska programistycznego firmy Microsoft. W tabeli 4.1 zebrane zostały informacje o poszczególnych wersjach, które były wydawane wraz z jakimi komponentami oraz kiedy.

Obiektowość Programowanie obiektowe (OOP, ang. *Object Oriented Programming*) jest to paradygmat programowania, w którym program definiuje się przy pomocy obiektów. Podstawowe cechy języka, który realizuje paradygmat programowania obiektowego przedstawił Alan Kay w odniesieniu do języka *Smalltalk*, pierwszego poprawnie zrealizowanego języka obiektowego, a tym samym jednego z poprzedników C#. Cechy te opisują czyste podejście obiektowe.

Wszystko jest obiektem. Obiekt można przedstawić jako specjalną zmienną, która nie tylko zawiera dane, ale może również realizować żądania, czyli wykonywać na swoich danych pewne ściśle określone operacje. Teoretycznie każdy element świata rzeczywistego, np. samochód, dom, zamek, pracownik, może być reprezentowany w programie przy pomocy tak skonstruowanego obiektu.

Aplikacja jest zbiorem komunikujących się między sobą obiektów. Przesyłanie komunikatów do obiektu to żądanie od niego wykonania pewnej operacji. Można to nazwać wykonaniem funkcji należącej do konkretnego obiektu. Przykładem może być następujący scenariusz: obiekt *Kierowca* wysyła do obiektu *Samochód* komunikat *przyspiesz*, co powoduje wykonanie operacji zwiększenia prędkości.

Każdy obiekt posiada własną pamięć, na którą składają się inne obiekty. Tworzenie nowego obiektu polega na łączeniu w jeden element grupy już istniejących obiektów. Powstaje w ten sposób wielowarstwowa aplikacja, która jednocześnie ukrywa swoją złożoność za prostymi obiektami. Przykładowo obiekt *Samochód* można zbudować z następujących obiektów: *Karoseria*, *Koło*, *Silnik* itd. Jednocześnie chcąc zwiększyć prędkość pojazdu odwołujemy się do

obiektu *Samochód*, a nie bezpośrednio do obiektu *Silnik*, który odpowiada za jego prędkość. Dopiero akcja *przyspiesz* wykonana przez obiekt *Samochód* wykonuje odpowiednie operacje na poszczególnych podzespołach, z których się składa.

Każdy obiekt posiada swój typ. W odniesieniu do obiektu słowo *typ* można zastąpić słowem *klasa*.

Każdy obiekt jest instancją pewnej klasy, której głównym zadaniem jest zdefiniowanie jakie komunikaty można wysłać do obiektu będącego jej egzemplarzem.

Wszystkie obiekty danego typu obsługują te same komunikaty. Każdy obiekt danej klasy obsługuje wszystkie komunikaty zdefiniowane w klasie. Dodatkowo jeśli np. obiekt typu *Student* jest jednocześnie obiektem typu *Człowiek*, to obsługuje wszystkie komunikaty zdefiniowane dla typu *Człowiek*. Umożliwia to pisanie bardziej uniwersalnego kodu, który będzie obsługiwał wszystkie obiekty pasujące do typu *Człowiek* [Lib06].

Współcześnie, aby język programowania został uznany za obiektowy musi charakteryzować się wymienionymi poniżej cechami.

Abstrakcja Każdy obiekt systemu jest widziany jako abstrakcyjny „wykonawca”, który może realizować pracę, określać i zmieniać swój stan oraz wysyłać komunikaty do innych obiektów w systemie. Jednocześnie nie ujawnia on jak zostały zaimplementowane dane cechy.

Hermetyzacja Czyli ukrywanie wewnętrznej implementacji obiektu. Gwarantuje to, że obiekt nie ma możliwości zmiany stanu wewnętrznego innego obiektu w nieprzewidziany przez programistę sposób. Tylko metody (funkcje) składowe obiektu mają prawo do modyfikacji jego stanu. Dodatkowo pozwala to na swobodną modyfikację kodu przez programistę o ile nie zmieniają się metody składowe. Przykładowo w pierwszej wersji klasy programista może zaimplementować listę jako tablicę o stałym rozmiarze, natomiast w kolejnej wersji zmienić tablicę na listę jednokierunkową.

Polimorfizm Referencje mogą wskazywać na obiekty różnego, ale zgodnego typu. Wywołanie metody dla referencji spowoduje wykonanie operacji odpowiedniej dla pełnego typu obiektu wywoływanego. Jeśli ma to miejsce w trakcie działania programu, nazywa się to późnym wiązaniem lub wiązaniem dynamicznym.

Dziedziczenie Umożliwia definiowanie specjalizowanych obiektów na podstawie ich ogólniejszych odpowiedników. Podczas definiowania obiektów specjalizowanych nie jest wymagana redefinicja całej funkcjonalności obiektu bazowego, ale tylko ta, której brakuje w obiekcie ogólniejszym, lub ta, której sposób działania chcemy zmienić.

4.1.4 Zapytania do obiektów czyli LINQ

Language INtegrated Query (*LINQ*) stanowi część technologii frameworka .NET, zaprojektowany i opracowany przez Andersa Hejlsberga, który jest znany z zaprojektowania języka Delphi oraz C#. Technologia LINQ stworzona została w celu uproszczenia wykonywania zapytań na

obiektach, a jego składnia przypomina SQL, przez co jest prostym narzędziem możliwym do wykorzystania przez programistów.

W świecie języków kompilowanych LINQ wydaje się być elementem nie pasującym do standardów, szczególnie przykładowe zapytanie poniżej przypomina bardziej SQL niż język programowania C#.

```
1 var adultNames = from person in people
                      where person.Age >= 18
3                      select person.Name;
```

Jednym z powodów, dla których oba języki wyglądają inaczej jest fakt, że LINQ jest *deklaratywny*, a nie *imperatywny* - zapytanie LINQ pyta o pewne właściwości obiektów, zamiast przedstawiać dokładne kroki do osiągnięcia celu. Takie podejście bliskie jest funkcjonalnemu podejściu do rozwiązywania problemów. Co ważne LINQ nie jest przeznaczone jedynie do zapytań wydawanych do baz danych - może ono być bowiem wykorzystywane w operacjach przeprowadzanych w pamięci operacyjnej, na obiektach programu. Dzięki temu właśnie LINQ zdobyło sobie rzesze wiernych użytkowników, którzy na codzień korzystają z dobrodziejstw tego nowoczesnego narzędzia programistycznego.

LINQ został wprowadzony do języka C# w wersji 3. Celem pracy nie jest opis tej technologii, ale ponieważ stanowiła ona jedną z częściej wykorzystanych została krótko wprowadzona. Po więcej szczegółów należy sięgnąć po [Ske10b], bądź poświęcić tylko tej technologii książkę [CK09].

4.1.5 Kwantyfikator *as parallel* czyli Parallel LINQ

W opisach procedur znajdowania zbiorów częstych używany jest kwantyfikator **as parallel**, który jest głównym elementem opisywanego algorytmu ParallelApriori.

W implementacji algorytmu Apriori technologia LINQ była jedną z częściej wykorzystywanych na obiektach. W roku 2006 Joe Duffy zapowiedział na swoim blogu wprowadzenie nowej technologii, rozszerzającej możliwości LINQ [Duf11]. PLINQ (lub też Parallel LINQ) zostało włączone w skład .NET 4, pozwalając budować programy równoległe w bardziej abstrakcyjny sposób.

Głównym założeniem PLINQ jest fakt, że biorąc wyrażenie LINQ, które wykonywane jest w stosunkowo długim czasie, można zapewnić szybsze jego wykonanie poprzez użycie wielu *wątków* (ang. *threads*), które wykorzystają możliwości wielu rdzeni na komputerze - z jak najmniejszą liczbą zmian w kodzie, jak to jest tylko możliwe. Jak w przypadku innych kwestii związanych z programowaniem równoległym - nie jest to aż tak proste podejście, jednakże nawet zwykła zmiana wyrażen prowadzi zazwyczaj do przyspieszenia działania aplikacji.

Punktem startu dla przeobrażenia wyrażenia LINQ do jego wersji równoległej jest metoda `AsParallel()`

Jej zastosowanie przekształca wyrażenie, które zostaje podzielone, a następnie będzie wykonywane w środowisku równoległym. Zastosowanie słowa kluczowego **as parallel** oznacza, że wy-

rażenie implementowane w przypadku klasycznego algorytmu Apriori (patrz część 2.2.1) w LINQ zostało przekształcone w miarę możliwości do nowej technologii, wprowadzonej w ostatniej wersji frameworka .NET.

4.1.6 .NET Framework 4.0

Wydany w roku 2010 w wersji 4.0 framework platformy stanowi podstawę implementacyjną stworzonego w ramach niniejszej pracy oprogramowania. Aplikacja stworzona została przy pomocy najnowszej wersji, czyli 4.0 w zintegrowanym środowisku programistycznym Visual Studio 2010 Professional.

W wersji 4.0 rozszerzono wiele funkcji dostępnych we wcześniejszych wersjach - m.in. wprowadzając PLINQ, czyli równoległe wykonywanie zapytań LINQ (ang. *Language-INtegrated Query*), które zostało wykorzystane w jednym z zaimplementowanych w ramach niniejszej pracy algorytmie. Poza tym w implementacji wykorzystano szereg wprowadzonych w 4. wersji usprawnień - m.in. nazywane parametry, czy też parametry opcjonalne w metodach.

Ponieważ nie jest celem niniejszej pracy opisywanie poszczególnych wersji języka, czy też platformy, pominięte zostały szczegóły odnośnie poszczególnych wersji tejże platformy. Doskonałym źródłem informacji o zmianach w poszczególnych wersjach, a także przyczynach zachodzących zmian jest książka Jona Skeeta [Ske10b].

4.2 CUDA

W realizacji projektu wykorzystana została uniwersalna architektura procesorów wielordzeniowych, która szerzej została omówiona w rozdziale 3. Użyto NVIDIA CUDA w wersji 3.0, dostępnej do pobrania pod adresem <http://developer.nvidia.com/cuda-toolkit-30-downloads> (zweyfikowana dostępność w dniu 26 września 2011). Językiem wybranym do implementacji metod optymalizujących algorytm został CUDA C, czyli rozszerzenie języka C.

4.2.1 CUDA na platformie .NET

Jak opisano w części 4.1 językiem, który wykorzystano do napisania aplikacji testowej był C# uruchamiany na platformie .NET. W nawiązaniu do rozdziału 3, w którym napisano, że NVIDIA wypuściła rozszerzenie języka C (tzn. CUDA C), w którym pojawiły się instrukcje bezpośrednio związane z uniwersalną architekturą wielowątkową CUDA.

Istnieje możliwość użycia mechanizmów współpracy kodu zarządzanego i natywnego, jednakże istnieją metody znacznie prostrze wykorzystania możliwości GPU do obliczeń. Można wykorzystać specjalną bibliotekę, stworzoną do umożliwienia programistom skorzystania z CUDA C wewnątrz programów zarządzanych.

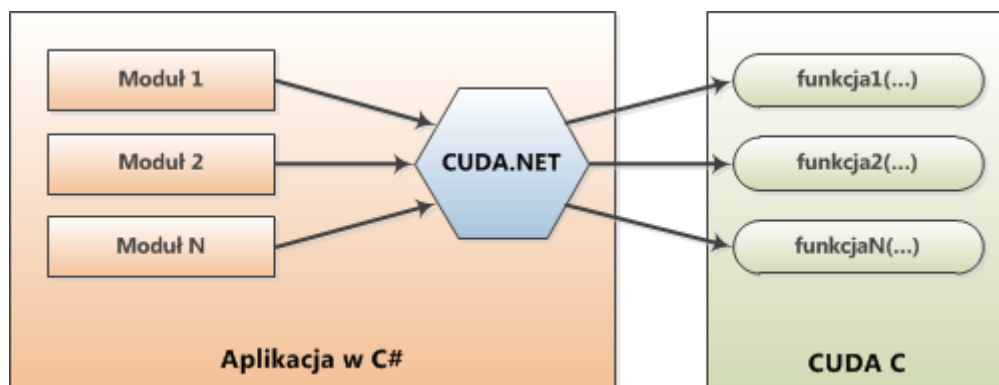
Aby skorzystać z karty graficznej wewnątrz programu przygotowanego na platformę .NET użyta została biblioteka o nazwie CUDA.NET [fASSL11]. Sposób wykorzystania CUDA wewnątrz aplikacji na platformę .NET przedstawione są poniżej.

1. Aplikacja oraz związana z nią logika biznesowa (np. przygotowanie i wyświetlanie danych) pisana jest w języku platformy .NET - w wypadku programu powiązanego z niniejszą pracą był to C# w wersji 4.0.
2. Funkcje obliczeniowe przygotowywane są w języku CUDA C - jako pliki z rozszerzeniem *.cu*. Jest to zgodnie z konwencją pisania programów na platformę CUDA [Cor11b].
3. Pliki z punktu poprzedniego powinny być skompilowane przy użyciu kompilatora przygotowanego przez firmę NVIDIA - w wyniku tej operacji powstaną pliki o rozszerzeniu *.cubin*. Poniżej zaprezentowane zostało polecenie dla kompilatora wykorzystywane w budowaniu aplikacji, która stanowi podstawę niniejszej pracy.

```
> nvcc *.cu --cubin
```

4. W aplikacji głównej programu należy dodać referencje do biblioteki CUDA.NET, a następnie dopisać kod pośredniczący, w którym przygotowane zostaną dane, uruchomione zostaną funkcje obliczeniowe na karcie graficznej, a następnie odczytane zostaną wyniki tych obliczeń.

Opisaną powyżej architekturę w sposób schematyczny zaprezentowano na rysunku 4.1.



RYSUNEK 4.1: Schematyczne przedstawienie architektury aplikacji wykorzystującej technologię CUDA przy pomocy biblioteki CUDA.NET

4.3 Środowisko programistyczne

W celu ułatwienia pracy podczas przygotowywania aplikacji wykorzystanych został szereg narzędzi ułatwiających pracę programiście, a także programistom (w przypadku, gdyby nad projektem pracowała więcej niż jedna osoba). Poniżej krótko opisane zostaną poszczególne narzędzia.

4.3.1 Visual Studio 2010

W pracy nad projektem głównym narzędziem wykorzystywanym do implementacji aplikacji było zintegrowane środowisko programistyczne (ang. *Integrated Development Enviroment*, *IDE*) firmy Microsoft w wersji 2010. Przez IDE rozumie się aplikację (lub ich zespół, czyli środowisko)

służącą do tworzenia, modyfikacji, testowania oraz konserwacji oprogramowania. Programy takie charakteryzują się tym, że udostępniają złożoną, wieloraką funkcjonalność obejmującą edycję kodu źródłowego, kompilowanie kodu źródłowego, tworzenie zasobów programu (tzn. formatek / ekranów / okien dialogowych, menu, raportów, elementów graficznych takich jak ikony, obrazy itp.), tworzenie baz danych, komponentów i innych. Wiele tych funkcji zależy od firmy tworzącej dany program, a także celu wykorzystania danego narzędzia.

Visual Studio [Cor09] jest używane do tworzenia różnego rodzaju oprogramowania na platformie .NET, na różnego rodzaju platformy: Microsoft Windows, Windows Mobile, Windows CE, Microsoft Silverlight.

Wykorzystywaną wersją było Microsoft Visual Studio 2010 Ultimate, czyli najwyższa dostępna wersja narzędzia dostępna dla studentów w ramach programu MSDNAA. Samo narzędzie zostało rozszerzone kilkoma dodatkami, o których przeczytać można w kolejnych punktach.

4.3.2 System kontroli wersji

W celu zarządzania kodem wykorzystany został rozproszony system kontroli wersji o nazwie *GIT*. Stworzył go Linus Torvalds, jako narzędzie wspomagające rozwój jądra systemu operacyjnego Linux. Git jest opublikowany na licencji GNU GPL v2, przez co stanowi wolne oprogramowanie z ogólnie dostępnym kodem źródłowym. Pierwsza wersja narzędzia Git została wydana 7 kwietnia 2005 roku, kiedy to zastąpiła używany wcześniej w rozwoju Linuksa system kontroli wersji BitKeeper, który w owym czasie zaprzestał darmowe wspieranie aplikacji tworzonych na zasadach wolnego oprogramowania.

Każdy katalog Gita jest pełnoprawnym repozytorium z kompletną historią oraz pełną bazą rewizji wstecznych, niezależnym od dostępu do sieci, czy też centralnego serwera. Dzięki takiemu rozwiązaniu programista jest niezależny od jakichkolwiek awarii, czy też centralnego serwera.

W tworzeniu aplikacji wykorzystano lustrzany serwer gita utworzony na serwerze GitHub [Inc11], będącym uzupełnieniem oraz rezerwą w wypadku awarii głównego repozytorium, zlokalizowanego na komputerze developera, a jednocześnie autora niniejszej pracy.

4.3.3 Budowanie projektu

Ponieważ w tworzeniu aplikacji wykorzystanych zostało kilka różnych technologii (o czym można było przeczytać w poprzednich rozdziałach) zdecydowano, że należy utworzyć jeden sposób do budowania aplikacji oraz zależności pomiędzy poszczególnymi modułami programu. Wykorzystanie samego Visual Studio pozostawało niewystarczające ze względu na wykorzystywaną technologię CUDA.NET, w ramach której należało kompilować pliki .cu, czyli źródła funkcji wykonywanych na karcie graficznej.

Z tychże powodów zdecydowano się wykorzystać psake [Kov11], czyli narzędzie do automatycznego budowania projektu napisane w PowerShell [Cor11a]. Poprzez wykorzystywanie w skryptach

funkcji oraz stylu wykorzystywanego w PowerShell ułatwia przygotowanie odpowiednich skryptów programiście przeznaczonych do wykorzystania w budowie projektu.

4.4 Testowanie

W ramach projektu wykorzystane zostały *testy jednostkowe* (*unit testing*), które miały zagwarantować poprawność pojedynczych części składowych systemu - np. metod, bądź obiektów, procedur, funkcji. Testowany fragment poddawany jest weryfikacji przez test, który wykonuje go i porównuje wynik (np. zwróconą wartość, stan obiektu, rzucone wyjątki) z oczekiwanymi wartościami - zarówno pozytywnymi, jak i negatywnymi (często testuje się odpowiednie zachowanie programu dla niepoprawnych danych).

Podstawową zaletą testów jednostkowych jest możliwość ich wykonywania na bieżąco na modyfikowanych elementach programu, co często umożliwia wychwycenie błędu natychmiast po jego pojawieniu się w kodzie programu oraz szybką jego lokalizację, zanim dojdzie do wprowadzenia błędnego fragmentu do ostatecznego kodu programu. Testy jednostkowe są również formą specyfikacji, ponieważ poszczególne metody muszą spełniać odpowiednie założenia.

4.4.1 Usprawnienie procesu testowania

W ramach tworzenia testów dla projektu wykorzystane zostały nowe narzędzia dostarczone przez firmę Microsoft, które jeszcze bardziej ułatwiają i usprawniają tworzenie aplikacji.

Code Contracts

Code Contracts [Res11] są jednym z najnowszych rozszerzeń platformy .NET przygotowane przez Microsoft Research, czyli jednostkę badań i rozwoju dla produktów firmy Microsoft. Code Contracts dostarczają metody wyrażenia założeń wobec metod (oraz innych elementów) w programach na platformie .NET. Kontrakty posiadają formy warunków wejściowych (ang. *preconditions*), wyjściowych (ang. *postconditions*) oraz ciągłych (ang. *object invariants*). Zachowują się one (kontrakty), jak dodatkowa dokumentacja dla wewnętrznego oraz zewnętrznego API aplikacji - tj. dostarczają dodatkowych informacji o metodach wewnątrz projektu. Kontrakty są wykorzystywane do udoskonalenia procesu testowania poprzez weryfikację na poziomie uruchomieniowym (ang. *runtime checking*), umożliwiając statyczną weryfikację kontraktu, a także poprzez generowanie dokumentacji kodu, która posiada dodatkowe informacje o metodach - dokumentacja nie bazuje jedynie na informacjach dostarczonych i napisanych przez autora projektu, a także poszczególnych kontraktach na i wewnątrz metod.

Biblioteka Code Contracts wprowadzają do każdego z języków platformy .NET korzyści wynikające z podejścia *projektowanie przez kontrakt* (ang. *design-by-contract*). Poniżej wymienione są korzyści wynikające bezpośrednio z wykorzystania Code Contracts.

- Udoskonalone testowanie

- Każdy z kontraktów działa na zasadzie wyroczni (ang. *oracle*), która powoduje przejście lub nie danego testu.
- Automatyczne narzędzia testowania (takie jak Pex, opisany w rozdziale 4.4.1) mogą wykorzystać wiedzę dostarczoną przez kontrakty do wygenerowania bardziej dokładnych testów jednostkowych, poprzez odfiltrowanie argumentów wejściowych, które nie spełniają warunków wejścia.
- **Statyczna weryfikacja** Microsoft przygotował zestaw narzędzi do statycznej weryfikacji kodu, dostarczając informacji o możliwych niedoskonałościach oprogramowania. Obecnie narzędzie może wykorzystać kontrakty do dostarczenia większej ilości i bardziej dokładnych informacji na temat błędów w programie.
- **Dokumentacja API** Dokumentacja API projektu często jest niepełna, bądź niewystarczająca. Te same kontrakty, które wykorzystywane są do testowania podczas uruchomienia, a także do statycznej weryfikacji mogą być również wykorzystane do wygenerowania lepszej dokumentacji projektu, np. poprzez dostarczenie informacji o tym, które parametry wejściowe nie mogą mieć wartości *null*.

Wykorzystanie biblioteki pozwala praktycznie każdemu językowi z platformy .NET skorzystać z możliwości, jakie niosą ze sobą kontrakty wewnątrz kodu aplikacji. Nie ma potrzeby pisania parsera, czy też kompilatora (lub dodatków do niego) języka. Co więcej, odpowiednie kompilatory języków w naturalny sposób dokonują weryfikacji poprawności poszczególnych kontraktów (sprawdzanie typów oraz nazw), a także produkują skompilowane formy kontraktów w języku MSIL. Tworzenie kontraktów w Visual Studio pozwala programiście skorzystać dodatkowo z możliwości *podpowiedzi* (ang. *intellisense*) dostarczanych przez serwisy języka. Poprzednie podejścia poprzez zastosowanie atrybutów .NET przegrywają z tym podejściem, ponieważ nie mogły w wystarczający sposób być wykorzystane w bardziej skompilowanych kontraktach, a także nie mogły one być wykorzystane w trakcie weryfikacji podczas kompilacji.

Kontrakty są wyrażone poprzez wywołania statycznych metod na początku metody. Narzędzia projektu biorą odpowiedzialność za interpretację zadeklarowanych kontraktów w odpowiednich miejscach programu. Biblioteka Code Contracts jest zlokalizowana w przestrzeni nazw *System.Diagnostics.Contracts* platformy w wersji 4.0.

Pex i Moles

Pex jest narzędziem rozszerzającym możliwości Visual Studio 2010 w kwestii testów jednostkowych na aplikacjach platformy .NET. Pex tworzy zestaw *testów białej skrzynki* (ang. *white box testing*), zwanych również testami strukturalnymi lub szklanej skrzynki. Testy takie polegają na poddaniu programu testowaniu poprzez podawanie na wejściu takich danych, by program przeszedł przez każdą ścieżkę wykonania programu. Zasady przejścia definiowane są przez kryteria pokrycia wszystkich pętli oraz wszystkich warunków w programie (metodzie programu). Testy białej

skrzynki nie są w stanie wykazać braku implementacji funkcji, którą powinien posiadać docelowy produkt. Testy te dokonują sprawdzenia operacji wykonywanych wewnątrz zaimplementowanych metod.

Co ważne Pex wykorzystuje informacje dostarczane przez Code Contracts, co pozwala temu narzędziu wygenerowanie bardziej dokładnych testów dla poszczególnych metod. Testy generowane są automatycznie, jednocześnie zapewniają one wysoki stopień pokrycia kodu testami. Istnieje również możliwość wygenerowania małych zestawów danych prosto z edytora Visual Studio i zachowania ich, jako małe zestawy testów z wysokim stopniem pokrycia (dla danej metody, czy też klasy).

Pex dostarczany jest w zestawie z Moles. Biblioteka ta umożliwia zastąpienie każdej metody w programie dowolnym delegatem. Dzięki temu możliwe jest wyizolowanie kodu z zewnętrznych zależności, w celu przetestowania działania kodu zgodnie z intencją programisty. Moles automatycznie generuje odpowiednie delegaty dla każdej zależności wobec zewnętrznej biblioteki.

Rozdział 5

Algorytm równoległy

Niniejszy rozdział można traktować, jako uzupełnienie rozdziału 2, w którym to dokonano opisu podstaw teoretycznych wraz z dwoma wiodącymi algorytmami do odkrywania reguł asocjacyjnych. W ramach niniejszej pracy opracowany został algorytm równoległy, oparty na algorytmie opisanym w sekcji ??.

Wcześniej zaś przedstawione zostanie zastosowanie algorytmu Apriori (opisanego w części 2.2.1 pracy), w którym wykorzystane zostały funkcje dostarczone przez najnowsze wydanie frameworka .NET.

W rozdziale tym wykorzystywane są oznaczenia wprowadzone w rozdziale 2.

5.1 Algorytm ParallelApriori

Pierwszym algorytmem wykorzystującym możliwości procesorów równoległych jest ParallelApriori - stworzony na potrzeby niniejszej pracy.

Jak wspomniano już wcześniej, zagadnienie odkrywania reguł asocjacyjnych można podzielić na dwa etapy [AIS93]:

1. Odkrywanie zbiorów częstych, których wartość wsparcia jest wyższa od wartości *minsup*.
2. Generowanie reguł asocjacyjnych na podstawie znalezionych zbiorów częstych.

Na tym etapie możliwe jest tworzenie reguł, w których w zbiorze *poprzedników* (X z oznaczeń z definicji 1) jest wiele elementów oraz jeden w *następniku* (zbiór Y z definicji 1) [AIS93] lub dopuszczana jest możliwość wielu elementów również w następniku [AS94]. Niniejsza praca analizuje algorytmy generujące reguł, w których oba zbiory mogą być zbiorami wieloelementowymi.

5.1.1 Generowanie zbiorów częstych

Podobnie jak klasyczny algorytm Apriori [AS94], ParallelApriori dokonuje analizy bazy danych DB , by w kolejnych iteracjach generować rodziny coraz to liczniejszych zbiorów, będących zbiorami częstymi dla zadanej wartości *minsup*. Algorytm zaczyna od znalezienia wszystkich zbiorów

jednoelementowych, które są zbiorami częstymi. Następnie w każdym kolejnym kroku generowane są zbiory częste na podstawie zbiorów wygenerowanych w kroku poprzednim. Proces ten jest kontynuowany do momentu aż nie zostaną znalezione żadne zbiory częste.

Algorytm generuje zbiory kandydatów jedynie na podstawie zbiorów częstych odkrytych w kroku poprzednim - co ważne generowanie ich odbywa się bez wielokrotnego przeglądania bazy danych transakcji. Każdy zbiór częsty zawierający k elementów może być wygenerowany na podstawie połączenia dwóch zbiorów posiadających $k - 1$ elementów, a na koniec kasując te zbiory, których jakikolwiek podzbiór nie jest częsty [AS94].

Procedura APRIORI FREQUENT SET GENERATION przedstawia pseudokod realizujący opisywany w tym rozdziale algorytm generowania zbiorów częstych.

APRIORI FREQUENT SET GENERATION

```

1   $L_1 \leftarrow \{1\text{-zbiory częste}\}$ 
2  for ( $k = 2; L_{k-1} \neq \emptyset; k++$ )
3      do  $C_k \leftarrow pAprioriGen(L_{k-1})$ 
4          for each transakcja  $t \in DB$  as parallel
5              do  $C_t \leftarrow subset(C_k, t)$ 
6                  for each kandydat  $c \in C_t$  as parallel
7                      do  $c.count++$ 
8           $L_k \leftarrow \{c \in C_k | c.count \geq minsup\}$ 
9  Answer  $\leftarrow \bigcup_k L_k$ 
```

Procedura pAprioriGen

Procedura *pAprioriGen* reprezentuje proces tworzenia zbiorów k -elementowych kandydatów na podstawie zbiorów wejściowych $(k - 1)$ -elementowych.

Jak łatwo zauważyć wynikiem działania PARALLEL JOIN STEP są zbiory k -elementowe, które powstały na podstawie zbioru zbiorów wejściowych L_{k-1} , a ich zawartość różni się tylko jednym, ostatnim elementem. Ważnym faktem jest to, iż elementy w zbiorach są uporządkowane leksyko-graficzne, co znacząco ułatwia implementację tej procedury.

PARALLEL JOIN STEP

```

1  insert into  $C_k$ 
2  select  $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$  as parallel
3  from  $L_{k-1}$   $p, L_{k-1}$   $q$ 
4  where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$ 
```

Warto zauważyć, że PARALLEL JOIN STEP jest ekwiwalentem rozszerzania zbioru L_{k-1} każdym elementem zbioru elementów I , a następnie kasowania tych $(k - 1)$ -zbiorów otrzymanych przez usuwanie $(k - 1)$ elementu, które nie są w L_{k-1} .

Widać, że procedura PARALLEL JOIN STEP jest rozszerzeniem procedury JOIN STEP z części ?? niniejszej pracy. Rozszerzona została ona o modyfikator **as parallel**, który opisany został w części 4.1.5 pracy.

W przypadku tego algorytmu zdecydowano się nie wprowadzać procedury PRUNE STEP obecnej w klasycznej implementacji algorytmu Apriori. Decyzja była podjęta na podstawie intuicji, że generowanie podzbiorów pewnego zbioru i porównywanie z innym zbiorem jest bardziej kosztowne niż sprawdzenie wartości *sup* dla danego zbioru k -elementowego, będącego jednym z elementów wyniku działania procedury *pAprioriGen*.

5.1.2 Generowanie reguł asocjacyjnych

Zgodnie z zasadą działania algorytmu Apriori po zakończeniu pierwszego etapu algorytm ParallelApriori przystępuje do etapu drugiego, czyli do budowania reguł asocjacyjnych na podstawie odkrytych zbiorów. Podobnie, jak w [AS94] algorytm będący przedmiotem analizy niniejszej pracy, generuje wszystkie możliwe reguły asocjacyjne dla zadanego zbioru - zupełnie jak ma to miejsce w implementacji algorytmu opisanej w 2.2.1.

Aby wygenerować reguły, dla każdego zbioru częstego l znajdowane są niepuste podzbiory - podzbiór taki oznaczony jest jako a . Dla takich oznaczeń wygenerowana zostanie reguła $a \Rightarrow (l - a)$, jeżeli spełniona jest nierówność $\frac{\text{support}(l)}{\text{support}(a)} \geq \text{minconf}$. Warto zauważyć, że dla każdego zbioru częstego generowane są wszystkie możliwe niepuste podzbiory - zapewnia to, że odkryte zostaną wszystkie możliwe reguły dla zadanego zestawu danych.

Procedura PARALLEL GENERATE FREQUENT ITEMSETS prezentuje generowanie reguł asocjacyjnych na podstawie odkrytych k -zbiorów częstych l_k będących elementami zbioru L_k ($l_k \in L_k$).

PARALLEL GENERATE FREQUENT ITEMSETS

```
1  for each zbiór częsty  $l_k$ ,  $k \geq 2$  as parallel
2      do call pGenrules( $l_k$ ,  $l_k$ )
```

W powyższym algorytmie wykorzystana została funkcja PGENRULES, która na podstawie dwóch zbiorów generuje reguły asocjacyjne. Zapis pseudokodu tej funkcji przedstawiony jest poniżej.

PGENRULES(l_k : k -zbiór częsty, a_m : m -zbiór częsty)

```
1   $A \leftarrow \{(m-1)\text{-zbiór } a_{m-1} | a_{m-1} \subset a_m\}$ 
2  for  $a_{m-1} \in A$  as parallel
3      do  $\text{conf} \leftarrow \frac{\text{support}(l_k)}{\text{support}(a_{m-1})}$ 
4      if  $\text{conf} \geq \text{minconf}$ 
5          then output reguła  $a_{m-1} \Rightarrow (l_k - a_{m-1})$  //ufność =  $\text{conf}$  oraz wsparcie=  $\text{support}(l_k)$ 
6          if  $m - 1 > 1$ 
7              then call pGenrules( $l_k$ ,  $a_{m-1}$ ) //generowanie reguł podzbiorów zbioru  $a_{m-1}$ 
```

Zupełnie, jak to miało miejsce w przypadku opisu generowania zbiorów częstych dla zadanego zbioru danych, można zauważyć stosowanie słowa kluczowego **as parallel**, którego znaczenie opisane zostało w części 4.1.5.

5.2 Algorytm CudaApriori

Opracowany algorytm wykorzystujący możliwości kart graficznych został oparty na algorytmie Apriori, celem udowodnienia, że następuje zysk czasowy dzięki wykorzystaniu mocy obliczeniowej współczesnych jednostek graficznych.

Jak wspomniano już wcześniej (zarówno w 2.2.1 oraz 5.1), zagadnienie odkrywania reguł asocjacyjnych można podzielić na dwa etapy [AIS93]:

1. Odkrywanie zbiorów częstych, których wartość wsparcia jest wyższa od wartości *minsup*.
2. Generowanie reguł asocjacyjnych na podstawie znalezionych zbiorów częstych.

Ponieważ metody te będą w części wykonywane na karcie graficznej. Z tego też powodu należy przygotować specjalne struktury danych, w których przechowywane będą dane o transakcjach oraz elementach w nich występujących.

5.2.1 Przygotowanie danych wejściowych do algorytmu

Należy podkreślić, że w trakcie obliczeń na karcie graficznej używane są inne struktury danych, ponieważ inne jest podejście do problematyki, a także inne jest przeznaczenie tego procesora. Jedną z podstawowych struktur jest bitmapa - w ogólności jest to plik wykorzystujący rastrowy sposób prezentacji dwuwymiarowej grafiki komputerowej, który polega na przypisaniu danemu pikselowi w tablicy (dwuwymiarowej) określonego koloru w danym trybie koloru. W specyficznych wypadkach bitmapa może być plikiem, w którym kolor zostanie zastąpiony wartością binarną.

Dodatkowo, dzięki takiemu przygotowaniu danych łatwiejsze jest dzielenie danych na mniejsze fragmenty, które mogą być analizowane, przekształcane i przekazywane dalej poprzez wątki na karcie graficznej.

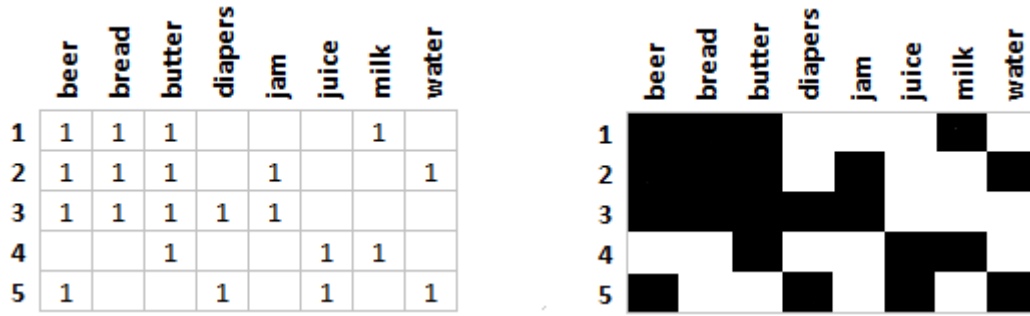
W celu wykonywania obliczeń na karcie graficznej zdecydowano się przekształcić dane wejściowe na bitmapę wartości binarnych. W rozdziale 2 wprowadzony został przykład bazy danych *DB* - zawarty w tablicy 2.1. Dla tych danych opracowana bitmapa ma postać prezentowaną przez rysunek 5.1.

Procedura BUILD BITMAP tworzy bitmapę na podstawie zestawu transakcji z bazy danych.

BUILD BITMAP

```

1  for each element  $e \in I$ 
2      do
3          for each transakcja  $T \in DB$ 
4              do if  $e \in T$ 
5                  then set pixel  $(e_{id}, T_{id}) \leftarrow true$  //element występuje w transakcji
```



RYSUNEK 5.1: Bitmapa przygotowana dla danych z tablicy 2.1

Jak łatwo zauważyć w procedurze BUILD BITMAP dokonywany jest jednokrotny skan wszystkich transakcji, a także wszystkich transakcji - na tej podstawie budowana jest "pionowa" bitmapa - w tym wypadku "pionowa" oznacza, że każdy wiersz reprezentuje jedną transakcję z bazy, natomiast każda kolumna reprezentuje element ze zbioru elementów I .

W tym miejscu wymagany jest także zastosowanie *mostu* (ang. *bridge*) pomiędzy obliczeniami na CUDA oraz przetwarzaniem logiki biznesowej w C# - należało przyjąć język identyfikacji poszczególnych elementów. Źródła danych stosują z reguły identyfikatory - dlatego też zdecydowano się za ich pomocą identyfikować poszczególne elementy. W przypadku gdy źródło danych nie wspierałoby identyfikatorów liczbowych, możliwe jest wprowadzenie takowych na etapie wczytywania danych - co powinno być odpowiednio obsługiwane przez importery danych wejściowych.

Przykładowy fragment bitmapy zbudowanej dla danych, na których oparte zostały eksperymenty został zaprezentowany na rysunku 5.2.

5.2.2 Ograniczenie bazy danych

Jak łatwo zauważyć na rysunku 5.2, gdzie czarnym kolorem zaznaczone są wystąpienia elementu w danej transakcji (biały to ich brak), bitmapa ta jest dość "żadka", tj. zdecydowanie mniej jest miejsc czarnych niż białych. Oznacza to, że już na tym etapie możliwe jest odrzucenie pewnych elementów, które na pewno nie spełniają wymagań w stosunku do *minsup*.

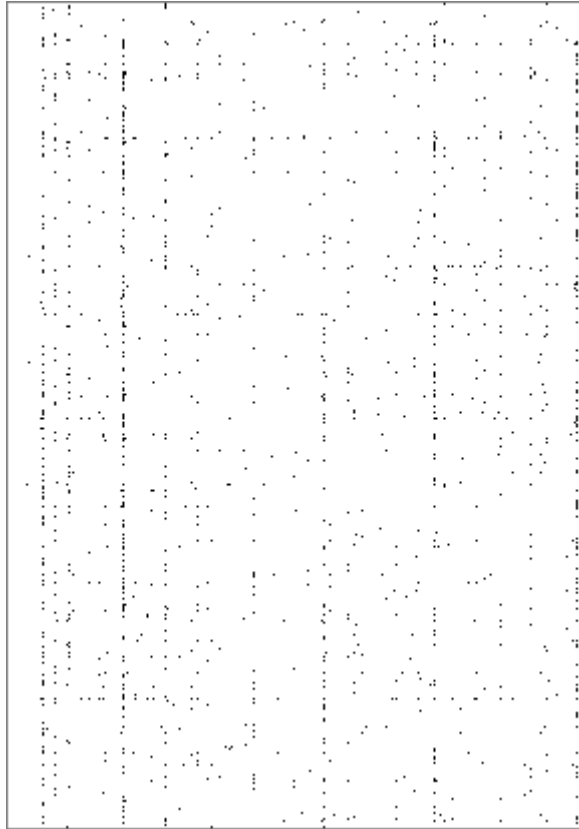
LIMIT BITMAP

```

1  for each kolumna  $c \in bitmap$  parallel
2      do
3           $sup_c \leftarrow \frac{wystapienia}{transakcje}$ 
4          if  $sup_c < minsup$ 
5              then delete kolumna  $c$ 

```

Procedura LIMIT BITMAP dokonuje analizy bitmapy zliczając wystąpienia danego elementu w transakcjach. Następnie oblicza wartość sup_c , czyli wsparcie dla danego elementu. Jeśli jego wsparcie jest mniejsze niż minimalne wsparcie *minsup*, to dana kolumna jest usuwana z bitmapy -



RYSUNEK 5.2: Fragment bitmapy z transakcjami

ponieważ element nie będzie brany pod uwagę w dalszej analizie. Dzięki tej procedurze ograniczona zostaje bitmapa, dzięki czemu możliwa jest szybsza komunikacja pomiędzy CPU a GPU.

5.2.3 Generowanie zbiorów częstych

Łatwo zauważyć, że ograniczona w punkcie 5.2.2 bitmapa składa się wyłącznie z jednoelementowych zbiorów częstych. Na tej podstawie można generować zbiory o krotnościach większych niż 1. W każdym kolejnym kroku generowane są zbiory częste na podstawie zbiorów wygenerowanych w kroku poprzednim, a proces ten jest kontynuowany do momentu aż nie zostaną znalezione żadne zbiory częste.

Procedura CUDA FREQUENT SET GENERATION przedstawia pseudokod realizujący opisywany w tym rozdziale algorytm generowania zbiorów częstych.

CUDA FREQUENT SET GENERATION

```

1   $L_1 \leftarrow \{1\text{-zbiory częste - czyli elementy z ograniczonej bitmapy} \}$ 
2  for ( $k = 2; L_{k-1} \neq \emptyset; k++$ )
3      do  $C_k \leftarrow \text{aprioriGen}(L_{k-1})$ 
4      for each kandydat  $c \in C_k$  parallel
5          do
6              for each transakcja  $t \in DB$ 
7                  if  $c \subseteq t$ 
8                      then      do  $c.\text{count}++$ 
9       $L_k \leftarrow \{c \in C_k | c.\text{count} \geq \text{minsup}\}$ 
10  $\text{Answer} \leftarrow \bigcup_k L_k$ 

```

Należy podkreślić, że w celu bardziej wydajnej implementacji procedury CUDA FREQUENT SET GENERATION dokonana została transpozycja macierzy (bitmapy) wystąpień elementów w transakcjach - bitmapa^T powstała z bitmapy poprzez zamianę jej wierszy na kolumny i kolumn na wiersze.

W wyniku tej operacji możliwe było wydajne zaimplementowanie wspomnianej procedury bez konieczności zapewniania semaforów przy dostępie do pamięci.

Procedura cudaAprioriGen Procedura *cudaAprioriGen* reprezentuje proces tworzenia zbiorów k -elementowych zbiorów kandydatów na podstawie zbiorów wejściowych $(k-1)$ -elementowych.

Jak łatwo zauważyć wynikiem działania CUDA JOIN STEP są zbiory k -elementowe, które powstały na podstawie zbiorów wejściowych L_{k-1} o licznosciach równych $k-1$, a ich zawartość różni się tylko jednym elementem - ostatnim. Ważnym faktem jest to, iż elementy w zbiorach są uporządkowane leksykograficznie, co wykorzystywane jest w implementacji tej procedury.

CUDA JOIN STEP

```

1  insert into  $C_k$ 
2  select  $p.\text{item}_1, p.\text{item}_2, \dots, p.\text{item}_{k-1}, q.\text{item}_{k-1}$  parallel
3  from  $L_{k-1}$   $p, L_{k-1}$   $q$ 
4  where  $p.\text{item}_1 = q.\text{item}_1, \dots, p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$ 

```

Warto zauważyć, że CUDA JOIN STEP jest ekwiwalentem rozszerzania zbioru L_{k-1} każdym elementem zbioru elementów I , a następnie kasowania tych (k) -zbiorów otrzymanych przez dodanie (k) -tego elementu, który nie był w L_{k-1} .

Warunek $p.\text{item}_{k-1} < q.\text{item}_{k-1}$ zapewnia, że nie będą generowane duplikaty. Dlatego też po etapie łączenia zachodzi zależność $C_k \supseteq L_k$.

Podobnie, jak w przypadku algorytmu równoległego implementowanego na CPU opisanego w części 5.1, algorytm CudaApriori nie implementuje etapu obcinania zbiorów, bazując na obserwacji, że zbiorów tych będzie stosunkowo niewiele, a koszt pamięciowy tej operacji byłby zbyt duży.

5.2.4 Generowanie reguł asocjacyjnych

Po zakończeniu pierwszego etapu algorytm przystępuje do drugiego, czyli do budowania reguł asocjacyjnych na podstawie odkrytych zbiorów częstych. Podobnie, jak w [AS94] algorytm będący przedmiotem analizy niniejszej pracy, generuje wszystkie możliwe reguły asocjacyjne dla zadanego zbioru.

Aby wygenerować reguły, dla każdego zbioru częstego l znajdowane są niepuste podzbiory - podzbiór taki oznaczony jest jako a . Dla takich oznaczeń wygenerowana zostanie reguła $a \Rightarrow (l - a)$, jeżeli spełniona jest nierówność $\frac{\text{support}(l)}{\text{support}(a)} \geq \text{minconf}$.

Procedura GENERATE FREQUENT ITEMSETS prezentuje generowanie reguł asocjacyjnych na podstawie odkrytych k -zbiorów częstych l_k będących elementami zbioru L_k ($l_k \in L_k$).

GENERATE FREQUENT ITEMSETS

```

1  for each zbiór częsty  $l_k$ ,  $k \geq 2$  parallel
2      do call genrules( $l_k$ ,  $l_k$ )

```

Widać, że w implementacji algorytmu użyta została metoda GENRULES, która powstała w ramach algorytmu Apriori, opisanego w części 2.2.1 niniejszej pracy. Postanowiono, że zrównoleglenie obliczeń wykonane zostanie na wyższym poziomie - tj. na poziomie poszczególnych zbiorów częstych, co znacząco przyspiesza operacje pamięci potrzebne na skopiowanie danych dla każdego zbioru zajęło by zbyt dużo czasu. Dlatego też pominięto opis wspomnianej procedury.

Rozdział 6

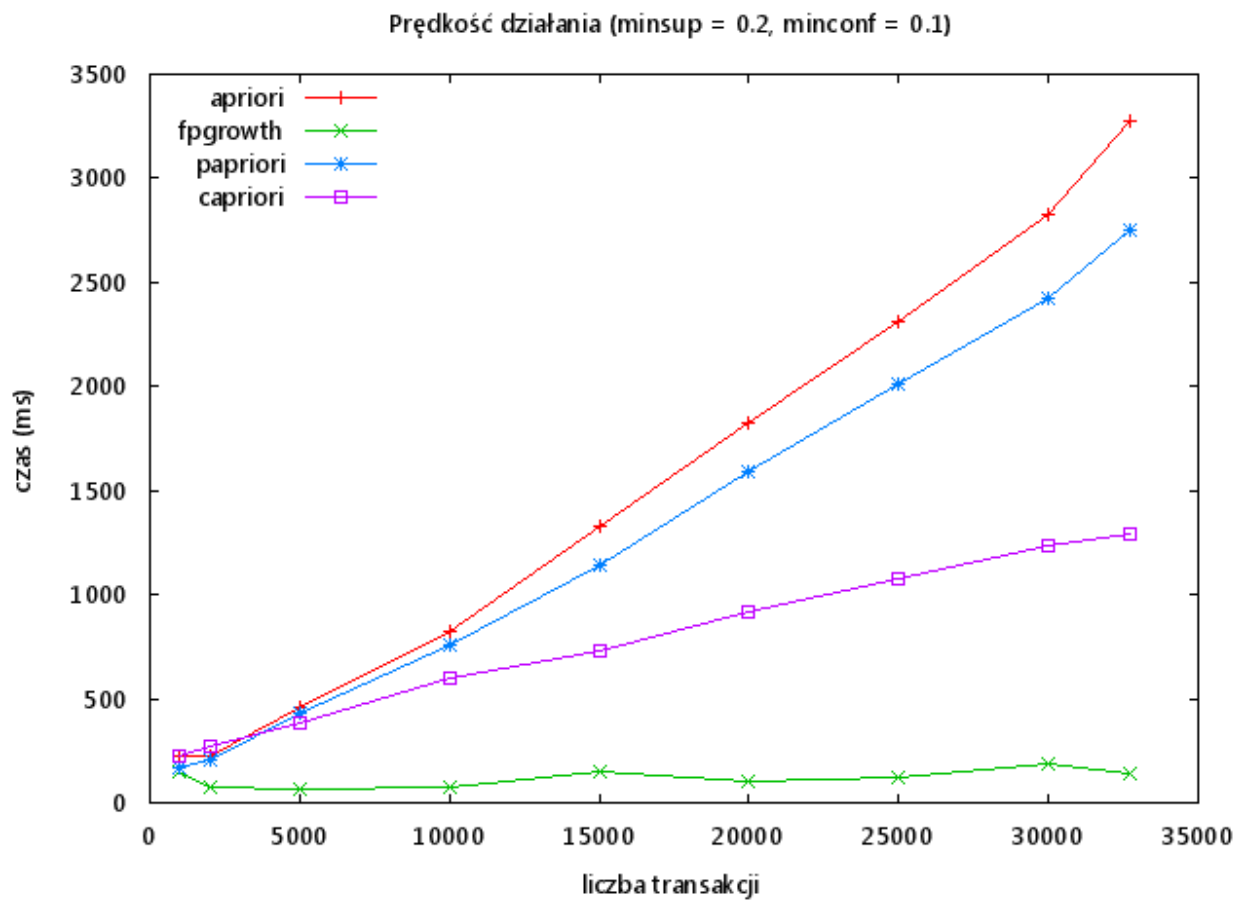
Eksperymenty

6.1 Parametry wejściowe oraz maszyna testowa

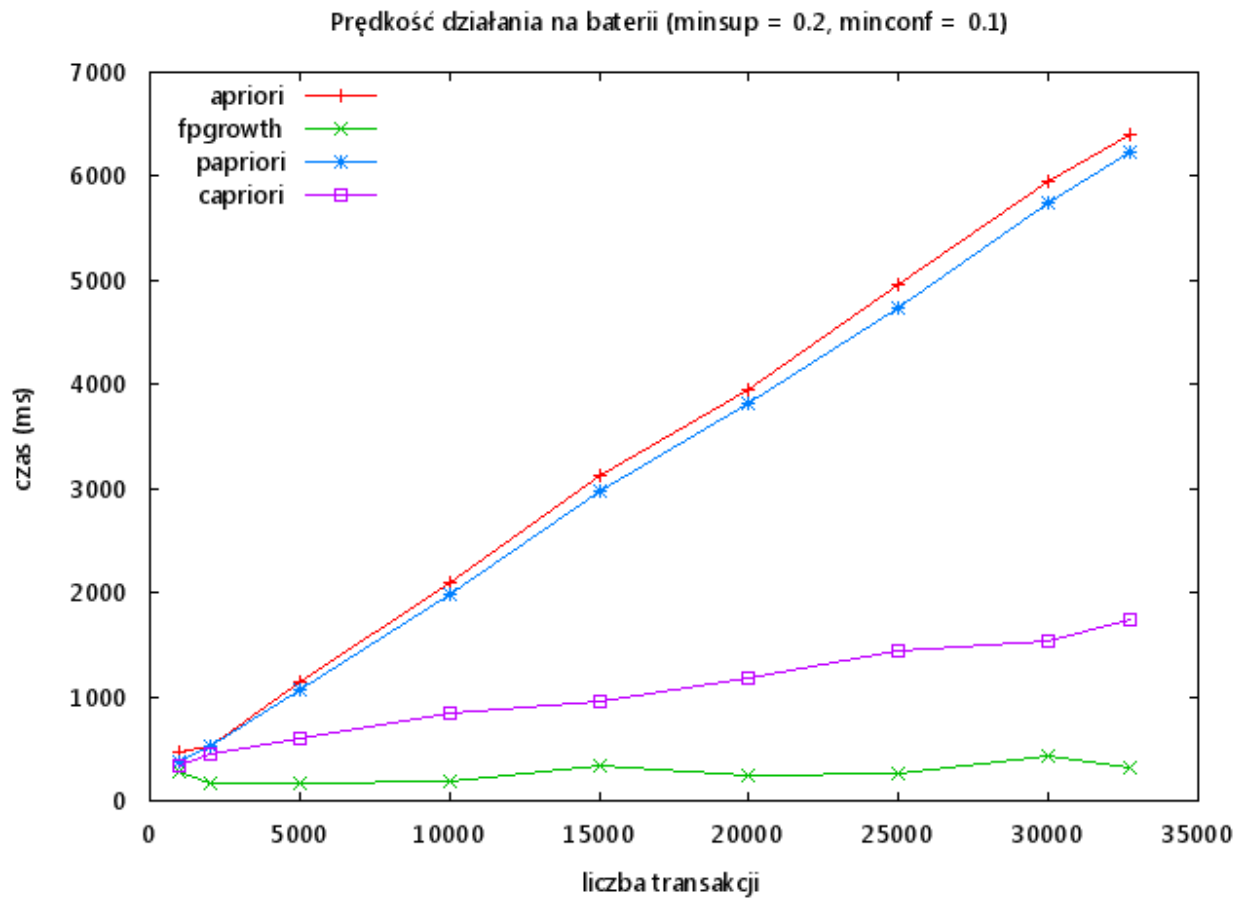
6.2 Wyniki eksperymentów

6.2.1 Wyniki pomiarów poszczególnych implementacji algorytmów

Wyniki dla minsup = 0.2, minconf = 0.1



RYСУNEK 6.1: Prędkość działania algorytmów



RYSUNEK 6.2: Prędkość działania algorytmów na baterii laptopa

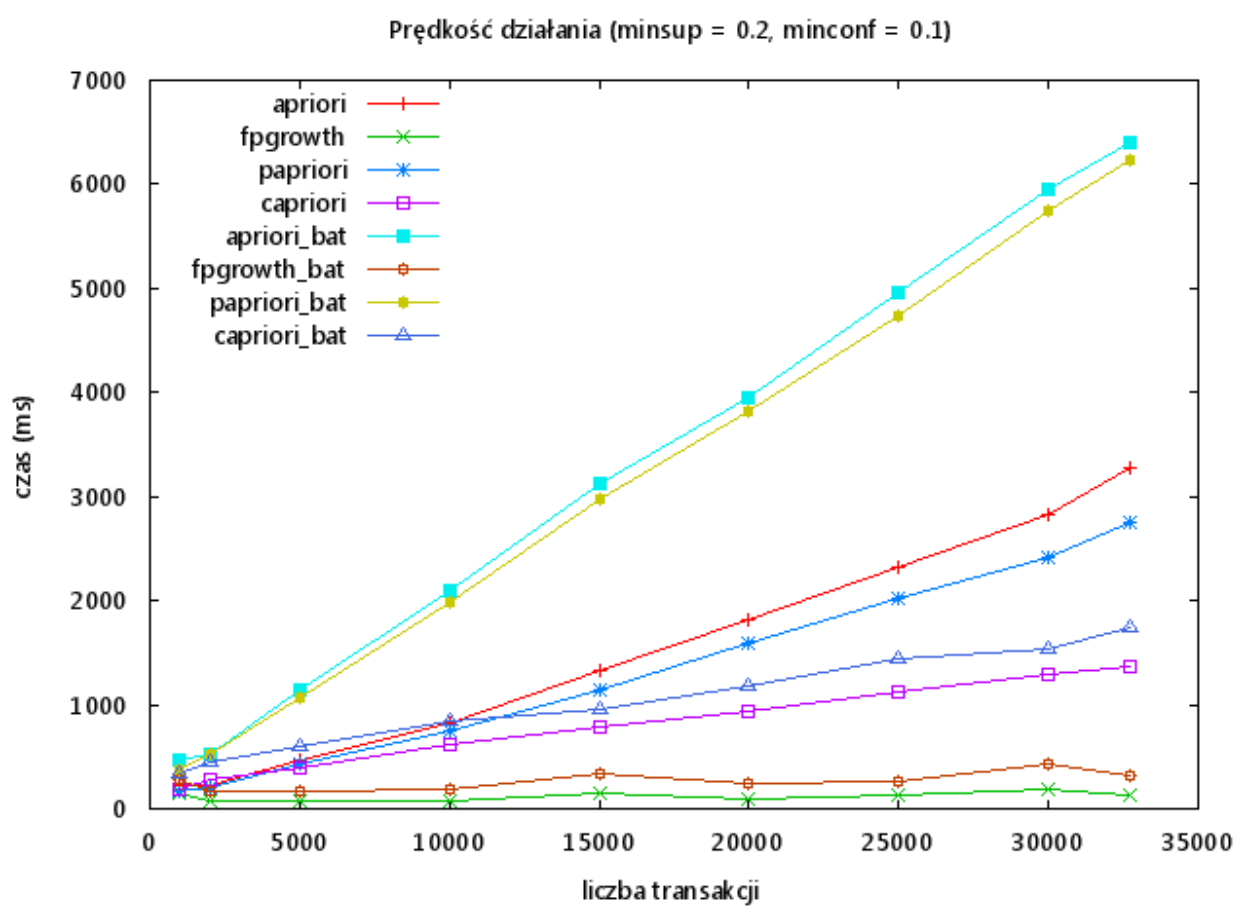
Wyniki dla minsup = 0.2, minconf = 0.4

Wyniki dla minsup = 0.05, minconf = 0.1

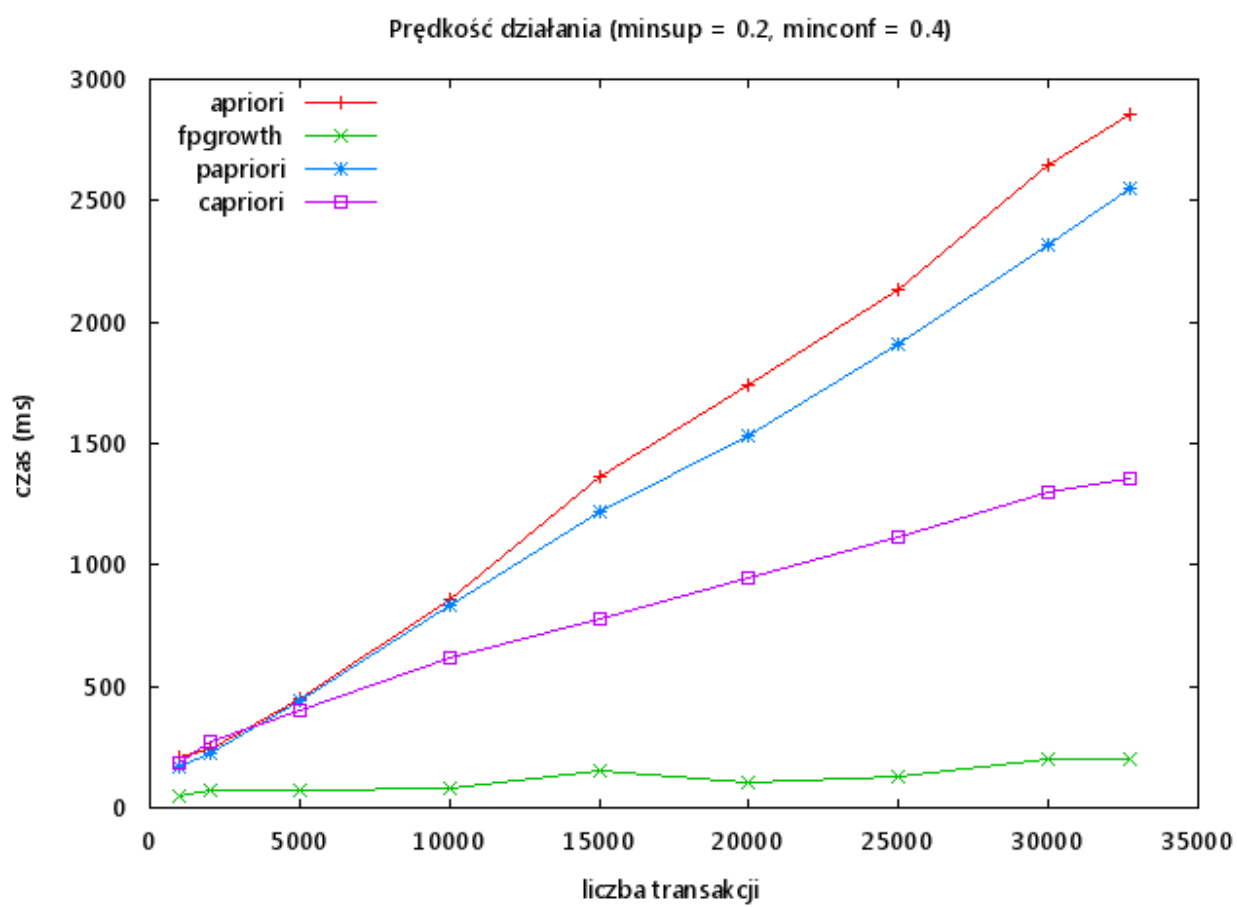
Wyniki dla minsup = 0.05, minconf = 0.4

6.2.2 Wyniki pomiarów dla algorytmu CudaApriori

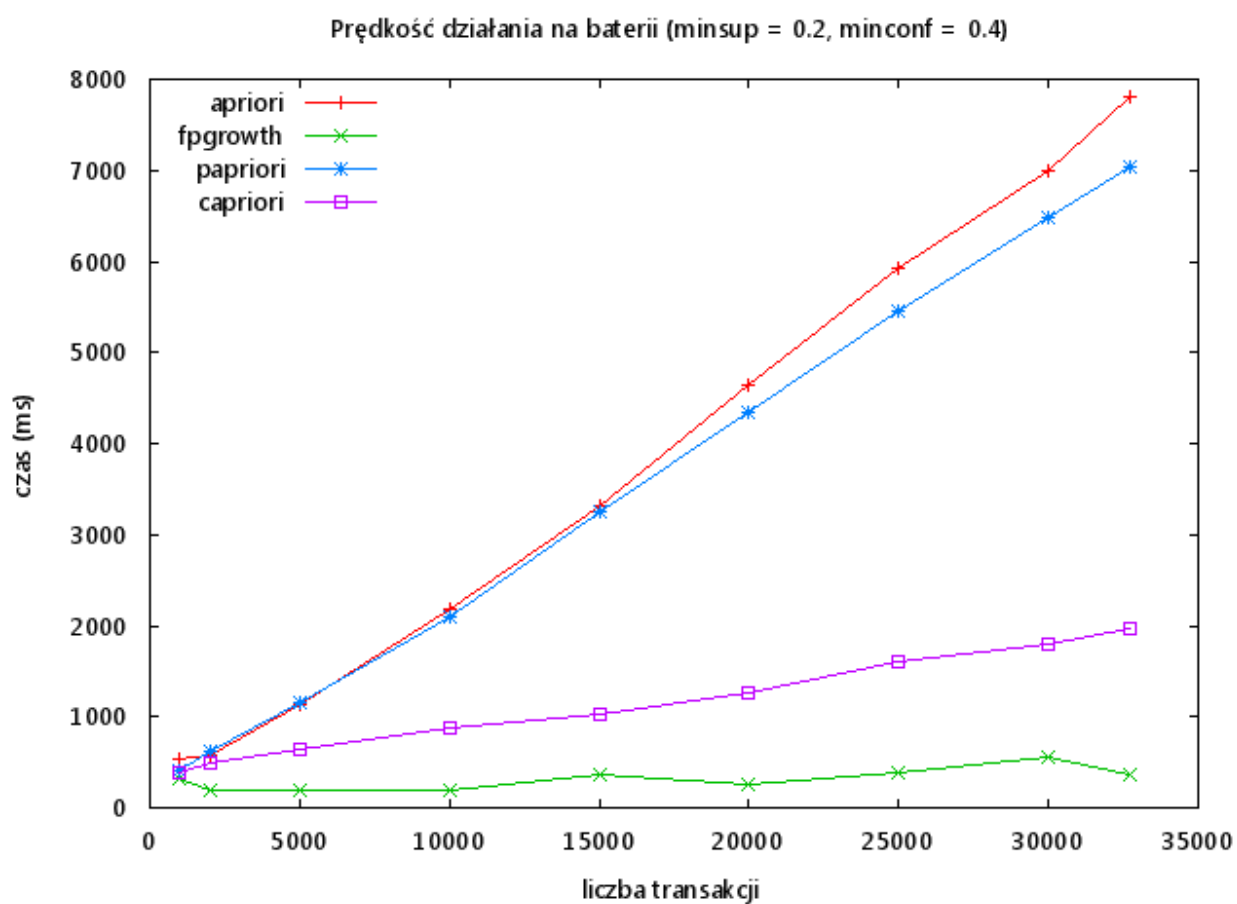
6.3 Podsumowanie eksperymentów i wnioski



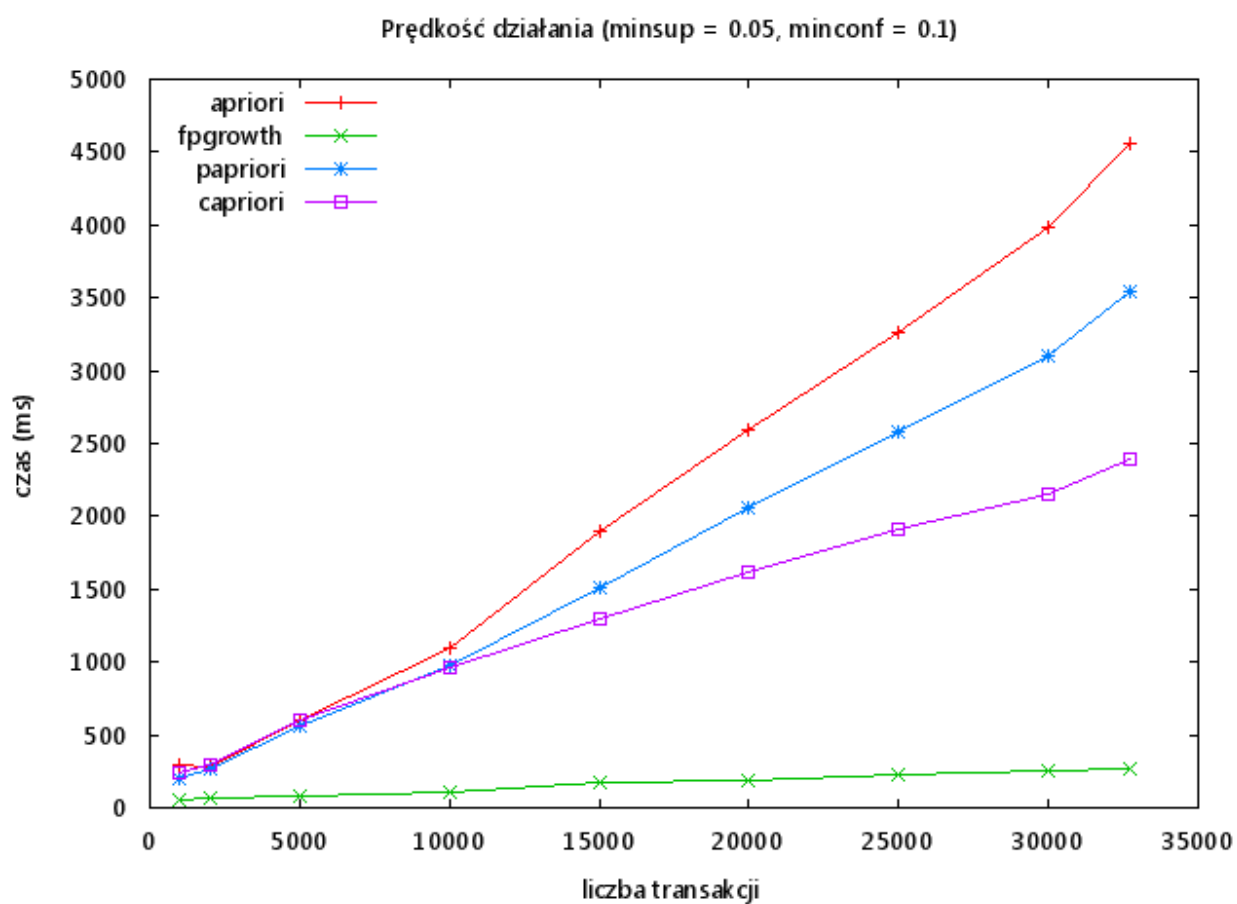
RYSUNEK 6.3: Prędkość działania algorytmów w dwóch środowiskach uruchomieniowych



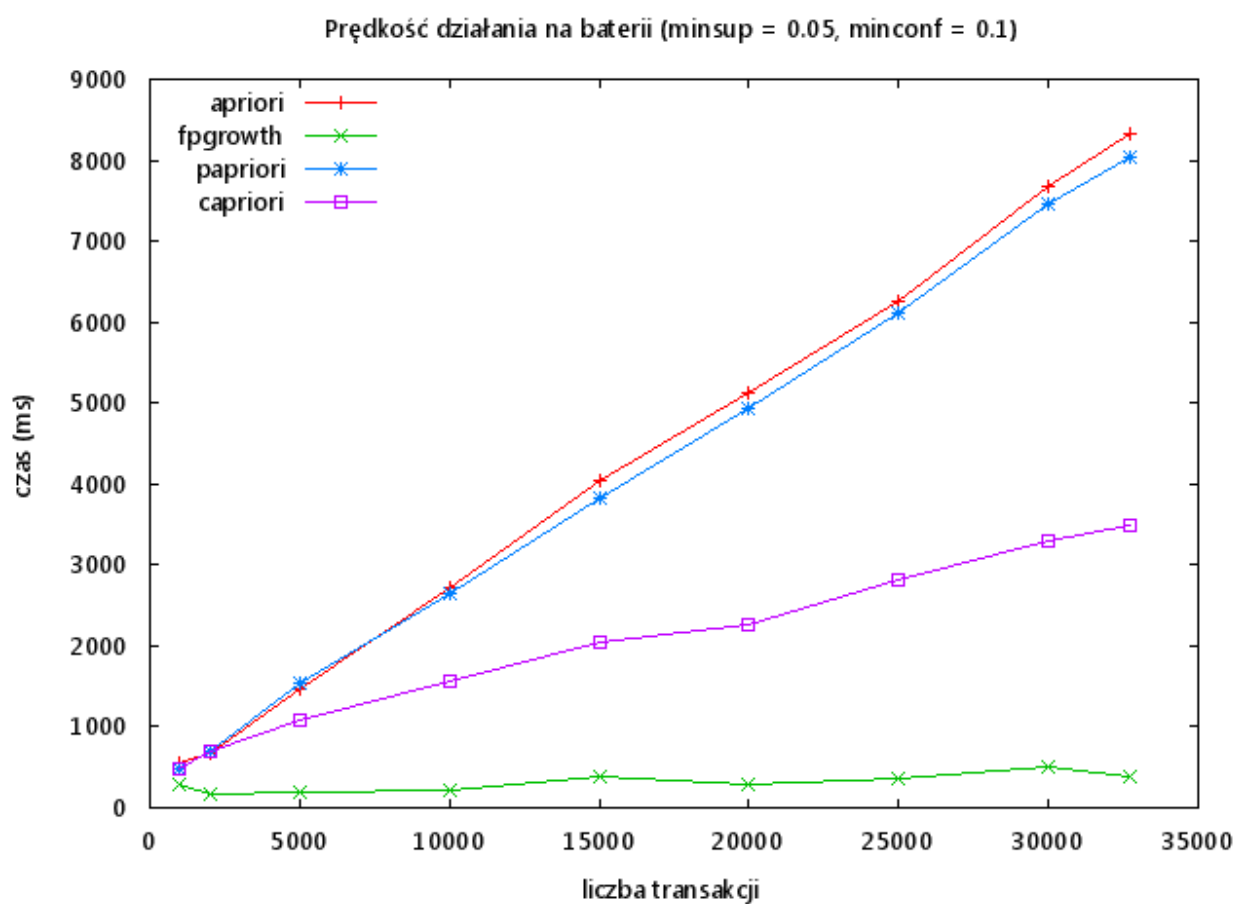
RYSUNEK 6.4: Prędkość działania algorytmów



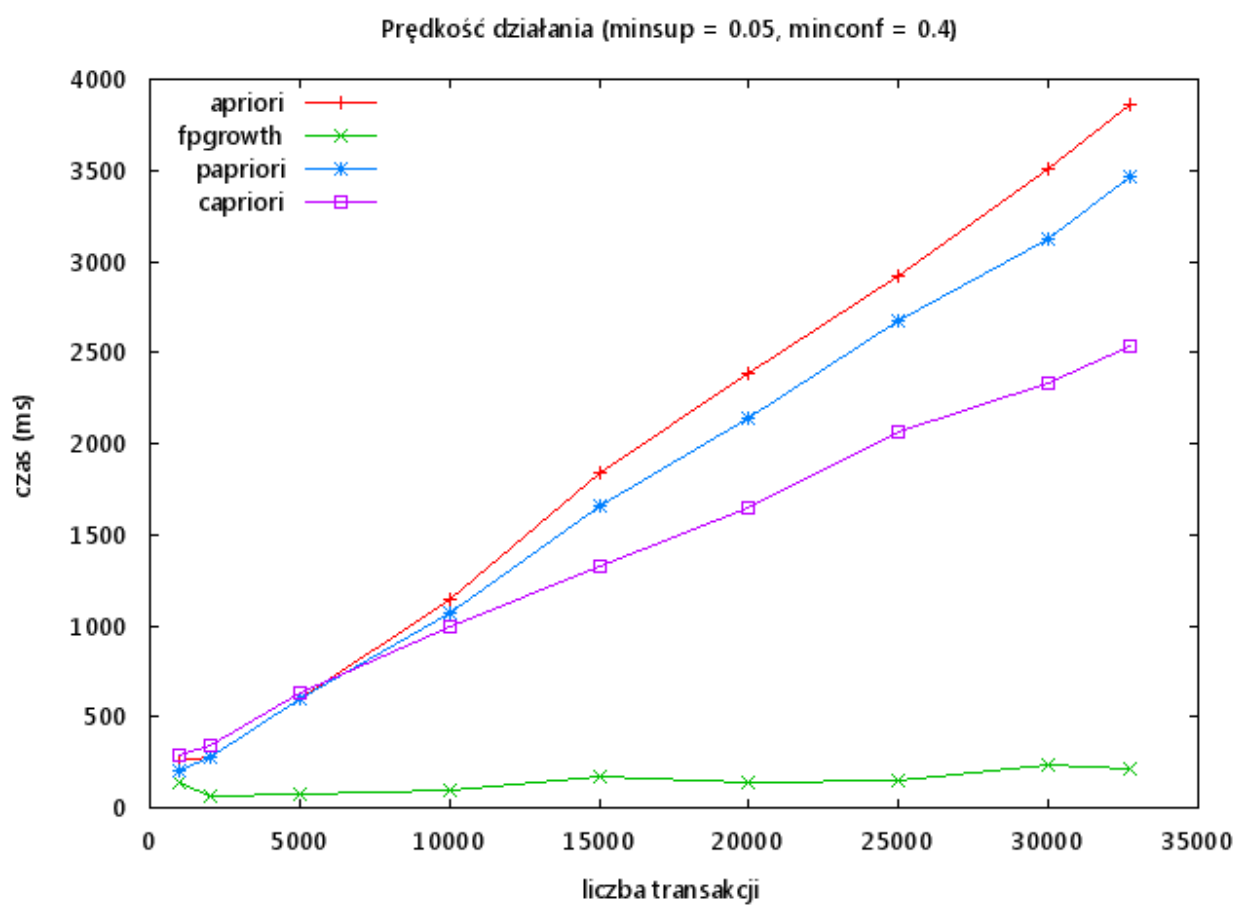
RYSUNEK 6.5: Prędkość działania algorytmów na baterii laptopa



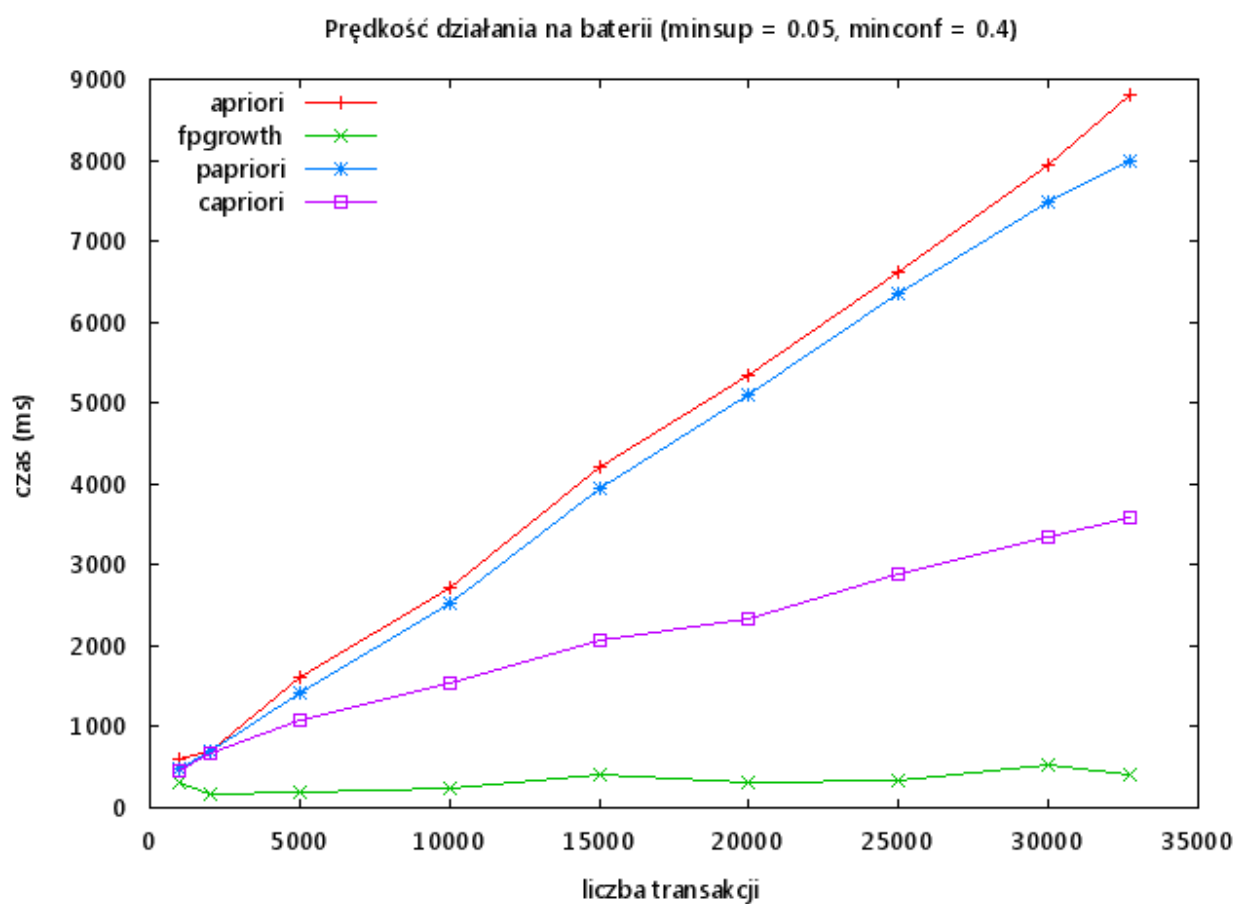
RYSUNEK 6.6: Prędkość działania algorytmów



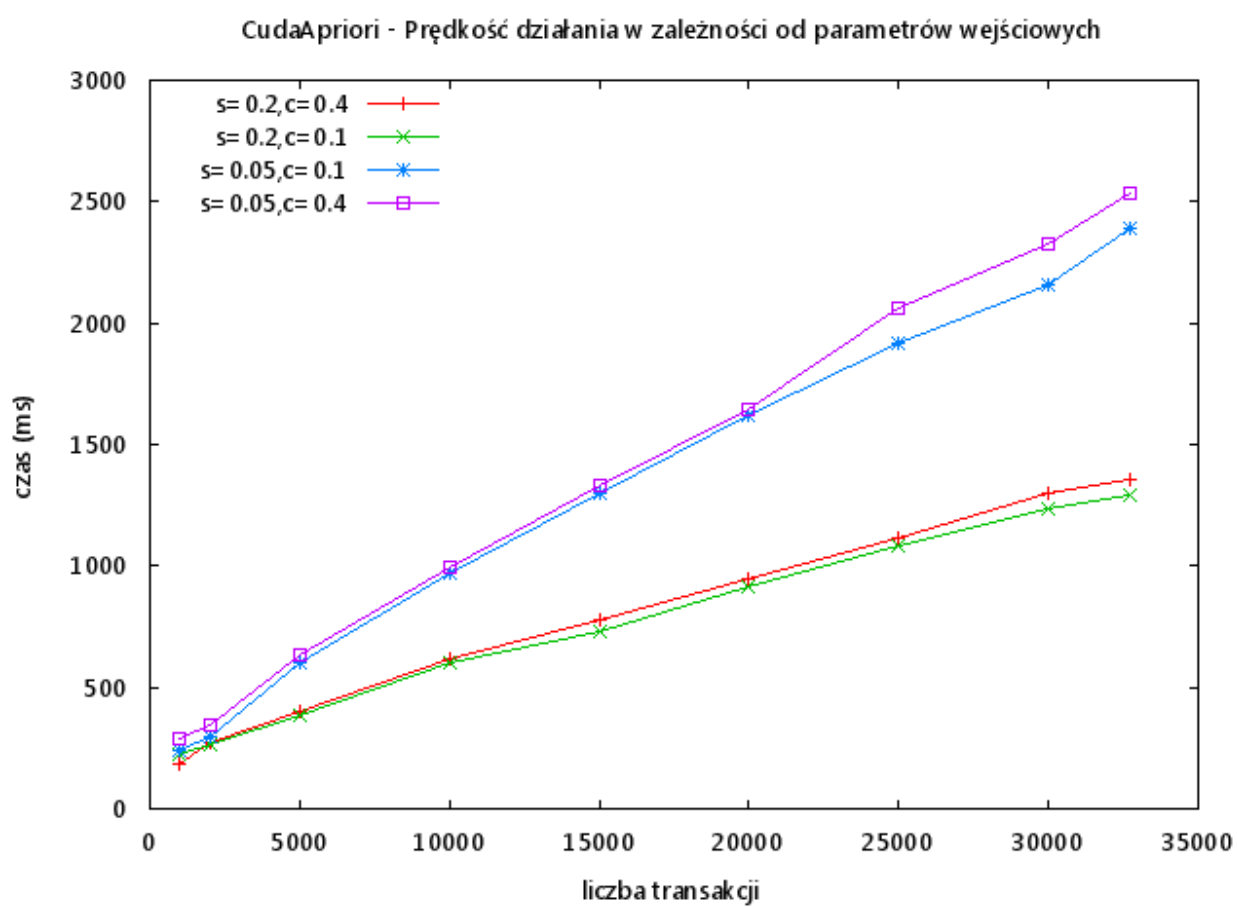
RYSUNEK 6.7: Prędkość działania algorytmów na baterii laptopa



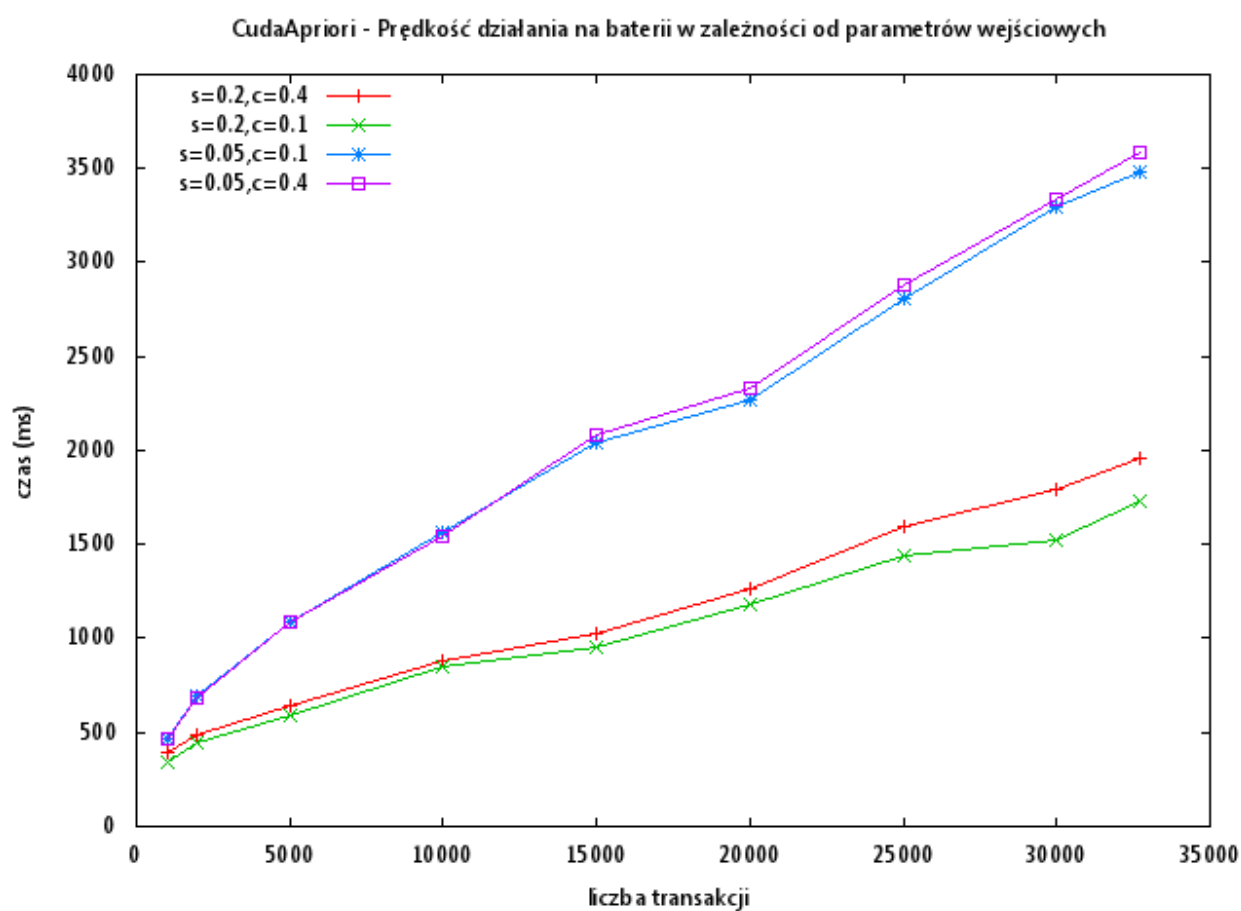
RYSUNEK 6.8: Prędkość działania algorytmów



RYSUNEK 6.9: Prędkość działania algorytmów na baterii laptopa



RYСУNEK 6.10: Prędkość działania algorytmu CudaApriori w zależności od parametrów wejściowych



RYСУNEK 6.11: Prędkość działania algorytmu CudaApriori na baterii laptopa w zależności od parametrów wejściowych

Rozdział 7

Podsumowanie i wnioski

N

Literatura

- [Agr94] Rakesh Agrawal. Quest: a project on database mining. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, strony 514–, 1994.
- [AIS93] Rakesh Agrawal, Tomasz Imielinski, Arun Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, strony 207–216, 1993.
- [AS94] Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithms for mining association rules. *VLDB*, 1994.
- [CK09] Charlie Calvert, Dinesh Kulkarni. *Essential LINQ*. Addison-Wesley Professional, 2009.
- [Cor99] NVIDIA Corporation. NVIDIA GeForce 256. [on-line]
<http://www.nvidia.com/page/geforce256.html>, 1999.
- [Cor00] Microsoft Corporation. Microsoft Delivers First .NET Platform Developer Tools for Building Web Services. [on-line]
<http://www.microsoft.com/presspass/press/2000/Jul00/PDCDeliversPR.mspx>, 2000.
- [Cor01] NVIDIA Corporation. NVIDIA GeForce 3. [on-line]
<http://www.nvidia.com/page/geforce3.html>, 2001.
- [Cor06a] Microsoft Corporation. *C# Language Specification*. ECMA, 2006.
- [Cor06b] NVIDIA Corporation. NVIDIA GeForce 8800 GTX. [on-line]
http://www.nvidia.pl/page/geforce_8800.html, 2006.
- [Cor07a] NVIDIA Corporation. CUDA - Zone. [on-line]
http://www.nvidia.pl/object/cuda_home_new_pl.html, 2007.
- [Cor07b] NVIDIA Corporation. NVIDIA Tesla 20. [on-line]
http://www.nvidia.pl/object/cuda_home_new_pl.html, 2007.
- [Cor09] Microsoft Corporation. Microsoft Visual Studio. [on-line]
<http://www.microsoft.com/poland/developer/>, 2009.
- [Cor10] Intel Corporation. New Dual-Core Intel Atom Processor-Based Netbooks Hit Shelves Today. [on-line] http://newsroom.intel.com/community/intel_newsroom/blog/2010/08/23/new-dual-core-intel-atom-processor-based-netbooks-hit-shelves-today, 2010.
- [Cor11a] Microsoft Corporation. Windows PowerShell Getting Started Guide. [on-line]
[http://msdn.microsoft.com/en-us/library/aa973757\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/aa973757(v=vs.85).aspx), 2011.

- [Cor11b] NVIDIA Corporation. *CUDA C Programming Guide Version 4.0*. NVIDIA Corporation, 2011.
- [Cor11c] NVIDIA Corporation. CUDA Downloads. [on-line] <http://developer.nvidia.com/cuda-downloads>, 2011.
- [Duf11] Joe Duffy. Hello PLINQ. [on-line] <http://www.bluebytesoftware.com/blog/PermaLink,guid,81ca9c00-b43e-4860-b96b-4fd2bd735c9f.aspx>, 2011.
- [EN05] Ramez Elmasri, Shamkand B. Navathe. *Wprowadzenie do systemów baz danych*. Addison-Wesley, Reading, MA, USA, 2005.
- [fASSL11] Company for Advanced Supercomputing Solutions Ltd. CUDA.NET. [on-line] <http://www.hoopoe-cloud.com/Solutions/CUDA.NET/Default.aspx>, 2008-2011.
- [FPSSU96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- [HPYM04] Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 2004.
- [Inc11] GitHub Inc. GitHub - social coding. [on-line] <http://www.github.com>, 2011.
- [Kov11] James Kovacs. psake v4.00. [on-line] <http://codebetter.com/jameskovacs/2010/04/12/psake-v4-00/>, 2011.
- [Lib06] Jesse Liberty. *Programowanie C#*. O'Reilly, 2006.
- [LQ05] Yong-Jie LAn, Yong Qiu. Parallel frequent itemsets mining algorithm without intermediate result. *School of Information and Electronic Engineering*, 2005.
- [Mon11] Mono. Mono Project. [on-line] <http://www.mono-project.com>, 2011.
- [Mue95] Andreas Mueller. Fast sequential and parallel algorithms for association rule mining: A comparison. *University of Maryland at College Park*, 1995.
- [PCY95] Jong Soo Park, Ming-Syan Chen, Philip S. Yu. Efficient parallel data mining for association rules. *strongy* 31–36, 1995.
- [Res11] Microsoft Research. Code Contracts. [on-line] <http://research.microsoft.com/en-us/projects/contracts/>, 2011.
- [SK10] Jason Sanders, Edward Kandrot. *CUDA by Example*. Addison-Wesley, Reading, MA, USA, 2010.
- [Ske10a] Jon Skeet. *C# in Depth*. Manning Publications Co., Stamford, CT 06901, wydanie pierwsze, 2010.
- [Ske10b] Jon Skeet. *C# in Depth*. Manning Publications Co., Stamford, CT 06901, wydanie drugie, 2010.
- [YC06] Yanbin Ye, Chia-Chu Chiang. A parallel apriori algorithm for frequent itemsets mining. *strongy* 87–94, 2006.

- [ZY08] Jiayi Zhou, Kun-Ming Yu. Balanced tidset-based parallel fp-tree algorithm for the frequent pattern mining on grid system. *International Conference on Semantics, Knowledge and Grid*, strongy 103–108, 2008.



© 2011 Tomasz Kujawa

Instytut Informatyki, Wydział Informatyki i Zarządzania
Politechnika Poznańska

Skład przy użyciu systemu L^AT_EX.

BibT_EX:

```
@mastersthesis{ key,  
  author = "Tomasz Kujawa",  
  title = "{Równoległe odkrywanie reguł asocjacyjnych zaimplementowane na procesory graficzne}",  
  school = "Poznan University of Technology",  
  address = "Pozna{\n}, Poland",  
  year = "2011",  
}
```